JOURNAL OF PHYSICAL ORGANIC CHEMISTRY
J. Phys. Org. Chem. 2003; **16:** 9–15
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/poc.556

Information theory description of synthetic strategies. A new similarity index

R. Barone,[1]* M. Petitjean,[2] C. Baralotto,[3] P. Piras[4] and M. Chanon[1]

[1]Laboratoire AM3 (UMR CNRS 6009), Faculté St Jérôme, case 561, 13397 Marseille cedex 20, France
[2]ITODYS (ESA CNRS7086), 1 rue Guy de la Brosse, 75005 Paris, France
[3]Provence Technologie, Technopôle Château-Gombert, BP 100, 13382 Marseille cedex 13, France
[4]ENSSPICAM (UMR CNRS 6516), Faculté St Jérôme, 13397 Marseille cedex 20, France

Received 20 April 2002; revised 21 June 2002; accepted 25 June 2002

ABSTRACT: Information theory was used to analyse and compare organic syntheses leading to the targets, daucene, longifolene and estrone. This paper expands the work of Bertz, who analysed syntheses from the complexity of molecular structures. Herein, a more complete model involving similarity was evaluated. We published previously a study in this direction limited to a skeletal level. In order to improve on this initial approach, we attempted to analyse some syntheses not only limited to the skeleton by including different definitions of similarity. Copyright © 2002 John Wiley & Sons, Ltd.

KEYWORDS: complexity; similarity; synthetic strategy

INTRODUCTION

In a previous study, we used information theory to describe by a semi-quantitative graphical representation the various strategies to reach a given target.[1] This approach was an extension of the initial work of Bertz,[2] who analysed syntheses from the complexity of the precursors. We proposed to complete this analysis by the calculation of the similarity of the precursors. This approach allows a more realistic description of syntheses by a three-dimensional model: complexity and similarity versus steps.

In our first study, similarity was calculated at a skeletal level from the number of atoms, number and types of bond (CH_3—CH_2, CH_2—CH_2, etc.) and information about the rings (number, size, connections, e.g. fused, spiro, bridged).[1] In this work we complemented this approach by using descriptions not limited to the skeleton.

Many definitions of similarity have been proposed.[3] In a review, Willett indicated that there are three main approaches to calculate similarity:[3a] fragment substructures (FS), topological indices (TI) and maximal common subgraph (MCS). Hence, in order to verify the legitimacy of our approach, we decided to test some of them. The measures based on the fragment substructures are the most used.[3a] Having the possibility of using the similarity index incorporated in MDL ISIS software,[4,5] which is computed according to this approach, we decided to use it as an FS descriptor. In ISIS, the similarity is estimated from a search of about 1000 fragments.[4] For topological indices, Randic's index is widely applied.[3] It is limited, however, to the skeleton of the structures. We therefore decided not to use TI in this study. For MCS, Petitjean's approach[6] was chosen. It calculates the 3D similarity between two molecules through the number of atoms of the largest fragment common to their structural formulae.

As we measure the similarity of precursors versus the target during a synthesis, we decided to introduce a new constraint for the comparison of the structures: the evolution and the position of the atoms during the synthesis. This is the third approach which was used in this work.

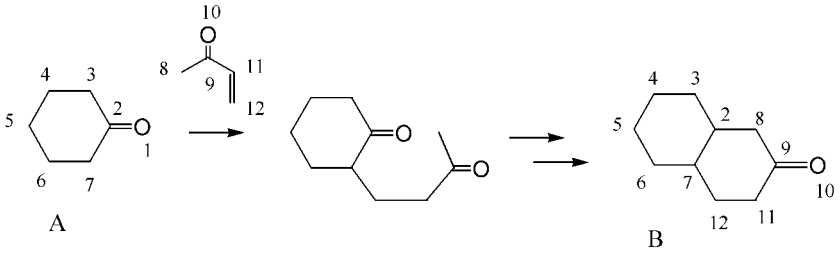
SIMILARITY FROM UNCHANGED REACTIONAL FRAGMENTS

In this new approach, to compare two structures with $n1$ and $n2$ atoms, respectively, we adopted the following rules: an atom $i1$ in molecule 1 is defined as equivalent to (or identical with) an atom $i2$ in molecule 2 when the following conditions are satisfied: (a) $i1$ and $i2$ have the same atomic number; (b) the number and type of bonds involving $i1$ are equal to the number and type of bonds involving $i2$; (c) numbering must be respected.

