Similarity Measures: Is It Possible To Compare Dissimilar Structures?

Guido Sello

Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano, via Venezian 21, 20133 Milan, Italy

Received February 12, 1998

The determination of the similarity between dissimilar compounds is a problem only partially solved. The necessity of defining unconnected description spaces to evaluate the diversity of chemical structures is an important drawback of many existing methodologies. Our objective is thus the definition of a method capable of making acceptable comparisons between any kind of structures. The definition of a virtual similarity index, indirectly calculated from standard similarity analyses, and the selection of a proper representation space allow the calculation of two new similarity indexes: the global and the local similarity indexes. The performance of the method is discussed by its application to three sets of compounds at different complexity levels.

INTRODUCTION

Similarity measures have become a standard tool in many fields of organic chemistry, from medicinal chemistry to synthesis design, from combinatorial library analysis to molecular property prediction.[1−7] Consequently, many approaches, both general and specialized, were developed using both geometrical descriptions and molecular properties. Most of the time the similarity prototype and its usage are strongly connected; thus, the advantages of each approach can be maximally exploited, while disadvantages can be minimized. However, the evaluation of molecular similarity represents only one of the two perspectives of the same attribute; the complementary one is molecular diversity. It is obvious that the two descriptions have, in principle, the same background; i.e., if one can measure the similarity between two compounds, s/he can also measure their diversity by just reversing the calculation. But this natural statement is not always directly demonstrable. As soon as the two compounds s/he is comparing have not common features, their similarity measure loses its mathematical meaning. Therefore the presently most common method for measuring diversity is not the reversal of similarity methodologies but the measure of the distance between the compounds in the multidimensional space of their descriptors; the more distant they are, the more diverse their representation is.[8−12]

A totally different aspect can be, however, considered. Let us take two compounds that are very different in terms of descriptors; let us suppose that the first one has values equal to 1, 1, 1, 0, 0, of the five selected descriptors and that the second one has values equal to 0, 0, 0, 1, 1; the result of the comparison is clear: they are completely diverse. But they have been measured in two different representation spaces, completely disjointed; thus, we have no guarantee about the existence of the possibility of validating their comparison.

We already met this impending characteristic in our previous work in the field of synthesis design.[13] There, the necessity of comparing very different objects, either simple structure transformations or complete syntheses, forced us to consider the opportunity of conjecturing an articulated methodology to manage the puzzle. The solution is based on the comparison of the separate analyses of the two objects. Indeed, this solution is not directly transferable to the comparison of molecular structures because their separate analysis is not easily conceivable.

	Characteristic	Similarity
Shape	Spherical	100%
Color	Orange	80%
Use	Food / Game	0%

Figure 1. Different similarity percentages depending on the compared attribute.

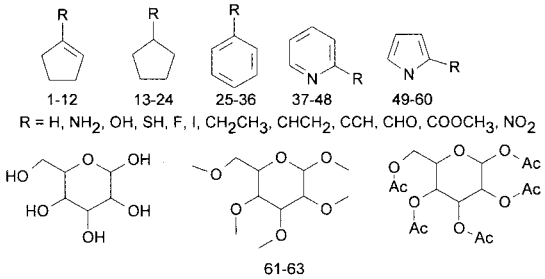

Figure 2. Structures inside the tool set.

The intent of the present work is the introduction of a new method of comparing diverse structures that implicitly contains a real connection between them through the dark world of the zero valued descriptions. In this work it will be possible to demonstrate the similarity between BLACK and WHITE passing through violet, blue, green, red, orange, and yellow.[14]

Finally, in consideration of the difficulties we are going to find during the attempt of making clear our ideas, we

S0095-2338(98)00180-2 CCC: $15.00
Published on Web 07/01/1998

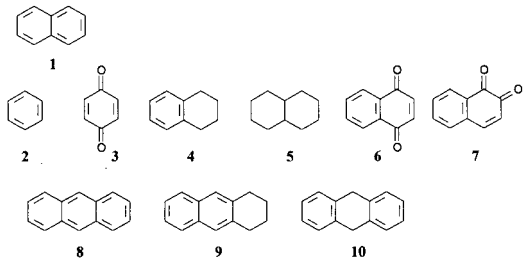

Figure 3. Set A compounds.

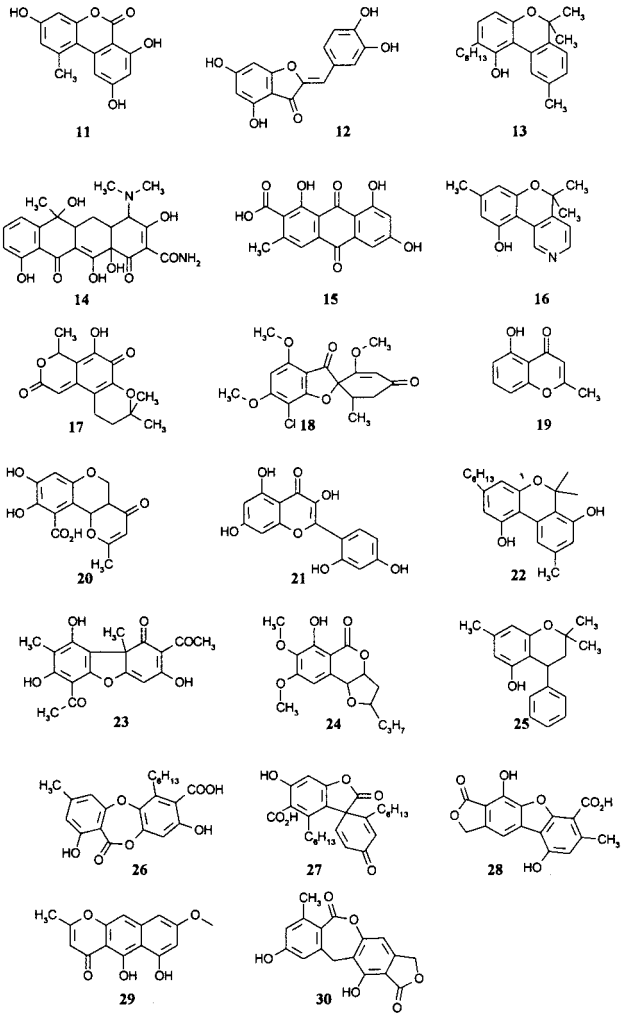

Figure 4. Set B compounds.

would like to assert that we are not in the position of warranting a productive use of our method (drug design, database analysis, or anything else). We will limit the presentation to the concepts and the operations; the next step is still open.

