An experiment on knowledge discovery in chemical databases

Sandra Berasaluce, Claude Laurenço, Amedeo Napoli, Gilles Niel

▶ **To cite this version:**

Sandra Berasaluce, Claude Laurenço, Amedeo Napoli, Gilles Niel. An experiment on knowledge discovery in chemical databases. PKDD 2004 - 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Sep 2004, Pisa, Italy. pp.39-51, 10.1007/978-3-540-30116-5_7 . hal-00162537

HAL Id: hal-00162537
https://hal.science/hal-00162537v1

Submitted on 19 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An Experiment on Knowledge Discovery in Chemical Databases

Sandra Berasaluce[1,2,3], Claude Laurenço[1,2], Amedeo Napoli[3], and Gilles Niel[1]

[1] LSIC – ENSCM, 8, rue de l'Ecole Normale, 34296 Montpellier,
[2] LIRMM, 161, rue Ada, 34392 Montpellier,
[3] LORIA, BP 239, 54506 Vandœuvre-lès-Nancy

Abstract. In this paper, we present an experiment on knowledge discovery in chemical reaction databases. Chemical reactions are the main elements on which relies synthesis in organic chemistry, and this is why chemical reactions databases are of first importance. From a problem-solving process point of view, synthesis in organic chemistry must be considered at several levels of abstraction: mainly a strategic level where general synthesis methods are involved, and a tactic level where actual chemical reactions are applied. The research work presented in this paper is aimed at discovering general synthesis methods from chemical reaction databases in order to design generic and reusable synthesis plans. The knowledge discovery process relies on frequent levelwise itemset search and association rule extraction, but also on chemical knowledge involved within every step of the knowledge discovery process. Moreover, the overall process is supervised by an expert of the domain. The principles of this original experiment on mining chemical reaction databases and its results are detailed and discussed.

Keywords: knowledge discovery, data mining, frequent level-wise itemset search, association rule, knowledge-based system.

1 Introduction

In this paper, we present an experiment on the application of knowledge discovery algorithms for mining chemical reaction databases. Chemical reactions are the main elements on which relies synthesis in organic chemistry, and this is why chemical reaction databases are of first importance. From a problem-solving process point of view, synthesis in organic chemistry must be considered at several levels of abstraction: mainly a strategic level where general synthesis methods are involved, and a tactic level where actual chemical reactions are applied. The research work presented in this paper is aimed at discovering general synthesis methods from chemical reaction databases in order to design generic and reusable synthesis plans. This can be understood in the following way: mining reaction databases at the tactic level for finding synthesis methods at the strategic level. This knowledge discovery process relies on the one hand on mining algorithms, i.e. frequent levelwise itemset search and association rule extraction, and, on the other hand, on domain knowledge, that is involved at every step of the knowledge discovery process.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 39–51, 2004.
© Springer-Verlag Berlin Heidelberg 2004

This research work is carried out within a long-term project for designing chemical information systems whose goal is to help a chemist building a synthesis plan [14, 19]. Actually, the general problem of synthesis relies on the design of a synthesis plan followed by an experimentation of this synthesis plan. Synthesis planning is mainly based on an analytical reasoning process, called *retrosynthesis*, where the first element of the plan is the *target* molecule, i.e. the molecule that has to be built (see fig. 1). This process can be likened to a goal-directed problem-solving approach: the target molecule is iteratively transformed by applying reactions for obtaining simpler fragments, until finding starting materials that are easy to build or to obtain (this constitutes a synthesis pathway). For a given target molecule, a huge number of starting materials and reactions may exist, e.g. thousands of commercially available chemical compounds. Thus, exploring all the possible pathways issued from a target molecule leads to a combinatorial explosion. Therefore the choice of reaction sequences to be used within the planning process is of first importance, and strategies are needed for efficiently solving the synthesis planning problem.

At present, reaction database management systems are the most useful tools for helping the chemist in synthesis planning. Other knowledge systems have been developed since the 70s for helping synthesis planning based on a retrosynthetic approach. The main problem in this kind of system is the constitution of the knowledge base. In our research work, we are designing a new kind of knowledge system for synthesis planning, combining the principles of knowledge systems, database systems, and knowledge discovery [4, 3, 19]. One aspect of this research is to study how data mining techniques may contribute to knowledge extraction from reaction databases, and beyond that, to the structuring of these databases and the improvement in their querying.