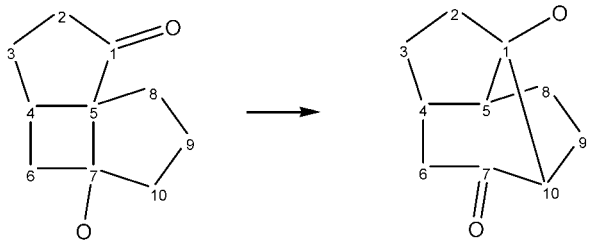
Let us consider the sequence in Scheme 1. In a classical approach, the common fragment between structures A and B should be the cyclohexanone system. With our new approach, atoms 1 and 10 are not equivalent, since atom 1 was removed. Hence the common fragment is composed of atoms 3, 4, 5, 6 and 7. Atom 2 is sp2 in A and sp3 in B, so they are not

**Correspondence to:* R. Barone, Laboratoire AM3 (UMR CNRS 6009), Faculté St Jérôme, case 561, 13397 Marseille cedex 20, France.
E-mail: rene.barone@univ.u-3mrs.fr

Copyright © 2002 John Wiley & Sons, Ltd.



Scheme 1



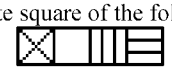
Scheme 2

Let us imagine two structures C and D :

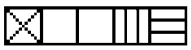
Structure C : n1 = 2

Structure D : n2 = 3 and n12 = 1 (the white square)

Tanimoto index : T = 1/(2+3-1)= ¼ = 0.25. This value is equivalent to the part of the white square of the following structure E :



Average index : A = 2*1/(2+3) = 0.4, it is equivalent to the part of the white squares of structure F :



Scheme 3

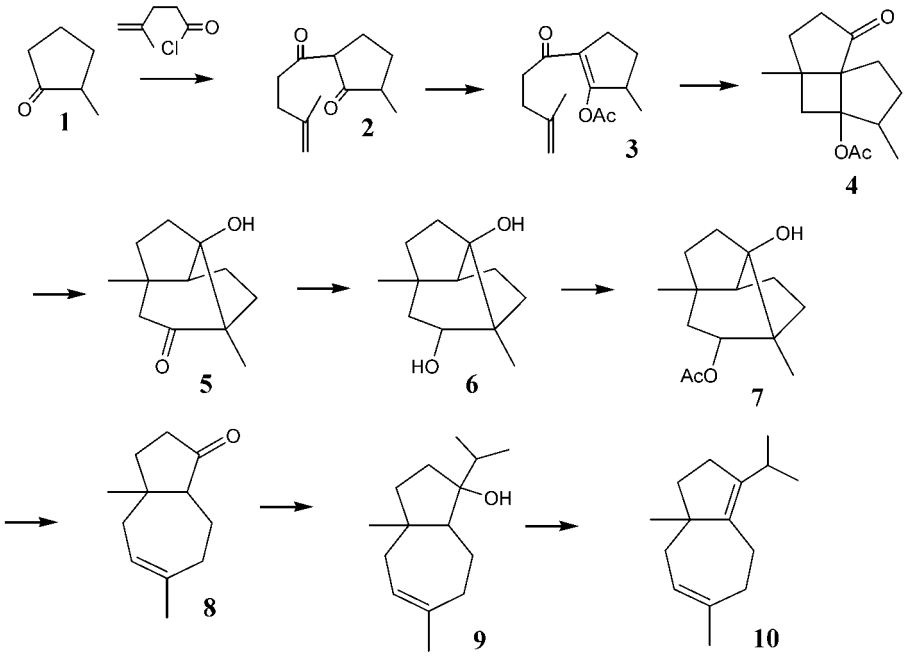


Figure 1. Synthesis of daucene.[8]

