Data mining in reaction databases: extraction of knowledge on chemical functionality transformations

Sandra Berasaluce, Gilles Niel, Amedeo Napoli, Claude Laurenço

Data mining in reaction databases: Extraction of knowledge on chemical functionality transformations

S. Berasaluce[a,b], G. Niel[c], A. Napoli[a], C. Laurenço[b,c]

[a]*Equipe Orpailleur,Laboratoire LOrrain de Recherche en Informatique et ses Applications, UMR 7503 du CNRS, Campus Scientifique, BP 239, 54506 - Vandoeuvre-lès-Nancy*
[b]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506 du CNRS, 161, rue Ada, 34392 - Montpellier Cedex 5,*
[c]*Laboratoire des Systèmes d'Information Chimique, Hétérochimie moléculaire et macromoléculaire, UMR 5076 du CNRS, ENSCM, 8, rue de l'Ecole Normale, 34296 – Montpellier Cedex 5*

Abstract

Some experiments on knowledge discovery in chemical reaction databases are presented. These databases are of first importance in organic synthesis, but their current exploitation remains limited to conventional queries. We assumed that the application of data mining techniques on such bases could bring to the foreground some knowledge elements about synthetic methods, which can be then reused to solve problems of chemical synthesis. The explored domain was focused on functional interchanges. To achieve data mining and problem solving, representing and exploiting of the domain knowledge is a necessary prerequisite. First results of a data mining experiment on reaction databases are presented herein. The frequent itemsets analysis furnishes useful statistical insights on the database's content whereas association rules enable to extract information at a generic level. Both techniques generate metadata which could be reused in designing new database query modes.

Keywords

Knowledge representation; modelling; functional groups; data mining; frequent itemsets; association rules.

1 – INTRODUCTION

This work is part of a long-term project of designing chemical information systems which aim at helping the synthetic chemist.[1-3]

Solving the problem of the synthesis of complex organic substances is a two-step process: make a synthetic plan, then experiment it. Mostly, the planning of synthetic sequences is done by retrosynthesis, i.e. the analytical reasoning by which a chemist starts from the target molecule, transforms iteratively its structure to a sequence of progressively simpler structures along a pathway leading to available starting materials. This method calls for numerous knowledge, notably on the reactions: their various categories, their synthetic uses and their limits, their known examples, etc. The synthetic plan is only a hypothesis which will be confirmed or invalidated at the bench. Such a problem is not a trivial one. For a given target molecule, a huge number of sets of starting materials can be constituted from hundreds of thousand commercially available chemical compounds. The choice of reactions to be used is mostly done by analogy with previously solved problems rather than resulting from theoretical considerations. Furthermore, a synthetic pathway can have several dozens stages. Exploring all the possible pathways would inevitably cause a combinatorial explosion and therefore a chemist needs heuristics in order to solve this problem efficiently. Heuristics are strategies that help to guide a search process in promising directions. Although they do not guarantee to find the best solution, they enable very often to find a good solution. For example, a common heuristic is to consider separately the target skeleton and its functionality, the former to define strategic goals based on molecule topology, the latter to seek out tactics enabling to reach the goals; more specific heuristics may help to find good tactics. At present, the most efficient and useful tools for the chemist in planning syntheses are molecule and reaction database management systems. Various organic reaction databases were created and have been updated regularly since the 80s, e.g. CASREACT, Beilstein or ChemInform. Some of them aim at being general and relatively exhaustive whereas most are specialized in a particular domain.[4] Their originality is to authorize effective structure and substructure searches. However, their management systems are not perfect and resulting queries often produce large sets of answers from which the relevant information is laboriously extracted.

Some other computer systems have been developed since the 70s to aid the synthetic chemist in retrosynthesis, prediction of reactions or search for starting materials.[5-11] However, based on expert knowledge, these systems are difficult to realize and still stay in the state of more or less elaborated prototypes. Therefore, they are of little usefulness. Major problems met in their implementation are the constitution and the update of their knowledge bases. Different approaches using machine learning techniques have been followed to constitute these knowledge bases from reaction databases[12-14] but this problem is still far from being solved.

