Automated Reaction Family Generation and Analysis Using Cheminformatics

Pieter P. Plehiers[1], Guy B. Marin[1], Christian V. Stevens[2] and Kevin M. Van Geem[1,*]
[1]Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Technologiepark 914 9052 Gent, Belgium
[2]SynBioC Research Group, Department of Sustainable Organic Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium

Abstract: The amount of known chemical reactions is ever increasing. By now, databases exist containing millions of reactions and related data. While being an invaluable source of knowledge, the sheer quantity of data has long surpassed the conceivability of the human mind. In order to put this data to good use, for example for reaction network generation or retro-synthetic analysis, tools and methodologies are required to interpret and access this data efficiently. The novel methodology implemented in the reaction network generator Genesys has been specifically designed to automatically analyze reactions and extract information on the reaction mechanism in order to construct reaction templates for future automated mechanism generation. The capabilities of the method are tested on a set of pyrolysis related reactions. The set of 110 reactions was analyzed and reduced to a set of 44 reaction families that was fully compatible with Genesys.

Keywords

Cheminformatics, Reaction network generation, Retrosynthesis, Reaction mechanism detection, Atom-atom mapping

I. INTRODUCTION

Our knowledge of chemistry has exploded over the past decades. Currently, the Reaxys® database contains over 40 million chemical reactions and 100 million compounds (Elsevier R&D Solutions 2016). This information is highly valuable for the study of known reactions and reaction networks, but also for the discovery of new pathways via retro-synthetic analysis. In both fields, it is necessary to know which transformations a molecule can undergo. The vastness of the available data has made it impossible to do this manually, as even smaller, open-source databases such as as NIST (National Institute of Standards and Technology 2016), KEGG (Kanehisha Laboratories 2016) and RMG (Green et al. 2017) contain too much information to process manually

One more specific example where the use of database information can be useful is in reaction network generation (Figure 1). Reaction network generation tools rely on a set of (user-generated) reaction templates or recipes which are repeatedly applied to all present species. Examples are Genesys and RMG (Van de Vijver et al. 2015). Composing these recipes is time-consuming and error-prone. Using (parts of) the aforementioned databases to generate the templates could eliminate this tedious step from the process. The method that is described here has been developed to interpret such database entries and translate them into a reaction template, which can consecutively serve as input for a reaction network generation tool. In what follows, the algorithm will be elucidated and the results of testing the method on a set of radical reactions will be discussed.

*Corresponding author: Kevin.VanGeem@UGent.be

Figure 1: Using a reaction database to generate a generalized reaction network.

II. ALGORITHM

Several steps are required to convert an entry into a valid reaction family scheme. More precisely, the method can be subdivided into four main conceptual blocks (illustrated by the colors in Figure 3). The first step is interpreting the format of the database. The second step is determining which reactant atoms correspond to which product atoms (the atom-atom mapping or AAM). This information is critical in discerning the reaction recipe, but is rarely available in the database. Besides the main AAM, which is performed using the reaction decoder tool (RDT) (Rahman et al. 2016), a post-mapping step attempts to find mappings for atoms that weren't mapped by RDT. The next module analyses the changes that occur during the reaction. The final step is the generation of the input file for the reaction generation tool. An important part of the reaction family analysis takes place in this final step, namely the comparison of the newly generated family to all previously generated ones.

The interpretation and conversion of an entry and conversion to a cheminformatics format is quite straightforward if a standardized identifier (Figure 2) is used in the database. This is not the case for the NIST database, which is therefore not further used in this work. Of the remaining two databases, KEGG uses chemical table files (Dalby et al. 1992), while RMG uses both SMILES (Weininger 1970) and InChIs (Heller et al. 2013).