Copyright © 2002 John Wiley & Sons, Ltd.

considered as equivalent. This approach is an intermediate between fragment substructure and maximal common subgraph and has never been described. We named it SURF (similarity from unchanged reactional fragments). (Optionally, for some applications, more conditions may be added, such as $i1$ and $i2$ have the same number of peripheral electrons; $i1$ and $i2$ have the same mass number; $i1$ and $i2$ have the same stereochemical configuration.)

Calculation of similarity based only on the atom and bond types is, however, not totally satisfactory. It neglects an important factor, as shown in our previous paper:[1] the presence of ring systems in the structures. Consequently, we settled for a new component: the ring factor.

This ring factor is composed of two factors: First, a factor indicating if atoms are involved in a ring; for instance, in the structures in Scheme 2, atoms 1–10 are all cyclic in both reactant and product so they are all equivalent; second, a factor characterizing the size of each ring; in Scheme 2, atoms 1–5 are in a five-membered ring in the two parts of the reaction, so they are equivalent, but atoms 4, 5, 6, 7, 8, 9 and 10 do not share the same ring and are not equivalent.

At the end of these comparisons, we compared $n1$ elements of structure 1 and $n2$ elements of structure 2, and there are $n12$ equivalent elements. The Tanimoto coefficient (T) is generally used to calculate similarity:[7]

$$T = n12/(n1 + n2 - n12)$$

Another way to calculate similarity is to use the average (A), given by

$$A = n12/[(n1 + n2)/2] = 2 \times n12/(n1 + n2)$$

Scheme 3 shows the differences between these two approaches. Let us imagine structures C and D. Structure C is composed of two elements ($n1 = 2$) and for structure D the number of elements is equal to three ($n2 = 3$). There is one common element (the white square) and $n12 = 1$. From these values, $T = 0.25$ and $A = 0.4$. They correspond to the part of the white squares of structures E and F.

From the two equations it follows that $T = A/(2 - A)$ or $A = 2T/(1 + T)$. Hence, as there is a linear relationship between A and T, one can select either of the two indifferently. Here we will use the Tanimoto index.

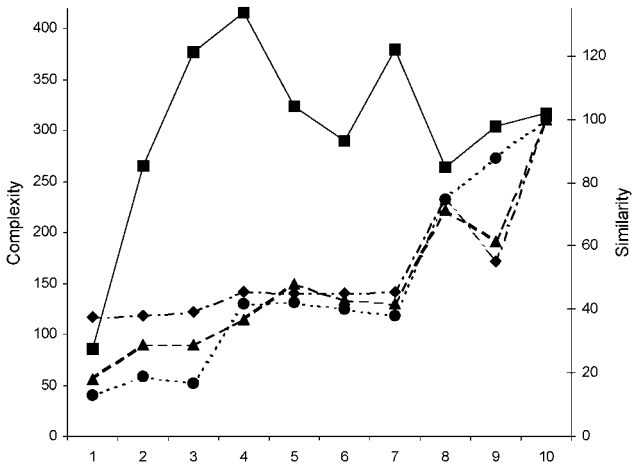


Figure 2. Complexity (continuous line) vs similarity (dotted lines) for daucene synthesis. The three similarity indexes are (■) Petitjean, (▲) MDL and (●) SURF

RESULTS

For this study we selected four syntheses which are displayed in Figs 1, 3, 5 and 7, namely syntheses of