This paper presents a preliminary experiment carried on two commercial reaction databases[1] using frequent itemset search and association rule extraction [2, 16]. This study is original and novel within the domain of organic synthesis planning, and is of first importance, with respect to chemical researches. Regarding the knowledge discovery research, we stress the fact that knowledge extraction in an application domain has to be guided by knowledge domain if substantial results have to be obtained. Indeed, the knowledge extraction process is performed under the supervision of a domain expert, but the computing process itself has to be guided by domain knowledge, at every step, i.e. cleaning and transforming data, and interpreting results. We claim that the role of knowledge within the knowledge extraction process is most of the time underestimated, and one of the goal of this paper is to show that taking advantage of the functionalities of a knowledge system within the knowledge discovery process may be of first importance for obtaining accurate and realistic results.

The paper is organized as follows. First, we introduce the chemical context, describing the synthesis problem. Then, we detail the selection and the preprocessing of data, i.e. organic synthesis reactions from reaction databases, and then, the application of data mining techniques to these data, namely frequent

[1] Supplied by Molecular Design Ltd – MDL (`http://www.mdli.com`).

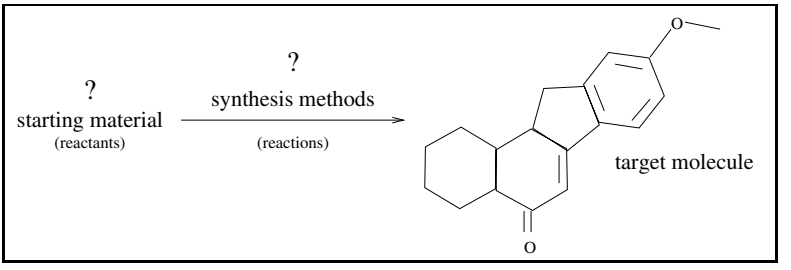

Fig. 1. The general schema of a synthesis problem.

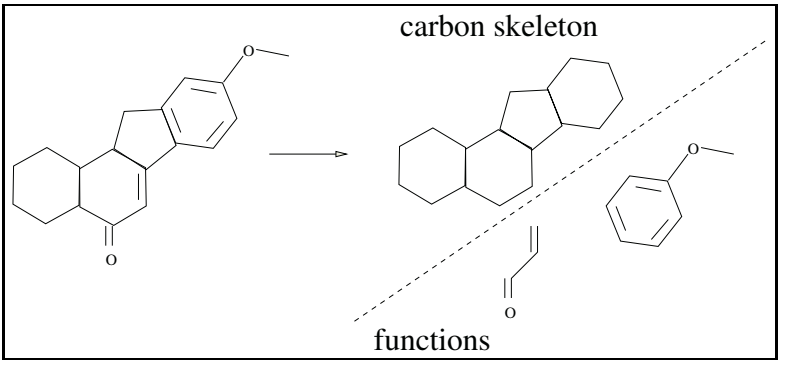

Fig. 2. Skeleton and functional groups of a target molecule.

itemset search and association rules extraction. We show how the results of such a knowledge discovery process give insights for information organization and retrieval within reaction databases. Moreover, the extracted knowledge units after been validated by a chemist may be useful in the search for efficient reactions in association with a synthesis problem. Then we conclude the paper with a discussion regarding the present research work and the research perspectives.

2 The Chemical Context

2.1 The Synthesis Problem

The information needs for a chemist solving a synthesis problem is related to a search in the literature for specific reactions solving synthesis problems considered to be similar to the current one. There is a very huge number of specific reactions described within articles in the literature, certainly more than 10 millions. Reaction documentation is complex and not yet standardized: many classification systems have been proposed, based on reaction mechanism, or electron properties, but they are not really useful for studying synthesis in the large. Actually, the main questions for the synthesis chemist are related to chemical families to which a target molecule belongs, and to the synthesis methods, i.e. a

reaction or a sequence of reactions building structural patterns, to be used for building these families. For the sake of simplicity, we will use hereafter only the term "reaction" for mentioning a basic reaction or a synthesis method as well.

