Pergamon

Tetrahedron, Vol. 53, No. 10, pp. 3729-3756, 1997
© 1997 Elsevier Science Ltd
All rights reserved. Printed in Great Britain
0040-4020/97 $17.00 + 0.00

PII: S0040-4020(97)00095-1

Organic Synthesis Planning: Some Hints From Similarity

Guido Sello* and Manuela Termini

Dipartimento di Chimica Organica e Industriale, Universita' degli Studi di Milano,
via Venezian 21, 20133 Milano, Italia.

Abstract: Despite of the present popularity of the new usages of similarity in chemistry, and of its counterpart dissimilarity, few approaches have used it in the design of organic synthesis. In addition most of the known applications refer to the mere comparison between structures of the synthetic tree. We would like to discuss the power of the use of similarity in organic synthesis planning and to give some examples. The possibilities of analysis of a synthesis at different levels, from the single step to the entire tree, are presented and criticised. Special attention is dedicated to the role of similarity, whilst the identification of good descriptors is postponed to future developments.

INTRODUCTION

After an initial long period of "artistic" approaches to organic synthesis the so-called "logic" approaches ruled the scene; this breakthrough of the ancient regime permitted the rationalisation of some fundamental principles of the field. The passage from the "instinctive" feeling of synthesis to a "regulated" planning was characterised by the contemporary development of concepts that tried to give a unitary view of synthesis. However, the use of molecular and plan analogies have been neglected for a long time in favour of other aspects of synthesis design (strategic bond definition, molecular complexity measures, etc.). As a consequence little has been said concerning the use of similarity in synthesis design.

Similarity in Organic Chemistry

In the past similarity has been used as a purely qualitative concept in organic chemistry; thus people used the words "similar reactivity" or "similar geometry" without attaching any particular meaning to them. Recently several works pointed to the quantification and the rationalisation of this concept.[1, 2] The application fields are various and spread over data management as well as quantitative biological activity prediction.[3, 4, 5]

The greatest part of the studies concern the determination of similarity by comparing structural characteristics, then giving different weights to the presence of common parts. Therefore, in developing approaches to structure search in databases it is natural to select structures containing the greatest number of the common parts; whereas in drug design the addition of a special "biological" weight to structural features is a requirement. In both cases the used concepts, and even the algorithms, can often be associated to the same aspect. In other less common cases the descriptors defining the similarity are more differentiated and include information coming from different sources.[6] even not related to structure description. In any case the use of similarity approaches are quite well represented in the general field of organic chemistry.[7, 8, 9, 10, 11]

Similarity in Organic Synthesis Planning

A completely different situation is found in the field of synthesis planning. In fact, searching the literature it is common to run across several references to similarity (e.g. similar transformations, similar synthetic pathways, similar synthetic plans), but it is very rare to find even an approximated description of what similarity means in context. There are obviously some clever exceptions; dividing them into two sections: the first implicitly using similarity as a tool;[12, 13, 14,15,16] the second doing it explicitly.[17, 18, 19, 20]

In order to make the citations in the following clearer let us make some general considerations on synthesis planning. We can look at synthesis planning from two directions: from the target (TGT) to starting materials (SM) (i.e. retrosynthetic approach), or from SMs to the TGT (i.e. forward synthetic approach). Both of them contain a piece of the synthesis space: a synthesis tree. Consequently having an optimal plan means to select (find, get, grasp) the BEST synthesis tree.

Let us divide the problems of synthesis and of similarity (Table 1).

It is clear from Table 1 that similarity can be of great value in simplifying synthetic analysis, not as much in the determination of the alternatives as in their ordering, grouping, and selection. We can point to three areas where similarity can intervene: 1) in strategy; by concurrent use of multiple approaches going to general strategies and tactics; 2) in transformation; by transformation generalisation going to new transform definition; 3) in reactivity interference; by analysis of reaction conditions and functional groups going to prototypical reaction conditions.

As an example of the first point we can cite the unconscious use of similarity present in LHASA.[21] This well-known approach contains at least five main strategies: transform based; structure-goal based; topology based; stereochemistry based; functional group based. Their concurrent use is implicitly suggested by the program and, even if an explicit reference to strategy generalisation is missing, the chemist is nearly automatically taken to compare alternative strategies.

The most common use of similarity in synthetic strategy concerns, however, the structural comparison of the TGT to the SMs where an accurate mapping of the SMs onto the TGT is required together with a method of quantifying the mismatchings. We can find well-known examples in LHASA[22] and in FORWARD.[23] In both programs the use of similarity is implicit only.

Table1. Problems of synthesis and similarity.

SYNTHESIS PROBLEMS		SIMILARITY PROBLEMS	
Strategy	Identification of strategic aspects of the TGT	**Representation**	Structure
	Identification of strategic aspects of the plan		Reaction (transformation)
	Evaluation and sorting and selection		Synthesis branch
			Synthesis tree
Tactics			
For each synthetic step:	Evaluation of the strategic weight against the realisation difficulties	**Comparison**	Structure
	Analysis of the alternative transformations		Reaction (transformation)
	Sorting and selection		Synthesis branch
			Synthesis tree
Refinement			
For each survived synthesis route:	Evaluation of the "fitness" with the "ideal" synthesis (synthetic distance)	**Evaluation**	Synthesis step
	Analysis of the integrity (number of less reliable steps)		Synthesis branch
	Addition of the secondary transformations (protection, FGI, etc.)		Synthesis tree
	Ordering and final selection		
		Selection	Synthesis step
			Synthesis branch
			Synthesis tree

Concerning the second point we can find examples of transform generalisation in many approaches to synthesis planning. In knowledge-bases, e.g. in LHASA[24], or in mechanistic approaches, e.g. EROS[25] or COMPASS.[26] The third point is usually part of the same transform structures, even if it is directly considered only in LHASA.[27] One approach remains to be mentioned that presents a uniform treatment of all the aspects reported, i.e. the SYNCHEM[17] program where a general view of synthesis is used, both for structure and transform description and management, giving rise to a complex expert system. This system uses similarity in many of its activities, but always limiting the analysis to structural generalisation, even in the reactivity field.

For what concerns the explicit use of similarity in synthesis planning we could easily cite all the known approaches because their number is very limited. COSYMA[18] is the only example of strategical similarity, more exactly of "genealogical" similarity in strategy. The idea is to generalise as much as possible the description of both structures and functions and to use the results in comparing strategic pathways. RAIN[28] uses the definition of Minimum Chemical Distance between molecules to order and organise structures in a synthetic tree. This artefact permits the comparison between alternative synthetic pathways just by summing the MCDs of each step, representing at the same time the degree of similarity between structures. Last but not least, WODCA[29] is another example of similarity application in synthesis planning. The approach basically consists of the coding of structure by a code containing many information concerning both the structure skeleton and its functionalisation. Then elaborating the codes WODCA connects very different structures through intermediates at increasing level of similarity, giving at the end a suggestion on the path between the TGT and its SMs. This program also contains some considerations on reaction similarity, but it seems to mainly refer to structure (or substructure) comparison.

Finally we can envisage the use of similarity in the field of combinatorial synthesis planning where the necessity of maximising the molecular diversity, on the one hand, and to limit the number of different reactions, on the other hand, can find an important contribution by the use of the synthetic similarity concept. However, we cannot cite any known experience in this area.

From the above citations we can conclude that in spite of the potential utility of similarity in synthesis planning, few attempts[18,28,29] have been made to introduce it as a novel tool and, moreover, the literature did not report an *a priori* analysis of the real importance of using similarity. In this paper we try to begin a discussion fully devoted to this particular application of similarity.

BACKGROUND

Utility of similarity in synthesis planning

At the beginning we would like to briefly discuss the importance of similarity use in synthesis design, where it can help and what it can be used for. We can distinguish two main uses of similarity: during the analysis of the synthesis of one TGT; when comparing the syntheses of different TGTs. In the first case we can characterise three activities: gathering alternative synthetic pathways; maximising the diversity; weighting the efficiency of each solution. These activities can concern both the structures and the transformations (TSF). In the second case we can imagine two applications: the comparison of the syntheses of similar TGTs; the evaluation of the difficulty level of a synthetic pathway (thus needing the definition of a scale). These applications may need the definition of new and unusual methods for similarity measuring.

Used descriptors

To continue our similarity analysis we will use a number of descriptors obtained from our preceding experience. This choice is by no means the best possible solution but it will make our work easier; however, the heart of the discussion can be easily transferred to any other set of descriptors on condition that they possess a correct meaning in the synthesis field.

The descriptors chosen can be divided into two sets: the first containing variables whose values are a single-valued representative of structure characteristics; the second containing miscellaneous data about similarity and/or synthesis. In the first group we include: molecular globularity,[30] atomic native polarity,[31] similar group interference;[32] the second is formed by sequences of similar atoms,[33] structure similarity indexes, number of synthetic steps, and transform molecularity.

Level of analysis

When combining such descriptors we can determine the object of the similarity analysis, being it the comparison of either structures (targets and products) or transformations (reactivity, reaction conditions). In any case we can consider different levels of comparison; we can compare objects at the same level on the synthetic tree, objects along a branch, objects at different level on the tree, complete trees (Figure 1), and combinations of them.