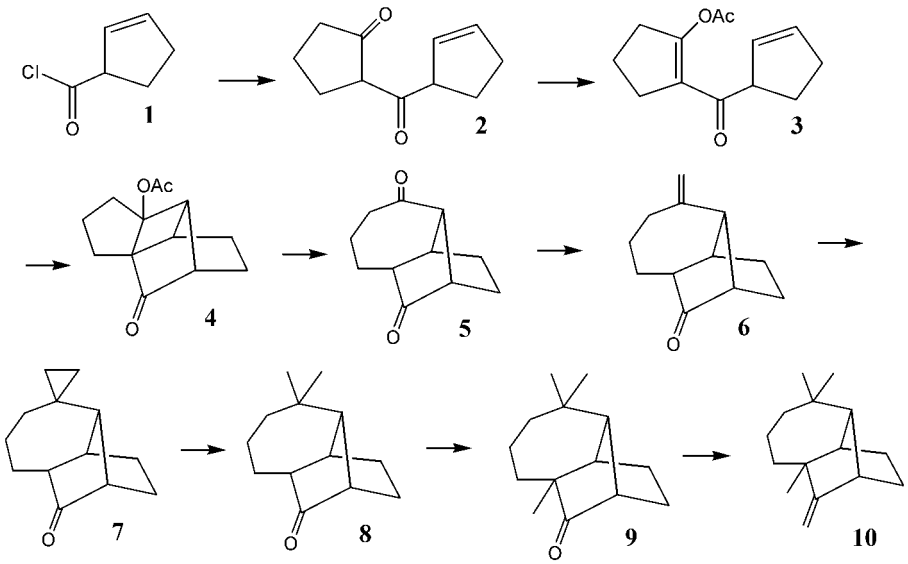


Figure 3. Oppolzer's synthesis of longifolene.[9]

Copyright © 2002 John Wiley & Sons, Ltd.

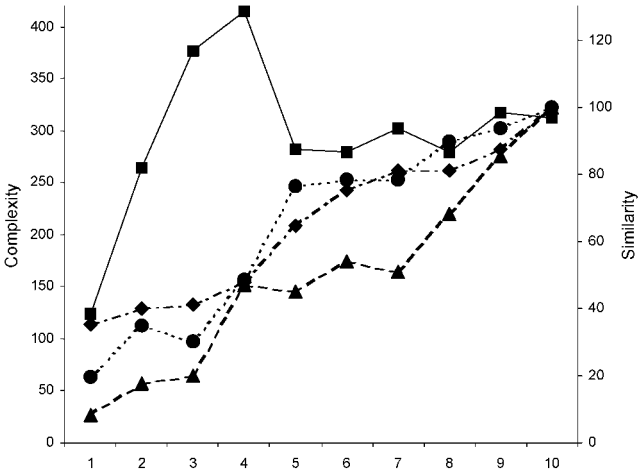


Figure 4. Complexity (continuous line) vs similarity (dotted lines) for Oppolzer's longifolene synthesis. The three similarity indexes are (■) Petitjean, (▲) MDL and (●) SURF

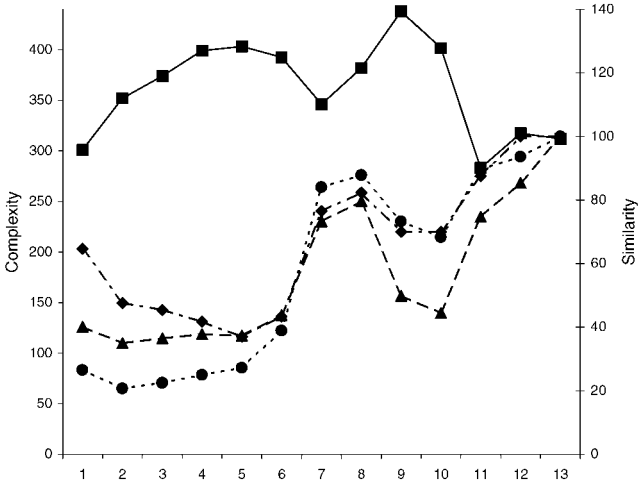


Figure 6. Complexity (continuous line) vs similarity (dotted lines) for Corey's longifolene synthesis. The three similarity indexes are (■) Petitjean, (▲) MDL and (●) SURF

daucene,[8] longifolene[9,10] and estrone.[11] Complexity versus similarity data are given in Figs 2, 4, 6, 8 and several comments can be made about these graphs.