Two main categories of reactions may be distinguished: reactions building the skeleton of a molecule –the arrangement of carbon atoms on which relies an organic molecule–, and reactions changing the functionality of a molecule, i.e. changing a function into another function (see fig. 2). In our framework, a function is mainly used for recognizing a given molecule as a member of a chemical family, for predicting and explaining the molecule reactivity. It is defined as a connected molecular substructure composed of multiple carbon-carbon bonds, carbon-heteroatom bonds –an heteroatom is an atom that is not a carbon atom–, and heteroatom-heteroatom bonds. Here, we are mainly interested in reactions changing the functionality, and in the following questions: (i) what is the starting function $\mathtt{F_i}$ for a given formed function $\mathtt{F_j}$? (ii) what are the reactions allowing the transformation of a function $\mathtt{F_i}$ into a function $\mathtt{F_j}$? (iii) what are the functions $\mathtt{F_i}$ remaining unchanged during the application of a reaction?

2.2 The Reaction Databases: Data Selection and Preprocessing

The experiment reported hereafter has been carried out on two reaction databases, namely the "Organic Syntheses" database ORGSYN-2000 including 5486 records, and the "Journal of Synthetic Methods" database JSM-2002 including 75291 records. The selection of these databases relies on size and quality criteria. In these databases, the filtering of the data related to functional transformations has been performed within a data preprocessing step, where only structural information about the reaction has been considered (details are given in 3.2).

The purpose of the preprocessing step of data mining is to improve the quality of the selected data by cleaning and normalizing the data. Reaction databases such as ORGSYN-2000 and JSM-2002 may be seen as a collection of records, where every record contains one chemical equation involving structural information, that can be read, according to the reaction model, as the transformation of an *initial state* –or the set of *reactants*– into a *final state* –or the set of *products*– associated with an atom-to-atom mapping between the initial and final states (see fig. 3).

In our framework, data preprocessing has mainly consisted in exporting and analyzing the structural information recorded in the databases for extracting and for representing the functional transformations in a target format that has been processed afterwards. The considered transformations are functional modifications, functional addition and deletion, i.e. adding or deleting a function. Moreover, no distinction has been made between one-step or multi-steps reactions. Errors in the atom-to-atom mapping have been neglected, as a reaction is considered at an abstract level, the so-called *block level*, as explained hereafter. The abstraction of a reaction from the atom level into the block level is carried out using the RESYN-ASSISTANT system [19, 3] (some details on RESYN-ASSISTANT are given in § 3.2).

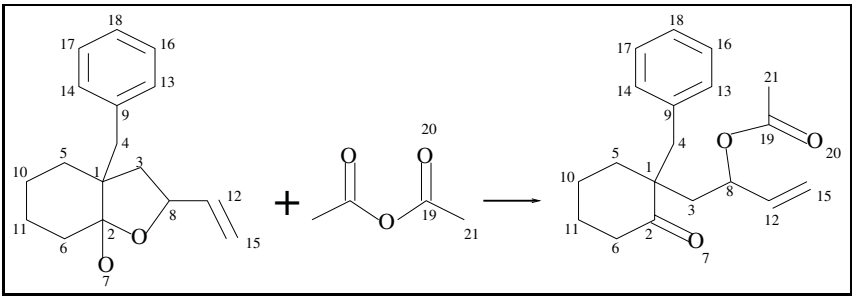

Fig. 3. The structural information on a reaction with the associated atom-to-atom mapping (reaction #13426 in the JSM-2002 database).

In the following, we discuss the whole process of chemical reaction databases manipulation, involving data transformation and data mining, for retrieving and organizing chemical reaction databases.

3 Knowledge Discovery in Reaction Databases

3.1 An Overview of the Knowledge Discovery Process

The knowledge discovery process in chemical reaction databases is considered as an interactive and iterative experimental process. An expert of the data domain, called hereafter the *analyst*, plays a central role in this process since he is in charge of controlling all the steps of the process (as discussed e.g. in [9, 5]). According to given synthesis objectives, the analyst selects first the data to be analyzed, applies data mining modules for extracting knowledge units from data, and finally interprets and validates the units having a sufficient plausibility for being reused. For carrying out this process with benefits, the analyst may take advantage of his own knowledge of the domain –he is an expert–, and, as well, of a set of modules including a knowledge system, ontologies, molecule and reaction databases[2].

