Machine Learning Approach to Discovering Cascade Reaction Patterns. Application to Reaction Pathways Prediction

Grażyna Nowak*,† and Grzegorz Fic‡

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

Received February 18, 2009

We propose a combinatorial learning procedure for discovering graph transformation patterns based on combining transformations that can be used in a consecutive fashion. Application of this kind of pattern to a specified chemical system allows to combine a sequence of consecutive transformations into a one-step operation limiting the complexity of a reaction tree. In a retrosynthetic sense, it provides a global strategy for bond disconnections to plan more efficient convergent syntheses. In a forward direction the approach reduces the number of iterations in order to exploring the courses of complex multistep processes. The procedure enhances the capabilities and applications of the CSB (Chemical Sense Builder) computer program to assist in organic synthesis, biochemistry, or medicinal chemistry. As an example, we present the automatic derivation of the transformation pattern for Ugi-type four-component reactions (reaction sequences) and its application to assist in the designing new cascade transformations for diversity-oriented synthesis (DOS).

INTRODUCTION

Chemical graph transformation patterns have become commonly used formalisms for computer-based generation, simulation and prediction of chemical and biochemical reactions in many research areas, ranging from computer assisted organic chemistry,[1−3] through modeling chemical processes, such as catalytic cracking, fermentation,[4] to reconstruction of metabolic pathways.[5] In these applications, each pattern describes the way the conversion of one set of molecular graphs into another one, by breaking, forming or changing the type of bonds represented by edges. Given an initial set of chemical species (as molecular graphs), and a set of graph transformation patterns, it is possible to generate a set of final graphs representing products of all mathematically conceivable chemical reactions. The iterative application provides a reaction network or reaction tree (when intermediates generated at different stages do not react with themselves and the input compounds). In pure mathematical approach, where all edges in molecular graphs of the reactants are equivalent (all the bond disconnections and formations are allowed) and all transformation patterns from a defined set have the same priority, a reaction space growths exponentially along with complexity of starting materials and a number of reaction stages.

This makes modeling the behavior of complex molecules and macromolecules in multistep processes, in particular, the formation of complex as natural products compounds, simulation of biochemical processes, transformations in natural environment, or design of combinatorial libraries, difficult.

In order to limit combinatorial productivity, a variety of tools to evaluate and select the respective candidate transformations have been proposed, including knowledge-based, semiformal and formal ones. The first category, commonly used in CAOS (for review involving *Computer Assisted Organic Synthesis*, see ref 6), includes a set of empirically derived constrains/heuristics describing reactivity and selectivity of the potential bond disconnections/formations in the context of structural, steric, and stereochemical conditions.[6−10] These types of constrains rely on searching of the applicability of particular transformations before they are generated. In result, only a part of a whole chemical space, including solutions that have analogs among known chemical reactions and molecules can be searched. So, the predictive capabilities in discovering of novel chemical reactions are limited and depend on the chemical knowledge, usually domain-specific, incorporated in a program's knowledge base (that must be continuously upgraded). Therefore, to exploit a real predictive power of graph-based reaction models, there still remains a need for alternative controlling tools, allowing to access unconventional or little investigated chemical reactions.

In semiformal approaches, sophisticated calculation models are used to limit reaction mechanisms, taking into account the global molecular characteristic of reactants or based on thermodynamic or kinetic properties of reactions. In *Computer Assisted Organic Synthesis* (CAOS) systems, the molecular orbitals calculations are used to estimate the relative reactivity, and probabilities of reactions on the basis of electronic effects, therein, inductive and resonance effects, polarizability, dipole moments, partial atomic charges.[11−13] Thermodynamic computations of reaction enthalpy, entropy, heat capacity, from literature data or quantum chemical energy calculations, constitute other category of constraints, imposed on reaction feasibility and directionality. These constraints are crucial for controlling the reaction network expansion, especially in category of specialized programs to modeling complex reaction systems, such as occur in biotechno-

* To whom correspondence should be addressed. Phone: (+48 17) 8651837. Fax: (+48 17) 8541519. E-mail: gnowak@prz.edu.pl.
† Department of Physical Chemistry.
‡ Department of Computer Chemistry.

10.1021/ci9000597 CCC: $40.75
Published on Web 05/15/2009

A-B + C-D → A-C + B-D

A-B + C=D → A-C-B + :D

A-B + D-C-E → A-C-B + D-E

:A + B-C → B-A-C

:A + B-C → A-B + :C

A-B + C-D + E=F → A-E-C + B-F-D

:A + B=C + D-E → D-B-E + :C-A

Figure 1. Examples of elementary transformation schemes in the library of the reaction generator LRG; subsequent symbols A, B, C, ..., G stand for nodes of reaction centers; they may represent individual atoms or groups of atoms in the molecular graphs.

logical production and chemical processes as combustion, pyrolysis, polymerization, fermentation.[14,15] As additional restriction, the reaction rate calculations are employed to discriminate between generated reactions and further refine the reaction model.[16] A variety of approaches combining the abstract graph transformation rules with stechiometry, thermodynamics, and enzymatic reaction rates are used for prediction or reconstruction of metabolic pathway/networks to accomplish a given metabolic function.

