Search for Complexity Generating Chemical Transformations by Combining Connectivity Analysis and Cascade Transformation Patterns

Grażyna Nowak*,† and Grzegorz Fic‡

Departments of Physical Chemistry and Computer Chemistry, Faculty of Chemistry, Rzeszów University of Technology, Al. Powstańców Warszawy 6, 35-959 Rzeszów, Poland

Received April 15, 2010

Retrosynthetic analysis involved in a backward search for strategic disconnections is still the most powerful strategy, recently advanced by topology-based complexity estimation, for discovering the shortest sequences of transformations and chemical synthesis planning. Therein, we propose an alternative strategy that combines backward and forward search embodied within a mathematical model of generating chemical transformations. The backward reasoning involves a new concept of the strategic bond tree for alternative multibond disconnections of a target molecule. In the forward direction, each combination of the resulted structural fragments is examined for reconstruction of the target structure by means of biomimetic transformation patterns that describe one-pot multibond forming reactions. The algorithm has been implemented into the CSB system, and its performance is illustrated by examples of published complex molecule syntheses for comparison and analysis. This paper describes the strategy for discovering the shortest synthetic pathways based on the multibond forming cascade transformations for application in synthesis design and generating synthetically accessible product libraries.

INTRODUCTION

Past years demonstrate continuing progress in the area of computer-assisted planning of organic synthesis (CAOS) to help chemists in discovering new strategies for a synthesis of valuable synthetic targets. This challenging task consists of the search of a huge chemical space resulting from all possible chemical transformations and species generated combinatorially, either in the retrosynthetic fashion or in the forward synthetic direction.[1−5] As structural target increase in size, a generated retrosynthetic tree or reaction network grows exponentially making unbiased exhaustive search impossible in real-time. This is a critical limitation particularly when mathematical-logical models are involved in generating the transformations to enable the search and discovering novel synthetic strategies or novel functional molecules for use in drug discovery, material science, and chemical biology applications. In an attempt to overcome this difficulty, a number of search strategies have been developed, capable of leading the search along potentially promising pathways. Their development over the year was driven by the rapidly growing complexity of synthesized compounds, with unusual structural architectures, shapes, and properties, as well as demand for design of new, more practical, step-economical, and environmentally friendly syntheses.[6]

The directed search through the space of possible transformations can be carried out backward from a target product applying chemical transformations in reverse or forward generating the reaction products given starting materials. The former is based on Corey's concept of retrosynthetic analysis of disconnections of a target structure during the tree search.[7] Following this concept, only those bonds are disconnected for converting the target into its constituent components that provide a greater decrease in topological or synthetic complexity. This guiding principle has provided a uniform logical basis for several disconnective strategies, recently advanced by using the tools of graph theory or topology. The most important of these are formal semiquantitative concepts based on an analysis of molecular complexity, estimation of similarity or potential symmetry. Strategies of this kind have emerged as one of the most promising approaches toward the "ideal synthesis", that is, achieving a maximal increase of structural complexity in a minimal number of steps.[8] They are useful to construct in a stepwised manner target-oriented complexity-generating synthetic pathways as well as to build and filtering combinatorial libraries of complex structural analogues that could be synthetically accessible.[9]

Alternative method of exploring a chemical space is a forward reasoning that involves the search for possible ways of combining/assembling the starting substrates or building blocks to predict the reaction products. The search is optimized with strategies controlling the bond formations between the potential reaction centers located in the reaction substrates. They include knowledge-based rules describing transformation of reaction centers, mathematical models for evaluation of chemical reactivity, or abstract graph-based rules describing chemical reactions in the context of topological similarity or complexity.[10−14]

This search technique can serve as a natural complement to backward tree search for synthetic precursors (or substrates), providing the possibilities for testing the feasibility of reactions converting those precursors to the desired target

* To whom correspondence should be addressed. Phone: (+48)17 8651810. E-mail: gnowak@prz.edu.pl.

† Department of Physical Chemistry.

‡ Department of Computer Chemistry.

10.1021/ci100146n
Published on Web 08/03/2010

product. Therefore, some of the CAOS systems embody both the backward and forward reasoning to effectively control the combinatorial space of chemical transformations.[15,16] Moreover, generating sequences of transformations in forward direction offers opportunities to search for collections of compounds that cover broader range of structural possibilities. This approach combined with diversity estimation is proposed as a general strategy for generating skeletally diverse collections of molecules that could be accessible by diversity-oriented syntheses (DOS).[9]

The strategies introduced above represent the type of structure-based exploration of chemical transformations (by disassembly/assembly or cutting/forming specific bonds between molecular subgraphs). An alternative is a function-based approach to search for transformations generating libraries of compounds with specific structural and biological properties. This innovative concept, known as function-oriented synthesis (FOS) has been introduced by Wender.[17] Recently, Schürer et al. described the computer implementation of this strategy for designing virtual syntheses of medicinally relevant molecules. The search for transformation sequences involves the iterative generation of transformations coupled with QSAR models for filtering generated products based on predicted activity.[18]

We propose here a novel strategy of the directed search for complexity generating transformations that can be useful both in CAOS area and in medicinal or biological chemistry applications to explore the space of synthetically accessible and structurally complex chemicals. It combines two complementary models of chemistry: a mathematical chemistry model with the concept of retrosynthesis and two modes of reasoning, backward and forward. The approach aims at finding the shortest sequences of transformations for a given target product in a 2-fold manner, by recognizing the alternative bond sets for strategic disconnections (backward) and by selecting those bond-sets that can be reconstructed (forward) in the target by one-pot cascade transformations. The first, backward search consists in generating the strategic bond tree based on analysis of molecular connectivity and breaking up the target molecule. In second forward stage, each set of the resulting structural units is tested for reassembly of the target molecule with biomimetic transformation patterns describing various one-pot-multistep reaction sequences.

RETROSYNTHETIC SEARCH FOR DISCONNECTIONS. A CONCEPTION OF THE STRATEGIC BOND TREE

A number of approaches have been developed for the retrosynthetic analysis of disconnections to optimize the search tree. According to Corey's logic and heuristic approach, useful bond disconnections are those that lead to the most accessible or simplest synthesis precursors. From this principle two the most successful strategies have been derived for finding strategic disconnections during the tree search. One is the recognition of guiding patterns inside the target structure corresponding to potentially available starting materials, building blocks, functional substructures. The second, complexity-based search consists in the recognition of disconnections giving the greatest decrease in structural complexity. Both strategies will allow conducting the search along the shortest and simplest synthetic pathways, the features of prime importance in devising innovative efficient syntheses.