Hence, in our approach, the knowledge discovery process is, first of all, guided by the analyst and domain knowledge. The knowledge discovery process itself is based on frequent itemsets search and association rules extraction. Practically, the Close and the Pascal algorithms have been used for data processing [16, 15]. Their application and the results that have been obtained are discussed in the next sections.

3.2 The Modeling of Reactions for Knowledge Discovery

Data on organic reactions are generally recorded in databases within structural and textual entries: the former describes the structural formulae of substances

[2] More generally, the Web in the large could be taken into account if necessary.

in terms of "molecular graphs" while the latter refers to reaction conditions, names and roles of implied substances, bibliographical references, keywords and comments. In our experiment, we have been mainly interested in the so-called functionality changes –or interchanges– occurring during a reaction. These interchanges can be recognized and represented by comparing the functionality of the reactants with that of the products: the removal of some (old) functions and the creation of some (new) functions can be made explicit. The comparison relies on the atom-to-atom mapping where functionality interchanges correspond to the substitution of an atom from one function to another.

Formally, the representation of a reaction equation relies on the atom-to-atom mapping relation between the graphs of the reactants and the graphs of the products, defining three bond sets: the set of the *broken* or *destroyed* bonds, of *formed* bonds and of *unchanged* bonds. Actually these three function modifications correspond to subgoals that are achieved during the synthesis, preparing a main objective [7].

The knowledge system RESYN-ASSISTANT has been designed for assisting the chemist in the design of organic synthesis problems. In particular, the RESYN-ASSISTANT system is able to recognize the building blocks of a molecule, and among these blocks, the functional blocks, or more simply functions. Every function is defined by a name, and is represented as a graph, called functional graph, modeling the structure of the function. The set of functional graphs is partially ordered by a subsumption relation based on a typed subgraph relation. In this way, the set of functional graphs constitutes a concept hierarchy, called $\mathcal{H}_f$, that is part of the knowledge base of the RESYN-ASSISTANT system. Recognizing a function $\mathtt{F_k}$ within a molecule say `M` means that the structure of `M` includes the functional graph $\mathtt{F_k}$ as a subgraph. This recognition process is based on the classification of `M` within the function hierarchy $\mathcal{H}_f$. At present, the $\mathcal{H}_f$ hierarchy includes about five hundred named functions.

The RESYN-ASSISTANT system has been extended to recognize the building blocks of reactions. Based on the atom-to-atom mapping, the system establishes the correspondence between the recognized blocks of the same nature, and determines their role in the reaction. The abstraction of the reaction introduced in figure 3 from the atom-to-atom level into the block level is shown in figure 4. A function may be present in a reactant, in a product, or in both. In the last case, the function is unchanged. In the two other cases, the function in the reactant is destroyed, or the function in the product is formed. During a reaction, either one or more reactant functions may contribute to form the functions in

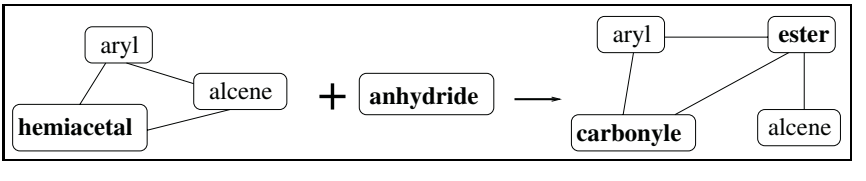

Fig. 4. The analysis of reaction #13426 in the JSM-2002 database in terms of blocks.

the products. At the end of the preprocessing step, the information obtained by the recognition process is incorporated into the representation of the reaction.