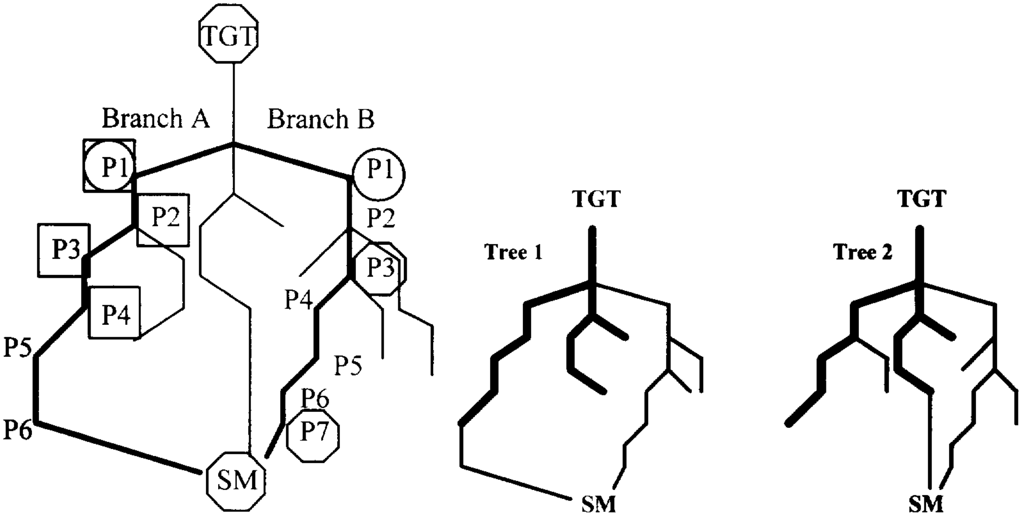


Fig. 1. Different levels of analysis: same level (circles), along branches (squares), jumping (octagons). Entire tree comparison (right half).

FOLLOWING AN EXAMPLE

In order to facilitate the discussion we will follow an example of synthetic analysis using the molecule of Picrotine (Figure 2). This is a medium sized molecule presenting a sufficient level of complexity, thus permitting the analysis of its synthesis.

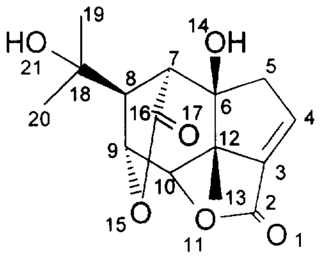


Fig. 2. Structure of Picrotine

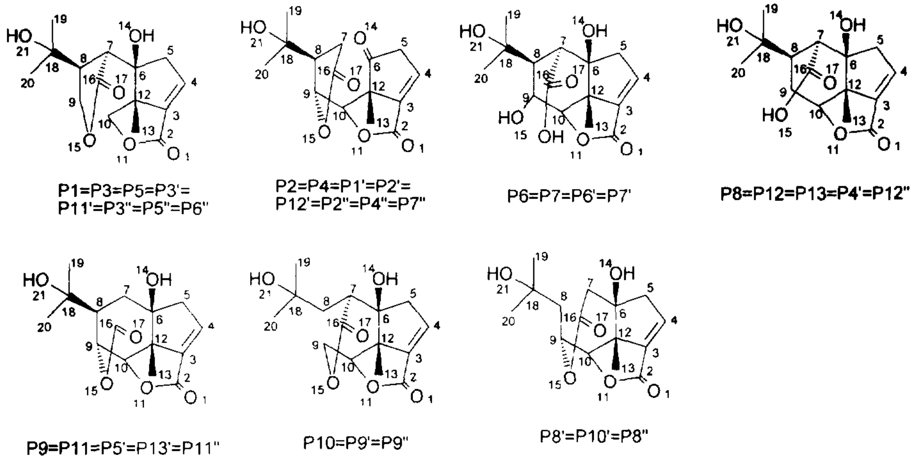
We submitted Picrotine to the most recent version of our program for synthesis planning[34] in order to have a synthesis tree available. The analysis suggests 13 synthetic routes of first order (i.e. giving two separate precursors), each one breaking 2 or 3 bonds. Using our system to enlarge the synthesis space we can change the bond breaking orders so to obtain 23 more solutions. The resulting tree is shown below (Table 2); in principle it contains 82 precursors (we will see in the following that some of them are duplicates).

Table 2. Tree of solutions from the synthesis analysis of Lilith[a].

Number of broken bonds	1	2	1	2	1	2	3
(Ordering)[b]	(1)	(1)	(2)	(2)	(3)	(3)	(1)
	P11	-	P11’	-	-	-	P13a,b
	P21	P22	P21’	P22’	P21”	P22”	P23a,b
	P31	P32	P31’	P32’	P31”	P32”	P33a,b
	P41	P42	P41’	P42’	P41”	P42”	P43a,b
	P51	P52	P51’	P52’	P51”	P52”	P53a,b
TGT	P61	P62	P61’	P62’	P61”	P62”	P63a,b
	P71	P72	P71’	P72’	P71”	P72”	P73a,b
	P81	P82	P81’	P82’	P81”	P82”	P83a,b
	P91	P92	P91’	P92’	P91”	P92”	P93a,b
	P101	-	P101’	-	-	-	P103a,b
	P111	P112	P111’	P112’	P111”	P112”	P113a,b
	P121	P122	P121’	P122’	P121”	P122”	P123a,b
	P131	-	P131’	-	-	-	P133a,b

[a] The last number of the precursor names corresponds to the number of broken bonds. Precursors are reported in Figures 5-10. [b] The number in parentheses corresponds to the order of bond breaks.

Product structures are reported in Figures 3, 4, 5, 6, 7, and 8. Because the structure generation is partly due to changes in the order of bond breaks, several structures are identical and only 44 original compounds remain to be examined. The tree levels are only 3 in agreement with the maximum number of bond breaks; the alternative routes are 13, two of them are two-level deep; the number of bonds affected is 11 on a total of 24 heavy atom - heavy atom bonds present in the structure. The modifications of the structures following a bond break have been restricted as much as possible in order to avoid the possible bias coming from different choices of special transformation (e.g. activating or leaving groups).



P1=P3=P5=P3'=P11'=P3"=P5"=P6"

P2=P4=P1'=P2'=P12'=P2"=P4"=P7"

P6=P7=P6'=P7'

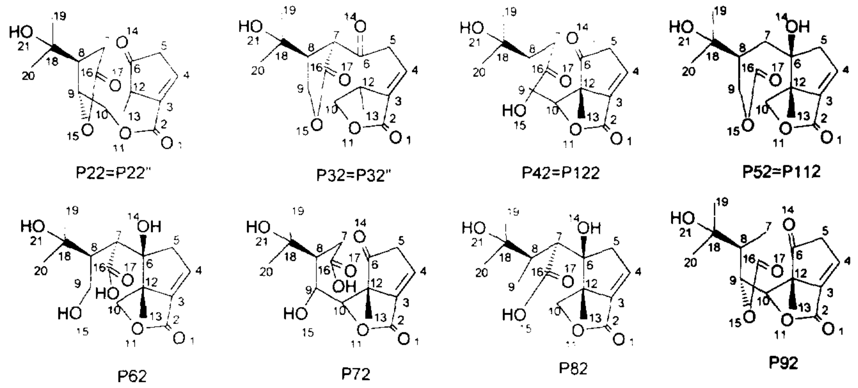
P8=P12=P13=P4'=P12"

P9=P11=P5'=P13'=P11"

P10=P9'=P9"

P8'=P10'=P8"

Fig. 3. Precursors of Picrotine at the first level



P22=P22"

P32=P32"