Many CAOS computer programs make possible to find the disconnections by locating the patterns of the starting materials within the structure of the synthetic target or intermediate.[19−21] WODCA provides substructure matching algorithm and different structural similarity criteria involved to search for appropriate starting materials.[22] LHASA uses the quantitative idea of synthetic proximity between two structures and enables to evaluate where in the target the starting material matched best.[23] The SST program employs hierarchical search and abstract subgraph definition to select the candidates from starting material library.[24,25] The CHIRON program is capable to recognize the chiral templates of starting materials that are apparent or hidden inside the target structure.[26,27] Here, the patterns recognition involves three options, carbon skeleton, functionality, and stereochemistry. Similar approach is embedded in the SESAM program that searches for nonobvious structural identities between the starting materials and targets described by their skeletons.[28] It can locate any substructure inside the complex synthetic target without the use of the starting materials database.

The second strategy to search for strategic disconnections is based on an analysis of molecular complexity during the iterative generation of transformations. In the simplest case, this strategy involves symmetrical disconnections that generate fragments of similar size and complexity, as required for maximal simplification and optimal convergent assemblage of the target structure. The convergency criterion was first introduced by Hendrickson's and implemented in the SYNGEN program for the strategic bonds selection.[29] Recently, this kind of the symmetry-based strategy has been significantly enhanced by two independed works. The first one, proposed by Bertz and Sommer consists in recognizing disjoint isomorphic substructures to enable the search for reflexive or two-directional routes.[30] In the second, described by Vismara et al., the detection of potential symmetry is formalized in graph-theoretical terms as the maximum symmetrical split of a molecular graph.[31] Other conception for symmetrical disconnections have been proposed by Tanaka et al., based on quantification of the bond centrality (BC) that increases toward the center of a molecule.[32]

For complex polycyclic systems, Corey has proposed the heuristics strategy to perceive strategic disconnections taking into account the synthetic difficulty.[33] Many other synthesis planning systems such as SYNCHEM, AIPHOS, SECS, WODCA, and LILITH have embedded this kind of knowledge-guided analysis of disconnections by using their original heuristics;[20,34,35] for reviews see refs 1 and 2. Although the directed heuristic search determines in many respects the reliability and practicality of the generated solutions, they will depend on the current state of knowledge incorporated in a programs knowledge base, which is constantly evolving.

An alternative is a formal approach based on the Bertz's idea to use the graph topology for calculating the complexity and complexity changes resulting from alternative disconnections.[36] Several quantitative and semiquantitative models have been offered to rank disconnections for use in synthesis planning. Bertz and Sommer proposed a series of complexity indices, based on constructing and counting subgraphs of a molecular graph. The known examples are the number of kinds of substructures, NS, and the total number of substructures, NT.[37] Rucker and Rucker developed another topological concept of molecular complexity by counting all walks of all lengths in a graph or molecule.[38] Among these kinds of graph-theoretical invariants are the total walk count, twc, and walk complexity, wcx, that are computationally simpler with respect to Bertz's measure.[39,40] Apart from topological characterization and discrimination of chemical structures, these kinds of formal theoretical measures have been applied to identifying the strategic disconnections in synthesis planning, indicating some degree of correspondence with LHASA heuristic rules.[41−43] It is important to mention, that mathematically rigorous method to strategic disconnections takes into account the topological complexity elements, without accounting for synthetic complexity/feasibility. Heuristic strategic rules assume structural features that affect the synthetic difficulty but depend on the state-of-the-art. Whitlock has proposed an empirical metric of structural complexity corresponding with chemical concept.[44] It is simply calculated from complexity values of locally (arbitrary) defined structural features, such as the number of heteroatoms, unsaturations, chiral centers, rings. According to the author, this chemical intuition measure was devised to provide a simple way of analyzing and comparing complex multistep chemical transformations by measuring complexity intermediates and constructing plots of complexity versus step. Barone and Chanon have proposed an extension of this approach by considering the connectivities of the atoms and the size of the rings.[45,46]

The use of the topological or structural complexity measures provides the most powerful tool to optimize the search for complexity-building synthetic pathways in terms of the number of steps and convergency. However, this strategy requires disconnecting of each bond in the target structure and generating all possible precursor candidates to calculate the corresponding changes of molecular complexities.

In this paper, we propose another approach that aims at finding strategic bonds without exhaustive transformation of the target structure at the successive levels. It is based on a new concept of a strategic bond tree that contains all the recognized strategic bonds represented by nodes. Identification of bonds and construction of the strategic bond tree is based on hierarchical analysis of molecular connectivity to partitioning the atoms and bonds of a product molecule. The bonds identified at subsequent levels of the tree are systematically removed from the initial structure decreasing its molecular complexity. Thus, the rank of a bond (node) corresponds to the tree level. The bonds (nodes) located on the same level get the same rank. A root node (first level) of the tree includes the best strategic bond identified by taking into consideration the full structure of the target product (if there are more bonds identified at first level then the respective tree for each one will be generated). The second level will contain bonds identified after removing from the initial structure the bonds belonging to the first level. The third level of the tree will contain bonds identified after removing from the initial structure all bonds belonging to the first and second levels, etc. In general, the bonds at LEV level will be identified after removing from the initial structure the combination of bonds belonging to levels from 1 to LEV-1 (they are placed on the patch going from 1 to LEV-1).

GENERATION OF THE STRATEGIC BOND TREE. IDENTIFICATION OF THE ALTERNATIVE STRATEGIC BOND SETS

The generation of the strategic bond tree and ranking of bonds is based on the partitioning of atoms into classes. The class of atom *i* (class identifier) is obtained by considering the number of non-hydrogen atoms directly connected to atom *i*, and the number of atoms connected to neighboring atoms. The partitioning is performed iteratively. Obtained in first iteration increments of atom correspond to the number of nearest neighbors (degree of atom), like in Morgan algorithm.[47] The partition obtained in one iteration is used for distinguishing atoms in the subsequent iteration. At the beginning, the first class will include atoms with the greatest number of neighbors. In subsequent iteration the atom partitioning is continued further, and lower (preferred) class will include atoms with the neighboring atom (at first sphere) belonging to lower class (assigned at the previous iteration). Unlike the original Morgan algorithm (in which the atoms are distinguished in result of the summation of increment values of neighboring atoms), here the atoms are classified by matching lists of the class numbers. If after this iteration, the partitioning of atoms is incomplete then their second and further neighborhood is taken into account. The iteration is continued until the number of different classes at one iteration is equal or smaller than that at the previous one. The strategic bond is that connecting atoms belonging to the lowest classes. If there are more such bonds, then all of them are classified as strategic ones. The algorithm is explained in more detail below.

