Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices

Steven H. Bertz*[a] and Toby J. Sommer[b]

[a] *Complexity Study Center, 88 E. Main Street, Suite 220, Mendham, NJ 07945, USA*
[b] *Sphinx Pharmaceuticals, A Division of Eli Lilly and Company, Cambridge, MA 02139, USA*

The enumeration of all possible subgraphs of the molecular graph is the basis for new complexity indices N_S (number of kinds of subgraphs) and N_T (total number of subgraphs) that are useful for synthetic analysis, *e.g.* the determination of topological strategic bonds.

The first bonds to be broken in the retrosynthetic direction (*i.e.* the last formed in the synthetic direction) have been termed 'strategic bonds' by Corey *et al.*[1] Topological strategic bond identifies one selected on the basis of purely mathematical techniques,[2] and heuristic strategic bond on the basis of current methodology.[3] (Note that it is possible for a bond to be both.) The methods introduced here to calculate the topological strategic bonds are based on the enumeration of all possible subgraphs of a molecular graph G.[4] Topological approaches such as graph theory[5] are of interest, as they focus on the fundamental connectivity relationships inherent in a target structure. Setting these in place lies at the heart of synthetic strategy, as emphasized by Hendrickson.[6]

The number of kinds of subgraphs $N_S(G)$ and the total number of subgraphs $N_T(G)$ are new graph invariants,† which are useful indices of the complexity of G.[7,8] The relationship between graphs and subgraphs is analogous to that between structures and substructures, which chemists grasp intuitively. For simplicity, only connected skeletal ('hydrogen-suppressed'[5b]) graphs are considered. Then, the molecular graph of methane is a point, ethane is two points connected by a line (the path of length 1), and propane is the path of length 2.[9,10] The molecular graph of 2-methylpropane (2-Me-propane) has 3 'propane' subgraphs in addition to 3 'ethane' and 4 'methane' subgraphs. The graph itself is counted as a subgraph to preserve mathematical rigor.[5]‡ Consequently, N_T(2-Me-propane) = 11 and N_S(2-Me-propane) = 4.

It is easy to verify that the total number of subgraphs increases monotonically with chain length, branching and cyclization (homologous series), the major criteria of a complexity index.[11,12] Multiple bonds are included by considering multigraphs, *e.g.* ethyne has 1 'ethyne' {●≡●} and 3 'ethene' {●=●} subgraphs in addition to 3 'ethane' and 2 'methane.' Heteroatoms are included in a natural way by finding all possible (labelled) subgraphs of the corresponding labelled molecular graph: propan-2-ol has 1 'propan-2-ol' {(●–)$_2$●–○}, 2 'ethanol' {●–●–○}, 1 'propane' {●–●–●}, 1 'methanol' {●–○}, 2 'ethane' {●–●}, 1 'water' {○} and 3 'methane' {●} subgraphs. Thus, N_T(propan-2-ol) = 11, the same as '2-Me-propane' (above). However, 'propan-2-ol' has more kinds of subgraphs; thus, N_S(propan-2-ol) = 7, whereas N_S(2-Me-propane) = 4. For connectivity N_T is a robust measure, and for the overall complexity N_S is a simple and useful index. When applied to substructures, the latter is also useful in the measurement of molecular diversity.[13]

Fig. 1 illustrates the 1-bond (**2**–**5**), 2-bond (**6**–**15**) and selected 3-bond (**16**–**17**) disconnections for the fused 6-membered ring system **1**,§ which has long been a paramount synthetic problem. The ordered pair N_S, N_T is given in parentheses for each structure. Based on $N_T(G)$, the order of increasing complexity in the 1-bond case is **5** < **4** < **3** < **2**, and in the 2-bond case it is **9**, **15** < **8** < **14** < **12** < **7** < **13** < **6** < **11** < **10**. Where the target is dissected into more than one piece, $N_T(G)$ is calculated for each and summed. Of all the possible 2-bond disconnections that break one bond in each ring, only the best one (**13**) is shown here, and it is not very effective.