For our part we are working on the design of new chemical information systems dedicated to problem solving in organic synthesis. These should ideally combine features of database management systems and knowledge-based systems, by using databases in which facts and hierarchically organized concepts are included together. One aspect of our research is to study how data mining techniques could contribute to the extraction of knowledge from reaction databases, and beyond that, the structuring of these databases and the improvement in their query modes.

This paper presents some results we obtained by mining frequent itemsets and association rules in two commercial reaction databases supplied by MDL®.[15] Our study concerns the functional interchanges occurring in these databases. We discuss briefly the synthetic chemist's needs for information on reactions. After a presentation of the problem at a conceptual level, we describe the selection of the data and their preprocessing, then the application of the data mining techniques we have selected. We show that frequent itemsets give an insight into the database

contents and that association rules validated by an analyst may be useful in the heuristic search of known reactions relative to a given problem.

2 – REACTION INFORMATION NEEDS OF THE SYNTHETIC CHEMIST

An information system is supposed to supply services to users. Before starting the construction of such a system, we must have defined exactly its following characteristics: what services could be helpful, in which domain and for whom? The needs of the synthetic chemist for information in devising a synthetic pathway were previously discussed.[16] The decisive factor to succeed in this task is often the retrieval from the literature of individual reactions which solve problems similar to the current one [17]. We do not know exactly how many individual reactions of the organic chemistry have been described until now in the literature, but their number is certainly higher than 10 millions. Reaction documentation is difficult and we can observe that no nomenclature has been developed to name them exactly. Many classification systems have been proposed but these are of little use for the synthetic chemist since they are based on the reaction mechanism, the molecularity of a reaction, or the number of electron pairs involved in a reaction[18] and not on synthetic criteria.

In fact, the basic questions the synthetic chemist asks himself are related to the chemical families to which a target belongs and the methods of getting molecules of these families. So, in designing a synthesis-oriented system we have to distinguish reactions from synthetic methods, two notions that are generally confused. The former are chemical properties of substances whereas the latter are synthetic applications of reactions to form substances from others by selectively building defined structural patterns. A synthetic method may apply a single reaction or a reaction sequence. For example, the so-called ‘Diels-Alder reaction’ and ‘Wittig reaction’ are synthetic methods which apply to a single reaction and a sequence of three (or four) reactions to produce cyclohexene derivatives and alkenes respectively.[19] Synthetic methods are concepts needed to design synthetic plans, while reactions are means to carry out these plans. Since synthetic methods are ways of reaching certain goals, they may be categorized according to this feature. As Ireland pointed out, two general categories of synthetic methods may be defined.[20] Their goals are to build the skeleton and to change the functionality respectively. Each of these categories has more specific subcategories organized in a multi-level hierarchy.

We have discussed elsewhere the synthetic methods for obtaining carbocyclic compounds, which are subcategories of the ‘skeleton construction methods’ category.[21] Here we are dealing with the methods which change the molecular functionality and more specifically with those which form a chemical function from another one. In this context, a chemical information system dedicated to organic synthesis should be able to answer the following queries : (i) From which functions is another given function obtained? (ii) What synthetic methods allow to transform a function into another one? (iii) What functions remain unchanged during the transformation of a function into another one? Furthermore, it should be worthwhile putting the answers in order of pertinence, according to the effectiveness of the methods used.

Some of these questions are usually formulated by substructure queries, others are better expressed via more abstract concepts such as names of methods, transformations or functions. In the latter cases, searching through hierarchies of such concepts that index reactions in a database could complement a substructure search or be an efficient alternative to it [22]. Thus the development of concept-based information retrieval tools is valuable for improving the access to reaction information. With these tools search for information will be based on the meaning of reactions rather than on keywords associated to them. Our approach to generate conceptual indexes of reactions consists in (i) modelling the domain, e.g. transformations of the chemical functions (ii) identifying concepts present inside the structural descriptions of reactions by a perception task (iii)

discovering patterns and rules by a mining task which is carried out over concepts extracted from reaction descriptions (iv) validating the results by an expert in organic synthesis.