Similarity and Diversity. Similarity is a word that includes several interpretations and that, for this reason, cannot be exactly defined. For example, we can compare a ball and an orange; they are similar in shape, they can be similar in color, but they are definitely diverse in use (Figure 1). The value of the example is that similarity can be defined only with respect to something and it is not an absolute property. As a consequence, if we choose *N* descriptors, we implicitly define *N* different similarities. As long as we are evaluating molecular similarity, the riddle of the interpretation of the results can be neglected because we are

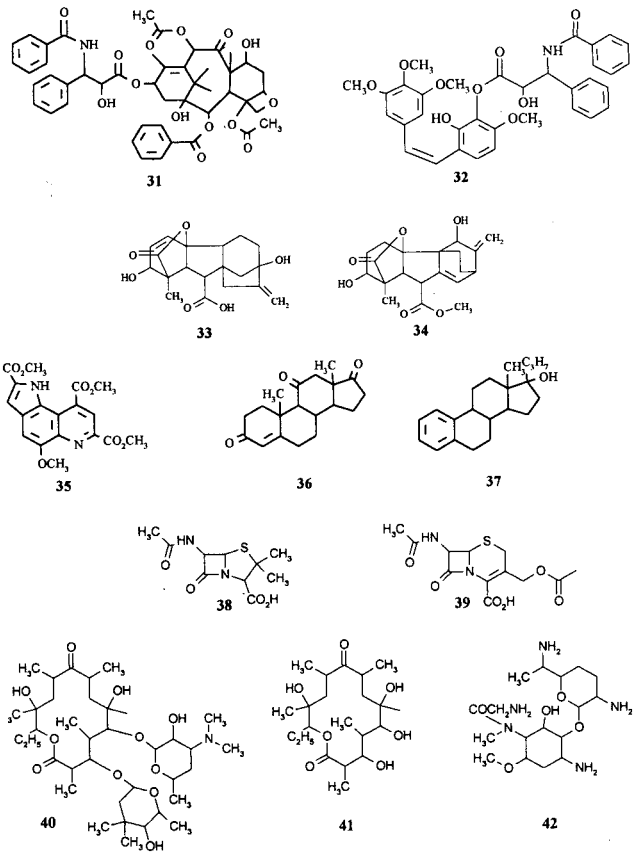

Figure 5. Set C compounds.

comparing similarities with the same reference and, at least for the descriptors that have nonzero values, we are working inside the same knowledge space.

Diversity, at a first glance, can have the same processing of similarity; i.e., we can define molecular diversity only with respect to a particular descriptor. Nevertheless, a discrepancy is evident when we explicitly perform the calculation. In fact, because we are searching compounds that are different, we are calculating descriptors that are different; thus, we are using a different definition of diversity (or of similarity). The main justification relies on the assumption that two compounds that have few, or even no, common descriptors are different. In a complete representation space this assumption can be accepted; in all other cases the assumption cannot be automatically authorized.

Another issue is connected with the use of molecular diversity. Even in a complete description space the comparison between two different objects can give mainly a yes/no answer: the two compounds are either different or not different. In addition, if you are calculating molecular diversity by counting the descriptors that are different, you can experience the following situation. Let us suppose that two objects have *M* descriptors, on the *N* potential descriptors, that are different. A third object will have M' and M'' different descriptors with respect to object one and two. Whatever will be the order of M, M', and M'', it is impossible to assign a corresponding order to the objects! It is impossible to order objects with respect to diversity by just counting the descriptors that are different! This statement should warn anybody wishing to measure molecular diversity.

Table 1. Global and Local Similarity Indexes of Set A Using the Complete Tool Set

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index	direct similarity index
1-2	33	0.89	25	1.00	0.7619	0.9014
1-1	33	0.89	33	0.89	0.7901	1.0000
1-4	33	0.89	28	0.82	0.5856	0.6000
1-5	33	0.89	19	0.71	0.1233	0.0000
			1	0.40		
1-8	33	0.89	33	0.73	0.4659	0.8975
			49	0.42		
1-10	33	0.89	28	0.67	0.4741	0.6009
			28	0.67		
1-9	33	0.89	33	0.73	0.1876	0.8333
			13	0.42		
1-7	33	0.89	32	0.80	0.5333	0.7273
1-6	33	0.89	32	0.80	0.1752	0.7273
			8	0.42		
3-6	32	0.80	60	0.63	0.0885	0.0000
	8	0.42				

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index	direct similarity index
1-2	33	0.76	33	0.76	0.7619	0.9014
1-1	33	0.79	33	0.79	0.7901	1.0000
1-4	33	0.59	33	0.59	0.5926	0.6000
1-5	33	0.24	19	0.24	0.2353	0.0000
1-8	33	0.64	33	0.64	0.6464	0.8975
1-10	33	0.48	33	0.48	0.4849	0.6009
1-9	33	0.64	33	0.64	0.6464	0.8333
1-7	33	0.62	46	0.62	0.6222	0.7273
1-6	33	0.62	46	0.62	0.6222	0.7273
3-6	46	0.38	60	0.38	0.3750	0.0000

Table 2. Global and Local Similarity Indexes of Set A Using the Tool Set without the Alkenyl Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	33	0.89	25	1.00	0.7619
1-1	33	0.89	33	0.89	0.7901
1-4	33	0.89	28	0.82	0.5856
1-5	33	0.89	19	0.71	0.1103
			13	0.40	
1-8	33	0.89	33	0.73	0.4659
			49	0.42	
1-10	33	0.89	28	0.67	0.4741
			28	0.67	
1-9	33	0.89	33	0.73	0.1876
			13	0.42	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	33	0.76	33	0.76	0.7619
1-1	33	0.79	33	0.79	0.7901
1-4	33	0.59	33	0.59	0.5926
1-5	33	0.24	19	0.24	0.2353
1-8	33	0.64	33	0.64	0.6464
1-10	33	0.48	33	0.48	0.4849
1-9	33	0.64	33	0.64	0.6464

Assumptions. To devise a system for comparing dissimilar structures, we should accept some statements that we cannot prove. The rationale behind the introduction of the following assumptions concerns the transfer of concepts from the classical similarity field to the field of dissimilarity. The idea is that it could be possible to measure the similarity between two compounds that are not similar at all. In other words, if you would like to measure the similarity of two compounds that are described by descriptors that have highly different values, you can adopt the standard procedures for similarity measuring inside a set of comparisons and not, as is usually done, in a single comparison. Consequently we will use the same terminology of standard similarity calculations, but we must remember that we are doing something different. Thus, we must agree on the following assumptions.