In the 1-bond case the largest simplification involves breaking the fusion bond (**1**⇒**5**). This disconnection was eliminated from the heuristic bondset by LHASA rule 4: 'to avoid the formation of rings having greater than seven members during antithetic [retrosynthetic] bond cleavage, any bond common to a pair of bridged or fused primary rings whose envelope is eight-membered or larger cannot be considered strategic.'[1] Progress in eight-membered ring synthesis renders this rule archaic,[14] and also rule 1: 'because of the relative ease of formation of common-sized rings, a strategic bond must be in a four-, five-, six-, or seven-membered 'primary' ring.'[1]

If we restrict the problem to annulation of a 6-ring onto a pre-existing bond, then the greatest retrosynthetic simplification is **1**⇒**4**, which corresponds to the ring closure step of the Robinson Annulation. This is consistent with LHASA rule 2A: 'a strategic bond must be directly attached to another ring (*exo* to another ring). . .'.[1] This is one of the most important rules, and it has a purely topological basis. In contrast, the exception is completely heuristic, *viz*. rule 2B: 'due to the paucity of ring closure methods in which bonds are formed to pre-existing three-membered rings, strategic bonds may not be *exo* to rings of that size.'[1] Rule 2B may change with more research, rule 2A

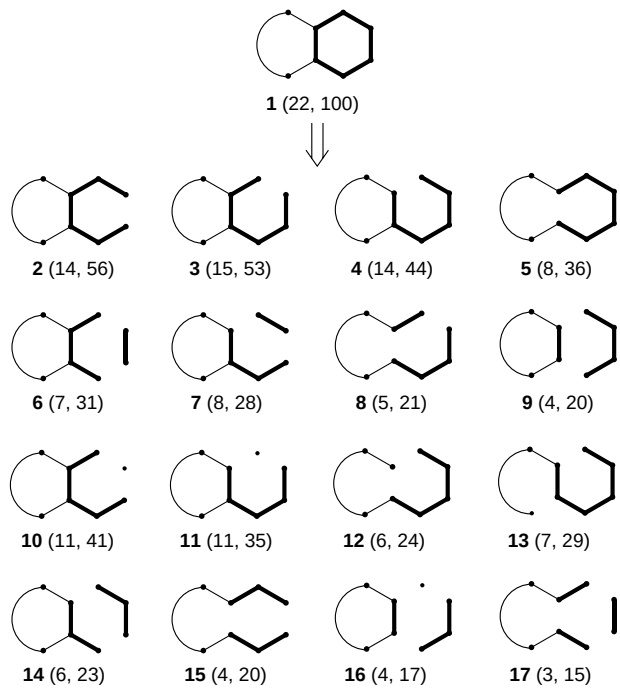

Fig. 1 1-bond (**2**–**5**), 2-bond (**6**–**15**), and selected 3-bond (**16**, **17**) disconnections of the fused six-membered ring system **1**; (N_S, N_T) is given below each structure

will not. Breaking the adjacent ('*exo*') bond minimizes the degrees of the points (valences of the atoms), which is also the basis for rule 3: '. . . strategic bonds should be in the ring (or rings) which exhibits the greatest degree of bridging.'[1]

Of the 2-bond 'Diels–Alder disconnections' resulting in daughters **6**–**9**, the one that affords two pieces of the same complexity (**1**⇒**9**)§ gives the greatest simplification, which is in harmony with the heuristic principle of convergence.[15] It is interesting to note that this 2-bond disconnection is also the overall disconnection for the Robinson Annulation. Furthermore, to minimize the total complexity of the synthetic route by maximizing the symmetry in the synthesis graph (reflexivity),[16] the two pieces should be identical if possible.

Particularly noteworthy is the impressive simplification provided by the 'bis-allyl disconnections' (*e.g.* **1**⇒**14**, **15**), which are comparable to the best Diels–Alder disconnections. This is clearly a reaction that would be worth developing.[17]¶ The 2-bond disconnections that produce an isolated 'methane' (*e.g.* **1**⇒**10**, **11**) are the least efficient of this class. (Synthetic equivalents include CCl_2, CN, CO, *etc*.) However, if the 1-carbon fragment is part of a process that forms other bonds as well, the overall result can be very efficient (next paragraph).