Program SBTG: Construction of a Strategic Bond Tree (SBT). *Input Data:* Molecular graph (list of atoms LA, list of bonds LB), LEVmax, maximum level of a tree (or the maximum depth of the tree or the path length from the root of the tree).

SBTG_1. LEV = 1 (initiation of first level of the SBT). Executing the **SBR procedure** presented below, recognizing strategic bonds at first level of the tree SBT.

SBTG_2. LEV = 2 (initiation of second level of the SBT).

SBTG_3. If LEV > LEVmax, then **SBTG_END** (the SBT construction is completed).

SBTG_4. Selecting subsequent node N at the LEV-1. If there is no other node, then LEV = LEV + 1 (initiation of the successive level of the SBT) and repeat SBTG_3.

SBTG_5. Finding subsequent combination of the strategic bonds at levels from 1 to LEV-1 (located at the path in the SBT starting in the initial node N at level LEV-1, and ending in one of the nodes of the first level). If there is no new combination, then repeat SBTG_4.

SBTG_6. Removing from the structure the bonds belonging to the combination generated in the preceding step (SBTG_5).

SBTG_7. Executing of the **SBR procedure**, recognizing the strategic bond(s) in the structure constructed in the preceding step (SBTG_6). This bond(s) is located in the node(s) at level LEV continuing the path generated in the SBTG_5. Repeat SBTG_4.

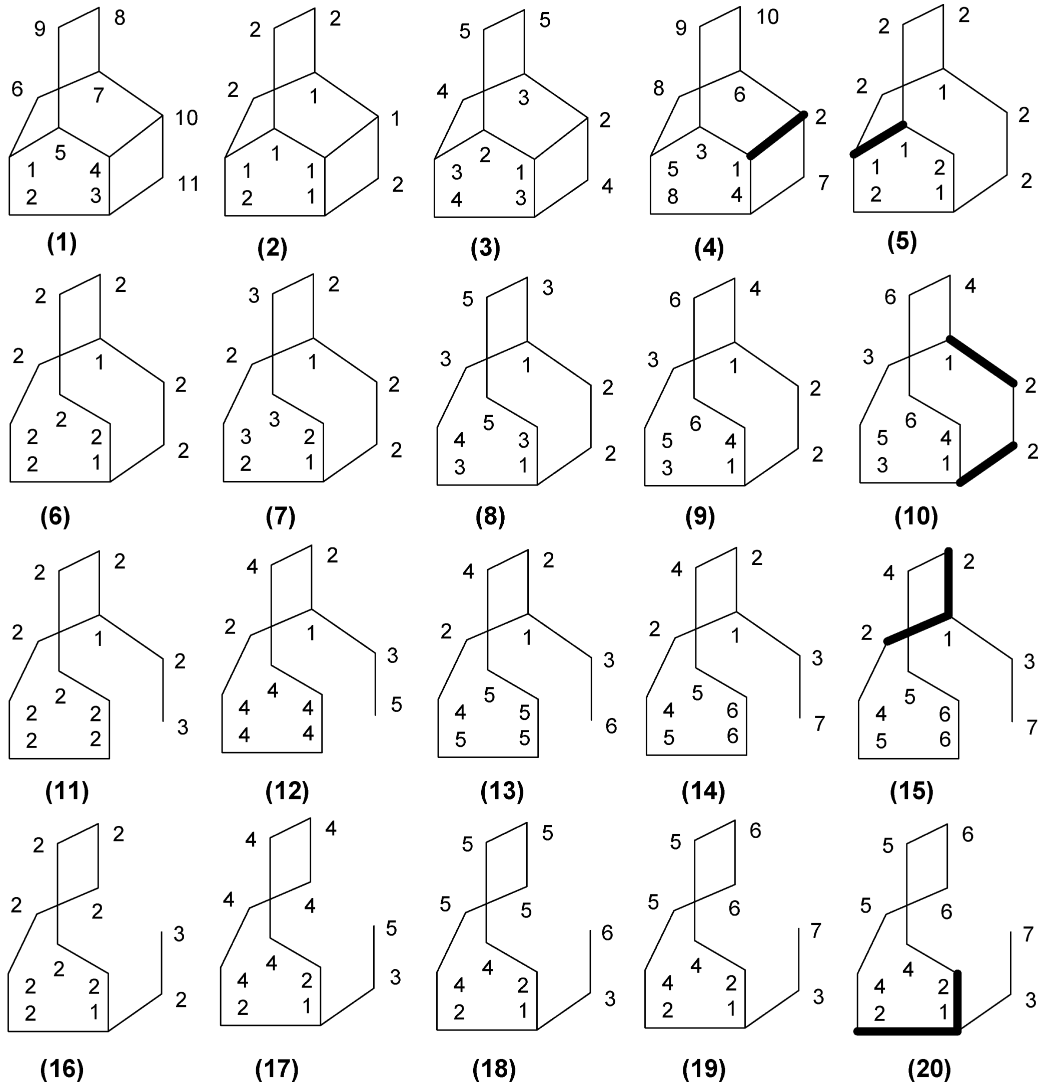

Figure 1. Generation of the strategic bond tree for the skeleton of the endiandric acid C ring system. The strategic bonds recognized in subsequent levels are denoted in bold.

Procedure SBR: Recognition of Strategic Bonds. *SBR_1.* Partitioning of atoms into (topological equivalence) classes according to the number of neighboring atoms: class (1) contains the atoms with the largest number of neighboring atoms; the highest class contains the atoms with the smallest number of neighboring atoms.

SBR_2. Counting of a neighbor vector NVi = {ci, NCV1, NCV2, NCV3, ..., NCVn} for each atom *i*, where ci is the class number of the atom *i* assigned at the previous iteration and NCV1 ..., NCVn are class numbers of all nearest neighboring atoms assigned at the previous iteration. [The elements NCV are sorted in ascending order, for example, the NV of the form {2,1,1,2} represents the atom belonging to the class 2 and linked to the atoms belonging to the classes 1, 1, 2, respectively (according to the previous iteration)].