P42=P122

P52=P112

P62

P72

P82

P92

Fig. 4. Precursors of Picrotine at the second level

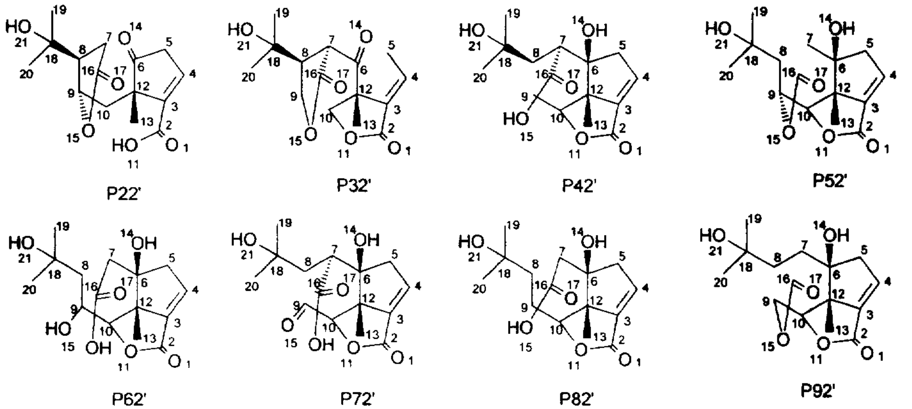


Fig. 5. Precursors of Picrotine at the second level

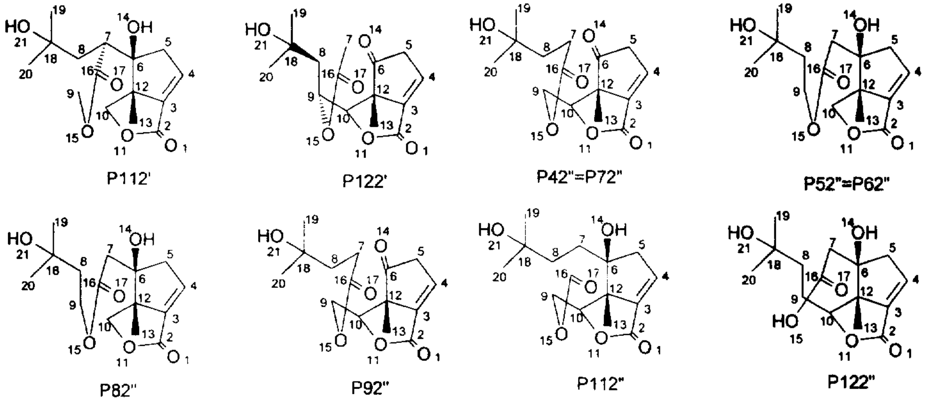


Fig. 6. Precursors of Picrotine at the second level

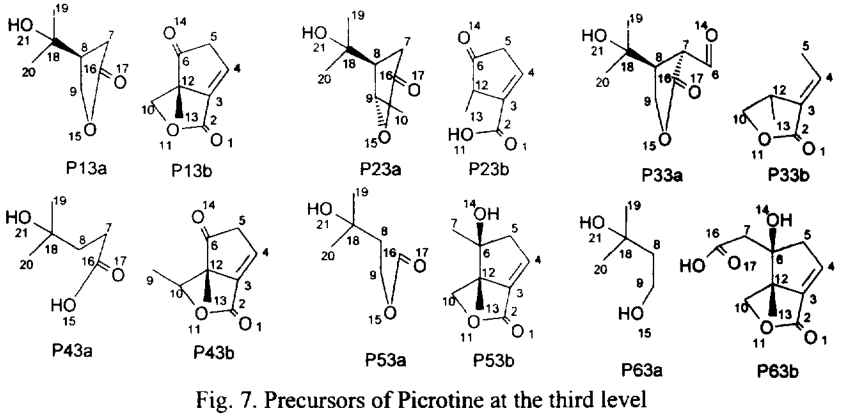


Fig. 7. Precursors of Picrotine at the third level

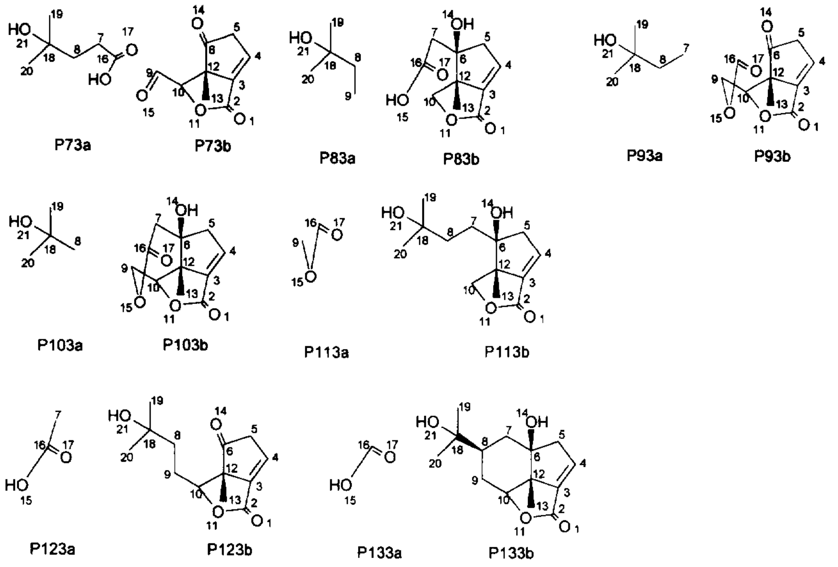


Fig. 8. Precursors of Picrotine at the third level

Same level comparisons

Considering our example we can compare structures belonging to the same level of the tree taking care or disregarding their origin; i.e. we can differentiate precursors obtained inside each bond break order or ignore the origin of all the compounds. Let us now comment on some descriptors. (Table 3, 4, 5)

Molecular globularity. It is an indirect measure of the topological complexity of a compound. It is calculated dividing the maximum complexity distance by the total molecular complexity; as a consequence the smaller is the globularity the less distributed is the molecular complexity. In the analysis of a synthetic route we expect that globularity increases going down a branch according to the simplification of the structure. In our example the TGT globularity is 0.524, for the compounds of the first level is within 0.524 and 0.595, and for the compounds of the second level is within 0.595 and 0.738. This course, in agreement with expectations, allows a preliminary selection of simplifying steps: all precursors showing an increase of globularity are simplifying. Then it is possible to group some precursors together just on the globularity base. Thus, as shown in Figure 9, both identical structures (e.g. **P1**, **P3**, **P5**) and different structures (e.g. **P1** and **P2**) are collected together; the same operation can be done on the whole tree. It is interesting to note that some first level products are not simpler than the TGT (e.g. **P6**, **P4'**, **P11"**), as well as some compounds of the second level with respect to some of the first level (e.g. **P1**, **P1'**, **P2"**, and **P6**, **P11'**). We can also use globularity to compare the overall simplification of a branch; for example taking the globularity difference between end products we can grasp the similarity, measured as relative complexity, between SMs and consequently the efficiency of the synthesis branch, measured as convergence. For example **P13** and **P23** are the best routes whereas **P133** is the worst.

Table 3. Descriptors calculated by Lilith for breaking of the first bond[a].

Solution	Globularity	Bond	Native Polarity	Native Polarity	Interference	Interference
			Atom1	Atom2	Atom1	Atom2
1	0.595	6-7	8.61	-10.29	64	0
	0.172	10-9	-4.01	-4.02	0	1
2	0.595	10-12	-2.90	-1.42	16	0
	0.643	10-11	3.36	-15.70	243	81
	0.011	6-7	8.63	-10.37	1024	16
3	0.595	6-12	8.61	-1.50	3	16
	0.643	6-5	9.00	-9.17	324	1
	0.011	10-9	-4.02	-4.03	16	0
4	0.595	7-6	-10.29	8.61	0	64
	0.643	9-15	3.32	-16.00	3	16
	0.304	9-8	0.00	0.00	0	0
5	0.595	9-10	-4.02	-4.02	81	0
	0.643	7-16	-3.76	5.75	0	0
	0.304	7-8	-0.01	0.01	16	3
6	0.595	9-10	-4.02	-4.02	81	0
	0.595	15-16	-11.57	10.54	16	0
	0.268	7-8	-9.18	-0.21	16	3
7	0.524	16-15	10.44	-11.52	0	16
	0.643	7-6	-10.63	8.64	0	64
	0.304	9-8	8.74	-1.23	16	3
8	0.595	9-10	-4.02	-4.02	1	0
	0.595	9-15	3.34	-16.14	0	16
	0.268	7-8	-15.44	-15.93	16	1
9	0.595	7-6	-10.29	8.61	0	64
	0.595	7-16	-3.69	5.68	1	0
	0.268	9-8	-2.65	-1.38	0	3
10	0.571	7-8	-8.83	-0.26	1	0
	0.404	9-8	-2.65	-1.37	16	0
11	0.524	7-16	-3.69	5.67	1	0
	0.643	9-10	-4.03	-4.02	1	0
	0.306	9-8	-2.65	-1.38	0	3
12	0.524	9-15	3.32	-15.99	0	16
	0.643	7-6	-10.40	8.60	0	64
	0.306	7-8	-8.85	-0.26	0	0
13	0.524	9-15	3.32	-15.99	0	16
	0.550	7-16	-3.60	13.12	1	4

[a] Atom numbers are reported on the corresponding Figures.

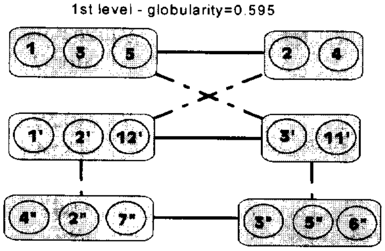


Fig. 9. Molecular globularity correlation at the same level on different trees

Table 4. Descriptors calculated by Lilith for breaking of the second bond[a].

Solution	Globularity	Bond	Native Polarity	Native Polarity	Interference	Interference
			Atom1	Atom2	Atom1	Atom2
1	0.595	10-9	-4.01	-4.02	81	0
	0.172	6-7	8.63	-10.37	324	1
2	0.595	6-7	8.61	-10.29	64	0
	0.643	10-11	3.33	-15.54	243	16
	0.011	10-12	-0.46	-9.21	243	16
3	0.595	10-9	-4.02	-4.02	81	0
	0.667	6-5	9.00	-9.16	48	81
	0.011	6-12	8.63	-1.48	1024	0
4	0.524	9-15	3.32	-15.99	3	16
	0.690	9-8	0.00	0.00	0	0
	0.304	7-6	-10.39	8.64	0	64
5	0.524	7-16	-3.69	5.67	0	0
	0.690	7-8	-0.02	0.02	0	0
	0.304	9-10	-4.02	-4.02	0	16
6	0.595	9-10	-4.02	-4.02	81	0
	0.738	7-8	-8.92	-0.25	81	0
	0.268	15-16	-11.56	10.54	1	0
7	0.524	16-15	10.44	-11.52	0	16
	0.690	9-8	8.72	-1.21	64	256
	0.304	7-6	-10.63	8.67	1	324
8	0.571	7-8	-8.83	-0.26	1	0
	0.738	9-10	-4.02	-4.02	1	0
	0.268	9-15	3.32	-16.00	3	1
9	0.571	9-8	-2.65	-1.37	81	0
	0.738	7-6	-10.37	8.63	81	324
	0.268	7-16	-3.69	5.68	1	0
10	0.571	9-8	-2.65	-1.37	16	0
	0.404	7-8	-8.91	-0.26	16	0
11	0.595	9-10	-4.02	-4.02	1	0
	0.595	9-8	-2.65	-1.38	3	256
	0.306	7-16	-3.81	5.66	81	0
12	0.595	7-6	-10.29	8.61	0	64
	0.595	7-8	-8.84	-0.25	0	0
	0.306	9-15	3.32	-16.01	3	0
13	0.524	7-16	-3.69	5.67	1	0
	0.550	9-15	3.33	-16.06	3	0

[a] Atom numbers are reported on the corresponding Figures.

Table 5. Descriptors calculated by Lilith for breaking of the third bond[a].

