COMMUNICATION

ChemComm
www.rsc.org/chemcomm

The role of isomorphism in synthetic analysis. Pruning the search tree by finding disjoint isomorphic substructures†

Steven H. Bertz* and Toby J. Sommer
Complexity Study Center, Mendham, NJ 07945, USA

Received (in Corvallis, OR, USA) 22nd January 2003, Accepted 5th March 2003
First published as an Advance Article on the web 25th March 2003

Isomorphism is an equivalence relation that is less stringent than identity (equality), and it is useful for synthetic analysis, since it allows one to find reflexive routes for targets even when they do not have an element of symmetry.

The prevailing paradigm for synthesis design is retrosynthetic analysis, which provides the intellectual framework for synthesis planning, as well as a basis for computerized approaches.[1] Using this method, a large number of synthesis plans can be generated for a molecule of even moderate complexity. Pruning the search tree is critically important, and several strategies have been introduced. Hendrickson confines SYNGEN to the shortest routes by using only construction reactions and to convergent syntheses by dividing the target into two pieces, which are matched with starting materials.[2] In HOLOWin Barone and Chanon further limit key reactions to those that form more than one bond, *i.e.*, rapidly increase molecular complexity (*cf.* holosynthons),[3a] and in SESAM they search for 'non-obvious' starting materials.[3b] Corey (LHASA),[1] Wipke (SECS),[4] Gasteiger (WODCA),[5] Hanessian (CHIRON)[6] and Funatsu (KOSP)[7a] also match targets and their precursors to starting materials. Funatsu and Sasaki (AIPHOS) use a 'forbidden substructure' list to limit the search.[7b] Sello employs a novel 'complexity distance' function to guide LILITH.[8] Chemistry-derived heuristics are used by Bersohn (SYNSUP),[9] and 'best-first' search algorithms, adapted from artificial intelligence programs, by Gelernter (SYNCHEM).[10] Herein we introduce a fundamentally different approach to pruning, which dramatically decreases the breadth of the search by generating only reflexive routes, which are found by identifying isomorphic substructures.

Our approach restricts routes to those with symmetry in the *synthesis digraph*, where points represent structures and arcs the reactions that interconvert them.[11] *Reflexivity* refers to the efficiency in a synthesis that results from this symmetry.[12] Since we do not wish to generate all possible synthetic routes in order to check them for symmetry, we need a method that will efficiently guide a breadth-first search to reflexive routes for in-depth development. The mathematical concept of isomorphism makes it possible to derive reflexive routes to a target, whether or not it has an element of symmetry.

Isomorphism is an equivalence relation that is less stringent than identity.[11] Identical structures M and N, M ≡ N, are superimposable in 3-dimensional space. Thus, they have the same bond lengths and angles in addition to the same connectivity relation. *Isomorphic structures* S and T, S ≅ T, have the same connectivity relation, *i.e.*, the same adjacency matrix for some labeling, but not necessarily the same bond lengths and angles.‡ All identical structures are isomorphic, but not *vice versa*. Thus, isomorphic structures can differ in their conformations, and this feature is critical for synthetic analysis. Many opportunities for efficiency would be missed by a computer (including the human brain) that only conducted searches based on identity. For example, the two five-membered rings in papuamine **1** (Fig. 1) are identical, since they are superimposable on each other by rotation about a C_2 axis. In contrast, the two in haliclonadiamine **2** have different conformations, owing to a different absolute configuration at one of the carbon atoms (*cf.* bold bond), and they are not superimposable. Nevertheless, **1** and **2** are equivalent at the level of isomorphism.

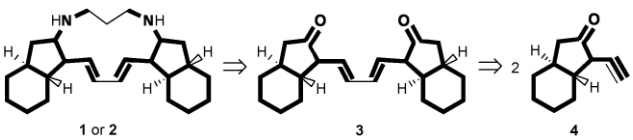
Fig. 1 Non-identical, isomorphic structures **1** and **2**.

Finding reflexive routes involves recognizing isomorphic substructures. A substructure S_i of structure M has all its atoms and bonds in M.§ Substructures in **1** and **2** include ammonia, cyclopentane and *N*-cyclopentyl-*N*-methylamine.§ For our method isomorphic substructures must also be *disjoint*, *i.e.*, they cannot share atoms. Then, the problem of finding reflexive syntheses is reduced to the problem of finding disjoint isomorphic (DI) substructures in a target or intermediates on the way to it (or both, if possible).

The 'all possible substructures' method was originally introduced for measuring similarity and complexity.[13] Since we are only looking for equivalent ones, it is not necessary to enumerate all of the substructures of a target with *n* atoms. Each of two DI substructures has at most $p = n/2$ atoms when *n* is even and $p = (n - 1)/2$ when *n* is odd, and analogously for three or more of them. Maximal DI substructures in **1** or **2** have $(27 - 1)/2 = 13$ non-H atoms, as shown by the bold bonds in Scheme 1. Finding DI substructures is easy when the target has an element of symmetry, as in the case of the significant number of natural products such as **1** with a C_2 axis. For the general case where there is no element of symmetry, the procedure involves enumerating all possible substructures with *p* atoms.