Bertz[2] showed that the sum of the complexities (ΣC) of the intermediates was inversely correlated to the yield. This assumption is verified when comparing the syntheses of longifolene from Corey's approach (Figs 5 and 6. $\Sigma C = 4701$, yield = 2%) and Oppolzer's approach (Figs 2 and 3: $\Sigma C = 2951$, yield = 25%).

The results obtained for the calculation of similarity from the three methods, although different in their approaches, display some analogies, particularly for daucene synthesis (Figs 1 and 2), except for one point (intermediate **9**).

Are similarity and complexity correlated? Similar structures should have similar complexities. In our previous paper we showed that this trend was not necessarily observed.[1] This can be verified from the first steps of the daucene synthesis (Figs 1 and 2): the variation of similarities is low, which indicates that the

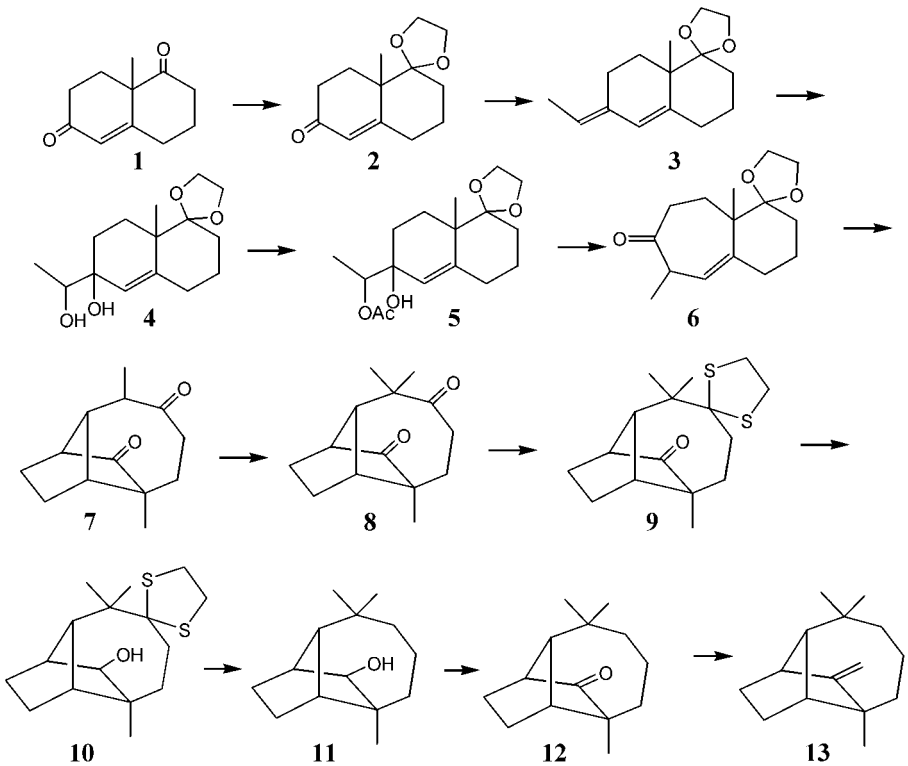


Figure 5. Corey's synthesis of longifolene.[10]

Copyright © 2002 John Wiley & Sons, Ltd.

