Discovery of Chemical Transformations with the Use of Machine Learning

Grzegorz Fic and Grazyna Nowak

Rzeszow University of Technology
6 Powstancow Warszawy Ave., 35-041 Rzeszow, Poland
{gfic, gnowak}@prz.rzeszow.pl

Abstract. Research activities in the computer reaction modeling and synthesis design have led to the development of a number of computer systems that are successively used as a research tools to study the organic synthesis problems. One of such systems is the CSB (Chemical Sense Builder), which combines four different logic- and knowledge-based models to generate possible structural modes of chemical conversions. In these paper we briefly discuss the architecture and specially the learning methods implemented in the CSB. They exploit different sources of knowledge and chemical reaction data for the classification, generalization, and derivation of rules in a form of the graph transformation schemes. Any transformation is general description of a collection of similar reactions from the reaction data base. Application of these learned rules in the course of reaction simulation enables us to predict – for a given set of reacting molecules – all possible reaction courses, having responses in real chemistry.

1 Introduction

A comprehensive study of fundamentals of formal-logic approaches (mathematical models) implemented in various systems for modeling chemical conversions has pointed out some problems dealing with the designing and exploitation of such system category. The most important is related to a compromise between two opposing interests: generating all the conceivable solutions (with wrong, trivial, and otherwise some valuable and novel one) or generating reliable solutions (avoiding combinatorial explosion). Thus, the solution space produced according to a given logic-based or knowledge based model for transformation of chemical structures may be either complete or reliable. The growing of the reliability implies reduction of completeness. Methods attempting to achieve a satisfactory compromise between completeness and reliability are based on:

- developing multimodel systems (CSB [1])
- developing a learning systems (SYNCHEM [2], CSB)
- using multilevel hierarchy of knowledge (CSB, in testing).

The above three methods are implemented in recent version of the CSB system. In result, the user has the possibility to select the most suitable model of simulation, matched to the synthetic problem being solved.

2 The CSB Multimodel System for Simulation of Chemical Transformations

In the current version of the CSB system, the following four models are applied for the generation of chemical reactions:

- *Ugi-Dugundji (U-D) model*, based on the *matrix model of constitutional chemistry* [3], which enables the generating complete but not reliable solution space
- *common-sense model*, combines the fundamentals of U-D model and approximate reasoning of reaction centers [4] enabling to avoid a combinatorial explosion and to reduce unreal chemical transformations
- *similarity model*, allows the simulating conversions that are similar to reactions in the training set (a source for supervised learning) [1]
- *purely empirical model*, reproduces known chemical knowledge from the Real Reaction Library.

3 The CSB Learning System for Searching, Classification and Generation of Rules from Data

Two methods of supervised learning are implemented in the CSB system:

- monitoring and reproducing the courses for these simulations that have lead to *good solutions*,
- comparative analyzing based on the concept of hierarchical levels of knowledge.

First of these methods was implemented in the similarity model of the CSB system, the second is tested yet, and will be involved in the architecture of the CSB as novel simulation model – i.e., *model of the reaction circle*.

3.1 Monitoring of the Simulation Course

In the learning process, the following information sources may be utilized to derive and generalize of new rules for the Similar Reaction Library (SRL):

- reaction entered by chemist, or
- reaction generated by the common-sense module (in the prediction process), and verified by searching the Real Reaction Library (RRL).

The essence of this method is an attempt to imitate the course of a given real reaction, using the common-sense model and the actual chemical knowledge stored in the Real Reaction Library and REACT knowledge bases. When the simulation process of the source reaction has been successfully finalized (i.e., entered graphs of substrates have been transformed into the graph of product, which is identical with a product of some real reaction), then the Knowledge Generator (using the track of this simulation

stored in the memory), induces the transformation scheme of the shape: cause $\Rightarrow$ consequence. It is a generalized description representing the source reaction and a collection of similar reactions (classifier). It contains the following information:

on the side of a cause: list of conditions needed for generating a real reaction as well as similar ones. It comprises: identifiers of active subgraphs that have to be present in graphs of substrates;

on the side of a consequence: transformation vector describing the course of the conversion of active subgraphs of substrates into product candidates. It contains: reaction model (number of reacting molecules); list of identifiers of edges (in active subgraphs) that are broken/made during the simulated transformation; a scheme (from Library of Reaction Generator [5]) allowing the formation of new edges in molecular graphs (products).