SBR_3. New partitioning of atoms: if NVj < NVk, then the class number of atom *j* is lower than the class number of the atom *k*; if NVj = NVk, then the atoms *j* and *k* belong to the same class. [The priority or equivalence of two vectors is assigned by comparing the vectors element by element, for example {2,1,2,2}is greater than {2,1,1,2}].

SBR_4. Counting a bond vector BVi = {c1, c2) for each bond *i*, where c1 and c2 are the current class numbers for atoms of the bond *i* and c1 ≤ c2.

SBR_5. Examination of first condition to complete the iteration process: If there is only one bond with BV = BVmin then the bond is strategic. [The contents of vectors BV are compared element by element as in SBR_3); BVmin = min{BV1, BV2, ..., BVn}]. **SBR_END**, iteration process is completed.

SBR_6. Examination of second condition to complete the iteration process: comparison of the number of different classes for two successive iterations. If the number of different classes remains unchanged, then the entire bond with BV = BVmin are strategic.**SBR_END**.

SBR_7. Return to SBR_2 and initiate the subsequent iteration.

APPLICATION OF THE PRESENTED ALGORITHM. GENERATION OF THE STRATEGIC BOND TREE FOR SKELETON OF THE ENDIANDRIC ACID C

In Figure 1, there is an example illustrating the generation of the strategic bond tree SBT for the endiandric acid C (more precisely, the skeleton of its ring system), the complex natural product possessing a broad array of biological activities. Part 1 shows the atom numbering (in arbitrary order). Part 2 gives initial partitioning of atoms on two classes. Here, class 1 contains atoms connected with three

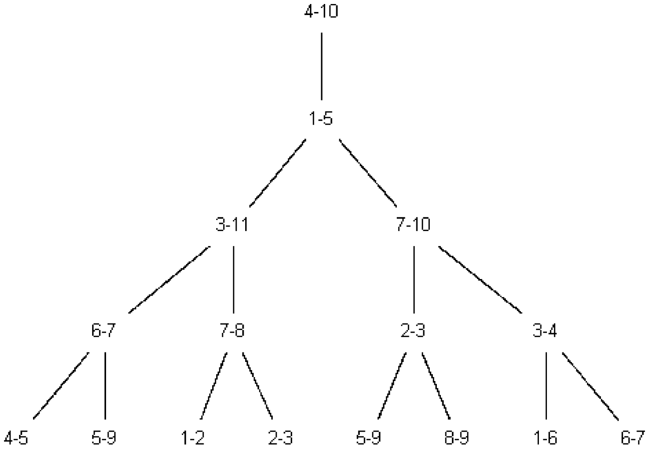

Figure 2. Strategic bond tree generated by the SBTG algorithm.

neighboring atoms and class 2 contains atoms connected with two neighboring atoms. This partitioning is continued in second iteration (part 3) resulting in the formation of five classes. Now, the atoms that in the previous iteration belong to class 1 are partitioned on three classes. The distinction between the atoms was achieved by comparing their neighbor vectors NV: NV = {1,1,2,2} for atoms 1, 3, and 7, NV={1,1,1,1} for atom 4, and NV={1,1,1,2} for atoms 5 and 10. According to the assigned priority: {1,1,1,1} < {1,1,1,2}< {1,1,2,2}, atom 4 belongs now to class 1, atoms 5 and 10 belong to class 2, and atoms 1, 3, 7 to class 3. After this iteration two bonds: 4−5 and 4−10 are topologically equivalent (possesses identical and the lowest bond vectors BV = {1,2}). Thus, the subsequent iteration (part 4) is needed. In this iteration one of the bonds, 4−10 gets the smallest vector BV = {1,2}. As a result, the bond 4−10 is identified as the strategic one. It is located in the root node of the SBT. Now, this bond is removed from the input structure, and a second level of the tree nodes (strategic bonds) is generated. As shown in part 5, a correct partitioning is achieved at first iteration, because only one bond 1−5 gets minimal vector BV = {1,1}. Thus, this bond is located at the second level of the tree.

The procedure for recognizing the bonds at the third level is illustrated by parts 6 and 10. It starts from deletions in the input structure of two bonds: 4−10 and 1−5. After 5 iterations, the second condition (SBR_6, taking the number of classes in two subsequent iterations), is satisfied to complete the generation of this level. In result, two bonds; 3−11 and 7−10 are recognized as strategically equivalent because both the assigned vectors BV are the same: BV = {1,2}. The identification of the strategic bonds located at the fourth level of the SBT is shown in parts from 11 to 20. The iterations 11−15 concern the structure obtained after removal the bonds 4−10, 1−5, and 3−11, and iteration 16−20 concern the structure obtained after after removal the bonds 4−10, 1−5, and 7−10. The five levels of the SBT generated by this procedure are shown in Figure 2. The tree will depend on the structure suppression method (Figure 3). For this purpose, three methods are involved into the CSB: in the first (whose the SBT is presented in Figure 2), only the skeleton of ring system is considered for the SBT generation; the second takes into account the skeleton of the ring system with substituent positions, and in the third, only hydrogen atoms are removed from the input structure (Figure 3).

The presented method offers a simple and quick search for the retrosynthetically best bonds in the target product structure. Given the strategic bond tree (STR) it enables to define alternative bond-sets for the generation of optimal synthetic pathways. Each bond-set corresponds to the succession of nodes in a path from the root (the bond with the highest priority) to some other node at level n (the bond with the lowest priority). If there are more bonds at the considered level then equivalent bond-sets can be defined. A value of n is equal to the number of bonds defined in the transformation pattern applied for recreating these bonds in the target product (see next chapter). For example, the application to the structure of endiandric acid C (the corresponding STR is shown in Figure 2) of the transformation pattern operating on three bonds (broken/formed during the reaction simulation), involves two alternative ways of the bond disconnections: 4−10, 1−5, 3−11, and 4−10, 1−5, 7−10.

FORWARD SEARCH FOR MULTI-BOND-FORMING TRANSFORMATIONS. GENERATION OF SYNTHETIC PATHWAYS

The approach we propose to search for the shortest sequences of transformations to the target product offers us two possibilities: the knowledge-independent or unbiased search for different types of bond-disconnections and the knowledge-based tree search to test the bonds for the reconstruction. The former, by analysis of different disconnection patterns stored in the strategic bond tree SBT, provides a tool for discovering novel concepts for synthetic strategies. The latter, by the application to preasumed disconnections the library of cascade transformations can be used to verify their synthetic accessibility in the product formation. Both the backward bond-disconnection and the forward bond-formation operate on multibond patterns.