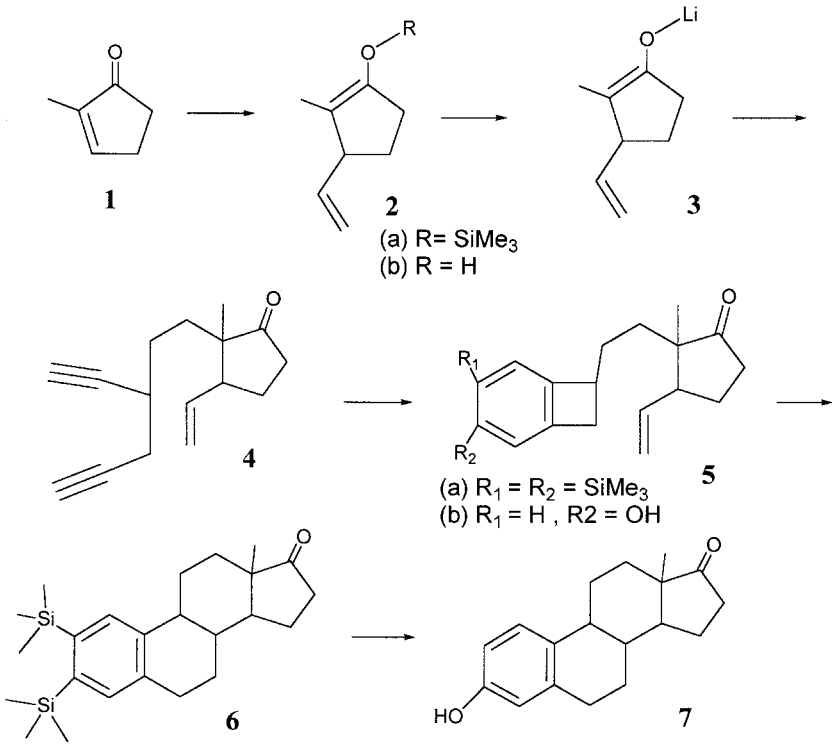


Figure 7. Vollhardt's estrone synthesis,[11] path (a)

structures remain similar. In contrast, the variation of complexity is very high. This is due to the presence of substituents which increase the complexity without affecting the similarity. The same point emerges from Oppolzer's and Corey's syntheses of longifolene (Figs 3, 4, 5 and 6) and in the last steps of the estrone synthesis (Figs 7 and 8).

Since the presence of protecting groups can alter the results, we studied among these syntheses two of them involving large protecting groups: Corey's longifolene synthesis and Vollhardt's estrone synthesis. The protecting groups were eliminated by generating simplified syntheses for which we studied the evolution of complexity vs similarity. To simplify the results, similarity was computed only with the SURF model. The results are shown in Figs 9 and 10. Now, for the simplified estrone synthesis [cf. Fig. 7: compounds **1**, **2** (R = H), **4**, **5** (R_1 = H, R_2 = OH) and **7**] complexity and similarity are more correlated, except for the last step

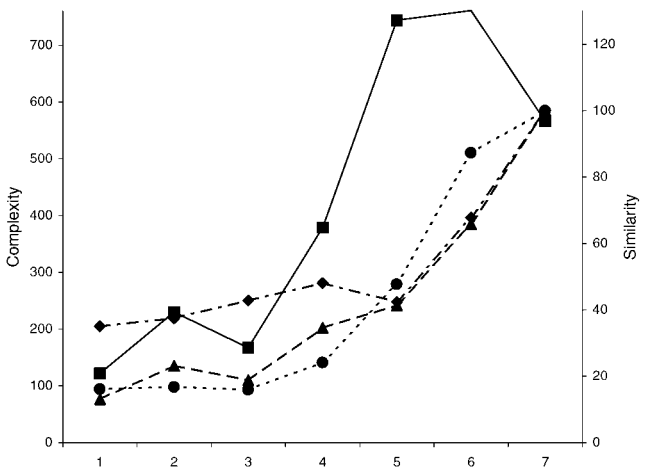


Figure 8. Complexity (continuous line) vs similarity (dotted lines) for estrone synthesis. The three similarity indexes are (■) Petitjean, (▲) MDL and (●) SURF