Table 3. Global and Local Similarity Indexes of Set A Using the Tool Set without the Alkyl Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	33	0.89	25	1.00	0.7619
1-1	33	0.89	33	0.89	0.7901
1-4	33	0.89	28	0.82	0.5856
1-5	33	0.89	1	0.40	0.0646
			1	0.40	
1-8	33	0.89	33	0.73	0.4659
			49	0.42	
1-10	33	0.89	28	0.67	0.4741
			28	0.67	
1-9	33	0.89	33	0.73	0.1816
			1	0.32	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	33	0.76	33	0.76	0.7619
1-1	33	0.79	33	0.79	0.7901
1-4	33	0.59	33	0.59	0.5926
1-5	33	0.13	4	0.13	0.1333
1-8	33	0.64	33	0.64	0.6464
1-10	33	0.48	33	0.48	0.4849
1-9	33	0.64	33	0.64	0.6464

Table 4. Global and Local Similarity Indexes of Set A Using the Tool Set without the Benzene Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	45	0.89	37	1.00	0.7619
1-1	45	0.89	45	0.89	0.7901
1-4	45	0.89	42	0.82	0.5856
1-5	45	0.89	19	0.71	0.1233
			1	0.40	
1-8	45	0.89	45	0.73	0.4314
			49	0.42	
1-10	45	0.89	42	0.67	0.4741
			42	0.67	
1-9	45	0.89	45	0.73	0.1876
			13	0.42	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	45	0.77	57	0.77	0.7658
1-1	45	0.79	45	0.79	0.7901
1-4	45	0.59	45	0.59	0.5926
1-5	45	0.24	19	0.24	0.2353
1-8	45	0.64	45	0.64	0.6464
1-10	45	0.48	45	0.48	0.4849
1-9	45	0.64	45	0.64	0.6464

First, we agree that the measure we are going to effect is not a classical similarity measure, or better, it cannot be directly compared with conventional similarity measures.

Second, the diversity measure is composed of similarity measures; each of them maintains its usual meaning, but none of their simple compositions can be used outside the scope of the complete methodology.

Third, we can compare objects even very different (like balls and shoes), but we should assume that, in the present representation space, they are comparable.

Table 5. Global and Local Similarity Indexes of Set A Using the Tool Set without the Pyridine Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	33	0.89	25	1.00	0.7619
1-1	33	0.89	33	0.89	0.7901
1-4	33	0.89	28	0.82	0.5856
1-5	33	0.89	19	0.71	0.1233
			1	0.40	
1-8	33	0.89	33	0.73	0.4659
			49	0.42	
1-10	33	0.89	28	0.67	0.4741
			28	0.67	
1-9	33	0.89	33	0.73	0.1876
			13	0.42	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	33	0.76	33	0.76	0.7619
1-1	33	0.79	33	0.79	0.7901
1-4	33	0.59	33	0.59	0.5926
1-5	33	0.24	19	0.24	0.2353
1-8	33	0.64	33	0.64	0.6464
1-10	33	0.48	33	0.48	0.4849
1-9	33	0.64	33	0.64	0.6464

Table 6. Global and Local Similarity Indexes of Set A Using the Tool Set without the Pyrrol Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	33	0.89	25	1.00	0.7619
1-1	33	0.89	33	0.89	0.7901
1-4	33	0.89	28	0.82	0.5856
1-5	33	0.89	19	0.71	0.1233
			1	0.40	
1-8	33	0.89	33	0.73	0.4439
			25	0.40	
1-10	33	0.89	28	0.67	0.4741
			28	0.67	
1-9	33	0.89	33	0.73	0.1876
			13	0.42	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	33	0.76	33	0.76	0.7619
1-1	33	0.79	33	0.79	0.7901
1-4	33	0.59	33	0.59	0.5926
1-5	33	0.24	19	0.24	0.2353
1-8	33	0.64	33	0.64	0.6464
1-10	33	0.48	33	0.48	0.4849
1-9	33	0.64	33	0.64	0.6464

Fourth, the combinations of the measures into similarity indexes are simple tools to permit the ordering of the objects, but it is impossible to directly assign any physical meaning to them.

Fifth and last, we know that there is a founded probability that the measure loses its reliability during the passage from one structure to another. However, the backbone idea of this work is that if we can transform one structure into a different one by small perturbations and if we can evaluate the cost of all of the applied perturbations, the cost can give a measure of the distance of the two structures in the current representation space.

This way, we recover the original interpretation of the term similarity; i.e., we can measure the similarity of two diverse structures with respect to some well-defined descriptor. Thus, we must only solve the problem of the procedure performing the small perturbations in a similarity space.

Table 7. Global and Local Similarity Indexes of Set A Using the Tool Set without the Alkenyl and Alkyl Derivatives

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2	33	0.89	25	1.00	0.7619
1-1	33	0.89	33	0.89	0.7901
1-4	33	0.89	28	0.82	0.5856
1-5	33	0.89			0.0000
1-8	33	0.89	33	0.73	0.4659
			49	0.42	
1-10	33	0.89	28	0.67	0.4741
			28	0.67	
1-9	33	0.89	33	0.73	0.6464

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	33	0.76	33	0.76	0.7619
1-1	33	0.79	33	0.79	0.7901
1-4	33	0.59	33	0.59	0.5926
1-5	33	0.24			0.0000
1-8	33	0.64	33	0.64	0.6464
1-10	33	0.48	33	0.48	0.4849
1-9	33	0.64	33	0.64	0.6464

Table 8. Global and Local Similarity Indexes of Set A Using the Tool Set without the Benzene and Pyridine Derivatives

entry tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index	
1-2	57	0.82	57	1.00	0.7602
1-1	57	0.82	57	0.89	0.6782
1-4	57	0.82	55	0.71	0.1510
			1	0.40	
1-5	57	0.82	19	0.71	0.1130
			1	0.40	
1-8	57	0.82	57	0.67	0.4453
			49	0.52	
1-10	57	0.82	55	0.57	0.3361
			55	0.57	
1-9	57	0.82	57	0.67	0.1569
			22	0.48	

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	local similarity index
1-2	57	0.76	57	0.76	0.7602
1-1	57	0.68	57	0.68	0.6782
1-4	57	0.48	57	0.48	0.4844
1-5	57	0.23	19	0.23	0.2325
1-8	57	0.55	57	0.55	0.5490
1-10	57	0.47	57	0.47	0.4706
1-9	57	0.55	57	0.55	0.5490

METHOD

Many manipulations can be envisaged for performing the stepwise transformation of one structure into another. But, to develop a simple, effective, and fast, method, we opted for a natural approach. We realized that when showing the connection between two objects to somebody inexpert, the most efficient way is to represent the objects with other objects that s/he has experienced, i.e., reasoning by examples. Consequently, we decided to use a set of well-known, completely described structures as a tool set through which new, unrelated structures can be connected. That is, we are going to describe the investigated structures using the tool set. Since the structures in the tool set have been completely

Table 9. Global Similarity Indexes of Set A Using the Tool Set without the Aromatic Derivatives[a]

entry	tool set member	virtual similarity index	tool set member	virtual similarity index	global similarity index
1-2					0.0000
1-1					0.0000
1-4			22	0.59	0.0000
1-5			19	0.71	0.0000
			1	0.40	
1-8					0.0000
1-10			21	0.29	0.0000
1-9			22	0.48	0.0000

[a] LSI cannot be calculated in absence of aromatic compounds in the tool set.

correlated one to the other, it is possible to follow a predetermined route connecting the investigated structures, and, more important, it is possible to weigh the total cost of the route just by combining the costs of each single tract.