Another approach to the problem of combinatorial complexity of chemical reaction networks, refers to the category of constraints independent of chemical knowledge, and thus offering possibility of exploring unknown regions of a chemical space. These types of constraints in CAOS systems are based on the graph theoretical description of chemical complexity or chemical similarity. Examples are RAIN and IGOR programs, that using the "Principle of Minimum Chemical Distance" allow to determine the shortest mechanistic pathway joining the given set of reactants and products.[17−19] In LILITH program, the concept of complexity distance is used to guide the flow of the reaction generation.[20,21] More recently Bertz, described the use of topological complexity indices, derived from graph theory and combinatorics, to evaluate the change of complexity or the degree of simplification that take place during a reaction or reaction pathway.[22,23] These calculations can be used for optimization of the number of steps to increase complexity in the product formation or to discriminate between alternative bond disconnections in synthesis planning. The searching for strategic bonds to perform disconnections is an approach to generate optimal retrosynthetic trees in synthesis planning. One of tactics is to cut the target (product) molecule into substructures, which are similar to starting materials.[9,24,25] Other concept, based on the similarity of reactions is implemented in our CSB (Chemical Sense Builder) system for the reaction prediction.[26] The learning module of the CSB creates the transformations (in a form of active subgraphs in substrate/product molecules) describing classes of similar reactions, using an original criterion of similarity and a reaction database.

Our current work concerns to another strategy aimed at construction and simplification of reaction pathway models by using a particular type of transformation pattern that imitate cascade processes, as that occur during biosyntheses of natural products, or their chemical analogues (one-pot cascade reactions, domino, tandem, multicomponent reactions (MCRs)[27−29]) applied in syntheses of organic molecules.

This kind of the biomimetic template involves a sequence of consecutive transformations (where, an initial transformation provides functionality for the next transformation scheme, which then leads to further consecutive transformations, and so on) that in the reaction generation process can be combined into a single step. For identification of such transformations and designing their biomimetic (biological) templates, we have applied a combinatorial learning algorithm that is already implemented in the CSB. Imitating cascade processes through the consecutively conducted simulation of the reaction steps, elementary transformation patterns are successively combined and resulting products generated to rich a final product. In result one can obtain the transformation templates describing mechanisms of sequences of reactions that can lead to the desired product from a given set of starting materials and the defined set of elementary transformations.

Using this type of pattern for construction of reaction trees will result in limiting the number of iterations to exploring the courses of complex multistep processes. This make it possible to obtain models for many applications as studying the formation of complex chemical compounds and their libraries, predicting biosynthetic pathways, designing novel cascade-based synthetic strategies. In retrosynthetic tree generation, applying these biomimetic templates to the product molecular graph allows to discover in one step, a cascade of bond disconnections to plan more efficient convergent syntheses. Another possibility of using the combinatorial algorithm is discovering the possible mechanisms for the sequences of reactions that can led to the desired product from a given set of reaction substrates. This option (described in a separate paper) can be helpful in discovering biogenetic schemes to correlate the identified structure of a product with their assumed precursors or for constructing pathways that can accomplish a given metabolic function.

GENERAL CONCEPT OF THE COMBINATORIAL ALGORITHM FOR DESIGNING CASCADE REACTION TEMPLATES

The primary goal of the combinatorial procedure is to recreate or discover mechanisms (in the sense of braking, forming, or changing the type of bonds or redistribution of valence electrons) for a sequence of consecutive reactions, that can lead to the desired product (target) from a given set of substrates. The key elements of this strategy, that is, generating all the possible candidate products that arise from a succession of elementary transformations (chosen from Library of the Reaction Generator, LRG) and matching the candidate products with the defined target, are iteratively performed to rich a desired goal. Here, we present the application of this procedure as a learning tool to designing the biomimetic transformation, giving as an example the known multistep reaction pathway (training example or source reaction).

The learning process consists in the attempt to recreate with the CSB system the course of this multistep sequence by applying a series of consecutive simulations. Starting

Step 1.

$$S_1 + S_2 + S_3 + S_4 \rightarrow FP$$

Step 2.

$$S_{i=1,2} + S_{j=i+1,3} + S_{k=j+1,4} \rightarrow P \xrightarrow{S_{m=1,4\ and\ m\neq i,j,k}} FP$$

Step 3.

$$S_{i=1,3} + S_{j=i+1,4} \rightarrow P \xrightarrow{S_{k=1,4\ and\ k\neq i,j} + S_{m=1,4\ and\ m\neq i,j,k}} FP$$

$$\xrightarrow{S_{k=1,4\ and\ k\neq i,j}} P \xrightarrow{S_{m=1,4\ and\ m\neq i,j,k}} FP$$

Step 4.

$$S_{i=1,4} \rightarrow P \xrightarrow{S_{j=1,4\ and\ j\neq i} + S_{k=1,4\ and\ k\neq i,j} + S_{m=1,4\ and\ m\neq i,j,k}} FP$$

$$\xrightarrow{S_{j=1,4\ and\ j\neq i} + S_{k=1,4\ and\ k\neq i,j}} P \xrightarrow{S_{m=1,4\ and\ m\neq i,j,k}} FP$$

$$\xrightarrow{S_{j=1,4\ and\ j\neq i}} P \xrightarrow{S_{k=1,4\ and\ k\neq i,j} + S_{m=1,4\ and\ m\neq i,j,k}} FP$$

$$\xrightarrow{S_{k=1,4\ and\ k\neq i,j}} P \xrightarrow{S_{m=1,4\ and\ m\neq i,j,k}} FP$$

Figure 2. Fragment of the learning algorithm based on the combinatorial method without repetitions.

$$FP = FP_p^0 \xrightarrow{[\text{if not } products]} FP_p^1 \xrightarrow{[\text{if not } products]} \ldots \xrightarrow{[\text{if not } products]} FP_p^n$$

Figure 3. Fragment of the learning algorithm based on the combinatorial method with final repetitions.

$$S_2 + S_3 \rightarrow P^1 \xrightarrow{S_1} P^{11} \xrightarrow{S_4} FP$$
$$\rightarrow P^{12} \xrightarrow{S_4} FP$$
$$\rightarrow \ldots \rightarrow P^{1n} \xrightarrow{S_4} FP$$
$$\rightarrow P^2 \xrightarrow{S_1} P^{21} \xrightarrow{S_4} FP$$
$$\rightarrow P^{22} \xrightarrow{S_4} FP$$
$$\rightarrow \ldots \rightarrow P^{2n} \xrightarrow{S_4} FP$$
$$\rightarrow \ldots \rightarrow P^n \xrightarrow{S_1} P^{n1} \xrightarrow{S_4} FP$$
$$\rightarrow P^{n2} \xrightarrow{S_4} FP$$
$$\rightarrow \ldots \rightarrow P^{nn} \xrightarrow{S_4} FP$$