Solution	Globularity	Bond	Native Polarity Atom1	Native Polarity Atom2	Interference Atom1	Interference Atom2
2	0.595	6-7	8.61	-10.29	64	0
	0.690	10-12	-3.10	-10.59	48	16
	0.011	10-11	3.33	-15.54	3	16
3	0.595	10-9	-4.02	-4.02	81	0
	0.690	6-12	8.63	-1.47	3	16
	0.011	6-5	9.00	-9.17	1024	81
4	0.595	7-6	-10.29	8.61	0	64
	0.738	9-8	-2.65	-1.37	16	0
	0.304	9-15	3.33	-16.01	3	1
5	0.595	9-10	-4.02	-4.02	81	0
	0.738	7-8	-8.92	-0.25	81	0
	0.304	7-16	-3.76	5.75	16	0
6	0.524	15-16	-11.52	10.44	16	0
	0.738	7-8	-9.18	-0.22	1	3
	0.268	9-10	7.35	-3.86	64	16
7	0.595	7-6	-10.29	8.61	0	64
	0.738	9-8	-2.65	-1.37	1	0
	0.304	16-15	7.94	-11.47	0	1
8	0.571	7-8	-8.83	-0.26	1	0
	0.738	9-15	3.32	-16.00	3	16
	0.268	9-10	-1.37	-2.66	3	16
9	0.571	9-8	-2.65	-1.37	81	0
	0.738	7-16	-3.76	5.75	0	0
	0.268	7-6	-1.45	8.89	1	324
11	0.524	7-16	-3.69	5.67	1	0
	0.738	9-8	-2.65	-1.38	16	0
	0.306	9-10	-4.02	-4.03	48	81
12	0.524	9-15	3.32	-15.99	0	16
	0.738	7-8	-8.94	-0.27	1	3
	0.306	7-6	-10.40	8.60	0	64

[a] Atom numbers are reported on the corresponding Figures.

Atom sequence. We have two different topological methods to calculate atom similarity in sequences; they both are calculated using electronic energy, but whilst the first selects similar atoms following bond paths without other restrictions, the second operates differently and selects atom sequences considering as discriminant energy trends (increase or decrease of energy) and substitution degrees. The results are different and we can, in principle, use both methods, but, for the sake of simplicity, we will comment only the first one. From the number of similar atoms we calculate a similarity index that is simply a normalisation of the absolute numbers and is obtained by the following equation (if we are going to analyse the synthesis of only one TGT the factor containing A and B is constant): **SF = N x (A + B) / AxB**, where N is the number of similar atoms, A and B are the numbers of significant atoms of molecule A and B. In this study SF varies in the range 0.67-2.0 corresponding to 7-21 atoms. Remembering the uses of similarity in synthesis design we can either group the most diverse compounds (e.g. **P2-P6**, **P32-P112'** (Figure 10), **P32-P52"**), or the most similar compounds (e.g.

P1-P9', P8'-P9', P9-P9', P32-P32', P112'-P52", P52'-P112"). Looking at the corresponding structures we note that some of these results are unexpected (e.g. **P32-P32'** (Figure 11)).[35]

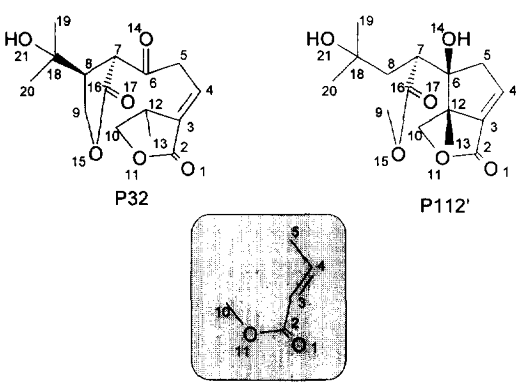


Fig. 10. Atom sequence similarity at the same level

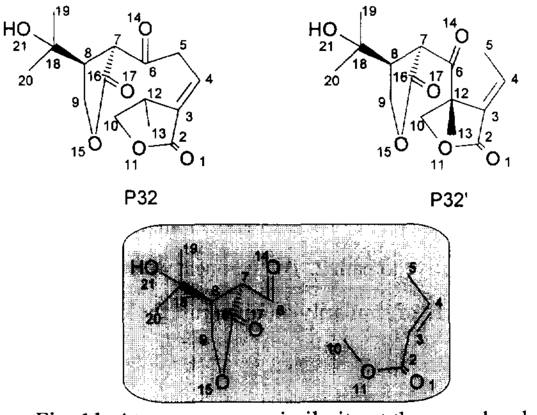


Fig. 11. Atom sequence similarity at the same level

Atomic native polarity. From a different viewpoint we can consider similarity of transformations. This can be done either studying reaction conditions[36, 37] or structural differences in compounds connected through a synthetic path. This last option is more meaningful during the activity of synthesis planning when the reactions that will be applied are still indefinite, whilst the reactivity characteristics of compounds are already defined.

Again we can choose different objects from our descriptors. Atomic native polarities (NP) are representative of the reactivity naturally present on the compound because they don't depend on the reaction effectively applied but on the electronic state of each bond that can be affected by the transformation.

Let now consider bonds that, at the same level of the synthesis, have similar NPs. For example bonds 6-7, 6-5, and 16-15, have well-defined NPs in all cases for both atoms of the bond. On the contrary, bonds 9-10 and 10-12, that have both contrasting NPs, show different dependence on solutions and levels. Bond 9-10 inverts NP only when broken at the third step in solution 6; bond 10-12, on the other hand, even if always showing contrasting NP changes definition when broken at the second or third step. Looking at the corresponding structures we can note that bonds 6-7 and 6-5 are fully comparable, thus their similarity is expected;[38] on the contrary bond 15-16 is a completely different example of reaction and its similarity to 6-7 or 6-5 is only related to the high definition level of the NPs thus suggesting a comparable ease of formation and a mechanism not too dissimilar. Concerning bonds 9-10 and 10-12 it is clear that if individually broken they are very similar (α or β to hydroxyl groups), but 10-12 becomes an α carbonyl atom if either bond 6-7 or 6-5 has been broken in advance. In this last case the transformation of 9-10 is no more comparable to that of 10-12.

Similar group interference. Besides NP we can consider the possibilities of interference that a particular TSF could experiment when applied to a compound. Also in this case the level of generalisation that the descriptor can reach is very important in order to permit a similarity analysis. SGI is an attribute of the atomic reactivity and thus its use is promising. Let consider again bonds 6-7 and 6-5. Their SGI levels are very similar (1024/16 and 1024/81) in solution 2 (third step) and solution 3 (third step), but quite dissimilar in solution 4 (third step, 64/0). To explain this result we must follow all the TSFs, because for what concerns the SGIs there is a relapse connected to the necessity of activation in the preceding steps (compare Figures 6 and 12). Without going into details it is however clear that both the anion formation on 7 or 5 and the condensation on 6 feel the presence of more carbonyl groups on the molecule. For what concerns similarity the discussion is more complicated; we can suppose that the application of TSFs 23 or 33 will require more attention to interferences than TSF 43, i.e. TSFs 23 and 33 are more similar thus it is probable that synthetic path 4 could be an alternative to path 2. To be honest we recognise that it is more difficult to understand, and consequently use, this type of similarity, but it is just the kind of new hints we are searching for.

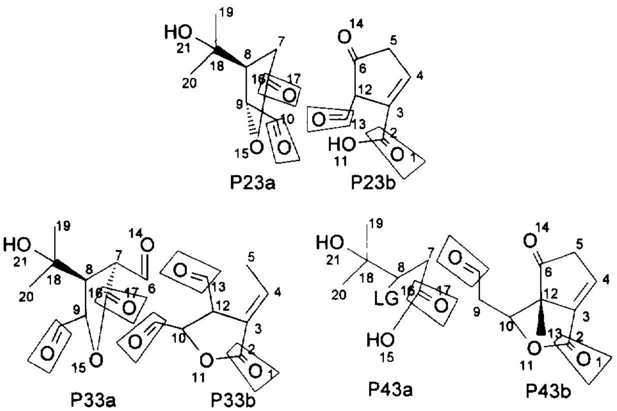


Fig. 12. Similar group interferences at the same level

Along a branch comparisons

The second possibility is the comparison of objects following a tree branch; this is equivalent to analysing short synthetic paths and can potentially give suggestions about alternative syntheses or synthetic shortcuts.

Molecular globularity. We consider one part of the tree from the viewpoint of molecular globularity (Figure 13). To emphasise the differences with the previous section let start from an *ad hoc* example where the comparison at the same level is not as informative. Considering solutions 2, 3, and 4, and their ordering alternatives, we observe that at the first level all the solutions (4' excluded) have the same globularity, thus they are very similar. Going down the branches the situation changes showing four different groups at the second level and two groups at the third level. It is clear that, along the branch, solutions 2, 2", and 3, 3", retain their similarity; 3' is also quite similar; 4 and 4" are dissimilar with respect of both 2 and 3, and of themselves; 2' is a special case because it is sufficiently dissimilar at the second level. In conclusion we can affirm that solutions 2 and 3 can be defined similar alternatives, whilst solution 4 is diverse. Looking at the corresponding structures this result is evidently justified.

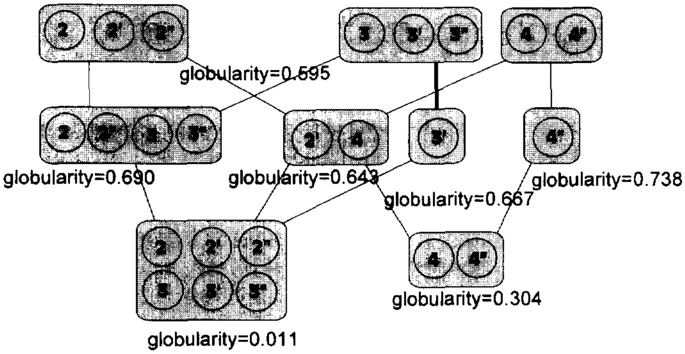


Fig. 13. Molecular globularity correlation along a branch

Atom sequence. Atom sequences can also be used to compare tree branches. For example we can compare solutions 2, 7, and 9, and their precursors 22, 23, 72, 73, 92, 93 (Figure 14). Looking at the similarity indexes we find the sequences: 1.52 / 1.52 / 0.74, 1.62 / 1.14 / 0.87, 2.0 / 1.43 / 1.26, for comparisons with the TGT; **P2** / **P7** = 1.05, **P2** / **P9** = 1.52, **P7** / **P9** = 1.52.

The hints that we can derive are the following: 1) solution 2 has a non-simplifying step (step 2); solution 7 shows a constant decrease in similarity along the branch thus all the steps are important in this sense; solution 9 remains similar to the TGT in all the steps, it is thus strategically near to the TGT. If we add that structure 2 is similar to structure 9 we can conclude that solution 9 is not an alternative to solution 2 as can be solution 7.