Similarity Measure. In agreement with the points previously mentioned the choice of the similarity measure is not critical in the present approach; it is sufficient to remember that any measure should continue to keep its limited definition. Thus, for our convenience, we again used our standard similarity measure, calculated by using atom electronic energy.[15]

$$W_i = |E_{\text{tot.}} - E_{(\text{tot.}-i)}| \quad (1)$$

where $E_{\text{tot.}}$ is the energy of the starting structure and $E_{(\text{tot}-i)}$ is the energy of the structure from which atom i has been eliminated; the energies come from the following formula

$$E = \sum_i E_{\text{i}} \quad (2)$$

where E_i is the energy of atom i.[16,17]

Because we are going to describe a method to measure similarity between dissimilar compounds based on true similarity measures, the method is independent of the actual choice. However, in favor of our choice there is the complete broadness of the measure that is a calculated value and can be extended to any organic compound.

Tool Set. Structural variety in organic compounds is high; thus, the selection of a complete representation set made by small representative molecules risks being a hard and time-consuming job. However, the best chance we have available is to begin with a limited training set and to check along the way if there is the necessity of making it bigger. The selected molecules should not be too many and too complex, but they should cover the field of structural variety. Moreover, to keep the set ordered, they must be grouped into classes. We selected six different compound classes, alkenyl, alkyl, benzene, pyridine, pyrrol, and sugar derivatives: all of them, sugars excluded, contain a representative set of functional groups; sugars are only glucose derivatives with different protective groups. The structures are sketched in Figure 2. Consequently, we have 63 structures in the tool set.

We calculated the corresponding electronic energies for all of the atoms in all of the structures, obtaining the desired molecular representations.

The next step expects the determination of the similarity between the structures of the set. This step is divided into two phases. In the first phase we calculate the similarity inside each group of the set, operating with our standard methodology[17] and obtaining the in-group similarity measures. Then, we apply the same procedure to all the structures of the set. The result is, as expected, an extremely limited number of similarity values different from zero.

The second phase is directed to the solution of the zeroes in the value matrix. In the same context of the general approach, a first level correlation between different structures is developed. It is based on the following idea: if we consider two groups of the set, we can always find at least two structures that have a nonzero similarity measure (because the sets were built in a way that ensures the similarity of at least two of their compounds), then we can use these two structures as the similarity connection between any other pair of structures in the two groups. Thus we proceed as follows: for each structure pair A and A′ with similarity measure equal to zero (nonzero measures keep their calculated value), (a) locate the two structures B and B′ of the corresponding groups that have the highest nonzero measure and (b) calculate the similarity measure between A and A′ using the measure of B and B′ and the corresponding in-group measures (A and B; A′ and B′).

This is the general approach; however, we still have to choose the mathematical operation to realize the connection. Tentatively, we propose the multiplication of the similarity indexes of each structure pair divided by the number of the connecting tracts (here, always equal to 2); we obtain eq 1, where VSI is the virtual similarity index, SI is the real similarity index, and NT is the number of tracts.

$$\text{VSI(A,A′)} = [\text{SI(A,B) SI(B,B′) SI(A′,B′)}]/\text{NT} \quad (3)$$

A comment on this calculation is necessary. First, because the SIs always range between 0 and 1, their products will keep this property; thus, the multiplication seems to be a good relation. Second, the division by NT introduces a bias to the operation, just marking that the VSI is less significant than the SIs (this clearly produces an overestimation of the compounds that are directly comparable, but it represents a simple way to introduce the bias). Third, it is well-known that if compound A is similar to compound B and compound B is similar to compound C, it is impossible to assert that compounds A and C are also similar. However, (a), inside each subset we calculate only true SIs that depend on the energy of the atoms (energy that is influenced by all of the atoms in the molecule); (b) the comparison between molecules of different subsets gives a true SI. Consequently, we can suppose that the VSIs maintain a physical meaning. We must acknowledge that this is only an assumption as clearly stated above. A similar procedure using different descriptors should be carefully verified.

The final result is a triangular matrix containing either the real or the virtual similarity indexes of all of the set structures. The matrix is internally represented by a vector whose index is calculated by the program, thus simulating the matrix.

Similarity Indexes. Having available a complete description set (i.e., the tool set, TS), we faced our principal problem: the comparison between complex structures outside the TS. We developed two calculation procedures, both

Table 10. Global Similarity Indexes of Set B

entry	tool set member	virtual similarity index[a]	tool set member	virtual similarity index[a]	global similarity index 1[b]	global similarity index 2[c]	direct similarity index
11-12	35	0.62	32	0.55	0.0762	0.482	0.6103
	51	0.40	32	0.55			
			11	0.20			
11-13	35	0.62	33	0.50	0.0688	0.448	0.5581
	51	0.40	14	0.33			
			52	0.33			
11-14	35	0.62	32	0.40	0.0204	0.378	0.3529
	51	0.40	8	0.21			
			20	0.21			
			3	0.16			
11-15	35	0.62	32	0.52	0.2049	0.515	0.5553
	51	0.40	32	0.52			
11-16	35	0.62	33	0.62	0.1937	0.515	0.6486
	51	0.40	51	0.42			
11-17	35	0.62	11	0.48	0.0165	0.438	0.1538
	51	0.40	11	0.34			
			11	0.28			
11-18	35	0.62	32	0.50	0.0521	0.445	0.3721
	51	0.40	8	0.26			
11-19	35	0.62	32	0.76	0.0643	0.535	0.6156
	51	0.40	23	0.36			
11-20	35	0.62	32	0.55	0.0643	0.483	0.4500
	51	0.40	8	0.36			
11-21	35	0.62	35	0.56	0.2166	0.528	0.5954
	51	0.40	32	0.53			
11-22	35	0.62	27	0.57	0.2346	0.545	0.3889
	51	0.40					
11-23	35	0.62	31	0.48	0.0233	0.400	0.4545
	51	0.40	8	0.31			
			3	0.19			
			11	0.18			
11-24	35	0.62	35	0.63	0.0556	0.485	0.4878
	51	0.40	15	0.29			
11-25	35	0.62	32	0.57	0.2269	0.540	0.5451
	51	0.40	32	0.57			
11-26	35	0.62	35	0.47	0.0445	0.423	0.7660
	51	0.40	27	0.40			
			14	0.29			
			3	0.18			
11-27	35	0.62	32	0.40	0.0176	0.390	0.3529
	51	0.40	8	0.36			
			13	0.27			
			3	0.16			
			3	0.16			
11-28	35	0.62	35	0.61	0.0809	0.470	0.7143
	51	0.40	27	0.47			
			3	0.21			
11-29	35	0.62	32	0.57	0.0709	0.460	0.5847
	51	0.40	51	0.38			
			23	0.28			
11-30	35	0.62	35	0.61	0.0729	0.467	0.5429
	51	0.40	27	0.47			
			11	0.19			