In contrast to other topology-based disconnective methods which analyze each bond individually (calculating the molecular complexity before and after disconnection), our strategy provides alternative multibond disconnection patterns without exhaustive transformation of the target structure. Similar approach based on a global perception of disconnections have proposed Barone and Chanon.[48] Their CONAN system based on connectivity analysis includes the breadth of options for different disconnection types, as for example the traditional one-bond, two-bond disconnections, formal [$x + y$] disconnections for rings, searching for linear precursors for polycyclic systems and others.

The second additional possibility given in our CSB system is re-examination of the preassumed disconnection patterns by tree search procedure conducted forward. The distinct feature of this procedure is the use of the multibond forming/braking transformations that imitate cascade reactions. These bioinspired processes provide a significant increase in molecular complexity by combining a series of consecutive bond-forming reactions in one synthetic operation. By analogy, their biomimetic transformation patterns enable us to form several bonds and construct considerable part of the product structure at one iteration. In contrast to other tree search strategies with transformations breaking/forming one or two

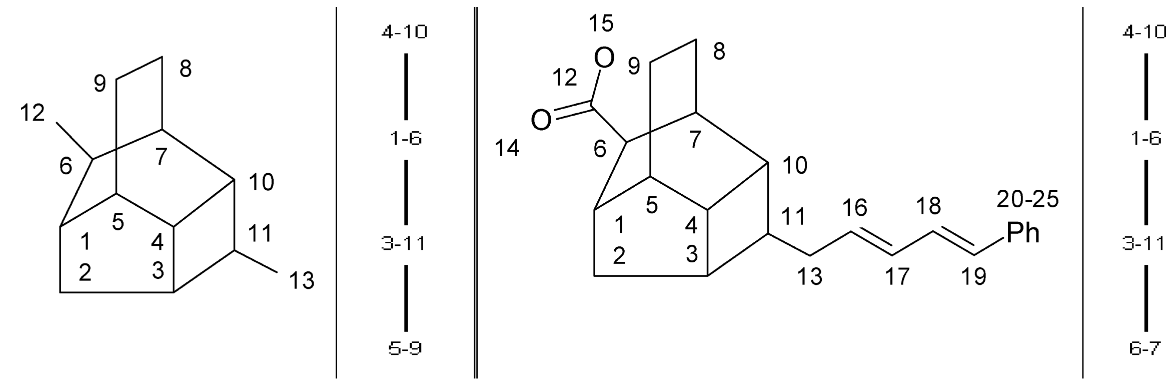

Figure 3. Alternative strategic bond trees generated for the endiandric acid C: left, taking as an input the carbon skeleton with substituent positions; right, hydrogen-suppressed structure of endiandric acid C.

bonds conducted over many levels of the tree, our strategy operating on multibond disconnection/formation patterns generates the same tree at two or three levels. The multibond-forming transformations are automatically created and stored in the library of the generator of the CSB system. Their concept and the learning procedure were recently described in ref 49. In this section, we will only discuss their application for testing the bond-formations in the product structure by means of cascade reactions. The idea of using the cascade transformations was also implemented in HOLOWin system to search for synthetic strategies.[50] In contrast to the Holotransforms that describe individual reaction steps in the cascade transformations that are generated in successive iterations, the proposed by us the transformation patterns combine several reaction steps to be generated in one iteration. In Figure 4, there are examples of cascade transformation patterns that will be used in the tree search process to test the formation of the skeleton of the endiandric acid C. As we can see all the functionality required to generate at one iteration the resulting products of this two- or three-step reactions are specified on the left-hand side of the transformation rules.

The tree search procedure starts from selecting with the strategic bond tree SBT a given disconnection pattern, that is, combination of the strategic bonds represented by nodes located at the path from the root to some node *n*. These bonds are removed from the target structure and resulting structural units can be seen on the screen. Figure 5 shows the alternative solutions obtained for the endiandric acid C core after removal equivalent disconnection patterns with four and five strategic bonds, in accordance with the SBT.

The resulting structural fragments will be tested for reassembly of the target molecule with cascade transformation patterns describing various one-pot-multistep reaction sequences. The general algorithm for the generation of the synthetic tree GST consists in folowing steps:

GST_1. Construct the strategic bond tree SBT for the target product.

GST_2. Select the level of the strategic bond tree SBT; select the respective subset of cascade transformations (for example, CASCADE_4 is the subset of transformations operating on the four strategic bonds disconnected/formed in the target product). If all levels of the SBT were tested then STOP.

GST_3. Select the pattern of bond disconnections (a combination of the strategic bond) from the SBT. If all disconnection patterns were tested then GOTO GST2.

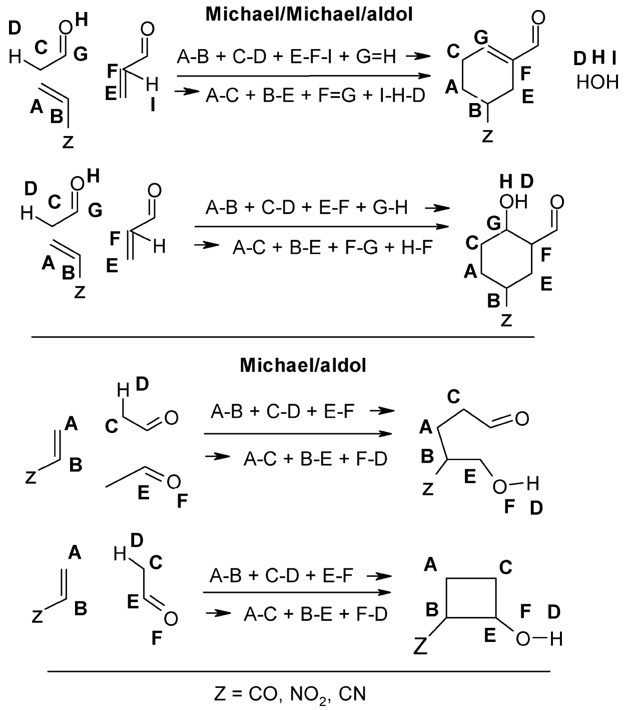

Figure 4. Examples of cascade transformation patterns. These are generalized reactions in a form of active subgraphs in substrate/product molecular graphs, automatically derived from reaction databases or user training sets. Active subgraphs describe both the reaction centers, as for example bonds broken/formed in the substrate/product structures (denoted here as A−B, C−D, etc.), and unchanged functional groups (Z, C=O in the first transformarion), which influence a reaction. Each transformation is stored in a form of the reduced reaction list that provides topologically unique numbers for nodes in molecular graphs.