The most dramatic simplifications calculated here involve 3-bond disconnections. A particularly interesting 3-bond process is **1**⇒**16**. It corresponds to the Dötz–Wulff reaction, in which the 1-carbon (CO) and 3-carbon units derive from a metal–carbene complex and the 2-carbon unit comes from an alkyne.[18] Disconnection **1**⇒**17** is realized in the cobalt-mediated alkyne trimerization developed by Vollhardt and coworkers.[19] This reaction renders obsolete LHASA rule 5: 'Bonds within aromatic rings are not considered to have potential strategic character.'[1]

An example of the use of N_S in synthetic analysis is provided by disconnections involving heteroatoms. The values of N_S for pentan-1-ol, butyl methyl ether, and ethyl propyl ether are 11, 13, and 12, respectively. These molecules may be considered to be the daughters from the 1-bond disconnections of oxacyclohexane (N_S = 18) at the C–O α-bond, the adjacent C–C β-bond, and the remote C–C γ-bond, respectively. These disconnections have $-\Delta N_S$(α-bond) = 7, $-\Delta N_S$(β-bond) = 5, and $-\Delta N_S$(γ-bond) = 6. (ΔN_S is the retrosynthetic change in complexity, daughter minus target; $-\Delta N_S$ is the change in complexity in the synthetic direction.) The greatest simplification involves breaking the C–O bond. This is consistent with the LHASA C-heterobond procedure: 'to the set of strategic bonds determined by application of rules 1–6 above is added the collection of bonds in the cyclic network between carbon and O, N and S. . .'.[1] Moreover, N_S also identifies **1**⇒**5** as the most efficient 1-bond and **1**⇒**9**, **15** as the most efficient 2-bond disconnections.

The daughter structures are all subgraphs of the target *T*, as are all the second generation daughters obtained by considering the first generation to be new targets *T*′. Consequently, by generating all possible subgraphs of *T*, we also generate all possible retrosynthetic intermediates and thereby all possible direct synthetic routes. Actual intermediates are usually synthetic equivalents of these retrosynthetic ones and typically contain more atoms, making them more complex. The various routes can then be evaluated by using the principle of minimization of excess complexity[20] or the related principle of maximization of target-relevant complexity.[14]

While our new approach is conceptually simple, for large problems it is important to have a simpler index that parallels $N_T(G)$, and the number of pairs of adjacent bonds η ('propane' subgraphs) does this well. Only one pair of structures (**12**, **14**) is ordered differently by $N_T(G)$ and η, which confirms the validity of our previous work on synthetic analysis.[2,8,11,20] (There are more degeneracies in the simpler index.) The number of spanning trees has also been proposed to be the 'complexity' of a planar graph,[21] but it is not useful for graphs without rings. A 'middle way' might be to count all paths[10] or walks.[22] Moreover, paths, walks or $N_T(G)$ include long-range 'interactions', which can be critically important, as emphasized by Lehn in his MI_2 approach to complexity.[23]

In conclusion, the total number of subgraphs $N_T(G)$ is a robust measure of connectivity ('topological complexity'), and the number of kinds of subgraphs $N_S(G)$ is a simple measure of the overall complexity of a molecular graph *G*. These new indices are useful for understanding strategic bond disconnections. Some of them correspond to powerful known reactions, while others suggest new reactions that would efficiently increase molecular complexity. The fact that we can derive certain of the 'heuristic' rules for strategic bonds mathematically establishes that they have a purely topological basis and suggests that topology may be the 'unseen hand' of synthesis. The other rules are revealed to be principally a function of the state of the art, for which our approach provides objective benchmarks.‖

We thank V. Shcherbukhin (Astra Hässle) and W. F. Wright (AT&T) for interesting discussions and useful references.