For allowing the application of the algorithms Close and Pascal for frequent itemsets search, the data on reaction have to be transformed into a Boolean table. Thus, the representation of a molecule as a composition of functional blocks cannot be used in a straightforward way. Moreover, a reaction can be considered from two main points of view, depending on the fact that the atom-to-atom mapping is taken into account or not (see fig. 5):

- a global point of view on the functionality interchanges leads to consider a single entry R corresponding to an analyzed reaction, to which is associated a list of properties, i.e. formed and/or destroyed and/or unchanged functions,
- a specific point of view on the functionality transformations that is based on the consideration of a number of different entries R_k corresponding to the different functions being formed, i.e. the atom-to-atom mapping gives the explicit correspondence between the blocks that are formed, destroyed, and unchanged (see figure 3).

Entries/Blocks	Destroyed blocks		Formed blocks		Unchanged blocks	
	anhydride	hemiacetal	carbonyle	ester	alcene	aryle
without correspondence entry **R**	x	x	x	x	x	x
with correspondence entry $\mathbf{R}_1$	x	x		x	x	x
entry $\mathbf{R}_2$		x	x		x	x

Fig. 5. The original data are prepared for the mining task: the Boolean transformation of the data can be done by not taking into account the atom mapping, i.e. one single line in the Boolean table, or by taking into account the atom mapping, i.e. two lines in the table.

For example, as shown in figure 5 (in association with the reaction introduced in figure 3), the block correspondence is taken into account implicitly in the first point of view, and explicitly in the second[3]. These two points of view on the analysis of the content of reaction provide two kinds of Boolean tables. The

[3] From a synthesis point of view, the first mode is suitable for studying the chemoselectivity of functionality interchanges, and the second mode is more suitable for comparing the relative reactivities of the studied functions.

rows correspond to the entries related to a single reaction, one row for the global (or implicit) point of view, and two or more for the specific (or explicit) point of view. The columns correspond to the three families of functions, destroyed, formed and unchanged (the same functions are repeated three times, one time per type of columns). Two remarks can be done: firstly, both correspondence have been used during the experiment, and, secondly, in both cases, spatial information on the graph structure of the molecules is lost.

3.3 The Search for Itemsets and the Extraction of Association Rules

The Close and the Pascal algorithms have been applied to Boolean tables (built as indicated just above) for generating first itemsets, i.e. sets of functions (with an associated support), and then association rules. The study of the extracted frequent itemsets may be done with different points of view. Firstly, studying frequent itemsets of length 2 or 3 enables the analyst to determine basic relations between functions. For example searching for a formed functions $\mathtt{F_f}$ ($\mathtt{\_f}$ for formed) deriving from a broken function $\mathtt{F_d}$ ($\mathtt{\_d}$ for destroyed) leads to the study of the itemsets $\mathtt{F_d} \sqcap \mathtt{F_f}$, where the symbol $\sqcap$ stands for the conjunction of functions. In some cases, a reaction may depend on functions present in both reactants and products that remain unchanged ($\mathtt{\_u}$ for unchanged) during the reaction application, leading to the study of frequent itemsets such as $\mathtt{F_f} \sqcap \mathtt{F_u} \sqcap \mathtt{F_d}$. This kind of itemsets can be searched for extracting a "protection function" supposed to be stable under given experimental conditions.

The extraction of association rules gives a complementary perspective on the knowledge extraction process. For example, searching for the more frequent ways to form a function $\mathtt{F_f}$ from a function $\mathtt{F_d}$ leads to the study of rules such as $\mathtt{F_f} \longrightarrow \mathtt{F_d}$: indeed, this rule has to be read in a retrosynthetic way, i.e. if the function $\mathtt{F_f}$ is formed then this means that the function $\mathtt{F_d}$ is destroyed. Again, this rule can be generalized in the following way: determining how a function $\mathtt{F_f}$ is formed from two destroyed functions $\mathtt{F_{d1}}$ and $\mathtt{F_{d2}}$, knowing say that the function $\mathtt{F_{d1}}$ is actually destroyed, leads to the study of the association rules such as $\mathtt{F_f} \sqcap \mathtt{F_{d1}} \longrightarrow \mathtt{F_{d2}}$. It must be noticed that for the sake of simplicity, the examples have been kept formal here. Concrete examples can be found either in [3] or in [4].