Figure 4. Fragment of the learning algorithm based on the combinatorial method with intermediary and final repetitions.

from the input molecular graph (representing substrate or ensemble of substrates) and a set of elementary transformation rules, all new graphs (representing candidate products) resulting from iteratively performed graph transformations are generated. Once this new set of molecular species is produced, the next simulation starts to recreate the second reaction stage. Now, each product generated in previous simulation step becomes a starting material for subsequent step and so on. The learning procedure is performed combinatorially; all combinations of the starting materials and intermediates are examined, and many simulation pathways are executed. When, a given simulation pathway results in the generation of end-product(s), which is identical to that in the training (source) reaction, and all starting materials of the source reaction have been consumed in respective steps of the simulated pathway, then the learning of the CSB has been successfully finalized, leading to creation and storage in the LRG of a new transformation pattern. It is composed from the combination of elementary transformation graphs that have led to achieve the product of a training reaction.

This template in the shape of the multiple bond forming/disconnecting sequence may then be involved in the reaction generation process, imitating the course of the one-pot multistep reaction cascades. Starting from reactants, all possible end-products can be generated in one iteration, in opposite direction, the product molecular graph is transformed in one-step toward all the possible reaction substrates. The repetition of this process will result in a reaction tree (or synthetic tree), in which every node represents a molecular graph, and every edge represents a sequential transformation.

DESCRIPTION OF COMBINATORIAL LEARNING METHODS

The learning procedure is based on the following three main components, implemented in the CSB (Chemical Sense Builder) system:

Set of Elementary Transformation Graphs. Each of these transformation graphs describes a class of chemical reactions by means of breaking/forming bonds (edges) or dislocating valence electrons in substrate/product molecular subgraphs (Figure 1). Alternatively, the transformation graph can be represented as the BE (bond-electron) matrix. This approach to the description of chemical reactions is based on the Dugundji−Ugi (DU) mathematical model of chemistry.[1] Transformation graphs in a canonical form with additional information for the reaction generating procedure are stored in the library of reaction generator, LRG.[30] It involves, among others, 23 transformations for closed-shell chemistry, 27 transformations for electron pair reactions, 5 transformations for electron gap reactions, and 3 transformations with participation of radicals.

Reaction Generator. The reaction generator generates all possible products (with the appropriate stoichiometry) that can be formed from a given set of reactants by the application of a set of transformation graphs. Details are given in ref 30.

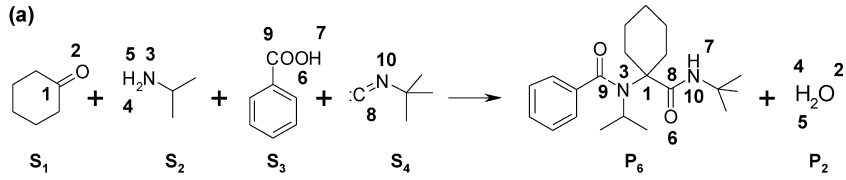


Figure 5. The Ugi four-component reaction U-4CR (a) used as a training example and the multistep simulation of the reaction mechanism (b).

(a)

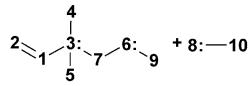
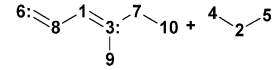
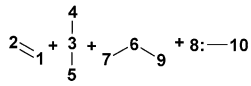
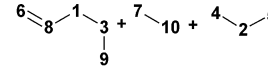
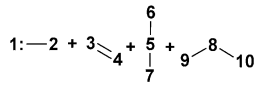
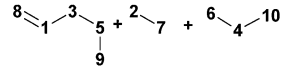
RL_1	A=B + D-C-E 1=2 + 4-3-5	→	A-C + D-B-E 1+3 + 4-2-5
RL_2	A: + B-C 3: + 6-7	→	A-B + C: 3-7 + 6:
RL_3	A-B + C: + D: 1-3 + 6: + 8:	→	C-D-A + B: 1-8-6 + 3:
RL_4	A-B + C-D + E-F 6-9 + 8-10 + 3-7	→	A-C + B-E + D-F 3-9 + 6-8 + 7-10
RL_t		→	
RL_r		→	
RL_c		→	
(b)	**:A—B + C=D + F—E—G + I—H—K** **:C-N + C=O + H-N-H + C-O-H**	→	**H=A—C—E—I + B—G + F—D—K** **O=C-C-N-C + N-H + H-O-H**

Figure 6. Course of the learning procedure for an example Ugi four-component reaction U-4CR (a), and a new transformation (b) learned from this reaction. RL_1, RL_2, RL_3, and RL_4 refer to reaction lists (transformation schemes) applied for the simulation of subsequent stages of the source reaction. Each of the transformation schemes is presented in a general form (above) and more detailed (below) where, subsequent symbols A, B, ..., K are replaced by identification numbers representing nodes of real atoms in graphs from Figure 5. RL_t, total reaction list; RL_r, reduced reaction list; RL_c, canonical reduced reaction list.

Transformation Builder. The transformation builder (new module) creates transformation templates describing cascades of bond disconnections/formations through the use of three combinatorial methods, already implemented in the CSB. This procedure starts from the combinatorial method without repetitions. The main part of this method is shown in Figure 2, related here to designing the transformation pattern, given on the set of four starting chemicals, denoted by S_1, S_2, S_3, and S_4 (equivalent set may include four molecules of the same substrate), interacting according to a cascade of transformations to form the end-products of the known reaction pathway (training example, source reaction).