3 – KNOWLEDGE DISCOVERY IN DATABASES AND DATA MINING

As seen in Figure 1, data mining is a step within the overall knowledge discovery in databases (KDD) process.[23] A wide range of data mining methods are available. These are often based on numerical techniques but symbolic methods developed in database research have been employed in the context of our work.[24, 25]

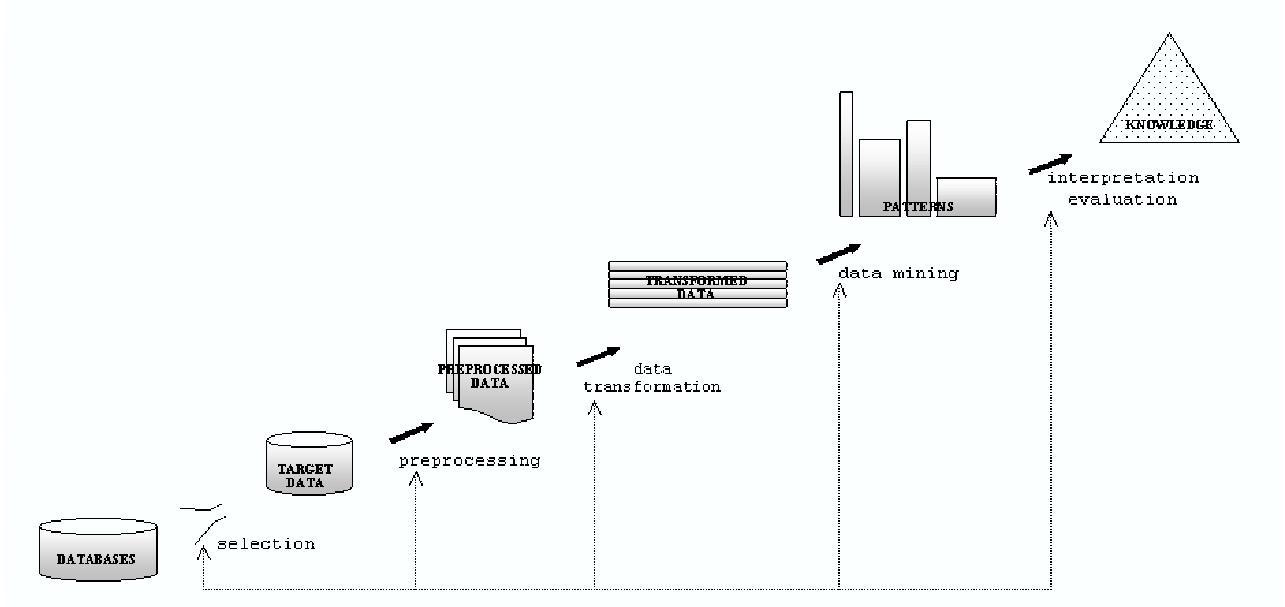

Figure 1. The KDD process.

3.1 – Analyst's role in the knowledge discovery process

Knowledge discovery in databases is an interactive and iterative experimental process. An expert of the data domain, called *the analyst*, plays a pivotal role in this process since he is in charge of controlling each of its steps.[26, 27] According to his objectives, the analyst selects first the data to be analyzed and uses mining tools to extract patterns describing these data. Secondly, the analyst secures and validates patterns which seem to be satisfactory for a further use. To achieve successfully such a process, he uses his knowledge as well as a set of tools and functionalities grouped together into a knowledge extraction system. Ideally this one should have four main components : (i) one or more databases with their management systems, (ii) a knowledge-based system relating to the data domain, (iii) a data mining system, (iv) an interface for interaction and visualization of intermediates and final results.

3.2 – Extraction of frequent itemsets and association rules

The main goal of the data mining algorithms is to search regularities in a large data set. To this purpose different approaches can be used : (i) data analysis methods, (ii) conceptual and lattices classification, (iii) extraction of frequent itemsets and (iv) extraction of association rules. We chose data mining programs which have enabled us to extract frequents itemsets and association rules generated from these frequent itemsets. The ‘Close’ and ‘Pascal’ algorithms [28, 29] we have used process data described as lists of objects and associated properties. Extraction of association rules is based on the concept formalization developed in the field of formal concept

analysis [30]. Some definitions and notations this mathematical framework offers are expressed below.