Methane – CH4

Smiles	InChI	Chem. Table File
Implicit H: C Explicit H: [C]([H])([H])([H])H	InChI=1S/CH4/h1H4	ACD/Labs03211713433D 5 4 0 0 0 0 0 0 0 0 1 V2000 20.7190 -4.6058 0.0760 H 0 0 0 0 0 0 0 0 0 0 0 0 19.9115 -4.7622 -0.6709 C 0 0 0 0 0 0 0 0 0 0 0 0 20.2982 -5.3781 -1.5106 H 0 0 0 0 0 0 0 0 0 0 0 0 19.5705 -3.7787 -1.0592 H 0 0 0 0 0 0 0 0 0 0 0 0 19.0582 -5.2864 -0.1897 H 0 0 0 0 0 0 0 0 0 0 0 0 1 2 1 0 0 0 0 2 3 1 0 0 0 0 2 4 1 0 0 0 0 2 5 1 0 0 0 0 M END

Figure 2: Illustration of different standardized identifiers for the simple case of methane

The second part of the method determines the AAM for the reaction and is the most time-consuming step. For the majority of the atoms, the AAM is determined using RDT, which is based on the maximal common subgraph approach (Rahman et al. 2014). The drawback of RDT is that it does not take radicals into account. For liquid phase reactions which take place at relatively low temperatures, this is not an issue. However, the goal of the method is to be applicable both for liquid- and gas phase reactions. For the latter, radical mechanisms are no exception. This problem is solved by converting the radicals into unique dummy atoms before the mapping and reverting the dummies to the original radicals after the mapping.

RDT is based on a maximal common subgraph algorithm. The success of the algorithm therefore depends on the presence of sufficient structural characteristics. For very small or symmetrical molecules, the amount of information gained from the structure can be limited, resulting in incorrect or incomplete mappings of the atoms. Incorrect mappings are filtered out via some deterministic rules, *e.g.* when a species contains radicals, at least one of these radicals should play some role in the mechanism. Incomplete mappings on the other hand can, in most cases, be completed quite easily, due to the type of molecules for which the mapping tends to fail. Of the non-mapped atoms, first those that are present only once are mapped to each other, *e.g.* if only one carbon and one oxygen atom are not mapped, the un-mapped reactant carbon atom can be mapped to the un-mapped product atom. This assumes that the initial mapping is correct. Then four cases are identified in which two reactant atoms can be mapped to two product atoms of the same type. Either the reactants in which the atoms are found are identical, or the reactant is symmetrical with respect to the position of the atoms, or the products are identical or the product is symmetrical. In these cases the two atoms are interchangeable, ensuring that the mapping will be correct.