In first stage, all possible reactions are generated to rich the desired product from the four starting materials and a given set of the graph transformations chosen by the user from the LRG. As a result, a set FP of *m* products (or *m* reactions) is created. If one of generated products is identical to the product of the source reaction, it means,

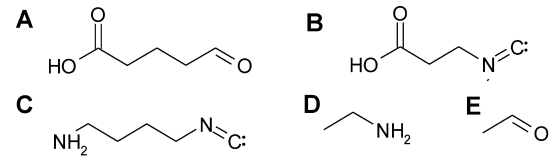


Figure 7. Initial set of reactants and the transformation template for generating skeletal diversity in the library of products.

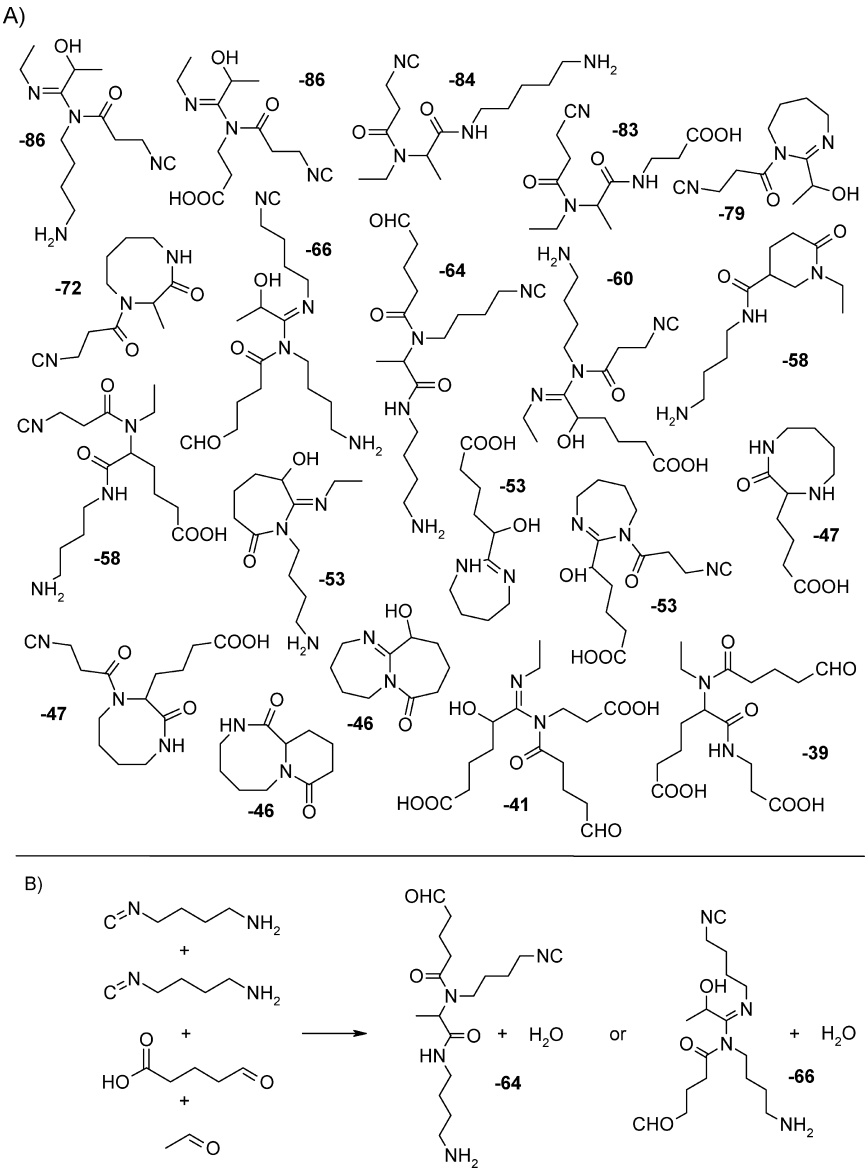


Figure 8. (A) The library of products constructed from the initial set of substrates and the U-4CR transformation pattern applied in the first iteration. The corresponding enthalpy change of generated reactions is given here in [kcal/mol]. The water accompanying the formation of each of the product candidates is here ignored. Reactions for all potentially applicable reaction models were generated, involving mono-, bi- three-, and four-molecular conversions with different or the same substrate molecules. The controlling tools involved into the reaction generating process have comprised: the chemical control, the heuristics, thermodynamic calculation of the reaction enthalpy, conversion of an unstable tautomeric form into the stable one. (B) The resulting product candidates of a single U-4CR reaction.

that the particular transformation used in one-step simulation of the tested reaction is already available in the LRG. In other case, the second stage of the learning procedure is realized to generate the successive transformations. In this stage, all the reactions proceeding only with three (among the four) starting materials are generated, and all possible combinations of these three reagents are examined. For each combination, a set P of *ml* resulting products (considered further as potential substrates), will be created. Each of these sets (P), complemented with the starting material S_m (not included in the previous combination) creates a new ensemble of substrates, that will be used for successive simulation. If one of the simulations has led to the same product as the product of the source reaction, then the learning process will be finalized successfully. In opposite case, third stage of this process will be performed, and in the case of a negative effect - the fourth stage. If even in this stage, the positive result can not be reached, then the second learning method, that is, combinatorial method with final repetitions (Figure 3) may be employed for designing the transformation pattern.

In contrast to the earlier method, now, for each set of end-products FP (Figure 2), sequences of simulations are consecutively performed, in which the next transformations are generated. Each cycle consists of *n* simulations that follow each other (parameter *n* is defined by the user). If one of the simulation steps has led to a product that is identical to the product of the source reaction, then a given sequence of transformations is completed, and as a result a new graph transformation pattern is created.