GST_4. Remove the strategic bonds and extract the structural fragments with open sites.

GST_5. Complete the structural fragments by adding successive combinations of functional groups (specified for a given transformation). If all combination were tested then GOTO GST3.

GST_6. Apply all transformations from the chosen CASCADE subset; compare the generated structures with the target product. If the target structure has been achieved, store successive set of intermediates in the synthetic tree. If all transformations were tested then GOTO GST5.

The structural fragments extracted after removal the strategic bond-set from the target product will contain open sites and cannot be immediately used in the reaction

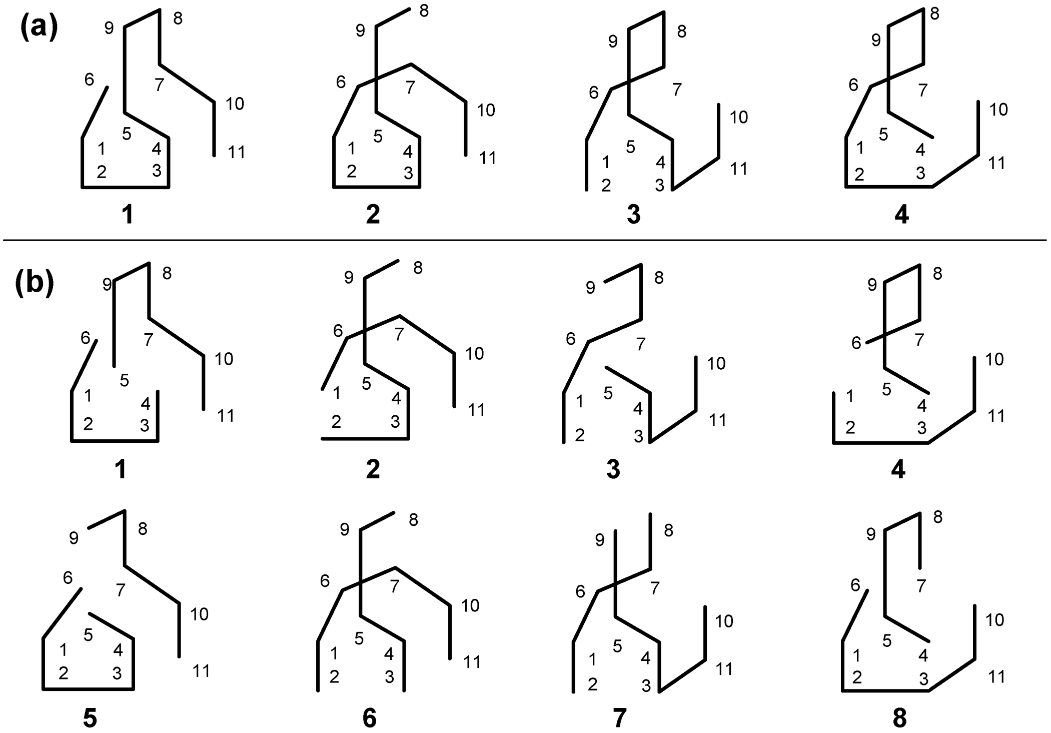

Figure 5. Structural fragments obtained after removal of (a) four strategic bonds and (b) five strategic bonds from skeleton of the endiandric acid C. The candidate disconnection pattern a1 represent the bond formed by using a cascade of perycyclic reactions in biosynthetic pathway and in the biomimetic synthesis reported by Nicolaou's group.[51]

generation process. Therefore the GST_5 stage consists in the addition of the respective functional groups for the construction of reactant or substrate candidates to be correctly matched with the cascade transformations in the library of the reaction generator. Utilized in this stage, the CASCADE subset contains for each transformation, a file of functional groups implying the recognition in substrate (reactant) structures of the sufficient reaction sites for generating the product structures. All combination of these groups attached to open sites are examined to obtain reactant(s) matching to the requirements of the transformations. Each formed reactant cosidered as a potential substrate is compared with a catalog of chemicals to search for real available starting materials. For the purpose of testing the performance of the presented method, we have used a set of structures from the PubChem database (http://pubchem.ncbi.nlm.nih.gov/). For structural fragment 1 shown in Figure 5a (obtained after removal 4 strategic bonds from the skeleton of the endiandric acid C), only 7 building blocks have been found. They have functionality compatible with the transformation pattern describing three-step cascade Michael/Michael/aldol. Applying this transformation to two of them (Figure 6a, example 1 and 2) allows us to build an essential part of the endiandric acid C skeleton with complete structural features for the next transformation (aldol) that finalizes the construction of the endiandric acid C skeleton. For pairs of structural fragments obtained after removal 5 strategic bonds (as for example the pair 5 in Figure 5b), the resulting sets of building blocks contained several hundred structures with desired functionality, and in this case additional filtering procedure was needed for their refinement. To test the reconstruction of the specified bond-set in the given product, we selected pairs of structurally simplest compounds. Some of the generated transformations that have allowed to reach the desired product are shown in Figure 6b. Each of the pathways contains tree stages required to construct the five skeletal bonds. The two first pathways proceed through the Michael reaction followed by two Michael/aldol cascades to achieve the product skeleton. In the third route, the pair of the starting materials is subjected to aldol/aldol condensation followed by Michael/Michael/Michael cascade.

The strategy optimizes the search through the space of combinatorially generated transformations and allows for rapid identification of the complexity-building cascade transformations for construction of the target product from the available starting materials.

CONCLUSIONS

Several strategical concepts and tactics have been developed over the years to optimize the search for efficient synthetic pathways during the combinatorially generated chemical transformations. One of the most promising is based on an analysis of the complexity during the iterative generation of transformations, taking into account of both the synthetic difficulty and molecular topology. This strategy allows conducting the search along the shortest and simplest synthetic pathways, the features of prime importance for construction of complex molecules from simple precursors. The latest trend in this area is focused on developing quantitative and semiquantitative models of molecular complexity to rank the bond disconnections and prioritize the retrosynthetic transformations.