If it is not possible to find DI substructures of maximal size *p*, then all possible substructures with $p - 1$ atoms are enumerated. This process can be repeated until a heuristic limit is reached, *e.g.*, *n*/4 is a reasonable one, since few natural products (outside of biopolymers) have four isomorphic substructures. If this procedure does not lead to the disconnection of the target into two or more pieces, then it can be repeated on the intermediates from the usual retrosynthetic analysis.

The final step involves finding an *isomorphic transform* that preserves the DI substructures in the disconnected precursors.



Scheme 1 Disjoint isomorphic substructures in **1–4**.

† Electronic supplementary information (ESI) available: Fig. S1, all possible 9-atom substructures of the steroid skeleton, and Fig. S2, all possible pairs of disjoint isomorphic steroid substructures with 9 atoms. See http://www.rsc.org/suppdata/cc/b3/b300935a/

DOI: 10.1039/b300935a

Currently, there is no synthetic reaction with a transform that will disconnect **1** or **2** and preserve the maximal DI substructures. Fortunately, the reductive amination transform affords a precursor **3** with maximal DI substructures that are preserved in the isomorphic transform **3** ⇒ **4** + **4**, which is the basis of Heathcock's reflexive synthesis.[14] (*N.B.*, protection is required.) This example illustrates the most general case, since there is no element of symmetry in **2**. The DI substructure method can then be iterated on the precursors with the aim of making the synthesis multiply reflexive (*cf.* next example).

As a more sophisticated example, the basic steroid skeleton has 19 atoms (Fig. 2), and the largest possible candidates for DI substructures have $p = 9$ atoms. The 54 possible steroid substructures with 9 atoms have been enumerated.† Of the 54, only 11 can be cut out of the steroid skeleton twice, as illustrated in Fig. 2, where one canonical example of each kind of DI pair is displayed. All possible variations of these 11 basic kinds are summarized in the ESI†, and none of them corresponds to a known steroid synthesis. Fig. 2k appears to be especially promising, since it is the only one that involves cyclic substructures. This disconnection is suggestive of the Diels–Alder transform, and the precursor skeleton has been prepared in the one-pot reaction of 2 equiv of methyl vinyl ketone and 1 equiv of malonate.[15] A synthesis based on this analysis would be multiply reflexive, as illustrated in Scheme 2, where X and Y are activating (control) groups. The carbon atom at the 17-position (along with a side-chain when it is present in the target) can be incorporated *via* several reaction sequences, *e.g.*, cyclopropanation-rearrangement.

Looking for maximal DI substructures also suggests the optimal approach to taxol **5** (Fig. 3). The taxol skeleton has 20 carbon and 8 oxygen atoms, which can be disconnected into a pair of 13-atom DI substructures (**6**), provided the oxygen atoms on C-4 and C-9 are left out. Alternatively, by adding oxygen atoms at C-12 and C-17, all of the skeletal carbons and oxygens can be included in a pair of 15-atom DI substructures (**7**). It is interesting to note that the LHASA analysis of taxol adds two

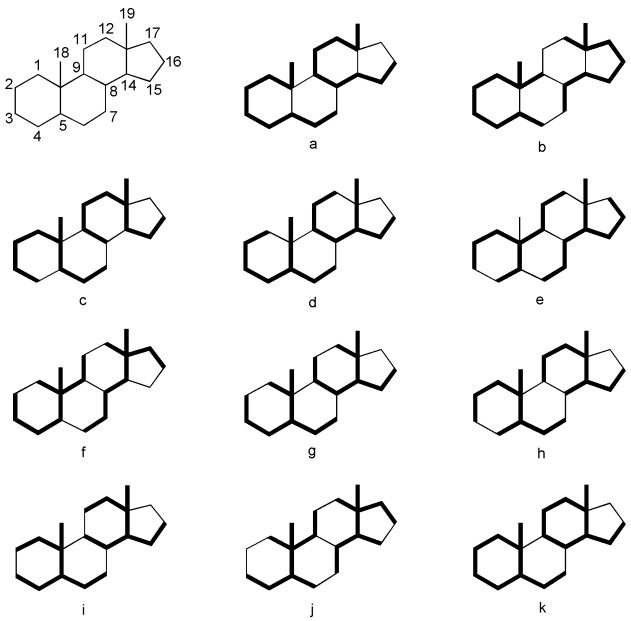
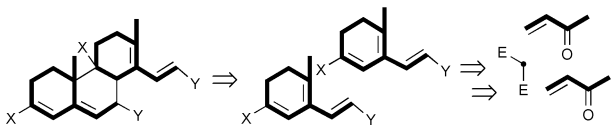


Fig. 2 Canonical examples of disjoint isomorphic steroid substructures. (The numbered atoms are left out in one or more cases; see Fig. S2†.)