[a] VSI values are calculated before connection through the TS. [b] Equation 2. [c] Equation 3.

using the previously described essential principles. In fact, we can operate two alternative comparisons: a global comparison, giving rise to a global similarity index (GSI), and a local comparison, calculating a local similarity index (LSI).

In the GSI case, the first structure A is compared with the complete TS, selecting the most similar tool structure. The comparison is repeated until one of two conditions is fulfilled: either molecule A has been fully described through the TS (i.e., all its atoms have been found similar to atoms in the TS compounds) or the last comparison has resulted in an empty result (i.e., the last check has not found new atoms similar to atoms in the TS compounds). The same procedure is applied to the second structure B. This way, we obtain the nearly complete description of A and B projected on TS. Then, using the precalculated SI or VSI, we obtain the VSI of A and B through TS. This calculation mode maximizes the separate representation of A and B. Consequently, the same molecule A gives the same representation disregarding the actual partner B. On the other hand, the GSI value depends on both A and B.

In the LSI case, both A and B are compared to the structures in the TS at the same time; the procedure selects the structure pair of the TS that maximizes the index. The comparison is repeated until one of two conditions is fulfilled: either the smaller molecule, A or B, has been fully

Table 11. Local Similarity Indexes of Set B

entry	tool set member	virtual similarity index[a]	tool set member	virtual similarity index[a]	local similarity index 1[b]	local similarity index 2[c]	direct similarity index
11-12	32	0.33	32	0.33	0.5017	0.550	0.6103
	27	0.19	27	0.19			
	59	0.03	51	0.03			
11-13	33	0.30	33	0.30	0.4794	0.440	0.5581
	42	0.09	39	0.09			
	23	0.05	23	0.05			
11-14	32	0.24	32	0.24	0.3103	0.270	0.3529
	11	0.02	60	0.02			
	16	0.01	27	0.01			
	0	0.00		0.00			
11-15	32	0.31	32	0.31	0.4841	0.490	0.5553
	27	0.18	27	0.18			
	0	0.00		0.00			
11-16	33	0.36	33	0.36	0.4487	0.470	0.6486
	52	0.11	51	0.11			
	0	0.00		0.00			
11-17	35	0.17	35	0.17	0.3145	0.230	0.1538
	59	0.06	59	0.06			
	0	0.00		0.00			
11-18	32	0.30	32	0.30	0.3600	0.375	0.3721
	59	0.07	59	0.07			
	12	0.005	27	0.005			
	0	0.00		0.00			
11-19	32	0.45	32	0.45	0.4331	0.530	0.6156
	59	0.08	59	0.08			
11-20	32	0.33	32	0.33	0.3769	0.400	0.4500
	59	0.06	59	0.06			
	6	0.01	39	0.01			
11-21	35	0.35	35	0.35	0.5046	0.540	0.5954
	27	0.19	27	0.19			
11-22	27	0.31	27	0.31	0.3141	0.310	0.3889
	0	0.00		0.00			
11-23	32	0.29	32	0.29	0.3696	0.350	0.4545
	59	0.05	59	0.05			
	8	0.01	27	0.01			
11-24	35	0.39	35	0.39	0.3153	0.420	0.4878
	7	0.03	57	0.03			
	0	0.00		0.00			
11-25	32	0.34	32	0.34	0.4905	0.510	0.5451
	27	0.17	27	0.17			
	0	0.00		0.00			
11-26	35	0.29	35	0.29	0.4612	0.440	0.7660
	27	0.15	27	0.15			
11-27	32	0.24	32	0.24	0.3448	0.290	0.3529
	4	0.04	33	0.04			
	16	0.01	27	0.01			
11-28	35	0.38	35	0.38	0.4995	0.550	0.7143
	51	0.17	51	0.17			
11-29	32	0.34	32	0.34	0.3937	0.497	0.5847
	51	0.15	51	0.15			
	23	0.004	27	0.004			
11-30	35	0.38	35	0.38	0.5098	0.560	0.5429
	27	0.18	27	0.18			

[a] VSI values are calculated after connection through the TS. [b] Equation 4. [c] Equation 5.

described through the TS or the last comparison has resulted in an empty result. The final index is obtained by combining the partial indexes. This calculation mode maximizes the comparison between A and B disregarding their best representation on the TS. In this case the same molecule can give different representation depending on its partner.

The use of GSI or LSI has clearly a different interpretation. The global projection of a structure on the TS gives its *complete representation* and thus permits a complete comparison of two compounds, even different in size, connecting all their possible substructures. On the contrary, the local projection, effecting a simultaneous decomposition of the two compounds, gives their *best comparison* on that TS. In this case the result is dependent on both the molecular dimension and the relative similarity.

Clearly, both approaches are strictly related to the representation space, the TS. However, it is also true that a sequence of comparisons made within the same TS is self-consistent.