Continuing this idea, we introduced a new simple method for the identification of strategic disconnections based on an analysis of molecular connectivity and the concept of the strategic bond tree. In contrast to other disconnective strategies based on one-bond or two-bonds disconnections

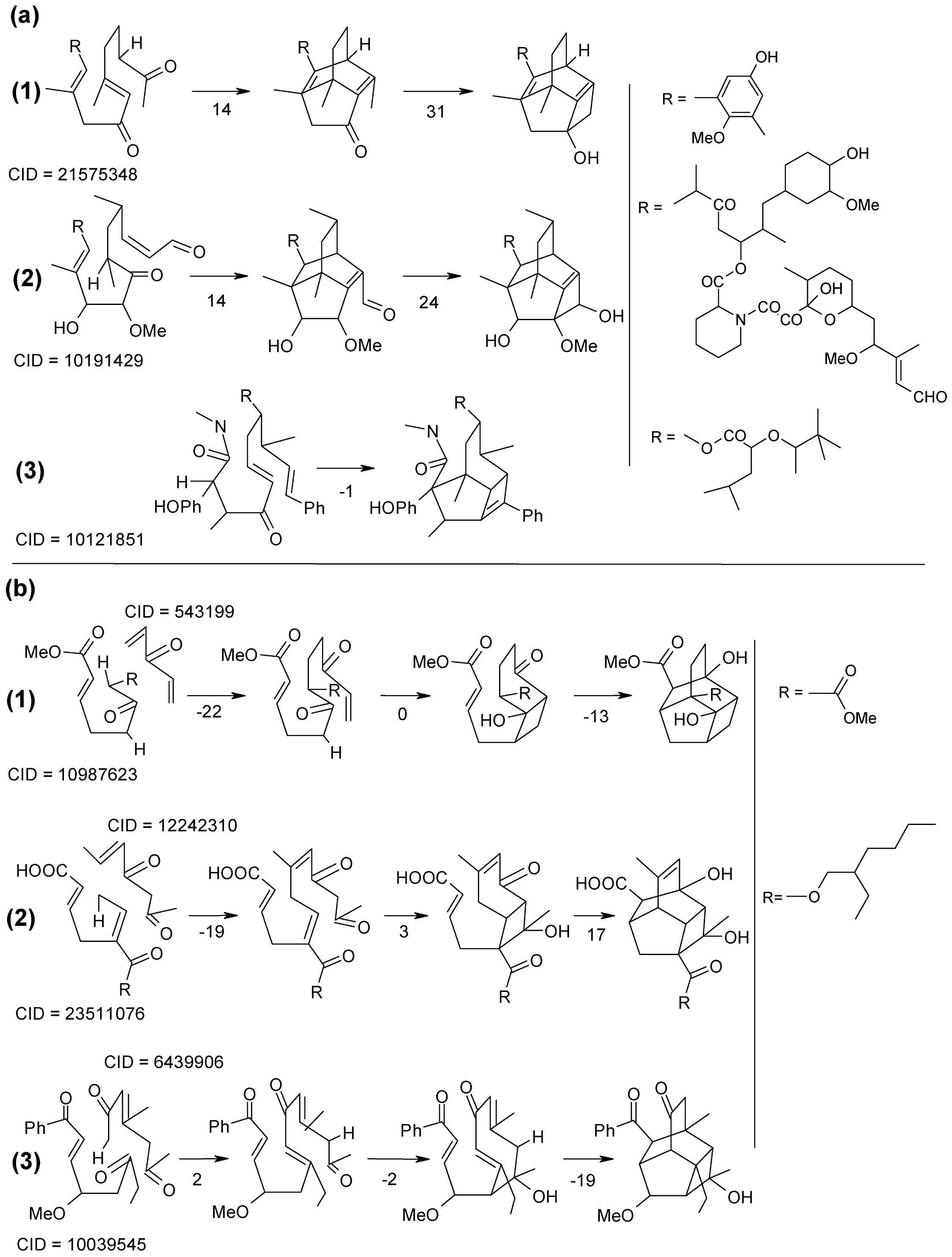

Figure 6. Examples of the generated reaction pathways for reconstructing the bond-set (a) with four and (b) five strategic bonds in the endiandric acid C skeleton. The corresponding enthalpy changes [kcal/mol] calculated by the CSB are given below the reaction arrows. CID = PubChem Compound Identifier.

exhaustively performed with the set of the transformation rules, the strategy we propose operates on multibond disconnection/formation patterns. The method enables leading the search toward the most promising pathways without exhaustive generating transformations throughout a huge solution space. The retrosynthetic search for disconnections we combined with forward generation of transformations with the use of the biomimetic multibond forming transformation patterns that describe the one-pot cascade reactions. Their application in the reaction generation process allows building a larger fragment of the target skeleton in one iteration step, instead of the traditional step-by-step procedure. In this way, each combination of bonds identified in the product structure by retrosynthetic analysis is re-examined to find the transformations for its reconstruction in a forward direction. In result a collection of building block precursors and reactions to build a product structure is suggested. This bidirectional strategy optimizes the search in terms of the number of steps and synthetic accessibility, and can be applied to a broad range of problems. The major focus is CAOS area to build the target-oriented synthetic pathways dedicated to a particular product. A further possibility is generating libraries of products that could be accessed by the same synthetic route from available building blocks. The approach can also serve as a tool to explore the space of novel functional molecules for use in drug discovery, material science, and chemical biology applications that could be effectively synthesized through one-pot biomimetic transformations.

REFERENCES AND NOTES

(1) Hanessian, S. Man, Machine and Visual Imagery in Strategic Synthesis Planning: Computer-Perceived Precursors for Drug Candidates. *Curr. Opin. Drug. Discovery Dev.* **2005**, *8*, 798–819.

(2) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.

(3) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed.* **1996**, *34*, 2613–2633.

(4) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201–227.

(5) Ott, M. A. Cheminformatics and Organic Chemistry. Computer-Assisted Synthetic Analysis. In *Cheminformatics Developments*; Noordik, J. H., Ed.; IOS Press: Amsterdam, 2004; pp 83−109.

(6) Wender, P. A. Introduction: Frontiers in Organic Synthesis. *Chem. Rev.* **1996**, *96*, 1–2.

(7) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178–192.

(8) Wender, P. A.; Handy, S. T.; Wright, D. L. Towards the Ideal Synthesis. *Chem. Ind. (London, U. K.)* **1997**, 765–769.

(9) Schreiber, S. L. Target-Oriented and Diversity-Oriented Organic Synthesis in Drug Discovery Design. *Science* **2000**, *287*, 1964–1969.

(10) Socorro, I. M.; Goodman, J. M. The ROBIA Program for Predicting Organic Reactivity. *J. Chem. Inf. Model.* **2006**, *46*, 606–614.

(11) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schultz, K.-P.; Herwig, A. Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482–494.