Scheme 2 Retrosynthetic analysis corresponding to Fig. 2k.

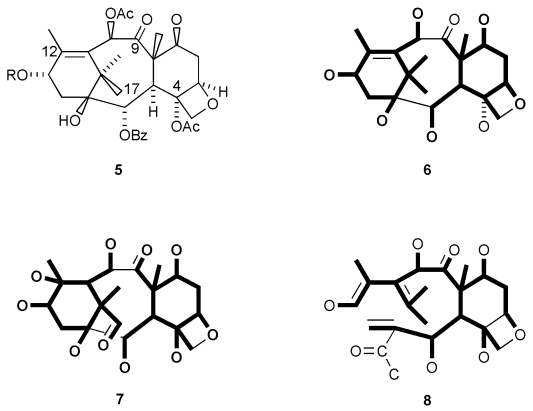


Fig. 3 Selected disconnections of taxol **5**, including disjoint isomorphic substructures in **6** and **7** (R = PhCH(NHBz)CH(OH)C(O) side-chain).

carbon atoms (labeled C in **8**) in order to use the intramolecular Diels–Alder reaction as the key step.[16] The Nicolaou total synthesis of taxol is based on the disconnection of the skeleton into the same DI pair as **6** and **7**, as far as the carbon atoms are concerned.[17] While not reflexive, it is highly convergent.

We thank J. B. Hendrickson (Brandeis U.) for his frequent advice and encouragement.

Notes and references

‡ Enantiomers are non-superimposable mirror images and therefore not identical, but they are isomorphic. They can be tracked by labeling.
§ H atoms are usually excluded. A structure is a substructure of itself; *e.g.*, when M is methane, $S_1 \cong M$ is the only substructure. A substructure is usually named after the stable molecule with the same skeleton.

1 (*a*) E. J. Corey and X.-M. Cheng, *The Logic of Chemical Synthesis*, Wiley, New York, 1989; (*b*) see also A. P. Johnson and C. Marshall, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 426–429.
2 (*a*) J. B. Hendrickson, *Chemtech*, 1998, September, , 35–40; (*b*) J. B. Hendrickson and P. Huang, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 145–151; (*c*) J. B. Hendrickson, *Angew. Chem., Int. Ed. Engl.*, 1990, **29**, 1286–1295.
3 (*a*) F. Barberis, R. Barone and M. Chanon, *Tetrahedron*, 1996, **52**, 14625–14630; (*b*) G. Mehta, R. Barone and M. Chanon, *Eur. J. Org. Chem.*, 1998, 1409–1412.
4 W. T. Wipke and D. Rogers, *J. Chem. Inf. Comput. Sci.*, 1984, **24**, 71–81.
5 W.-D. Ihlenfeldt and J. Gasteiger, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2613–2633.
6 S. Hanessian, J. Franco, G. Gagnon, D. Laramée and B. Larouche, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 413–425.
7 (*a*) K. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 316–325; (*b*) K. Funatsu and S.-I. Sasaki, *Tetrahedron Comput. Meth.*, 1988, **1**, 27–37.
8 L. Baumer, G. Sala and G. Sello, *Anal. Chim. Acta*, 1990, **235**, 209–214.
9 M. Takahashi, I. Dogane, M. Yoshida, H. Yamachika, T. Takabatake and M. Bersohn, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 436–441.
10 D. Krebsbach, H. Gelernter and S. M. Sieburth, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 595–604 and references cited therein.
11 K. A. Ross and C. R. B. Wright, *Discrete Mathematics*, 4th ed., Prentice-Hall, Upper Saddle River, NJ, 1999.
12 S. H. Bertz, *J. Chem. Soc., Chem. Commun.*, 1984, 218–219.
13 (*a*) S. H. Bertz, *Chem. Commun.*, 2001, 2516–2517; (*b*) S. H. Bertz and T. J. Sommer, *Chem. Commun.*, 1997, 2409–2410.
14 T. S. McDermott, A. A. Mortlock and C. H. Heathcock, *J. Org. Chem.*, 1996, **61**, 700–709.
15 I. L. Shih and J. B. Hendrickson, *J. Chinese Chem. Soc.*, 1997, **44**, 133–140.
16 E. L. M. van Rozendaal, M. A. Ott and H. W. Scheeren, *Recl. Trav. Chim. Pays-Bas*, 1994, **113**, 297–303.
17 K. C. Nicolaou, P. G. Nantermet, H. Ueno, R. K. Guy, E. A. Couladouros and E. J. Sorensen, *J. Am. Chem. Soc.*, 1995, **117**, 624–633.