Combinatorial Method with Final and Intermediary Repetitions. In the combinatorial method with final and intermediary repetitions (Figure 4), a cycle of consecutive simulations is executed, both for the end-products and for all sets of intermediates (in Figure 2 denoted as P). Here, S_1, S_2,

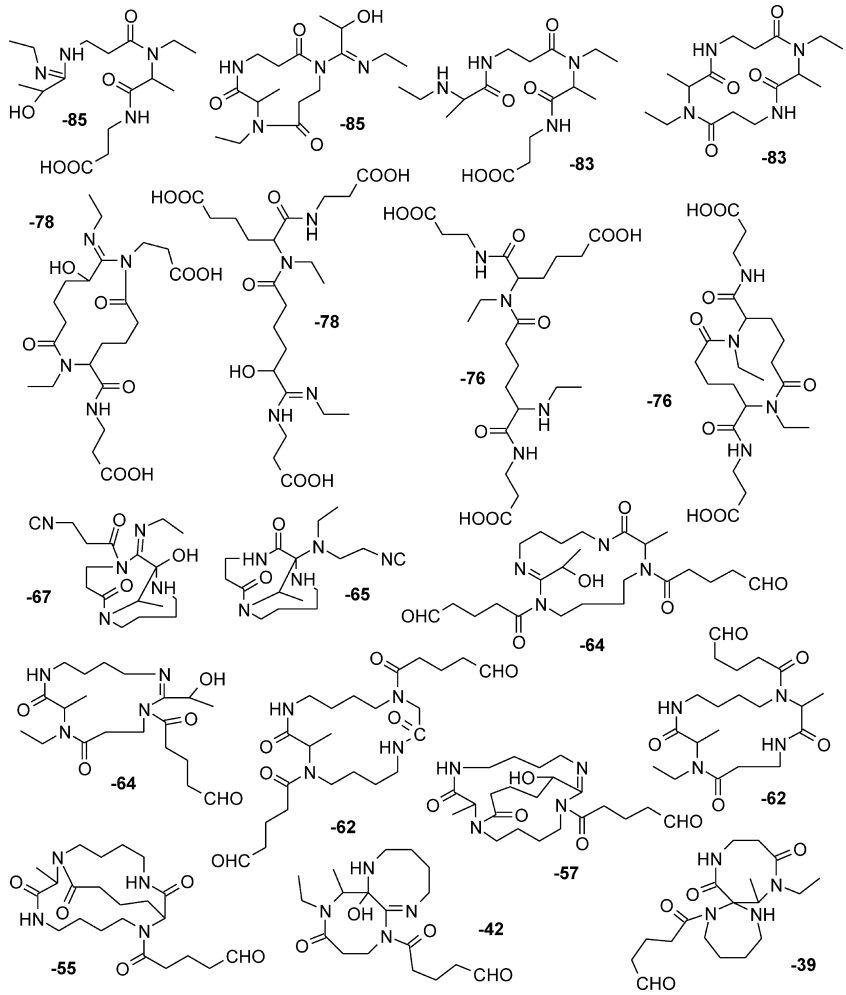


Figure 9. Collection of products formed in two-step sequential transformation. To generating this collection, each of the products of the fist generation (Figure 8A) was further consecutively transformed to mimic the second step of the cascade transformation, using the same U-4CR reaction pattern. The reaction generating process involved mono-, bi- or three-molecular transformation models (four-molecular transformations were excluded). The reaction enthalpy refers to the second transformation step.

S_3, and S_4 represent starting materials of the source reaction, P^1, one of the product sets resulting from the reaction $S_1 + S_2$, P^2, one of the product sets resulting from subsequent transformations of P^1, P^{11}, one of the product sets generated from $P^1 + S_1$, P^{12}, one of the product sets achieved by subsequent transformations of P^1, and so on. After every set FP of end-products is being generated, the conditions needed for finalizing the learning procedure are examined, and as necessary, the cycle of consecutive simulations is executed according to the scheme in Figure 3.

EXAMPLE OF DESIGNING A CASCADE REACTION TEMPLATE

The above procedure is illustrated with the use of the Ugi four-component reaction (U-4CR) as the training example (source reaction), where four starting materials are assembled according to a cascade of elementary (monoand bimolecular) chemical reactions to give an α-acyloamino amide (Figure 5a). To devise the reaction mechanism and generate the respective product, the four-step combinatorial simulation with final repetitions has been executed. Figure 5b shows the proper sequence of steps that have allowed to reach the desired product.

The first stage of this path refers to the simulation of the reaction, in which two starting materials S_1 and S_2 react leading to products P_1 and P_2; the second stage is the reaction simulation with the use as the starting materials the previously generated products P_1, P_2, and the substrate S_3, and the third stage is the reaction simulation starting from the last generated products P_3, P_4 and the substrate S_4. The fourth stage is the reaction simulation starting from products of the third stage. The product P_2 generated in the first stage (molecule of water) has been kept for the next simulations.