Atomic native polarity. Let consider one bond interested by the synthetic plan; it can be broken in the first, or in the second, or in the third step, so its NP can be always equal or can change. As a consequence we can speak about similar or dissimilar transformations. For example, bond 6-7, that is broken in solutions no. 1, 2, 4, 7, 9 , 12, always keeps the same NP. On the contrary, bond 8-9, that is broken in solutions no. 4, 7, 9, 10, 11, changes its NP in solutions 4 and 7 if broken at the second or third steps. The immediate information we can realise is that all the TSFs touching bond 6-7 are very similar (in fact, they are all aldol type reactions), whilst the character of the TSFs interesting bond 8-9 depends on the branch level of application (in fact a

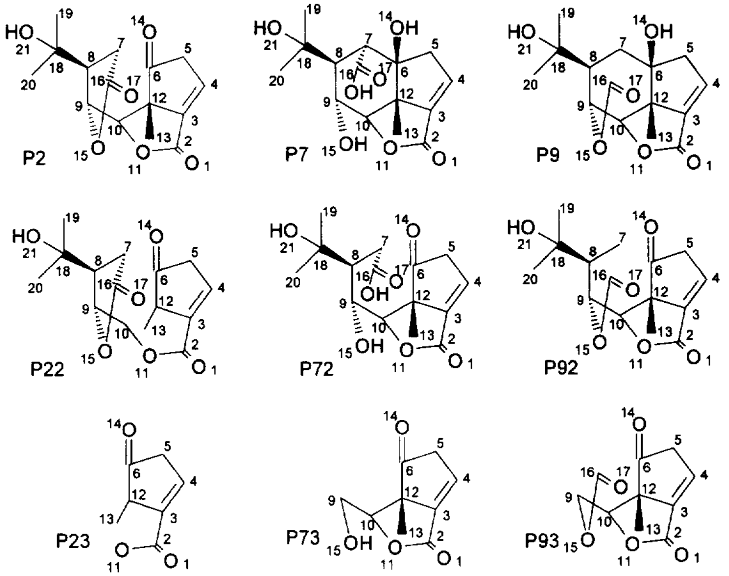


Fig. 14. Atom sequence similarity along a branch

powerful reacting group is missing on those atoms). Besides, we can also suppose that this result suggests a limited influence of the bond breaking order on 6-7 TSFs.

Similar group interference. Continuing our example we can be curious to know if bond 6-7 is strongly determined also for what concerns the SGIs. Looking at the values we find two cases with different SGIs (solutions 1 and 9) when 6-7 is broken during the second step (but for solution 1 at this step we separate the structure into two parts) and three cases (solutions 2, 7, and 9) at the third step. Considering bond 8-9 we can locate four cases (solutions 4, 7, 10, and 11) at the second step and four cases (solutions 4, 7, 9, and 11) at the third step. Therefore we can conclude that also the SGI suggests a more constant behaviour for TSFs applied to bond 6-7 than to bond 8-9. For the sake of completeness, we would like to point to bonds 6-5 and 6-12, both generating a carbonyl group on atom 6 as bond 6-7. In these cases the NP remains constant throughout the steps and the orders, as expected; on the contrary, the SGI changes very often (four cases on five). This result shows that we must be very accurate when speaking about similarity among TSFs because they depend on many factors and their comparison can be risky if not done with care.

Jumping on different levels

A third possibility is represented by comparisons at different levels of the tree. In some cases such an approach can give useful information concerning possible shortcuts or fundamental alternatives that, even if present on the tree, are not directly detectable because they are not contiguous and thus difficult to locate. The value of this type of analysis is its originality and it can give unexpected benefits.

Molecular globularity. On a short tree like our example it is not easy to make many jumps having only three levels available. Anyway, it could be interesting to compare the globularity of the TGT to that of the structures at the second level considering, at the same time, the changes occurred at the first level. TGT globularity is equal to 0.524, consequently the most different precursors (with g = 0.738) are **P62’**, **P82’**, **P92’**, and **P42”**, **P52”**, **P62”**, **P72”**, **P82”**, **P92”**, **P112”**, **P122”**. The heavy presence of precursors coming from the third ordering of the bond breaks is expected because the first order of bond breaks is mainly based on breaking last the most central bond. More interesting are solutions 6, 8, and 9, that score at g = 0.738 in two orderings and at g = 0.595 in the other; or even solution 5 that scores at g = 0.643, 0.690, 0.738. From these values we can predict the maximal dissimilarity with the TGT. But looking at the first level calculations (using again the TGT as a probe) we can conclude that it is the break of the 6-7, 7-8, 8-9, and 9-10, bonds that makes the difference.

Atom sequence. More informative can be, in this case, the use of atom sequences. In fact, knowing both the similarity indexes and the sequences themselves we can obtain more hints on our synthetic tree. Let look at the data concerning the compounds at the second level compared to the TGT. We find: **P62**, **P52’**, **P92’**, and **P112”**, at 1.90; **P82**, **P82’**, **P112’**, and **P52”**, at 1.81; these represent the most similar precursors, the nearest to the TGT, and are potentially the least effective. On the contrary, **P122** at 1.05 is the most dissimilar precursor and the most promising (Figure 15).

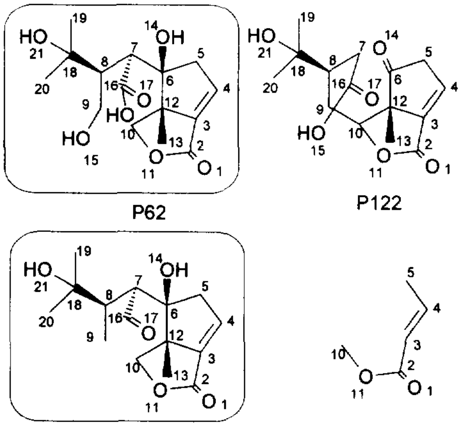


Fig. 15. Atom sequence similarity jumping on levels

This result is not visible at the first level where **P2** (1.52), **P12** (1.81), and **P6** (1.62), have very similar index values.[39] Another suggestion that can be derived is that, to try a synthesis of the TGT, it can be sufficient to analyse either solution 2 or 12.

Atomic native polarity. Because native polarity is a pointer to reactivity it represents, in our discussion, TSF similarity. One of the most important problem of NP is the presence on the same bond of atomic polarities of the same sign, i.e. where there is the necessity of umpolung of one atom. In the example we can find 4 bonds presenting this characteristic, bonds 9-10, 9-8, 7-8, and 10-12. They cleanly group in the majority of cases (solutions 1, 2, 3, 5, 6, 8, 9, 10, 11, and 12); but there are two exceptions: 1) bonds 9-8 and 7-8 become contrasting in solutions 4 (before breaking bond 9-16) and 5 (before breaking bond 7-15) only when broken at the second step; 2) bonds 9-8 and 9-10 loose their contrast in solutions 6 and 7 when broken after bond 15-16. The first exception is thus due to the presence of the ester group, the second to its absence, i.e. the ester group influences the polarity of the atoms adjacent to it in the sense of causing a contrasting polarity case. It is worth noting that this conclusion has been reached without explicitly considering the ester group.

Similar group interference. A possible application of SGI would be the location of the TSFs most sensitive to the molecular environment. Looking at the values we find at the very top of the list all the bond breaks involving atom 6 (i.e. bonds 6-7, 6-5, and 6-12, in solutions 1, 2, 3, 7, and 9) with particular emphasis when they appear at the last synthetic step. This result is in agreement with the characteristic of the TSFs involved because they all concern the reaction of a carbonyl group. The only two exceptions are solutions 4 and 12 where the SGI levels are much more similar to the other bond breaks. In this case too the result has its own logic: when breaking bond 6-7 in the presence of ester group 10-16-17 the SGI level is high. Solution 7 is a particular case among the regular cases; in fact in this solution the ester group is absent when breaking bond 6-7, but a new carbonyl group is here appeared (see Figure 7). Has this result the meaning of limiting the use of the solutions containing atom 6? The answer is absolutely negative, because other factors participate to the decision (e.g. definition of polarities, strategical weight); however, the choice of the bond break order can be influenced by the high level of interference shown by atom 6.

Complete tree comparisons

We can extend the analysis to the comparison of entire synthetic trees. It is clear that this operation requires the simultaneous consideration of many aspects and thus could be more easily done using an automatic methodology. The comparison of complete trees using many descriptors can represent the last and most valuable use of similarity, perhaps permitting the choice of the most promising syntheses. In our example we can compare only trees generated by order permutation; as a consequence the comparison will be biased by their inherent similarity.

Molecular globularity. The molecular globularity decreases going down the tree and, in principle, we can expect to select those tree branches with the greatest decrease and thus the best simplification. Let us compare the trees generated by the permutations, grouping together solutions with the same final level of simplification.

Solutions 1, 10, and 13, are present on two trees only and are consistently similar, being all dissimilar within them. Solutions 2 and 3 are always similar and represent two similar branches of all the trees. Solutions 8 and 9 are similar on two trees (the second and the third). Solutions 4, 5, and 7, solution 6, and solutions 11 and 12, show different similarity depending on the level, therefore cannot be considered comparable. Adding all the results together the three trees are similar for branches 1, 2, 3, 10, and 13; two of them also for branches 8 and 9 (Table 6).

Table 6. Tree comparison: corresponding branches.