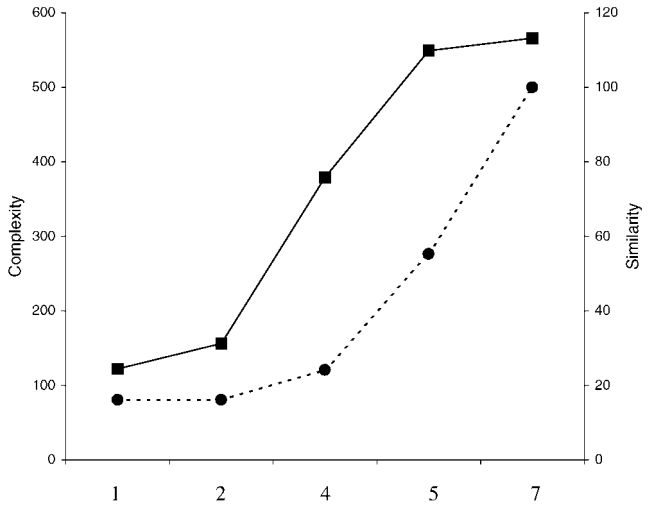


Figure 9. Complexity (continuous line) vs similarity (dotted line, SURF model) for simplified Vollhardt's estrone synthesis [cf. Fig. 7 (path b), **1**, **2**, **4**, **5**, **7**]

Copyright © 2002 John Wiley & Sons, Ltd.

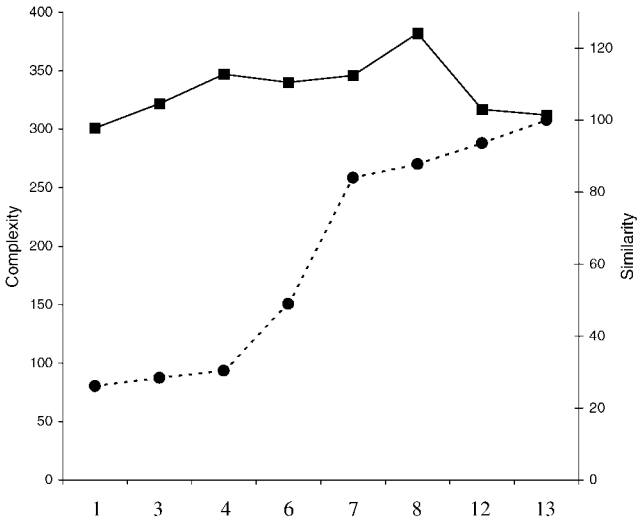


Figure 10. Complexity (continuous line) vs similarity (dotted line, SURF model) for Corey's longifolene simplidied synthesis (cf. Fig. 5, compounds **1**, **3**, **4**, **6**, **7**, **8**, **12** and **13**, with C═O instead of —OCH_2CH_2O—)

where complexity increases slightly whereas similarity increases considerably (Fig. 9). For Corey's simplified longifolene synthesis [cf. Fig. 5, **1**, **3**, **4** and **6** (the protecting group is replaced by C═O) **7**, **8**, **12**, and **13**], however, they are not correlated (Fig. 10). Between **4** and **7** there is a large jump in similarity, whereas complexity stay constant. For **8** –**12** the opposite occurs: similarity remains almost constant whereas complexity decreases.

Finally, as a subsequent result of this study, the key step of a synthesis can be directly highlighted from the evolution of similarity. The jump in similarity clearly shows where this (these) key step(s) is (are) located. For example, in the case of daucene synthesis (Figs 1 and 2), the three approaches indicate very clearly that it is situated between intermediates **7** and **8**. In Corey's longifolene synthesis (Figs 5 and 6) the jump is between intermediates **6** and **7**. In the case of estrone synthesis (Figs 7 and 8) the key step is from **5** to **6**. For Oppolzer's longifolene synthesis (Figs 3 and 4), the key step in the MDL approach is between intermediates **3** and **4**, whereas for the two other approaches the key step is between **4** and **5**.

One referee remarked that the key step hinted at by complexity is not always the same as the key step based on similarity. One may return to the two-dimensional representation given previously (Ref. 1, Fig. 2) to clarify this point. In the space of construction of targets starting from simple structural units, one may select a route because one of the steps allows a drastic change in complexity. The route may be such that this large gain in complexity is reasonable also in terms of similarity. In some cases, however, the jump in complexity leads in the direction of a target dissimilar to the one aimed at. Then one has to return towards the right target. This operation may be made slowly or rapidly. In this second hypothesis, a key step with a large change in similarity may be needed. Another factor is the influence of substitutents and protecting groups: it is particularly clear in Fig. 10. Similarity indicates that the key step is between **6** and **7**, whereas complexity varies only from 340 to 346, but for **7** and **8** there is a jump from 346 to 382 (formation of a quaternary carbon) then from **8** to **12** there is a large decrease in complexity (382 to 317) due to the loss of C═O and of a quaternary carbon.