Let $\mathcal{O}$ be a finite set of objects, $\mathcal{P}$ a finite set of properties (items) and $\mathcal{R}$ a binary relation between these two sets. A *database or formal context* is the boolean table $\mathcal{O} \times \mathcal{P}$. In this table x $\mathcal{R}$ y = 1 if the object x and the property y are related through the relation $\mathcal{R}$, then the object x is said to *possess* the property y. An *itemset* is a subset of $\mathcal{P}$. An itemset P is *included* in the object O if P and O are in relation. The support of an itemset sup(P) is defined as the ratio between the number of occurrences of this itemset and the number of objects in the database:

$$\text{sup(P)} = \frac{\text{card}(f(\text{P}))}{\text{card}(\mathcal{O})} \qquad \text{with } f(\text{P}) = \{\text{O} \in \mathcal{O} / \text{ O includes P}\} \qquad \text{Eq. 1}$$

An itemset P is called *frequent* if sup(P) ≥ minsup, where minsup is a threshold defined by the analyst. The data mining algorithms aim at calculating *association rules* from the extracted frequent itemsets. An *association rule* is a conditional implication comprising one premise and one conclusion and is characterized by a *confidence*. The premise and the conclusion of a rule are itemsets, i.e. property conjunctions. Such a rule may be represented by "R : A ⇒ B". The confidence conf(R) is calculated from the supports of the premise itemset sup(A) and of the conjunction of the premise and conclusion itemsets, sup(A ∪ B), according to the following formula:

$$\text{conf(R)} = \frac{\text{sup(A} \cup \text{B)}}{\text{sup(A)}} \qquad \text{Eq. 2}$$

The support of the rule is: sup(R) = sup(A ∪ B) ≥ minsup for, in theory, any rule is derived from a frequent itemset A ∪ B.

4 – KNOWLEDGE DISCOVERY IN CHEMICAL REACTION DATABASES

4.1 - Modelling

Published results on organic reactions are generally archived in databases by means of two classes of data: structural and textual data. The latter refer to reaction conditions, the names and roles of implied substances, bibliographical references, keywords and comments. The former describe the structural formulae of substances in terms of 'molecular' graphs.[30] The representation of a reaction equation relies on the role of the substances involved in it as well as on the atom-to-atom mapping relation between the graphs of the reactants and those of the products.

This relation defines the reaction centre, i.e. the set of all the bonds which are destroyed, created or whose type is changed during the reaction.[31, 32] Such a representation of reactions doesn't get beyond the atom and bond level. It supplies no explicit information on the molecular functionality and its modification during the reaction.

Corey and coworkers have previously formalized the functionality changes within the framework of retrosynthesis by introducing the notions of functional group interchange (FGI), addition (FGA) and removal (FGR).[33, 34] These notions allow to denote subgoals to be achieved before a main goal, e.g. the application of a simplifying transform.[30] They also allow to categorize the non-simplifying transforms which realize the subgoals. This seems to be a good starting point for our modelling, however the prospect of having to formulate queries in natural

language raises particular problems. For example, we can wonder what sort of answer a chemist expects by asking a question such as: *can we obtain an ester from a hemiacetal*?

This sentence can be analyzed in various ways because it expresses some concepts ambiguously. What are denoted by ester and hemiacetal? Substances belonging to the classes of compounds named 'ester' and 'hemiacetal' or the functional groups which characterize these classes? In the first case the reaction shown in Figure 2, Equation 3 is an acceptable answer because the product which belongs to the ester class, among other classes, is obtained from a reactant which belongs to the hemiacetal class, among others. Nevertheless, such an answer is probably not expected by the chemist. In the second case, the sentence will be interpreted as a question on the interconversion of a functional group into another one. Equations 5-7 present relevant answers to this question.