Footnotes and References

* E-mail: sbertz@ispcorp.com

† Invariants are independent of isomorphism, *i.e.* how the graph is drawn.[5]

‡ By including the graph *G* as a subgraph of itself, $N_T(G)$ is given by simple formulas, *e.g.* $N_T(K_{1,n}) = 2^n + n$ for the star graphs and $N_T(P_{n-1}) = n(n + 1)/2$ for the *n*-alkanes.

§ Strictly speaking, the new 6-ring is annulated onto the middle bond of butane, as the arcs are not included in the subgraphs. Functionality need not be included at the strategic level; *cf*. the antepenultimate paragraph for tactical considerations.

¶ An intramolecular example has been reported;[17] however, it does not increase complexity significantly, as it is not a construction reaction.[6]

‖ A list of all the subgraphs of **1**–**17** is available upon request.*

1 E. J. Corey, W. J. Howe, H. W. Orf, D. A. Pensak and G. Petersson, *J. Am. Chem. Soc*., 1975, **97**, 6116.
2 S. H. Bertz and T. J. Sommer, in *Organic Synthesis: Theory and Applications*, ed. T. Hudlicky, JAI Press, New York, 1993, vol. 2, p. 67.
3 W.-D. Ihlenfeldt and J. Gasteiger, *Angew. Chem., Int. Ed. Engl*., 1995, **34**, 2613.
4 S. H. Bertz and W. C. Herndon, in *Artificial Intelligence Applications in Chemistry*, ed. T. H. Pierce and B. A. Hohne, ACS, Washington, D.C., 1986, p. 169.
5 (*a*) F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969; (*b*) N. Trinajstić, *Chemical Graph Theory*, 2nd edn., CRC Press, Boca Raton, FL, 1992.
6 J. B. Hendrickson, *Angew. Chem., Int. Ed. Engl*., 1990, **29**, 1286.
7 S. H. Bertz, *J. Chem. Soc., Chem. Commun*., 1981, 818.
8 S. H. Bertz, *J. Am. Chem. Soc*., 1981, **103**, 3599.
9 M. Gordon and J. W. Kennedy, *J. Chem. Soc., Faraday Trans. 2*, 1973, **69**, 484.
10 M. Randić, *J. Chem. Inf. Comput. Sci*., 1996, **36**, 1092.
11 S. H. Bertz, in *Chemical Applications of Topology and Graph Theory*, ed. R. B. King, Elsevier, Amsterdam, 1983, p. 206.
12 S. H. Bertz, *Discrete Appl. Math*., 1988, **19**, 65.
13 C. Cheng, G. Maggiora, M. Lajiness and M. Johnson, *J. Chem. Inf. Comput. Sci*., 1996, **36**, 909.
14 P. A. Wender and B. L. Miller, ref. 2, p. 27.
15 L. Velluz, J. Valls and G. Nominé, *Angew. Chem., Int. Ed. Engl*., 1965, **4**, 181.
16 S. H. Bertz, *J. Chem. Soc., Chem. Commun*., 1984, 218.
17 W. W. Win, K. G. Grohmann and L. Todaro, *J. Org. Chem*., 1994, **59**, 2803.
18 K. H. Dötz, *Angew. Chem., Int. Ed. Engl*., 1975, **14**, 644; W. D. Wulff, in *Comprehensive Organometallic Chemistry*, Pergamon, Oxford, 1995, vol. 12, p. 469.
19 J. K. Cammack, S. Jalisatgi, A. J. Matzger, A. Negrón and K. P. C. Vollhardt, *J. Org. Chem*., 1996, **61**, 4798.
20 S. H. Bertz, *J. Am. Chem. Soc*., 1982, **104**, 5801.
21 E. C. Kirby, R. B. Mallion and P. Pollak, *Mol. Phys*., 1994, **83**, 599.
22 G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci*., 1993, **33**, 683.
23 J.-M. Lehn, *Supramolecular Chemistry: Concepts and Perspectives*, VCH, Weinheim, 1995, p. 201.

Received in Corvallis, OR, USA; 12th September 1997; 7/06192G