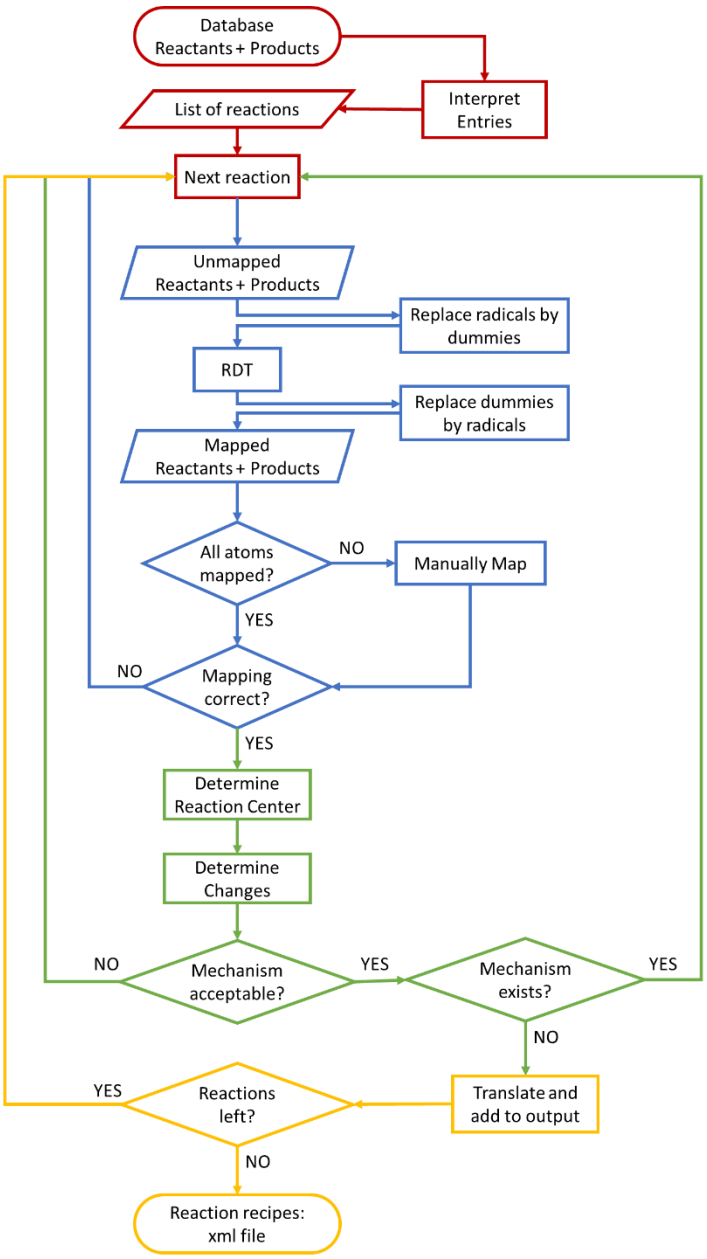

Figure 3: Schematic representation of the reaction recipe generation and analysis method. The colors correspond to the four main blocks in the algorithm.

In the third block, the changes that take place during the reaction are detected, based on the mapping from the previous block. These changes are breaking and formation of bonds, changes in bond order, gaining or losing charges and radicals. Hetero atoms that are connected to the reaction center are also included. The latter makes the determined reaction families slightly more specific, but can be necessary not to over-generalize. To determine the changes to the reactants, each atom is compared to its mapped counterpart. The neighbors and connecting bonds are compared. New neighbors indicate the formation of bonds and missing neighbors breaking of bonds. The perceived recipe is compared to all formerly generated recipes. If an identical recipe has not yet been found, the changes are finally translated to the fitting xml document format used in Genesys.

III. Method Testing and Validation

The approach described in the previous section is tested as follows. A set of reactions is first narrowed down to a set of reactions with unique recipes and filtering out those reactions that are incorrectly mapped. The remaining reactions are then processed one by one as a single reaction family in Genesys. However, instead of using the reactants and products as defined by the database entry, random reactants are generated with the same reaction center as deduced from the database entry (Figure 4). The corresponding products are then constructed using the detected mapping. These reactants are then passed to Genesys and a "reaction network" is generated based on the reactants and the single reaction recipe. The products that are formed by Genesys are then compared to those formed based on the mapping. If they match, the generated reaction recipe is considered to be consistent. For each recipe 25 random molecules are tested in this way and the separate success are reported.

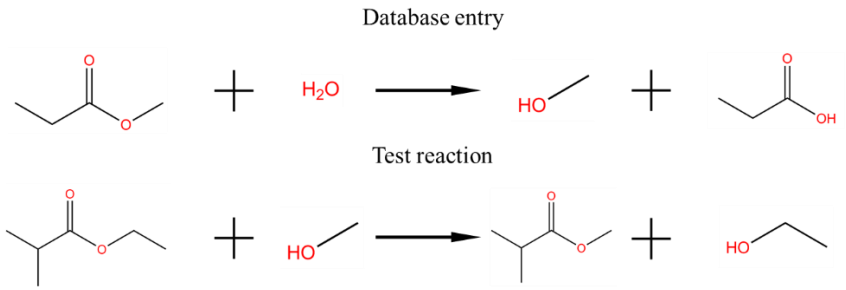

Figure 4: Illustration of the construction of test reactions based on the database entry.

The described testing method has been performed on a set of reactions related to the pyrolysis of hydrocarbons. The reaction set consists of two databases in the RMG kinetics library (Green et al. 2017). The first system is the "C3" (37 reactions) database and the second one the "vinylCPD_H" (73) database. Both are related to cyclopentadiene pyprolysis.