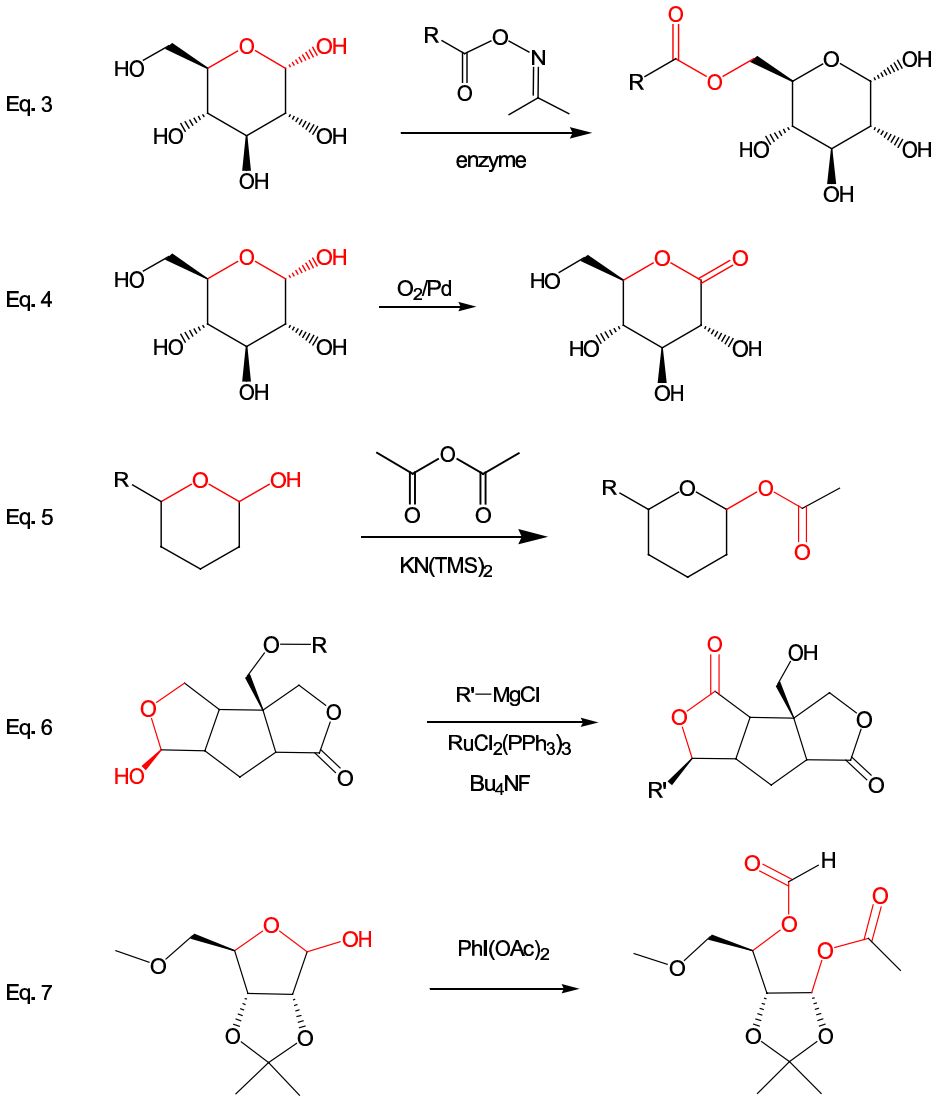

Figure 2. Distinction between substances and functional groups. Equations 3-7 refer respectively to references.[35-39]

We notice that each reaction of this answer set exhibits a different structural relation between the reactant and product functional groups. So four different queries should be submitted to retrieve this set by performing substructure searches in a reaction database. We also notice that only the reaction shown in Equation 4 corresponds to a 'FGI transform' whereas the one shown in Equation 7 corresponds to a 'disconnective FGI transform', i.e a FGI transform associated with a carbon-carbon single bond disconnection.