As usual, the number of itemsets and of association rules to be considered depends on:

- the way of considering the block correspondence, either implicit (one entry per reaction) or explicit (two or more entries per reaction),
- the minimal value of the support of the itemsets to be considered,
- the confidence level chosen for considering and interpreting the association rules.

The results obtained by the application of the data mining algorithms are discussed in the two next sections, firstly from a chemical point of view, and then from a knowledge discovery point of view.

4 Chemical Interpretation of the Knowledge Extraction Results

A whole set of results of the application of the data mining process on the ORGSYN-2000 and JSM-2002 databases is given in [4]. These results show that both reaction databases share many common points though they differ in terms of size and data coverage, i.e. among 500 functions included in the $\mathcal{H}_f$ hierarchy, only 170 are retrieved from ORGSYN-2000 while 300 functions are retrieved from JSM-2002. The same five functions are ranked at the first places in both databases with the highest occurrence frequency. However, some significant differences can be observed: a given function may be much more frequent in the ORGSYN-2000 database than in JSM-2002 database, and reciprocally. These differences can be roughly explained by different data selection criteria and editor motivations for both databases.

A qualitative and statistical study of the results has shown the following behaviors. Some functions have a high stability, i.e. they mostly remain unchanged, and, in the contrary, some others functions are very reactive, i.e. they are mostly destroyed. All the reactive functions are more present in reactants than in products, and some functions are more often formed. Some functions, that are among the most widely used functions in organic synthesis, are more often present and destroyed in reactants, e.g. `alcohol` and `carboxylic acid`. For example, among the standard reactions involving functions, it is well-known –for chemists– that the `ester` function derives from a combination of two functions, one of them being mostly an `alcohol`. The search for a second function relies on the study of rules such as $\texttt{ester}_\texttt{f} \sqcap \texttt{alcohol}_\texttt{d} \longrightarrow \texttt{F}_\texttt{d}$. The main functions that are retrieved are `anhydride`, `carboxylic acid`, `ester`, and `acyl chloride`. If the chemist is interested in the unchanged functions, then the analysis of the rule $\texttt{ester}_\texttt{f} \sqcap \texttt{alcohol}_\texttt{d} \sqcap \texttt{anhydride}_\texttt{d} \longrightarrow \texttt{F}_\texttt{u}$ gives functions such as `acetal`, `phenyl`, `alkene`, and `carboxylic acid`.

These first results provide a good overview on the function stability and reactivity. They also give partial answers to the questions that have been posed in section 2.1. However, some questions remain open, such as the classification of reactions with respect to a given point of view, e.g. reactivity, stereochemistry,...

Working on functionality interchanges within organic synthesis is a complex problem, and the data mining experiment presented in this paper raises the question of the selection of the reaction databases. The choice of the ORGSYN-2000 and the JSM-2002 databases has been guided by their coverage relevance. The ORGSYN-2000 database provides an electronic version of the entire series of Organic Syntheses, and offers an access to new general synthesis methods. The principle followed by the editors of Organic Syntheses is particularly interesting since each synthesis method has been checked by experts in laboratories. This practice confers to the data a high confidence value, because they have been verified in compound preparations. On the other hand, the JSM-2002 database is a document-based organic reaction database presenting a high coverage in organic synthesis (from 1975). To be selected and recorded, a reaction must be novel or have a particular advantage over an existing one. In addition the reac-

tion must have a clear experimental method, and must be repeatable. For these reasons, the ORGSYN-2000 and the JSM-2002 databases appear to be suitable for a global study of chemical functionality. Both databases contain fine-grained selected data, and thus the information retrieval process is necessarily focused. The ORGSYN-2000 and the JSM-2002 databases have proven to be useful sources for exploring organic synthesis knowledge rather than for providing exhaustive information about particular reactions.