Tree	1	1	2	2	3	3	1,2,3
TGT	P11	-	P11’	-	-	-	P13a,b
	P21	P22	P21’	P22’	P21”	P22”	P23a,b
	P31	P32	P31’	P32’	P31”	P32”	P33a,b
	P41	P42	P41’	P42’	P41”	P42”	P43a,b
	P51	P52	P51’	P52’	P51”	P52”	P53a,b
	P61	P62	P61’	P62’	P61”	P62”	P63a,b
	P71	P72	P71’	P72’	P71”	P72”	P73a,b
	P81	P82	P81’	P82’	P81”	P82”	P83a,b
	P91	P92	P91’	P92’	P91”	P92”	P93a,b
	P101	-	P101’	-	-	-	P103a,b
	P111	P112	P111’	P112’	P111”	P112”	P113a,b
	P121	P122	P121’	P122’	P121”	P122”	P123a,b
	P131	-	P131’	-	-	-	P133a,b

Atom sequence. To compare complete trees we choose again the TGT as the reference state. The analysis goes through the same steps as in the preceding sections using the SF index as a measure of the relative similarity. When the TGT separates into two pieces all the solutions in all the trees are obviously coincident, thus we have to check two levels only. The result is the similarity between branches corresponding to solutions 3, 5, 7, 11, and 13. It is noteworthy that these branches, excluding solution 3 and 13, are different from those obtained using the molecular globularity as a measure; solution 3 is a special case because the trees coincide completely with the exclusion of **P32”**. This is an artefact of our algorithm for reordering bond breaks that could generate the same solution more than once if that solution is favourite. To give an idea of the similarity between branches two examples, solutions 5 and 11, are reported in Figures 16 and 17.

Atomic native polarity. The comparison of native polarities of complete trees presents some more difficulties, as we will also see in the SGI case. In fact, it is very uncommon that the same sequence of TSFs is used in two different trees, and, as a consequence, the NPs will be not easily analysed. One possibility is offered

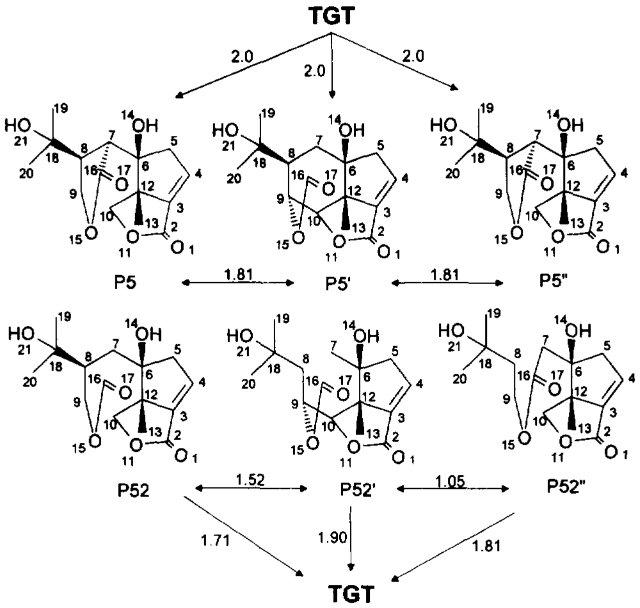


Fig. 16. Atom sequence similarity at entire tree level

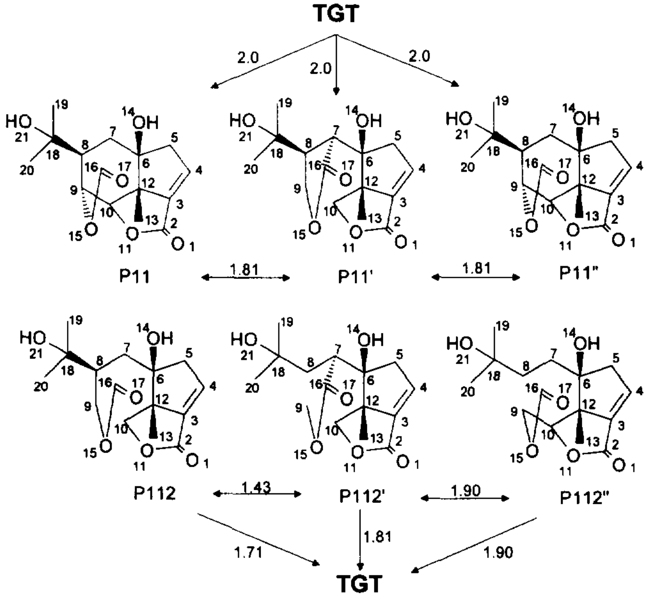


Fig. 17. Atom sequence similarity at entire tree level

by the comparison of their steadiness, i.e. by the number of changes the NP of each atom involved in each TSF suffers throughout the synthesis. The depth of the discussion can site at different levels; we will keep it as general as possible. Being our trees made by different ordering of the same strategic solutions the hypothetical differences are limited. In fact, solutions 1, 3, 9, 10, 11, 12, and 13, always show the same NPs; solutions 2, 6, 7, and 8, have one inversion; solutions 4, and 5, have two inversions. The differences in NP are:

- solution 2 - in **P23'** the contrasting polarity of bond 10-12 (minus on both atoms) is solved with the minus on atom 12;
- solution 6 - in **P63"** the contrasting polarity of bond 9-10 (minus on both atoms) is solved with the plus on atom 9;
- solution 8 - in **P83** the defined polarity of bond 7-8 (minus on atom 7) becomes a contrasting polarity (minus on both atoms);
- solution 7 - in **P72"** the defined polarity of bond 9-8 (plus on atom 9) becomes a contrasting polarity (minus on both atoms);
- solution 4 - in **P42"** the undefined polarity of bond 9-8 (both zero) becomes a contrasting polarity (minus on both atoms);
- solution 5 - in **P52"** the undefined polarity of bond 7-8 (both zero) becomes a defined polarity with the minus on atom 7.

We can thus conclude that the native polarities are very similar in all those trees, i.e. reordering of bond breaks does not affect too much the TSFs involved in the syntheses.

Similar group interference. SGIs can be treated in the same way as NPs. The principal difference is the greater sensitivity of this factor, interference, to structure changes. Again the result, even if less clear, is limited by the overall similarity of the trees. We will speak of SGI changes when the value relative to a particular atom changes of at least one order of magnitude. In this view we can note: solutions 8, 10, and 12, show 0 change; solutions 4 and 13, show 1 change; solutions 1 and 6, show 2 changes; solutions 7 and 9, show 3 changes; solutions 5 and 11, show 4 changes; solutions 3 shows 5 changes; solution 2 shows 6 changes. Adding all changes together we arrive at a total of 31 changes that, considered the number of bond breaks, is a small quantity. We can thus affirm again that this synthesis analysis suggests quite similar trees for what TSFs are concerned.

COMPARING TWO TARGETS

We would like to add one more example of application of similarity to synthesis design. This application concerns the comparison of the syntheses of two targets that are evidently correlated: Gibberellic and Antheridic acids (Figure 18). The synthetic pathways derived from the literature[40] are reported, in short form, in Figures 19, 20, and 21, and we would restrict our analysis to few aspects, only.

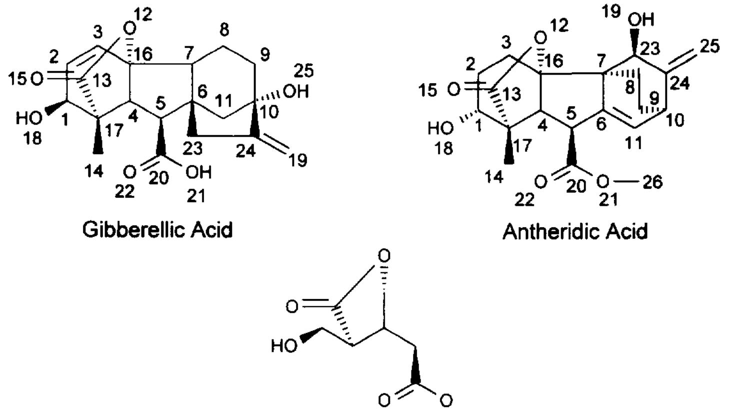


Fig. 18. Structures of Antheridic and Gibberellic acids and their common part

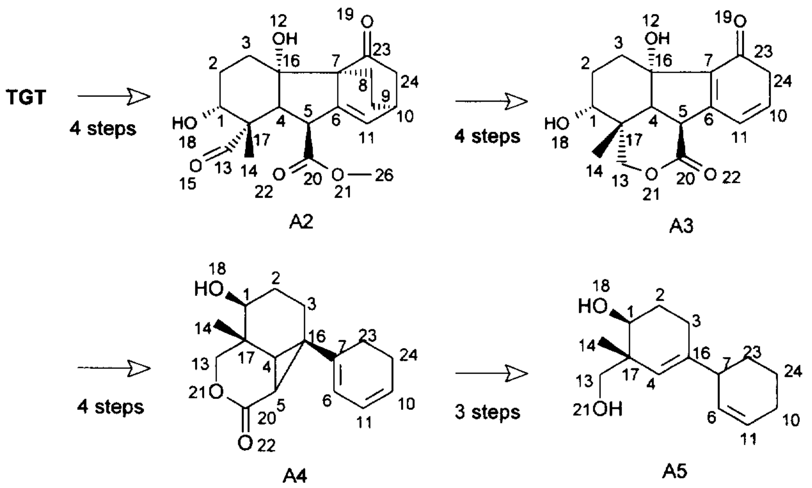


Fig. 19. Sketch of the literature synthesis of Antheridic acid

The comparison between the two TGTs will consider only two descriptors of the previous four: atomic sequences and native polarities.

Atomic sequence. There are many possible comparisons that we can envisage but we will limit our analysis to two of them.

An interesting aspect is represented by the search for common sequences between compounds at corresponding levels of the syntheses. For example, we can compare: the two TGTs, compounds **A2** and **G3**, **A3** and **G4**, **A4** and **G7**, **A5** and **G8**. The corresponding sequences are long 12, 4, 4, 0, and 4 atoms, giving rise to 1.02, 0.42, 0.42, 0, and 0.42, similarity indexes. This result indicates that despite of the apparent structural similarity the analysis made using our method is sensitive enough to distinguish the compound pairs. However, it is still possible to define the two TGTs as similar compounds because their similarity index is not negligible.