However, the *functional group* concept is badly defined. It gives place to different interpretations according to the context in which it is used, e.g. to recognize a given molecule as a member of a chemical family, to give it a systematic name, to predict or explain its reactivity, etc. The proposed empirical[40] and theoretical[15, 41] definitions are loose or subject to debate and\or difficult to implement. Usually, this concept is described in extension, by listing the individual groups it covers though not any real consensus has been reached on their structural definitions. The corresponding sets, established according to the chemists' experience, can include some dozens,[42] several hundreds[41] or thousands of elements depending on whether only basic functional groups or also composite or complex ones are considered.[22]

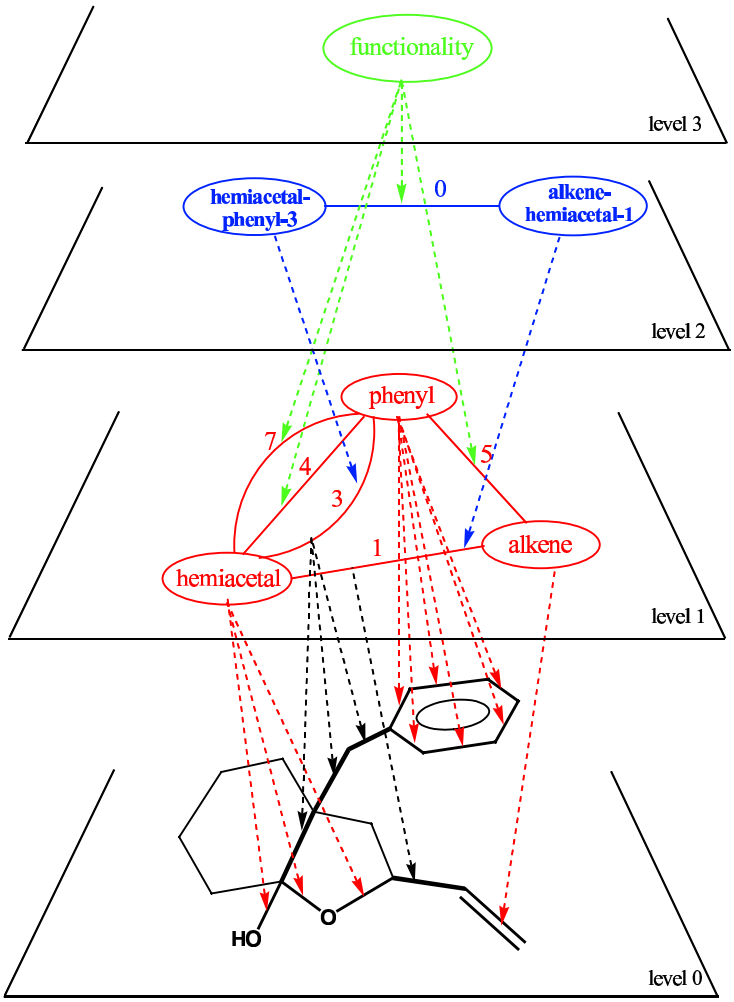

Figure 3. Hierarchical representation of molecular functionality.

In order to avoid conceptual ambiguity and to determine easily the functionality changes occurring during a reaction, the functional group notion has been replaced in our model by a more formal one that we have named *Function*. We define a Function as **a connected molecular substructure which comprises exclusively carbon-carbon multiple, including aromatic, bonds and\or carbon-heteroatom bonds and\or heteroatom-heteroatom bonds**. This definition determines an edge-partition of the molecular graph from which we can build a representation of the functionality. This representation organizes information at several levels of abstraction in a hierarchical graph (Figure 3). By passing from level 0 where the molecular graph is located, to level 1 which contains the graph of functional blocks, we aggregate the edges of each subgraph representing a function into a node labelled by the name of this function and each path representing

a carbon-carbon single bond or a chain of such bonds joining two functions into an edge labelled by the path length.

Like in the example shown in Figure 3, a graph of functional blocks may be a multigraph. Some associations of functions make up remarkable functional systems (for example: amine-carboxyl, dicarbonyl, glycosyl, etc.) which are represented at level 2. By passing at this level we aggregate the edges of the corresponding subgraphs in level 1 into nodes labelled by the names of the systems. When several functional systems are present, the corresponding nodes are joined by edges labelled by the lengths of the paths connecting them.