Elementary transformations (described in the original algorithm as the reaction lists RL_{1-56}) that have been used in this four simulations, and the new transformation template derived as result of the learning procedure are given in Figure 6. Here, subsequent symbols denote respectively: RL_t, the total (fused) reaction list describing all the four reaction steps; RL_r, the reduced reaction list, obtained by the elimination from the RL_t of the same structural changes (i.e., edges representing bonds broken/formed) in molecular graphs of substrate/products (for example, the edge 3−7 broken in substrate graph in scheme RL_4, and formed in product graph in RL_4 are removed); RL_c, unique form (after canonization that provides topologically unique numbers for nodes in molecular graphs) of the reduced reaction list RL_r, and at the end, the general form of the devised transformation template. It describes the overall mechanism of the Ugi four-

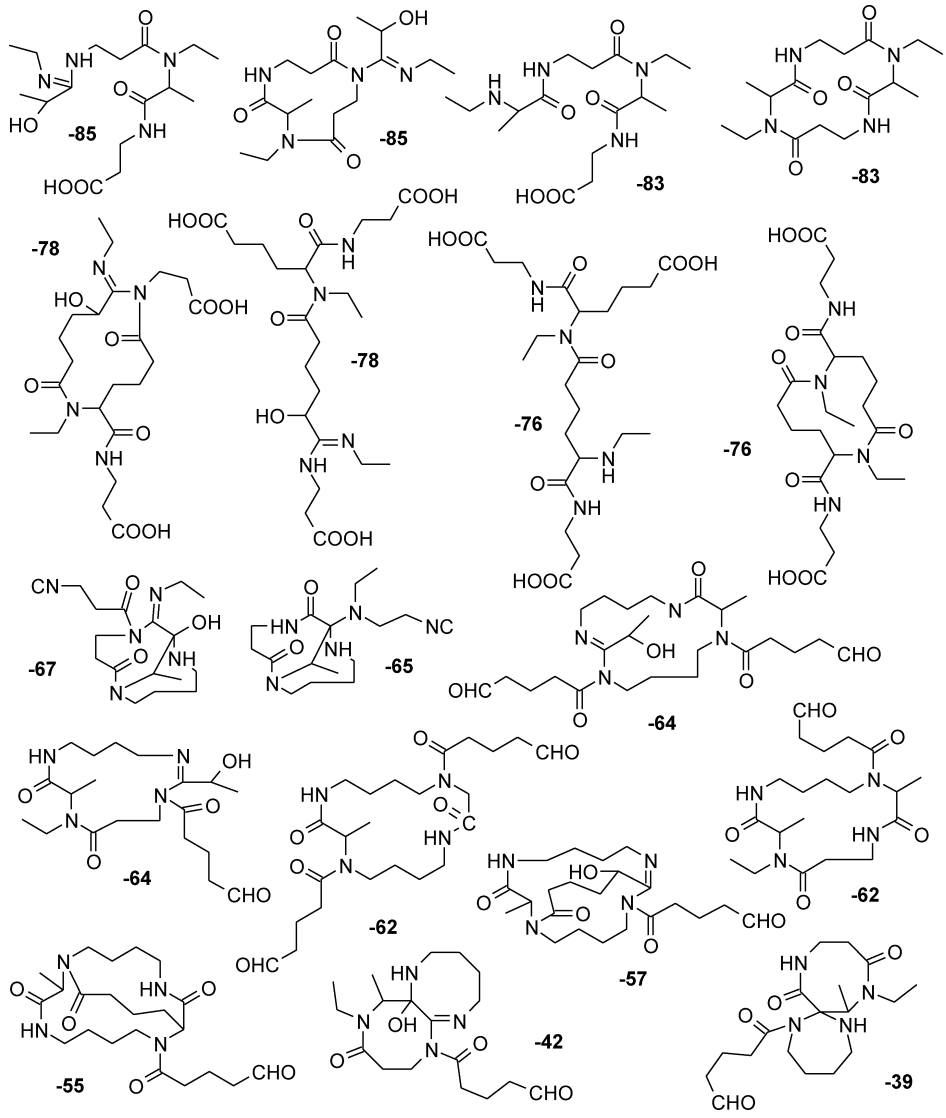


Figure 10. Subsequent products of the second generation.

step sequential reaction[31,32] or a class of similar reactions, in the context of breaking/forming bonds or shifting valence electrons.

APPLICATION OF CASCADE REACTION TEMPLATES FOR DIVERSITY-GENERATING TRANSFORMATIONS

The biomimetic-type transformation template can be applied to a specified chemical system (a set of starting chemicals or a reaction product), inducing a cascade of bond braking/forming transformations (if structural requirements for application of the transformation template will be satisfied) that can be combined into a one-step operation. In a retrosynthetic sense, it provides a global strategy for bond disconnection to plan more efficient convergent syntheses. In a forward direction the approach reduces the number of iterations in order to exploring the courses of complex multistep processes.

As an example, we present here the generation of the reaction pathway models to predict and analyze a product space from the viewpoint of usefulness in diversity-oriented synthesis (DOS). The primary goal of DOS is designing synthetic strategies that allow to convert a collection of similar substrates into a collection of more diverse and complex products. The systematic approach to planning such syntheses was developed by Schreiber to gain access to broad, unexplored regions of chemical space, for the identification of drug candidates and medicinal chemistry studies.[33,34] One of the proposed strategies relies on the union of two or more reactions or reaction sequences into another sequence of reactions, available to use in a consecutive fashion (tandem, domino, MCRs). This approach involves either the identification of such reactions (reaction sequences), where structural requirements for the subsequent reaction are formed in the product of the first, or the design of starting materials with all structural requirements for the both reactions (functional groups that serve as a starting point for a secondary reaction have to be compatible with whose participating in the primary reaction). In the simplest case, the diversity-generating sequence may be constructed by multiple applications of the same reaction type (in the same reaction vessel) for suitable multifunctional substrates. We have employed this strategy in computer experiments applying the same transformation pattern (describing the U-4CR reaction sequence discussed in previous section) in subsequent iterations. The respective input set of mono- and bifunctional substrates with an amine, aldehyde, isocyanide