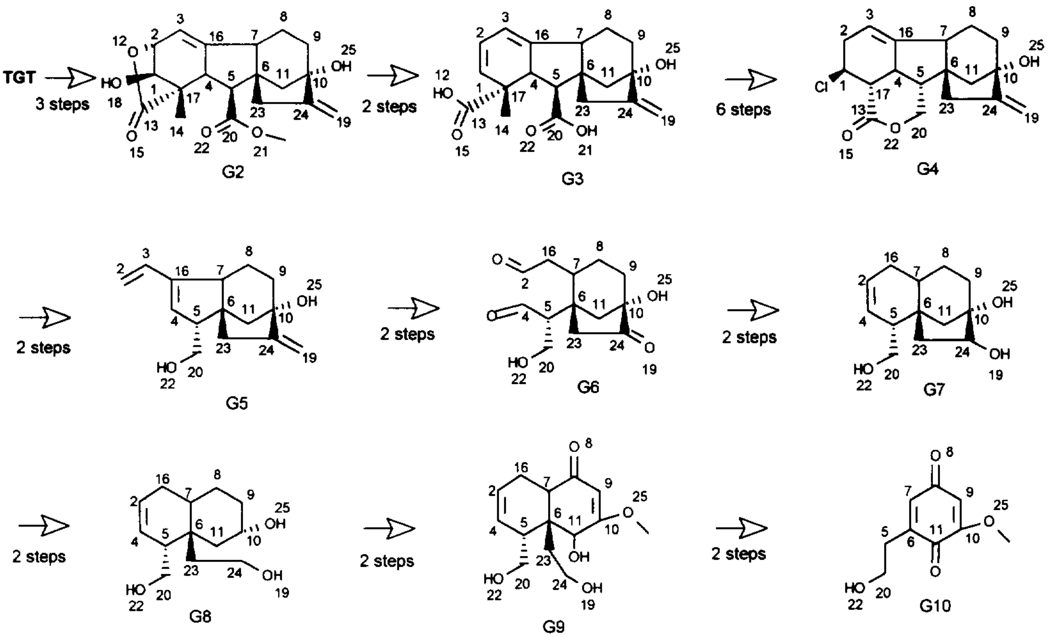


Fig. 20. Sketch of the literature synthesis of Gibberellic acid

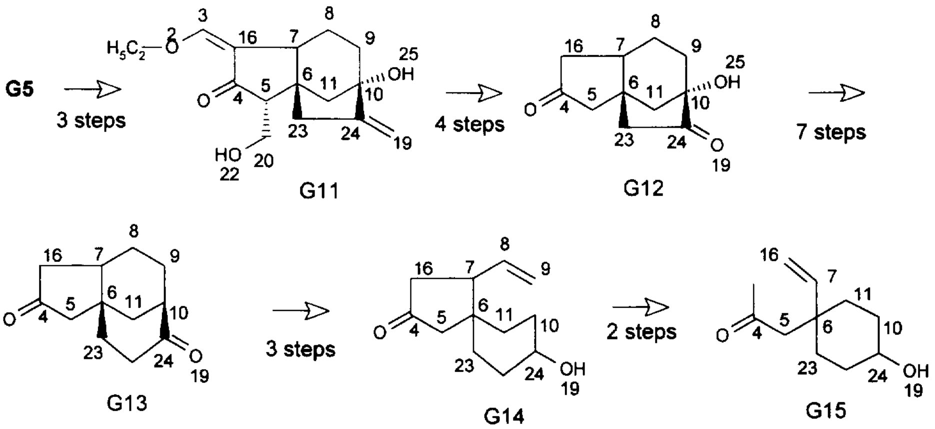


Fig. 21. Sketch of an alternative literature synthesis of Gibberellic acid

By the same descriptor we can also compare short synthetic paths, i.e. we can compare **G5**, **G6**, **G7**, **G8**, **G9**, and **G10**, with **G11**, **G12**, **G13**, **G14**, and **G15**; these two paths start from structures of similar complexity and end with structures of similar complexity. The corresponding similarity indexes are: 0.67, 0.83, 1.88, 1.15, and 0.65; 1.21, 1.29, 1.14, and 1.63. It is immediately evident that the two syntheses are very different for what the course of the similarity concerns: the first beginning with a sharp decrease followed by one step characterised by a high similarity and ending with a sharp change, the second showing a different course that, after a fast start, continues with a constant similarity ratio in each step. The overall similarity can be evaluated using the geometrical mean that gives 0.95 and 1.30, respectively, thus suggesting a different speed of the structural simplification. In conclusion we can affirm that the two paths are comparable, they simplify the TGTs at the same degree, the first is one step longer but it arrives at a starting material less similar to the TGT (SF = 0.58 and 1.17 respectively); in agreement they are proposed as alternative syntheses in the literature.

Atomic native polarity. In Table 7 are reported the NP values for the compounds considered, i.e. **A** through **A5** for Antheridic acid, and **G** through **G8** for Gibberellic acid.

Comparing the NP of the two TGTs we can find 8 differences only (8 on 25 heavy atoms); the two TGTs are consequently quite similar by this descriptor. More impressive is the similarity of the two proposed syntheses. In fact, summing all the changes along the paths, there are respectively 35[41] and 25 NP changes that, considered the evident differences shown by the precursors, keep much of the similarity from the TGTs to the SMs. Moreover, if we simplify the synthesis of Gibberellic acid taking off those precursors that cannot be compared to their partners of the Antheridic acid synthesis, we reduce its NP changes to 19 thus getting an even better similarity.

Looking at the syntheses we note:

I. Gibberellic acid.

   A. **G3**. The lactone is opened and one OH is eliminated.

   B. **G4**. The acid group on 20 is reduced to alcohol and lactonised with 13. HCl is added.

   C. **G7**. A retro Diels Alder, a retro aldol reaction, and a retro ozonolysis have been applied.

   D. **G8**. The glycol group is opened.

II. Antheridic acid.

   A. **A2**. The lactone is opened and its acidic group reduced to aldehyde, the alcohol group on 23 is oxidised and dealkylated

   B. **A3**. A retro Diels Alder has been applied, the aldehyde on 13 is reduced and lactonised with 20

   C. **A4**. The keto group on 23 is eliminated, the alcohol group on 16 is eliminated together with ring shrinking.

   D. **A5**. The lactone is opened, a retro diazocarboxylation applied, and a double bond eliminated.

Summarising we have an overlap of reactions around 75% and 60% for Gibberellic and Antheridic acids, respectively; these values can be compared with the corresponding NP overlap of 75% derived from the previous calculation. Despite of the fact that the two syntheses are represented at the second level of abstraction (first level: reactions to structures; second level: structures to NPs) the formal and practical agreement with the experimental data is good. However, we would like to emphasise that the mere comparison of the number of changes does not imply the similarity between the syntheses but only the similarity between their synthetic complexity. A better hint could come from the comparison of groups of NPs corresponding to atoms restricted to small transformations; we would not comment this possibility.

Table 7. Native polarities of compounds on Antheridic and Gibberellic acid trees.

	Atom
Compound	

	1	2	3	4	5	6	7	8	9	10	11	12	13
G	7.52	-4.23	-4.35	-0.69	-7.71	0.21	-0.51	-0.35	-0.71	7.92	-0.80	-13.83	5.32
G2	7.10	-3.46	-4.39	-0.27	-7.72	0.21	-0.08	-0.35	-0.71	7.92	-0.80	-13.76	5.33
G3	-4.00	-3.82	-3.83	-0.28	-7.72	0.21	-0.08	-0.35	-0.71	7.92	-0.80	-15.52	10.44
G4	-1.37	-0.56	-3.89	-0.14	-0.48	0.27	-0.08	-0.35	-0.71	7.92	-0.80	-	5.67
G5	-	-5.66	-3.73	-3.97	-0.45	0.26	-0.08	-0.35	-0.71	7.92	-0.80	-	-
G6	-13.1[c]	2.82	-13.2[d]	2.76	-8.14	0.19	-0.11	-0.35	-0.73	0.11	-0.80	-	-
G7	-	-3.78	-	-3.84	-0.30	0.19	-0.08	-0.35	-0.73	7.66	-0.80	-	-
G8	-	-3.78	-	-3.84	-0.30	0.19	-0.08	-0.35	-0.73	7.69	-0.80	-	-
A	7.49	-0.65	-0.74	-0.70	-7.63	-2.01	-0.43	-0.42	-0.37	-0.07	-3.97	-13.82	5.32
A2	7.49	-0.65	-0.74	-0.70	-7.63	-2.06	-7.73	-0.42	-0.37	-0.07	-3.97	-11.17	2.60
A3	7.56	-0.65	-0.74	-0.70	-14.83	-1.62	-8.99	-	-	-3.45	-8.66	-11.17	-1.70
A4	7.56	-0.65	-0.43	-0.30	-7.71	-3.83	-1.88	-	-	-3.79	-3.82	-	-1.70
A5	7.56	-0.65	-0.31	-4.02	-	-3.85	0.08	-	-	-0.24	-3.82	-	-1.70

	14	15	16	17	18	19	20	21	22	23	24	25	26
G	-0.77	-13.53	-0.80	-7.56	-11.10	-6.32	10.51	-15.52	-13.51	-0.42	-2.15	-11.17	
G2	-0.77	-13.53	-1.87	-7.57	-11.11	-6.32	5.41	-13.64	-13.52	-0.42	-2.15	-11.17	-2.00[a]
G3	-0.78	-13.52	-1.87	-7.32	-	-6.32	10.51	-15.52	-13.51	-0.42	-2.15	-11.17	
G4	-	-7.92	-1.86	-13.88	-	-6.32	-1.61	-	-13.90	-0.42	-2.15	-11.17	-1.05[b]
G5	-	-	-1.72	-	-	-6.32	7.34	-	-11.06	-0.42	-2.15	-11.17	
G6	-	-	-8.13	-	-	-13.35	7.25	-	-10.97	-8.13	5.54	-11.12	
G7	-	-	-0.28	-	-	-11.11	7.34	-	-11.06	-0.75	7.28	-11.17	
G8	-	-	-0.28	-	-	-11.06	7.34	-	-11.06	-0.75	7.40	-11.11	
A	-0.77	-13.53	-0.91	-7.49	-11.10	-11.11	5.40	-13.65	-13.52	7.49	-2.14	-5.77	-2.00
A2	-0.77	-13.04	7.77	-7.68	-11.10	-13.24	5.40	-13.65	-13.52	5.83	-8.05	-	-2.00
A3	-0.71	-	7.77	-0.30	-11.10	-13.90	5.38	-13.70	-13.53	5.67	-6.48	-	-
A4	-0.71	-	0.20	-0.30	-11.10	-	5.38	-13.70	-13.53	-0.32	-0.37	-	-
A5	-0.71	-	-1.82	-0.30	-11.10	-	-	-11.06	-	-0.37	-0.31	-	-

[a] It is the methyl of the ester group. [b] It is the chlorine atom. [c,d] They are the oxygen atoms of the aldehydic groups.