IV. Results and Discussion

Of the 110 reactions in the two considered databases that make up the pyrolysis test case, 27 (25 %) were either not (fully) mapped or expected to be incorrectly mapped. Of those 27 reactions, 13 originated from the "C3" database and 14 from the "vinylCPD_H" database. This implies a success rate of only 65 % for the "C3" database. The cause hereof is twofold. First of all, using dummy atoms to represent radicals implies that in order for the radical to be transferred to another atom, an additional combination of breaking and forming a bond is required. Hence the common substructure algorithm will initially attempt to find a mapping in which the radical remains on the same atom. This problem is aggravated by the fact that in the considered database, the reactants and products tend to be small and symmetrical or have only limited branches. Such branches act as structural markers, increasing the probability of a correct mapping. This is illustrated by following example.

Reactions 7 and 32 are nearly identical, but 7 is mapped correctly, while 32 is not (Figure 5). In the latter, both reactants have an identical maximum common substructure, which is also found in the product. Hence the algorithm cannot discern from which of the reactants the fragment originates. RDT applies some optimization to increase the probability of choosing the correct option in such cases. The conversion of the radical to a dummy atom deranges this optimization as transferring it correctly (as in reaction 7) requires more bonds to be broken than keeping it on

the original atom. This problem is also noticed in several cyclisation reactions in which the algorithm prefers breaking and forming a bond with hydrogen above the dummy atom.

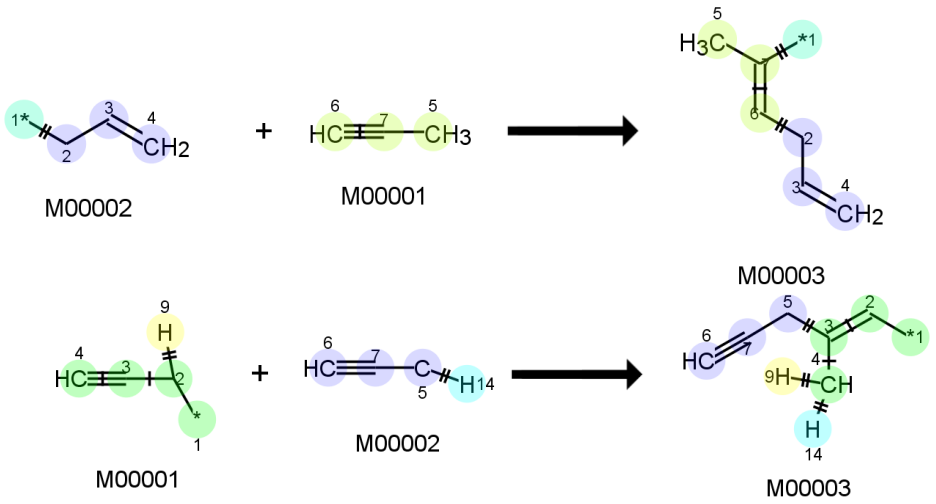

Figure 5: Reactions C3.7 (top) and C3.32 (bottom), with mapping. Coloring shows which molecular reactant fragments appear as which moleculare product fragments.

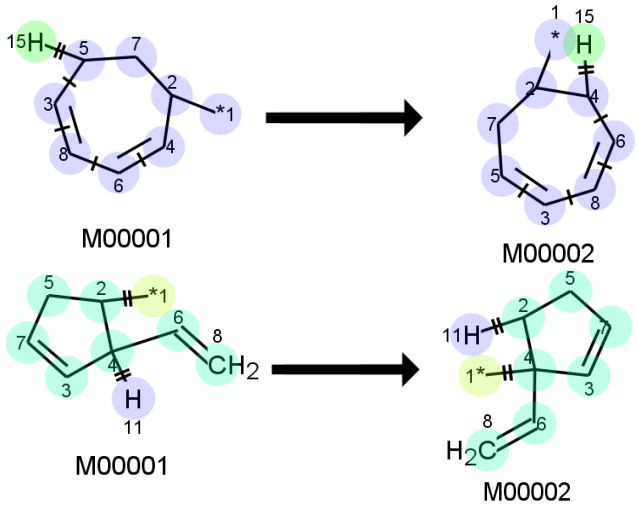

Figure 6: Reactions vinylCPD_H.9 (top) and 10 (bottom), with mapping. Coloring shows corresponding molecular fragments.