Finally, we are going to expose the calculations of GSI and LSI. Considering that the bias for the comparison through the representation space has been already introduced, we opted for two ways of index calculation: a geometrical mean and a summation. We will comment on the value of

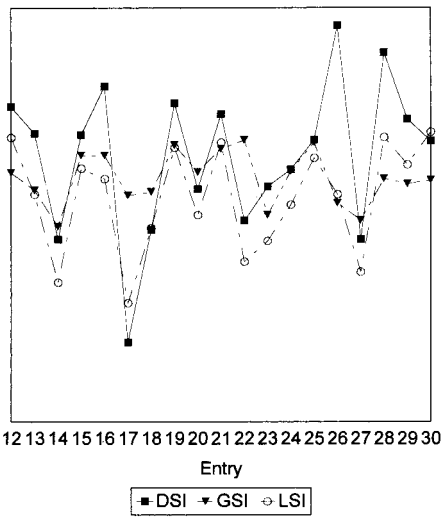

Figure 6. Correlation of similarity indexes (DSI with GSI and LSI) of set B.

the two ways in the Results. We obtained two pairs of equations:

$$\mathrm{GSI} = (\prod_i \mathrm{VSI}_i)^{**}(1/(\sum i_\mathrm{A} \sum i_\mathrm{B})) \quad (2)$$

$$\mathrm{GSI} = ((\sum_i \mathrm{VSI}_{i_\mathrm{A}}/\sum i_\mathrm{A}) + (\sum_i \mathrm{VSI}_{i_\mathrm{B}}/\sum i_\mathrm{B}))/2 \quad (3)$$

$$\mathrm{LSI} = (\prod_i \mathrm{VSI}_i)^{**}(1/(\sum i_\mathrm{A} \sum i_\mathrm{B})) \quad (4)$$

$$\mathrm{LSI} = \sum_i \mathrm{VSI}_i/2 \quad (5)$$

Equations 2 and 4 are mathematically equivalent, but the VSIs are differently calculated. Equations 3 and 5 are not mathematically equivalent, but, because in the case of LSI the VSIs of A and B are equal and $\sum i$[18] in eq 3 is usually equal to 1 or 2, the two calculations are often in the same value range.

RESULTS

The best way to comment on a new methodology is the discussion of some representative results. To this aim, we selected three sets of compounds: (1) set A (Figure 3), containing very simple, easy to follow examples, whose objective is to make the calculation method clear; (2) set B (Figure 4), containing a complete series of complex compounds sufficiently similar to show the power of the method and its differences with respect to a standard similarity measure; (3) set C (Figure 5), containing complex diverse compounds, to discuss the similarity measures obtained using our method.

Set A. The structures of this set are very simple, and their similarity can be ascertained without difficulties. Indeed, they have been chosen in order to make the method presentation clear. When comparing naphthalene with all of the structures of the set (Table 1), including naphthalene itself, we expect distinct results because the set contains structures that are very similar to **1** (**1**, **2**, **8**, **9**), structures that are sufficiently similar (**4**, **6**, **7**, **10**), and structures that are scarcely similar (**3**, **5**). The direct similarity indexes (DSI), calculated with our standard procedure,[19] give the following natural order: **1**, **2**, **8**, **9**, **6** and **7**, **4** and **10**, **5**.

The GSI, calculated using eq 2, already shows some diversity. First, inherent to the method, the self-similarity of naphthalene is not complete. In fact, the comparison is made through projection on the TS and cannot give a result equal to unity unless the examined structure is contained in the TS; nevertheless, the GSI of **1**-**1** is the greatest. On the other hand, the ratio **2**/**1** has increased because benzene is included in the TS. Conversely, the position of **8** has gone down to the sixth. Structure **5** is still the last in the order, even if **6** is now very near, while **7** has moved up to the fourth position. In general, the structures containing complete benzene rings have moved up, giving rise to an order more consistent with the representation.

The LSI, calculated using eq 4, gives a different ordering, too: **1**, **2**, **8** and **9**, **6** and **7**, **4**, **10**, **5**. This order is extremely similar to the natural order, even if the values are different. The search for the best simultaneous comparison enhances the similarity between the compared compounds, while it decreases the goodness of their representation (e.g., compound **6** is best represented by the GSI that uses two references from the TS, 8 and 32).

Indeed, the real improvement of the method is shown by the nonzero value of compound **5**. In this respect, we also made a comparison between structures **6** and **3** that in the direct mode give an index equal to 0. Because compound **3** is very badly described by the TS, the GSI is the lowest; on the contrary, the LSI is greater than the LSI of **1**−**5**, as expected.

We used set A also to analyze the sensitivity of the method to the TS composition. We ran more investigations, alternatively excluding from the TS: or the alkenyl, or the alkyl, or the benzene, or the pyridine, or the pyrrol, or the alkenyl plus the alkyl, or the benzene plus the pyridine, or all the aromatic derivatives (Tables 2−9). The results clearly show that the GSI and the LSI values are affected by the changes, but if there is at least one representation chance, they are still meaningful (compare Tables 2−6 to Tables 7−9).

Set B. Set B includes 20 compounds that have similar biogenetic origin. Consequently, they share some characteristics that permit their comparison. All comparisons were made using compound **11** as the reference. The DSI spans from 0.1538 (**11**−**17**) to 0.7660 (**11**−**23**). The DSI (Table 10) can be roughly considered as similarity percentages; thus, in set B we have DSI in the range 15% < DSI < 77% with good coverage of the range.

The use of the GSI (Table 10) always gives the same representation for **11** (35 and 51 in the TS); therefore we can consider the obtained values as a measure of the ability of the TS to represent each compound. Figure 6 reports the course of the GSI compared with the LSI and the DSI. It is clear that the GSIs are the most different, containing a different aspect of the comparison. We can separate, using eq 3, very interesting groups of structures: (a) GSI > 0.500, i.e., compounds **15**, **16**, **19**, **21**, **22**, and **25**; (b) 0.500 < GSI < 0.450, i.e., **12**, **20**, **24**, **28**, **29**, and **30**; (c) 0.450 < GSI < 0.400, i.e., **13**, **17**, **18**, and **26**; (d) GSI < 0.400, i.e., **14**, **23**, and **27**.

The LSI, calculated by eq 5, behaves differently (Table 11). Its course is much more similar to that of the DSI, as