On the basis of these observations, we can propose that such developments can play an important role in the tasks aimed at automatically detecting the key step of a synthesis from the search of large reaction databases.

CONCLUSION

We proposed to visualize the evolution of a synthesis by studying the complexity and similarity of intermediates. A new approach to calculate similarity has been developed which seems to be in good agreement with other well-established methods. Indeed, similarity has been shown to be a valuable key component for the analysis and description of a synthesis, as demonstrated by the results of this work.

Along with the development of these tools, we have seen that this information may help in identifying the key step(s) of a given synthesis, since a rapid method aiding the automatic identification of the determinant steps of syntheses could also provide straightforward information regarding their organization. This is an important perspective for future work as we can conceive the integration of such a methodology in a reaction database. Accordingly, we believe that the improvement of the model may result in a more successful selection of synthetic strategies and enhance the rate at which a given target may be obtained.

These calculations of complexity and similarity may be easily implemented in a computer organic synthesis program and could be used as a complementary tool for helping in answering the question of which among the *N* possible routes to this target should one select. This question is not really critical for really experimented chemists because they rely strongly on good intuition to go through this decisive step. If one aims, however, to quantify this intuition to a young chemist, this tool has the advantage of displaying graphically what is behind the intuition.

REFERENCES

1. Chanon M, Barone R, Baralotto C, Julliard M, Hendrickson JB. *Synthesis* 1998; **11**: 1559–1583.
2. (a) Bertz SH. *J. Am. Chem. Soc.* 1982; **104**: 5801–5803; (b) Bertz SH, Sommer TJ. In *Organic Synthesis: Theory and Applications*, Hudlicky T (ed). vol. **2**. JAI Press: Greenwich, CI, 1993; 67–92.

Copyright © 2002 John Wiley & Sons, Ltd.

3. (a) Willett P. In *The Encyclopedia of Computational Chemistry*, Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds). Wiley: Chichester, 1998; 2748–2756; (b) Gasteiger J, Ihlenfeldt WD, Fick R, Rose JR. *J. Chem. Inf. Comput. Sci.* 1992; **32**: 700–712; (c) Jochum C, Gasteiger J, Ugi I, Dugundji J. *Z. Naturforsch., Teil B* 1982; **37**: 1205–1215.
4. Moock TE, Grier DL, Hounshell WD, Grethe G, Cronin K, Nourse JG, Theodosiou J. *Tetrahedron: Comput. Methodol.* 1998; **1**: 117–128.
5. Roussel C, Pierrot-Sanders J, Heitmann I, Piras P. In *Chiral Separation Techniques. A Practical Approach* (2nd edn), Subramanian G (ed). Wiley-VCH: Weinheim, 2000; 95–125.
6. Petitjean M. *J. Comput. Chem.* 1995; **16**: 80–90.
7. Willett P, Winterman V, Bawden D. *J. Chem. Inf. Comput. Sci.* 1986; **26**: 36–41.
8. Seto H, Fujimoto Y, Tatsuno T, Yoshioka H. *Synth. Commun.* 1985; **15**: 1217–1224.
9. Oppolzer W, Godel T. *J. Am. Chem. Soc.* 1978; **100**: 2583–2584.
10. Corey EJ, Ohno M, Mitra RB, Vatakencherry PA. *J. Am. Chem. Soc.* 1964; **86**: 478–485.
11. Funk RL, Vollhardt KPC. *J. Am. Chem. Soc.* 1977; **99**: 5483–5484; 1979; **101**: 215–217.

Copyright © 2002 John Wiley & Sons, Ltd.