(12) Satoh, H.; Funatsu, K. Further Development of a Reaction Generator in the SOPHIA System for Organic Reaction Prediction. Knowledge-Guided Addition of Suitable Atoms and/or Atomic Groups to Product Skeleton. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 173–184.

(13) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921–1932.

(14) Ugi, I.; Wochner, M.; Fontain, E.; Bauer, J.; Gruber, B.; Karl, R. Chemical Similarity, Chemical Distance, and Computer-Assisted Formalized Reasoning by Analogy. In *Concepts and Applications of Chemical Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons, Inc.: New York, 1990; pp 239−288.

(15) Fontain, E.; Reitsam, K. The Generation of Reaction Networks with RAIN. 1. The Reaction Generator. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 96–101.

(16) Matyska, L.; Koča, J. MAPOS: A Computer Program for Organic Synthesis Design Based on Synthon Model of Organic Chemistry. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 380–386.

(17) Wender, P. A.; Verma, V. A.; Paxton, T. J.; Pillow, T. H. Function-Oriented Synthesis, Step Economy, and Drug Design. *Acc. Chem. Res.* **2008**, *41*, 40–49.

(18) Schürer, S. C.; Tyagi, P.; Muskal, S. M. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space. *J. Chem. Inf. Model.* **2005**, *45*, 239–248.

(19) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A Logic-Based Program for Synthesis Design. *J. Am. Chem. Soc.* **1985**, *107*, 5228–5238.

(20) Funatsu, K.; Sasaki, S.-I. Computer-Assisted Organic Synthesis Design and Reaction Prediction System, AIPHOS. *Tetrahedron Comput. Method.* **1988**, *1*, 27–37.

(21) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Design* **1995**, *3*, 34–50.

(22) Gasteiger, J.; Ihlenfeldt, W. D.; Fick, R.; Rose, J. R. Similarity Concepts for the Planning of Organic Reactions and Syntheses. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 700–712.

(23) Johnson, A. P.; Marshall, C.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 1.General Description. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 411–417.

(24) Wipke, W. T.; Rogers, D. Tree-Structured Maximal Common Subgraph Searching. An Example of Parallel Computation with a Single Sequential Processor. *Tetrahedron Comput. Method.* **1989**, *2*, 177–202.

(25) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An Application of Superstructure Search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.

(26) Hanessian, S.; Franko, J.; Larouche, B. The Psychobiological Basis of Heuristic Synthesis Planning—Man, Machine and the Chiron Approach. *Pure Appl. Chem.* **1990**, *62*, 1887–1910.

(27) Hanessian, S.; Botta, M.; Larouche, B.; Boyaroglu, A. Computer-Assisted Perception of Similarity Using the Chiron Program: A Powerful Tool for the Analysis and Prediction of Biogenetic Patterns. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 718–722.

(28) Mehta, G.; Barone, R.; Chanon, M. Computer-Aided Organic Synthesis—SESAM: A Simple Program to Unravel "Hidden" Restructured Starting Materials Skeleta in Complex Targets. *Eur. J. Org. Chem.* **1998**, 1409–1412.

(29) Hendrickson, J. B. Systematic Synthesis Design. 6. Yield Analysis and Convergency. *J. Am. Chem. Soc.* **1977**, *99*, 5439–5450.

(30) Bertz, S. H.; Sommer, T. J. The Role of Isomorphism in Synthetic Analysis. Pruning the Search Tree by Finding Disjoint Isomorphic Substructures. *Chem. Commun.* **2003**, 1000–1001.

(31) Vismara, P.; Tognetti, Y.; Laurenco, C. Maximum Symmetrical Split of Molecular Graphs. Application to Organic Synthesis Design. *J. Chem. Inf. Model.* **2005**, *45*, 685–695.

(32) Tanaka, A.; Kawai, T.; Fujii, M.; Matsumoto, T.; Takabatake, T.; Okamoto, H.; Funatsu, K. Molecular Centrality for Synthetic Design of Convergent Reactions. *Tetrahedron* **2008**, *64*, 4602–4612.

(33) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General Methods of Synthetic Analysis. Strategic Bond Disconnections for Bridged Polycyclic Structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116–6124.

(34) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492–504.

(35) Baumer, L.; Sala, G.; Sello, G. Organic Synthesis Planning: An Algorithm for Selecting Strategic Bond Forming Sequences. *Tetrahedron* **1989**, *45*, 2665–2676.

(36) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.

(37) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices. *Chem. Commun.* **1997**, 2409–2410.

(38) Rücker, G.; Rücker, C. Substructure, Subgraph, and Walk Counts as Measures of the Complexity of Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457–1462.

(39) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788–802.

(40) Rücker, G.; Rücker, C. Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99–106.

(41) Bertz, S. H.; Wright, W. F. The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity. *Graph Theory Notes of New York* **1998**, *35*, 32–48.

(42) Bertz, S. H.; Rücker, C.; Rücker, G.; Sommer, T. J. Simplification in Synthesis. *Eur. J. Org. Chem.* **2003**, 4737–4740.

(43) Rücker, C.; Rücker, G.; Bertz, S. H. Organic Synthesis—Art or Science. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386.

(44) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982–7989.

(45) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272.

(46) Barone, R.; Petitjean, M.; Baralotto, C.; Piras, P.; Chanon, M. Information Theory Description of Synthetic Strategies. A New Similarity Index. *J. Phys. Org. Chem.* **2003**, *16*, 9–15.

(47) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(48) Barone, R.; Chanon, M. Search for strategies by computer: the CONAN approach. Application to steroid and taxane feameworks. *Tetrahedron* **2005**, *61*, 8916–8923.

(49) Nowak, G.; Fic, G. Machine Learning Approach to Discovering Cascade Reaction Patterns. Application to Reaction Pathways Prediction. *J. Chem. Inf. Model.* **2009**, *49*, 1321–1329.

(50) Barberis, F.; Barone, R.; Chanon, M. HOLOWin: A Fast Way to Search for Tandem Reactions with Computer. Application to the Taxane Framework. *Tetrahedron* **1996**, *52*, 14625–14630.

(51) Nicolaou, K. C.; Petasis, N. A.; Zipkin, R. E. The Endiandric Acid Cascade. Electrocyclizations in Organic Synthesis. 4. Biomimetic Approach to Endiandric Acids a−G. Total Synthesis and Thermal Studies. *J. Am. Chem. Soc.* **1982**, *104*, 5560–5562.

CI100146N