Table 12. Global Similarity Indexes of Set C

entry	tool set member	virtual similarity index[a]	tool set member	virtual similarity index[a]	global similarity index 1[b]	global similarity index 2[c]	direct similarity index
31-32	35	0.29	32	0.31	0.0082	0.221	0.3333
	40	0.22	32	0.31			
	63	0.20	32	0.31			
	15	0.13	25	0.24			
	1	0.10					
	63	0.09					
	19	0.09					
	63	0.07					
33-34	15	0.32	15	0.32	0.0135	0.217	0.5200
	2	0.19	11	0.25			
	3	0.19	11	0.18			
	5	0.19	63	0.12			
	20	0.19					
35-36	32	0.52	8	0.41	0.0427	0.352	0.1333
	11	0.38	5	0.36			
	11	0.25	5	0.29			
	11	0.25					
36-37	8	0.41	28	0.48	0.0201	0.363	0.2273
	5	0.36	15	0.43			
	5	0.29	16	0.21			
38-39	8	0.29	3	0.25	0.0115	0.213	0.3626
	5	0.22	5	0.25			
	5	0.22	63	0.13			
	63	0.13					
38-40	3	0.25	62	0.27	0.0037	0.193	0.0000
	5	0.25	11	0.19			
	63	0.13	61	0.13			
			15	0.11			
38-41	3	0.25	4	0.29	0.0095	0.232	0.0000
	5	0.25	23	0.27			
	63	0.13	61	0.20			
38-42	3	0.25	62	0.30	0.0042	0.209	0.2667
	5	0.25	14	0.24			
	63	0.13	61	0.21			
			14	0.18			
			63	0.11			
40-42	62	0.27	62	0.30	0.0046	0.192	0.2584
	11	0.19	14	0.24			
	61	0.13	61	0.21			
	15	0.11	14	0.18			
			63	0.11			
41-42	61	0.26	4	0.29	0.0085	0.232	0.1091
	14	0.24	23	0.27			
	14	0.18	61	0.20			
	63	0.16					

[a] VSI values are calculated before connection through the TS. [b] Equation 2. [c] Equation 3.

expected (Figure 6). In fact, the LSI performs the best concerted comparison and, thus, reproduces the overall standard similarity. Because the structures are more complicated, we can observe seven cases where the analysis is unable to furnish a full final result, stopping the search before the number of still unassigned atoms is smaller than the fixed threshold (equal to three atoms). This event happens when the two molecules have pieces that cannot be connected through the TS. In our experience this inability is due to an excessive fragmentation of one of the molecules, fragmentation that is caused by the search for the connection optimization.

Set C. Set C clearly contains compounds that are very different. It is sufficient to look at the values of the DSI to observe that only the comparison of **33** and **34** gives a 52% similarity (Table 12). All of the other comparisons range from 0 to 36%. Set C is thus a good test of the application field of our new similarity measure for two reasons: first, because the structures are different; second, because they span a wide area of organic compounds. It is also clear that other similarity measures could hardly describe such a situation. If we look at Figure 7 we can have an impression

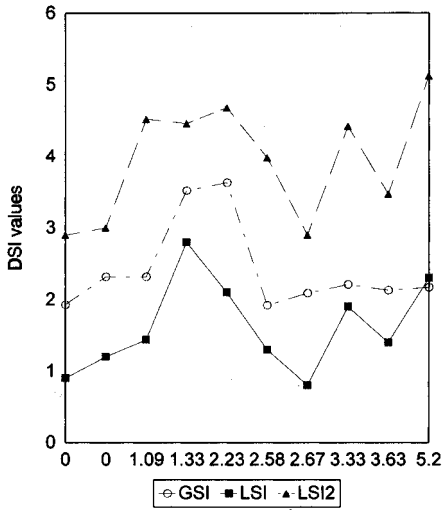

Figure 7. Comparison of DSI with GSI and LSI of set C.

of the diverse behavior that the GSI and the LSI have in comparison with the DSI: no obvious correlation is present. This result was both expected and rewarding; in fact, the

Table 13. Local Similarity Indexes of Set C

entry	tool set member	virtual similarity index[a]	tool set member	virtual similarity index[a]	local similarity index 1[b]	local similarity index 2[c]	direct similarity index
31-32	35	0.09	35	0.09	0.4418	0.190	0.6103
	35	0.06	35	0.06			
	35	0.02	35	0.02			
	35	0.02	35	0.02			
	0	0.00		0.00			
33-34	15	0.10	15	0.10	0.5123	0.230	0.5581
	11	0.04	11	0.04			
	20	0.04	19	0.04			
	3	0.04	3	0.04			
	23	0.01	63	0.01			
35-36	11	0.15	11	0.15	0.4456	0.280	0.3529
	5	0.07	5	0.07			
	23	0.06	23	0.06			
	0	0.00		0.00			
36-37	8	0.09	8	0.09	0.4681	0.210	0.5553
	18	0.06	18	0.06			
	22	0.04	22	0.04			
	16	0.02	23	0.02			
38-39	3	0.06	3	0.06	0.3469	0.140	0.6486
	5	0.06	5	0.06			
	63	0.02	63	0.02			
38-40	4	0.04	3	0.04	0.2901	0.090	0.1538
	23	0.03	5	0.03			
	63	0.01	63	0.01			
	0	0.00		0.00			
38-41	4	0.06	3	0.06	0.2986	0.120	0.3721
	23	0.05	5	0.05			
	61	0.01	63	0.01			
38-42	3	0.05	3	0.05	0.1916	0.080	0.6156
	63	0.03	63	0.03			
	0	0.00		0.00			
40-42	62	0.08	62	0.08	0.3978	0.130	0.4500
	11	0.03	11	0.03			
	15	0.01	14	0.01			
	63	0.01	63	0.01			
41-42	61	0.06	61	0.06	0.4524	0.144	0.1091
	4	0.05	4	0.05			
	23	0.02	14	0.02			
	61	0.01	62	0.01			
	61	0.004	63	0.004			
	0	0.00		0.00			

[a] VSI values are calculated after connection through the TS. [b] Equation 4. [c] Equation 5.

objective of the present approach was not the reproduction of an existing measure but the realization of a method suited for performing structure similarity evaluation in the cases not covered by conventional approaches. Because this set is representative of the real aim of the work, we will accurately examine each entry.

Compounds **31** and **32** are a pair of structures that we already examined in a previous work in order to find a possible explanation of the similar activity that these two very diverse molecules have.[20] The DSI is approximately 30% and results from the common lateral chains. The representation by the TS is good with **31** described by 8 pieces and **32** by 4 (Table 12). Structure **32** is mainly composed by benzene rings, and this results from the description (3 times TS 32 and once TS 25, i.e., all four benzenes), while **31** is described by both aromatic and nonaromatic reference structures (TS 35 and 40, TS 10, 15, and 19, TS 63). The calculated indexes are small because the through TS connections have low values. When calculating the LSI (Table 13), the common description only takes care of the aromatic parts, evidencing the incomplete representation of **31** for what the aliphatic part is concerned. However, the LSI maintains its much greater than zero value, giving the correct impression that, at least for some parts of the compounds, i.e., locally, **31** and **32** are similar.

Compounds **33** and **34** are the most similar structures in set C. They are well-described by the TS, but not all the connections are so good, resulting in a GSI smaller than other GSIs in the set (Table 12). This result emphasizes a similar feeling we got during the examination of the DS analysis. The LSI, as in previous cases, better reproduces the DSI; it is worth noting that the LSI requires a big number of TS structures to give a good description (Table 13).