Similar observations are made for the "vinylCPD_H" database, though overall the success rate is higher. Here reactions 9 and 10 are excellent examples of how branched molecules result in higher success (Figure 6). Both reactions describe an identical mechanism (1,2 hydrogen shift), but this is only correctly detected in reaction 9. In the unbranched reaction 10, it is again preferred to not move the radical.

Considering the reactions that were correctly mapped, 44 unique recipes could be extracted from them. For each recipe, the expected products were generated by Genesys in all 25 cases. A summary of these results can be found in Figure 7.

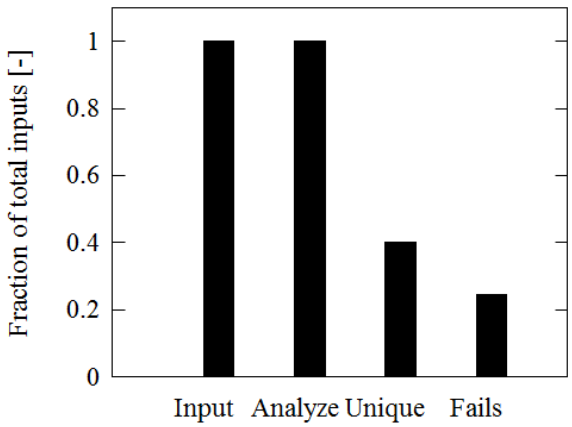

Figure 7: Summary of the reaction recipe analysis for the test set of pyrolysis reactions as fraction of the total number of inputs. Analyze: fraction that could be analyzed (i.e. two or less reactants). Unique: number of unique reaction families generated with respect to the total number of inputs. Fails: fraction of the input for which no correct mapping was generated.

V. Conclusions

To facilitate the generation of reaction networks based on extensive chemical reaction databases, a method has been developed that analyzes a given set of reactions from a database and extracts unique reaction families in the form of recipes that are compatible with a given reaction network generator. Detection of the reactive centers of the reaction is done based on an AAM by the open-source tool RDT.

The method consistently succeeds at constructing correct recipes. 100 % of the generated recipes for a set of 110 pyrolysis reactions were correctly processed by Genesys. This indicates that the method performs excellently, under the condition that the accuracy of the atom-atom mapping is decent.

VI. Acknowledgements

P.P.P acknowledges financial support from a doctoral fellowship of the Research Foundation – Flanders (FWO).

VII. References

Dalby, A., et al. (1992). "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited." *Journal of Chemical Information and Computer Sciences* **32**(3): 244-255.
Elsevier R&D Solutions (2016). Reaxys Fact Sheet.
Green, W. H., et al. (2017). "RMG Kinetics Libraries." 2017, from http://rmg.mit.edu/database/kinetics/libraries/.
Heller, S., et al. (2013). "InChI - the worldwide chemical structure identifier standard." *Journal of Cheminformatics* **5**(1): 7.
Kanehisha Laboratories. (2016). "KEGG: Kyoto Encyclopedia of Genes and Genomes." 2017, from http://www.kegg.jp/.
National Institute of Standards and Technology. (2016). "NIST Chemical Kinetics Database." from http://kinetics.nist.gov/kinetics/.
Rahman, S. A., et al. (2014). "EC-BLAST: a tool to automatically search and compare enzyme reactions." *Nat Meth* **11**(2): 171-174.
Rahman, S. A., et al. (2016). "Reaction Decoder Tool (RDT): extracting features from chemical reactions." *Bioinformatics*.
Van de Vijver, R., et al. (2015). "Automatic Mechanism and Kinetic Model Generation for Gas- and Solution-Phase Processes: A Perspective on Best Practices, Recent Advances, and Future Challenges." *International Journal of Chemical Kinetics* **47**(4): 199-231.
Weininger, D. (1970). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Proc. Edinburgh Math. SOC.*