5 Discussion: Frequent Itemsets, Rules and Chemical Reactions

First of all, it can be mentioned that only a few research works hold on the application of data mining methods on reaction databases (see for example [8, 6, 12, 13]). Moreover, these studies have different objectives, and are mainly concerned with molecular graph manipulation rather than reaction database mining. Another study on the lattice-based classification of dynamic knowledge units has been a valuable source of inspiration for the present work [10], leading to the division of functions in three categories, formed, destroyed, and unchanged. The work in [10] is more focused on formal concept analysis and lattice construction rather than on data mining concerns. A number of topics can be discussed here regarding the experiment presented in this paper:

- The abstraction of reactions within blocks and the separation in three kinds of blocks, namely formed, destroyed, and unchanged blocks. Indeed, this is one of the most original idea in that research work, that is responsible of the good results that have been obtained. This idea of the separation into three families may be reused in other contexts involving dynamic data. However, the transformation into a Boolean table has led to a loss of information, e.g. the connection information on reactions and blocks. This loss of information on the connection of the entries introduces a bias in the data mining process, that is quite difficult to take into account.
- Frequent items or association rules are generic elements that can be used either to index (and thus organize) reactions or to retrieve reactions. Termed in another way, this means that frequent itemsets or extracted association rules may be in certain cases considered as a kind of meta-data giving meta-information on the bases that are under study. For example, questions that the chemist wants to be answered are the following: if $\mathtt{A} \longrightarrow \mathtt{B}$ is true, and $\mathtt{B} \longrightarrow \mathtt{C}$ is also true, then it can be deduced that $\mathtt{A} \longrightarrow \mathtt{C}$ is also true, meaning that we can have access to three reactions if needed (or the access to two reactions allows the access to a third inferred reaction).
- Knowledge is used at every step of the knowledge extraction process, e.g. the coupling of the knowledge extraction process with the RESYN-ASSISTANT system, and domain ontologies such as the function ontologies, the role of the analyst,... Indeed, and this is one of the major lesson of this experiment: the knowledge discovery process in a specific domain such as organic synthesis

has to be *knowledge-intensive*, and has to be guided by domain knowledge, and an analyst as well, for obtaining substantial results.

- The role of the analyst includes fixing the thresholds, and interpreting of the results. The thresholds must be chosen in function of the objectives of the analyst, and in function of the content of the databases. A threshold of 1% for an item support means that for a thousand of reactions, ten reactions may form a family: this is not a bad hypothesis. Moreover, if ten thousand reactions are considered, then 1% means that a hundred reactions may form a family, and in this case, this is a very realistic hypothesis. This shows that the thresholds are linked in a very close way to the knowledge of the domain. Here, we see again the influence of the domain: the value of a threshold here is very different from the values that can be used for a threshold in marketing analysis.

  Another remark may be done on what could be called "exceptions", i.e. a reaction that appear only once in a database; this means that there is no other reaction of the same kind in the database, or, that the item associated with this reaction is a unique one. The notion of exception has no substantial meaning here: one unique reaction in one database may be found under several examples in another database. A unique exemplar is rather a matter of point of view taken by the editors of the considered database.
- Other research directions have to investigated, namely sequential patterns [1, 17], or working with closed itemsets and icebergs [18]. Regarding closed itemsets and icebergs, it must be noticed that closed itemsets are the longer itemsets, and those that potentially bring the most of information for the current mining problem. Thus, it could be interesting to consider only these closed itemsets, and to work more in the spirit of formal concept analysis, where a concept lattice –based on closed sets of properties– is built [11].

  Moreover, the use of data mining methods such as frequent itemsets search or association rule extraction has proven to be useful, and has provided encouraging results. It could be interesting to test other (symbolic) data mining methods, e.g. OLAP technology, relational mining, cluster analysis, or Bayesian network classification, knowing that numerical methods such as hidden Markov models or neural networks are not really adapted to the kind of data that are considered in our experiment.

6 Conclusion

In this paper, we have presented an experiment on knowledge discovery in chemical reaction databases. Two databases have been deeply studied and mined, namely the ORGSYN-2000 and the JSM-2002 databases, using frequent levelwise itemset search and association rule extraction. The main topic of interest in the reactions is related with functionality interchanges. Thus, the reactions in the databases have been abstracted in terms of three kinds of building blocks for molecules involved in the reactions, namely formed, destroyed and unchanged blocks. This categorization of blocks has been the basis for building the Boolean