DISCUSSION

Many hints coming from similarity analysis on organic synthesis planning have been presented in this paper; some of them are direct formalisation of well-known theoretical principles; others are more meddlesome application of less clear and less defined possibilities of comparison. However, the general feeling coming from all the results is that a great potential is buried inside similarity use even when applied to synthesis design. The work required to extract the wanted references is more than justified by the quality and diversity of the results.

In order to give the flavour of the power of similarity use in synthesis design we presented several bits of different applications. In this section we would like to reconsider what has been shown and, where possible, to suggest some directions.

Recalling Scheme 1 we will try to connect the similarity results with the synthesis problems.

Strategy. The use of similarity descriptors (e.g. globularity and atomic sequence) can be very helpful in solving two of three strategy problems. The third one, identification of strategic aspects of the TGT, is only marginally touched. However, the importance of similarity, for the identification of the strategic aspects of the plan and for the evaluation and sorting and selection, is evident. It is particularly worth of note the chance offered to locate similar and diverse strategies for synthesis either of one or of two TGTs. Our examples have shown that it is possible to group uncommon similar solutions, both on single steps and on short paths, looking at simplification and convergence. It is also feasible to locate strategical shortcuts avoiding those syntheses that pass many times through similar precursors.

Tactic. This aspect takes into account both structural and reactional descriptors. Many of the previous considerations can be applied here too. But, besides them, the role of the TSFs here becomes important. Once a strategy has been chosen the success of its application relies on the tactics chosen, whose power is mainly due to their ability in reactivity management. The use of reactivity descriptors (as native polarity and similar group interference) inside a similarity analysis specially helps when locating alternative paths, either jumping on similar TSFs or changing TSF application orders. However, the chances of success increase if it is possible to know where to go after bumping into an unforeseen obstacle. This kind of information can arrive from a single TGT analysis or from the comparison with a preceding experience. The union between structural and reactional analysis can also aid to evaluate the weight of the difficulties against that of the simplification and, as a consequence, to choose the best tactic and/or strategy. Finally, the possibility of obtaining the best ordering of alternatives is important.

Refinement. The availability of a tool capable of weighting many of the characteristics of the synthesis tree can be helpful also when considering small alterations of the plan in order to optimise as much as possible the solution. Reasoning by analogy or, better, using similarity methods is the very chance that can go beyond the common practice of remembering what seen in the literature or what learnt by experience. Some of the results presented in this paper are undoubtedly questionable but there are hints that cannot come from different approaches (e.g. let's consider the Gibberellic-Antheridic acid case as analysed by NP; the benefit is not immediate (the syntheses and the TGTs are different) but is not trivial: if the transformation of compound **G8** into Gibberellic acid presented many difficulties we can expect that the passage from **A5** to Antheridic acid will be as much difficult).

CONCLUSION

We have presented the results we obtained by the application of similarity concepts to synthesis design. The scope of the presentation is obviously limited and doesn't aim to be the solution of the problem. However, the main objective of this paper was to assess the power of the similarity approach also in helping synthesis planning. By selecting a bunch of our own descriptors and using them in a novel perspective we hope to have convinced the readers on the many potentialities offered.

Acknowledgements. Partial financial support by the Consiglio Nazionale delle Ricerche, and by the Ministero dell'Universita' e della Ricerca Scientifica e Tecnologica, is gratefully acknowledged.

REFERENCES AND NOTES

1. *Concepts and Applications of Molecular Similarity.* M.A. Johnson, G.M. Maggiora, Eds.; Wiley Interscience: New York, 1990.
2. *Molecular Similarity and Reactivity: from Quantum Chemical to Phenomenological Approaches.* R. Carbo' Ed.; Kluwer Academic Publishers: Dordrecht, 1995.
3. Bath, P.A.; Poirrette, A.R.; Willett, P.; Allen, F.H.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141-147.
4. Judson, P.N.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 148-153.
5. 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.; *Chem. Struct. 2 Proc. Int. Conf. 2nd 1990.* W.A. Warr, Ed.; Springer-Verlag, Berlin, 1993; pp. 399-407.
6. Benigni, R.; Andreoli, C.; Giuliani, A.; *Environ. Mol. Mutagen.* **1994**, *24*, 208-219.
7. Satoh, H.; Funatsu, K.; *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 34-44.
8. Ugi, I.; Bauer, J.; Blomberger, C.; Brandt, J.; Dietz, A.; Fontain, E.; Gruber, B.; v.Scholley-Pfab, A.; Senff, A.; Stein, N.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 3-16.
9. Zefirov, N.S.; Baskin, I.I.; Palyulin, V.A.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 994-999.
10. Hendrickson, J.B.; Sander, T.; *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 251-260.
11. Rose, J.R.; Gasteiger, J.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74-90.
12. Johnson, A.P.; Marshall, C.; Judson, P.N.; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 310-316.
13. Hendrickson, J.B.; *Anal. Chim. Acta* **1990,** *235*, 103-114.
14. Gordeeva, E.V.; Lushnikov, D.E.; Zefirov, N.S.; *Tetrahedron* **1992**, *48*, 3789-3807.
15. Barone, R.; Arbelot, M.; Chanon, M.; *Tetrahedron Comput. Method.* **1988**, *1*, 3-11.
16. Azario, P.; Arbelot, M.; Baldy, A.; Meyer, R.; Barone, R.; Chanon, M.; *New J. Chem.* **1990**, *14*, 951-956.
17. Gelernter, H.; Rose, J.R.; Chen, C.; *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492-504.
18. Hamm, P.; Jauffret, P.; Kaufmann, G.; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 317-322.
19. Fontain, E.; *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 748-752.
20. Gasteiger, J.; Ihlenfeldt, W.D.; Rose, P.R; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270-290.
21. Long, A.K.; Kappos, J.C.; Rubinstein, S.D.; Walker, G.E.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 922-933.
22. Johnson, A.P.; Marshall, C.; Judson, P.N.; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 310-316.
23. Hendrickson, J.B.; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 323-334.
24. Long, A.K.; Kappos, J.C.; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 915-921.
25. Gasteiger, J.; Hondelmann, U.; Rose, P.; Witzenbichler, W.; *J. Chem Soc. Perkin Trans. 2* **1995**, 193-204.
26. Gordeeva, E.V.; Lushnikov, D.E.; Zefirov, N.S.; *Tetrahedron* **1992**, *48*, 3789-3807.

27. Corey, E.J.; Long, A.K.; Lotto, G.I.; Rubinstein, S.D.; *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 304-309.
28. Wochner, M.; Brandt, J.; v.Scholley-Pfab, A.; Ugi, I.; *Chimia* **1988**, *42*, 217-225.
29. (a) Ihlenfeldt, W.; Gasteiger, J.; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2613-2633. (b) Gasteiger, J.; Ihlenfeldt, W.D.; Rose, J.D.; *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 700-712
30. Baumer, L.; Sala, G.; Sello, G.; *Anal. Chim. Acta* **1990**, *235*, 209-214. The equation for the calculation of molecular globularity is: $G = MAXD / COMP_{TOT}$ where MAXD is the greatest of the smallest distances of atom pairs measured as atom complexity, and $COMP_{TOT}$ is the molecular complexity measured as the sum of all the atom complexity (Baumer, L.; Sala, G.; Sello, G.; *Tetrahedron* **1988**, *44*, 1195-1206).
31. Baumer, L.; Sala, G.; Sello, G.; *J. Am. Chem. Soc.* **1991**, *113*, 2994-2500.
32. Baumer, L.; Sala, G.; Sello, G.; *Tetrahedron* **1993**, *49*, 3367-3386.
33. Leoni, B.; Sello, G. A Proposal Toward the Identification of Substructure Electronic Similarity. In *Molecular Similarity and Reactivity: from Quantum Chemical to Phenomenological Approaches*; Carbo', R.; Ed.; Kluwer Academic Publishers: Dordrecht, 1995. pp. 267-289.
34. Sello, G.; *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 120-129.
35. If we consider our second method of similarity calculation P32 and P32' result much less similar (only 12 similar atoms). A second important aspect is represented by the number of sequences containing the similar atoms. In the present case we have two sequences (7+13) that can weight differently compared with a single 20 atom sequence.
36. Schulz, K.P.; Gasteiger, J.; *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 395-406.
37. Blurock, E.S.; *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 607-616.
38. Despite of this we can, however, emphasise that bond 6-5 is a vinylogous example of bond 6-7.
39. It is interesting to note that the index value for solution 6 increases going from the first to the second level. This apparently wrong result is, however, completely correct; in fact, breaking bond 9-10 the energy of atom 9 changes enough to make it different from atom 9 in the TGT.
40. Corey, E.J.; Cheng, X. In *The Logic of Chemical Synthesis*. John Wiley & Sons, Inc.: New York, 1989; pp. 205-214.
41. This number is comprehensive of two useless changes passing from the TGT to G2 to G3 for a total of 4 changes.

(*Received in UK* 14 *November* 1996; *accepted* 23 *January* 1997)