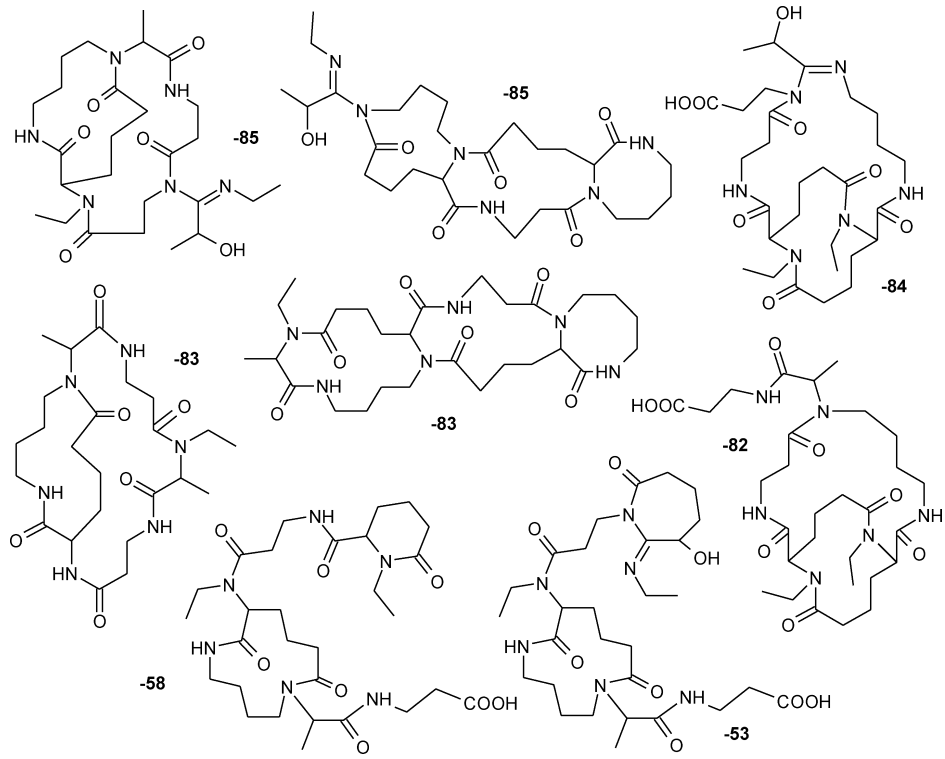


Figure 11. Library of end-products constructed with three-step consecutive transformation with the U-4CR reaction pattern in the each step. The reaction enthalpy refers to the third step.

and carboxylic acid functionality is shown Figure 7. This set was subsequently transformed employing

- One-step transformation with the U-4CR reaction pattern,
- Two-step cascade transformation with the same U-4CR reaction pattern in each step, and
- Three-step transformation based on the consecutive use of the U-4CR reaction pattern.

One iteration mimics all of the four steps of the Ugi four-component reaction sequence U-4CR to form the final product (and possible coproducts resulting from the same reaction mechanisms), without generating the reaction intermediates. By variation in each of the four components one can obtain a library of end-products having diverse molecular skeletons. In the example, the input set of starting chemicals for generating structural diversity combinatorially consists of only five building blocks, among them, three are bifunctionalized representants to increase the possibilities for structural variation. In consequence, four different reaction models were involved in the generating process (i. e., four-molecular reactions (A + B + C + D →), three-molecular (A + B + C →), two-molecular (with two bifunctional substrates, A + B →), and intramolecular conversions, with different or the same molecules) producing a large variety of combinations to obtain diverse products. The library of products generated in one-step transformations (with the U-4CR reaction pattern) from the input set of five building blocks, is shown in Figure 8A. This collection, in addition to the main (real) products of U-4CRs, also contains some of the possible coproduct candidates. However, the number of the candidates generated for a single U-4CR is limited in comparison to the step-by-step generation procedure, without the learning option. For example, the pair of molecules in Figure 8A ($\Delta H = -64$ kcal/mol and $\Delta H = -66$ kcal/mol) refer to the main product and coproduct of a single U-4CR, generated from the following molecules: A+ 2C + E in Figure 7. The generated product candidates for this single U 4CR are presented in Figure 8B.

To achieve more diverse skeletons, each of the products of the fist generation was further transformed with the same U-4CR reaction pattern to mimic second (and then third) reaction step of the projected cascade transformation. The successive transformations were conducted consecutively for each starting ensemble (of four, three, two) of the substrates chosen from the input set of five molecules. So, after the generation of the product of the first U-4CR reaction, it was taken as a potential substrate and involved with the remain substrates (of this first reaction) for the second stage. After generation of the products of third step, the next ensemble of starting substrates is chosen to generate products of first, second, and third generation (reaction stage), and so on. The increase in complexity and diversity of products achieved after the two-step consecutive transformation is shown in Figures 9 and 10. The collection of products (reduced by us, now there is no filtering function in the CSB) resulting from three-step cascade transformation is given in Figure 11.

This simple computer experiment makes possible a systematic exploration of chemical reaction space in result of changing both, the set of initial reactants and diversity-generating transformations. The strategy of combining reactions into cascade sequences, already implemented in the CSB, can be used to automatically contrust other diversity-generating transformations and test them for utility in DOS.

CONCLUSIONS

We have described a new approach of generating complex multistep reaction pathway models that allow avoidance of the explosive growth of a reaction tree or a network. The approach is based on the application of the biomimetic-type templates that combine the sequences of consecutive trans-

formations into a single step (mimicking one-pot cascade reactions), without generating the reaction intermediates. So, the application of such templates to a given set of initial species results in the limitation of the size of a reaction tree, providing models for studying complex multistep processes in various fields, as synthetic and medicinal chemistry, biochemistry, environmental chemistry.