tables on which data mining algorithms such as Close and Pascal have been applied. From a chemical point of view, the results are very encouraging, and provide a set of meta-data for organizing and retrieving chemical reaction according to given synthesis objectives. From a knowledge discovery point of view, a number of questions can be discussed, such as the value of the thresholds (usually lower than in marketing analysis), on the processing of the data, and on the interpretation of the results. Moreover, two major elements have to be pointed out, and can be reused in other contexts : the categorization of dynamic data such as reactions into three families, here, formed, destroyed and unchanged functions, and the use of knowledge at every stage of the knowledge discovery process. Indeed, in a domain such as organic synthesis, the knowledge discovery process has to be fully guided by domain knowledge, and the analyst, an expert of the domain, as well. There are a number of research perspectives following the present work, including the adaptation of sequential pattern algorithms to chemical reactions, taking actually into account the structures of the molecules involved in reactions, and working in the spirit of concept analysis for lattice-based classification of the data.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In P.S. Yu and A.L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering, (ICDE-95), Taipei, Taiwan*, pages 3–14. IEEE Computer Society, 1995.
2. Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *ACM SIGKDD Explorations*, 2(2):66–75, 2000.
3. S. Berasaluce. *Fouille de données at acquisition de connaissances à partir de bases de données de réactions chimiques.* Thèse de chimie informatique et théorique, Université Henri Poincaré Nancy 1, 2002.
4. S. Berasaluce, C. Laurenço, A. Napoli, and G. Niel. Data mining in reaction databases: extraction of knowledge on chemical functionality transformations. Technical Report A04-R-049, LORIA, Nancy, 2004.
5. R.J. Brachman and T. Anand. The Process of Knowledge Discovery in Databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 37–57, Menlo Park, California, 1996. AAAI Press / MIT Press.
6. R. Chittimoori, L. B. Holder, and D. J. Cook. Applying the Subdue substructure discovery system to the chemical toxicity domain. In *Proceedings of the Florida AI Research Symposium*, pages 90–94, 1999.
7. E.J. Corey and X.M. Cheng. *The Logic of Chemical Synthesis.* John Wiley & Sons, New York, 1989.
8. L. Dehaspe, H. Toivonen, and R.D. King. Finding frequent substructures in chemical compounds. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 30–36, 1998.
9. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Proceedings of the Second International Conference on Knowledge Discovery & Data Mining (KDD-96), Portland, Oregon*, pages 82–88, 1996.

10. B. Ganter and S. Rudolph. Formal Concept Analysis Methods for Dynamic Conceptual Graphs. In H.S.Delugach and G. Stumme, editors, *Conceptual Structures: Broadening the Base – 9th International Conference on Conceptual Structures, ICCS-2001, Stanford*, Lecture Notes in Artificial Intelligence 2120, pages 143–156. Springer, Berlin, 2001.
11. B. Ganter and R. Wille. *Formal Concept Analysis*. Springer, Berlin, 1999.
12. A. Inokuchi, T. Washio, and H. Motoda. An Apriori-based algorithm for mining frequent substructures from graph data. In D. Zighed, J. Komorowski, and J.M. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD-2000, Lyon, France, 2000, Proceedings*, Lecture Notes in Computer Science 1910, pages 13–23. Springer, 2000.
13. M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. Technical Report 02–026, Department of Computer Science, University of Minnesota, 2002. To be published in IEEE Transactions on Knowledge and Data Engineering.
14. A. Napoli, C. Laurenço, and R. Ducournau. An object-based representation system for organic synthesis planning. *International Journal of Human-Computer Studies*, 41(1/2):5–32, 1994.
15. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In C. Beeri and P. Buneman, editors, *Database Theory - ICDT'99 Proceedings, 7th International Conference, Jerusalem, Israel*, Lecture Notes in Computer Science 1540, pages 398–416. Springer, 1999.
16. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Pruning closed itemset lattices for association rules. *International Journal of Information Systems*, 24(1):25–46, 1999.
17. M. Sena and G. Karypis. SLPMiner: An algorithm for finding frequent sequential patterns using length-decreasing support constraint. Technical Report 02–023, Department of Computer Science, University of Minnesota, 2002.
18. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with titanic. *Journal of Data and Knowledge Engineering*, 42(2):189–222, 2002.
19. P. Vismara and C. Laurenço. An abstract representation for molecular graphs. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 51:343–366, 2000.