The comparisons of **36** with **35** and **37** can be discussed together. In this case the DSI result is reversed by the GSI analysis exchanging **35** and **37**; however, the DSI results are near their application limit (10−20%) (Table 12). On the contrary, in set C **35**−**36** and **36**−**37** are the two best correlated pairs. The result of the LSI is again more in line with the DSI, even if the order of the solutions is highly affected by the method (Table 13).

Last, there are five compounds that have been selected in the wide world of antibiotics, i.e., penicillins, macrolids, and azasugars. Their structural diversity is clearly evident and so is the near impossibility of their analysis by standard similarity methods. Consequently, it is interesting to observe

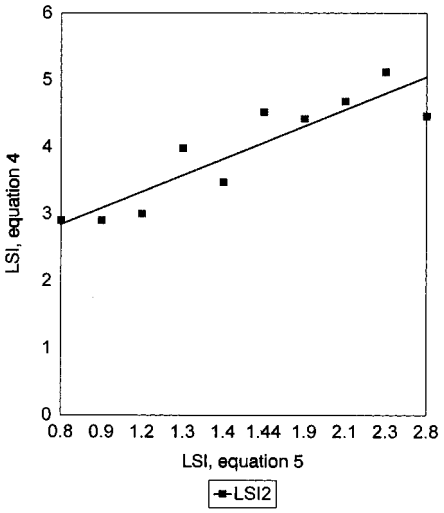

Figure 8. Comparison of LSI of set C calculated by eqs 4 and 5.

the order imposed by GS and LS analyses (Tables 12 and 13). GSIs (eq 2) give **38**−**39**, **38**−**41**, **41**−**42**, **40**−**42**, **38**−**42**, and **38**−**40**. LSIs (eq 4) give **41**−**42**, **40**−**42**, **38**−**39**, **38**−**41**, **38**−**40**, and **38**−**42** (Figure 8). Because here we are comparing diverse structure pairs without using a reference structure, it is hard to decide what order makes more sense. But the whole result is anyway interesting; in fact, we have demonstrated that the method can compare highly diverse structures using a measure that is fully based on similarity analysis.

CONCLUSION

Some final remarks seem worthy. We have presented an original method to calculate the structural similarity between dissimilar compounds. The method is firmly based on the stepwise comparison of connected reference structures, and it can be extended to more efficient transfer of similarity. The potential danger of the stepwise operation is limited by the linearity of the procedure and by the power of the selected similarity measure. We think we have demonstrated that the results are consistent with the premises and that the method is widely applicable. It remains to assess the possible uses of the results; we think that in all the areas where diversity measures have been applied it is possible to adopt either the GSI or the LSI as a more reliable procedure. Moreover, we can think to different modes of application of similarity to dissimilar structures using meaningful values.

ACKNOWLEDGMENT

Partial financial support by the Consiglio Nazionale delle Ricerche (Progetto Strategico "Modellistica Computazionale di Sistemi Molecolari Complessi") and by the Ministero dell'Università e della Ricerca Scientifica e Tecnologica is gratefully acknowledged.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley Interscience: New York, 1990.
(2) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995.
(3) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: London, 1996.
(4) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(5) Judson, P. N. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 148−153.
(6) For some commercial packages using molecular similarity: (a) Grethe, G.; Moock, T. E. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 511−520. (b) Grethe, G.; Hounshell, W. D. In *Chemical Structures 2. Proceedings of the 2nd International Conference*, 1990; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1993; pp 399−407.
(7) Benigni, R.; Andreoli, C.; Giuliani, A. *Environ. Mol. Mutagen.* **1994**, *24*, 208−219.
(8) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(9) McGregor, M. J.; Pallai, P. V. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(10) Lewis, R. A.; Mason, J. S.; McLay, I. M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.
(11) Agrafiotis, D. K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(12) Gillet, V. J.; Willet, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(13) Sello, G.; Termini, M. *Tetrahedron* **1997**, *53*, 3729−3756.
(14) Basak, S. C. 7th Conference on Mathematical Chemistry, Girona, Spain, 1996.
(15) Sello, G.; Termini, M. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: London, 1996; pp 213−242.
(16) The atomic energy is calculated as follows. Remembering the equation correlating the chemical potential with the electronic energy, $\mu = (\partial E/\partial N)_Z$, and considering eq 1 for calculating μ (eq 1: $\mu = -k_1 Z_{star} Z_{star}'/(Z^0{}_{star}' R^0{}_{cov}) + k_2$), we obtain eq 2: $E = k_3{}^*(A + B + C) - k_2{}^* N_3$, where $k_3 = -k_1/(Z^{0*\prime} R^0{}_{cov})$, k_1 and k_2 are constants depending on the atom type, $Z^{0*\prime}$ is the effective nuclear charge of the isolated atom for a complete electronic shielding, $R^0{}_{cov}$ is the atomic covalent radius of the isolated atom, $A = (N^2 + aN - 2NN_1 - 2bNN_2 + N_1{}^2 + 2bN_1N_2 - aN_1 + b^2N_2{}^2 - abN_2)N_3$, $B = 0.5(-2aN + 2aN_1 + 2abN_2 - a^2)N_3{}^2$, $C = (a^{2/3})N_3{}^3$ where *a*, *b*, *c* are Slater's coefficients, *N* is the atomic number, and N_i are the shell occupation numbers.
(17) Mechanism of W calculation is based on a very simple idea: if we think that the electronic energy of a molecule is in relation with the energy of its atoms as result from their reciprocal interactions, we can affirm that the presence/absence of an atom in a specific position determines an energetic variation which can be thought of as the quantification of the difference between that molecule and a hypothetical molecule where that atom is isolated (annihilation principle). Therefore if we consider two identical molecules and we work on atoms in complete correspondences the energy variation will be identical; on the contrary, if we consider two different molecules and we work on similar atoms or if we consider two identical molecules and we work on different atoms, the calculated energy differences will be different and they will represent a measure of the "importance" of the atom in the molecule. Thus, we have available a similarity measure that, comparing the energy variations corresponding to specific substructures, evaluates the degree of their similarity.
(18) $\sum_i$ is the number of compounds of the TS used to describe *A* and *B*.
(19) DSI is calculated by the following equation: $\mathrm{DSI} = \sqrt{\sum \mathrm{SF}_i{}^2}$ with $\mathrm{SF}_i = 2N_i/(A + B)$, where N_i is the number of similar atoms in fragment *i* and *A* and *B* are the total numbers of atoms of molecules A and B.
(20) Sello, G.; Termini, M. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: London, 1996; pp 243−266.

CI980180K