Each chemical reaction may be a source for generating at least one transformation scheme. Set of learned transformations kept in the Similar Reaction Library, may then be applied to generate (for a given ensemble of molecular graphs representing substrates) the reaction being the source for the learning process as well as all similar reactions.

The learning process is conducted here on expert-prepared rules for the knowledge base REACT. Uncompleted and subjective nature of this knowledge (reflecting the opinions of the individual expert) may cause some limitations of this learning method.

3.2 Hierarchical Levels of Knowledge – The Model of Reaction Circle

In the learning method discussed above, the cause-consequence correlation is created as result of a comparative analysis of molecular graph for substrates (cause) with molecular graphs of products (consequence). This operation involves the following six stages:

Input data: reaction graph = {molecular graphs of starting materials, molecular graphs of products}

step_1: topological analysis of molecular graphs

step_2: atom-atom mapping (recognizing in the graph of products the nodes which are equivalent to nodes in graph of substrates and creating the reaction map)

step_3: recognition of the molecular subgraph of reaction center {1st level of knowledge}

result: **transformation describing the 1st level of knowledge**

step_4: L=2 {2nd level of knowledge}

step_5: expansion of the L-1 level of knowledge (structural environment) by next bound length of the reaction center

result: **transformation of n-th level of knowledge**

step_6: checking whether all levels were tested

if yes: **STOP**

if not: L = L+1 and realization of the **step_5**.

The learning method considered here, in contrast to the first one, enables to avoid the influence of the subjective knowledge and intuition of the expert on results of computer simulation of chemical conversions. The main advantage is the capability to analyze and generalize the data from the Real Reaction Library, leading to derivation of rules for Library of Similar Reactions, arranged into multi-level, hierarchical structure. Application of various knowledge formalisms (reflecting various levels of generality/specificity) for generating chemical conversions, brought out the possibility of creating various solution spaces, depending on the goal of computer simulation. This method is capable to discover new chemical reactions (here the application of the most general knowledge, i.e., on the lowest level, is preferred), from the other hand, it is useful for the classification of chemical reactions (assuming as a criterion the highest level of knowledge). Additional possibility is the bi-directional generation of chemical transformation, i.e., forward for the simulation of organic reactions, and backward for the design of chemical syntheses.

4 Conclusion

The methodology implemented in the CSB system, integrates the concept of multi-model simulation with the knowledge processing and learning tools, supporting the process of the generation of chemical conversions. Recent advances in the development of the CSB provide new machine learning capabilities. Among them is the ability to learn different type of rules from various knowledge sources, and automatic application of learned or acquired rules in the reaction prediction process. The CSB can be applied as a tool to wide variety of problems in different disciplines of chemistry. Some of possible fields of CSB-applications are: the prediction of the reaction courses proceeding in different environment (natural atmosphere, soil, water, living cells, etc.), research on metabolic pathways of various drugs, research in drug-drug interactions, the simulation and prediction of reactions in the natural environment (as for example, the degradation reaction and their products, the study of the fate of chemicals in agriculture).

References

1. Fic, G., Nowak, G.: Implementation of similarity model in the CSB system for chemical reaction prediction. *Comput. Chem.* 25 (2001) 177-186.
2. Gelernter, H., Rose, J.R., Chen, C.: Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput. Sci.* 30 (1990) 492-504.
3. Dugundji, J., Ugi, I.: An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* 39 (1973) 19-64.
4. Nowak, G., Fic, G.: Implementation of Commonsense Reasoning in Organic Reaction Prediction System CSB. *Comput. Chem.* 22 (1998) 147-152.
5. Fic, G., Nowak, G.: Extended Concept of Reaction Generators in an Organic Reaction Prediction System CSB. *Comput. Chem.* 22 (1998) 141-146.