We present the learning procedure that automatically creates the discussed type of the graph transformation template from examples of known cascade reactions by combining elementary transformation rules applied in successive steps. Constructed in this way the transformation patterns embody the feature of "real chemistry" that allow more plausible predictions to be obtained. Thus, the generated solutions space is refined additionally from unreal or less probable solutions. In result, a number of possible reaction pathways are eliminated and only a part of a whole chemical space, including solutions that have analogs among known chemical reactions can be searched. On the other hand, a variety in the types of elementary transformation rules stored in the library of the reaction generator LRG, in conjunction with generality of graph based formalism, allow considerable flexibility in the designing of the biomimetic transformation patterns. Therefore, they can be used to generating models and studying the formation of natural compounds, predicting biosynthetic pathways, discovering novel cascade-based transformations. As an example, we presented here the approach based on combining cascade reaction patterns to generating and studying diversity in libraries of products. The computer experiment resembles the synthetic strategy that assumes the identification and incorporation into the planning synthetic pathway of pairs or combinations of reactions that may be used in consecutive fashion (domino or tandem). This type of research, conducted with the use of CSB-system, can be helpful in the development of new diversity-generating transformations for DOS or for planning cascade-based synthetic strategies.

REFERENCES AND NOTES

(1) Dugundji, J.; Ugi, I. The Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.

(2) Kvasnička, V.; Kratochvil, M.; Koča, J. Reaction Graph. *Collect. Czech. Chem. Commun.* **1983**, *48*, 2284–2304.

(3) Fujita, S. Description of Organic Reactions Based on Imaginary Transition Structures. Part 6−9. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 99–120.

(4) Benkó, G.; Flamm, C.; Stadler, P. F. A Graph-Based Toy Model of Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1085–1093.

(5) Rosselló, F.; Valiente, G. Graph Transformation in Molecular Biology. Formal Methods in Software and System Modeling. *Lect. Notes Comput. Sci.* **2005**, *3393*, 116–133.

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

(7) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163–168.

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

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

(10) Salatin, D. T.; Jorgensen, W. L., J. Computer Assisted Mechanistic Evaluation of Organic Reactions. l. Overview. *J. Org. Chem.* **1980**, *45*, 2043–2051.

(11) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Top. Curr. Chem.* **1987**, *137*, 19–73.

(12) Baumer, L.; Sello, G. New Method for the Calculation of Bond Native Polarity Using Molecular Electronic Energy. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 125–130.

(13) Baumer, L.; Sala, G.; Sello, G. Computer Assisted Organic Planning: Effective Bond Polarity As a Guideline to Reactivity. *J. Am. Chem. Soc.* **1991**, *113*, 2494–2500.

(14) Truong, T. N.; Cook, T.; Nayak, M.; Boonyasiriwat, C.; Tran, L. T.; Zhang, S. Computational Science and Engineering On-line: An Integrated Web-Based Environment for Multi-Scale Modelling of Complex Reaction Systems. *Mol. Phys.* **2004**, *102*, 353–360.

(15) Ratkiewicz, A.; Truong, T. N. Application of Chemical Graph Theory for Automated Mechanism Generation. *J. Chem. Inf. Comput. Sci.* **2000**, *43*, 36–44.

(16) De Witt, M. J.; Dooling, D. J.; Broadbelt, L. J. Computer Generation of Reaction Mechanisms Using Quantitative Rate Information: Application to Long-Chain Hydrocarbon Pyrolysis. *Ind. Eng. Chem. Res.* **2000**, *39*, 2228–2237.

(17) Jochum, C.; Gasteiger, J.; Ugi, I.; Dugundji, J. The Principle of Minimum Chemical Distance and the Principle of Minimum Chemical Structure Change. *Z. Naturforsch.* **1982**, *37b*, 1205–1215.

(18) 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.* **1993**, *32*, 201–227.

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

(20) Baumer, L.; Sala, G.; Sello, G. The LILITH Approach to Organic Synthesis Planning. *Anal. Chim. Acta* **1990**, *235*, 209–214.

(21) Baumer, L.; Sala, G.; Sello, G. Organic Synthesis Planning: An Algorithm for Selecting Strategic Bond Forming Sequences. *Tetrahedron* **1989**, *45* (9), 2665–2676.

(22) Bertz, S. H.; Rücker, C. In Search for Simplification: the Use of Topological Complexity Indices to Guide Retrosynthetic Analysis. *Croat. Chem. Acta* **2004**, *77*, 221–235.

(23) Bertz, S. H. Compexity of Synthetic Reactions. The Use of Complexity Indices to Evaluate Reactions, Transforms and Disconnections. *New J. Chem.* **2003**, *27*, 860–869.

(24) 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.

(25) 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**, *7*, 1409–1412.

(26) Fic, G.; Nowak, G. Implementation of Similarity Model in the CSB System for Chemical Reaction Prediction. *Comput. Chem.* **2001**, *25*, 177–186.

(27) Nicolau, K. C.; Edmonds, D. J.; Bulger, P. C. Cascade Reactions in Total Synthesis. *Angew. Chem., Int. Ed.* **2006**, *45*, 7134–7186.

(28) Tieze, L. F. Domino Reactions in Organic Synthesis. *Chem. Rev.* **1996**, *96*, 115–136.

(29) Zhu, J., Bienaymé, H., Eds. *Multicomponent Reactions*. Wiley-VCH: Weinheim, Germany, 2005.

(30) Fic, G.; Nowak, G. Extended Concept of Reaction Generators in an Organic Reaction Prediction System CSB. *Comput. Chem.* **1998**, *22*, 141–146.

(31) Dömling, A.; Ugi, I. Multicomponent Reactions with Isocyanides. *Angew. Chem., Int. Ed.* **2000**, *29*, 3168–3210.

(32) Ugi, I.; Werner, B.; Dömling, A. The Chemistry of Isocyanides, Their MultiComponent Reactions and Their Libraries. *Molecules* **2003**, *8*, 53–66.

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

(34) Burke, M. D.; Schreiber, S. L. A Planning Strategy for Diversity-Oriented Synthesis. *Angew. Chem., Int. Ed.* **2004**, *43*, 46–58.

CI9000597