This hierarchical representation of molecular functionality is one of the parts of the molecule representation we use in computer-aided synthesis planning. The latter, which also comprises topological [3] and stereochemical [43] points of view, will be described in a more formal and detailed way elsewhere.[44] In the context of the present work we have used essentially level 0 and level 1 of our representation. From it the functionality changes occurring during a reaction can be recognized and represented. By comparing the functionality of the reactants with that of the products we observe the disappearance of some functions and the appearance of new ones whereas the others remain unchanged. This comparison relies on the atom-to-atom mapping which particularly allows to recognize the functional interconversions. For perceiving them it is sufficient to notice that at least one atom of a missing function corresponds to one atom of a formed one. The functionality transformations are represented at level 1 by a node mapping between reactant and product functional blocks which is not necessarily a one-to-one mapping. In the example shown in Figure 4, the carboxylate function is formed from both anhydride and hemiacetal functions, this last one being also at the origin of the carbonyl function, while the alkene and phenyl functions remain unchanged.

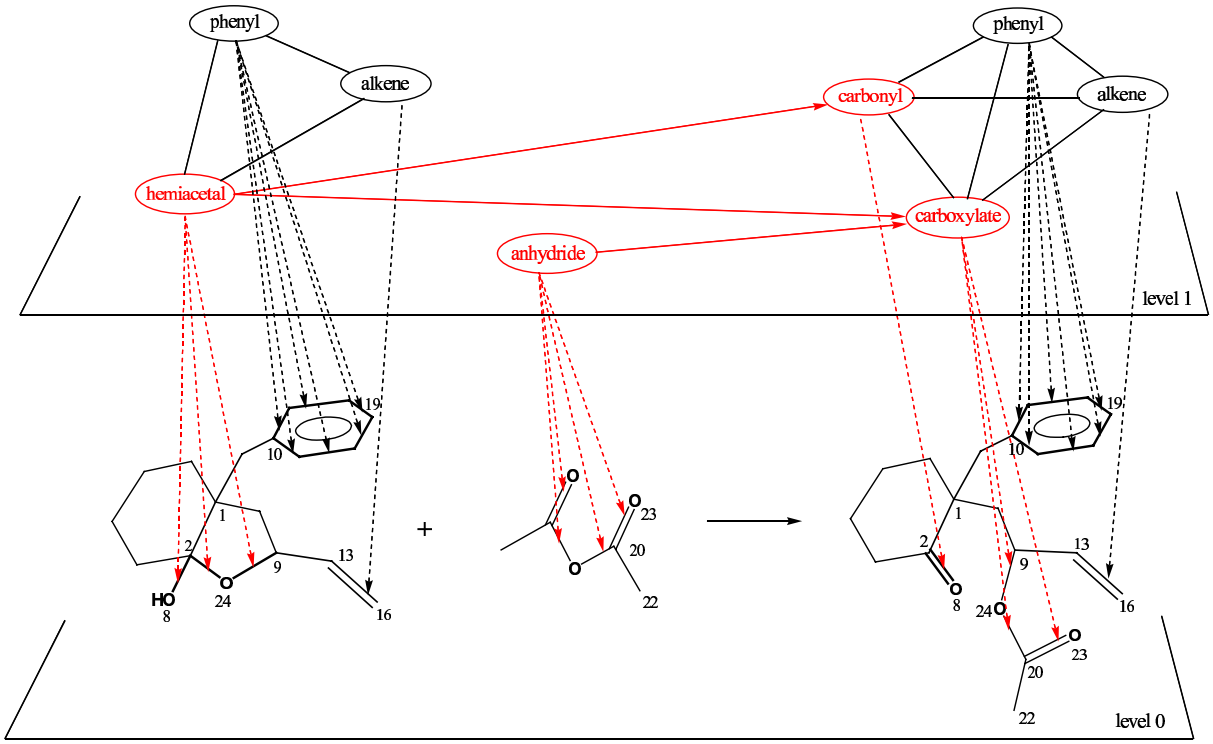

Figure 4. Functional blocks representation of the JSM2002-record-13426.

4.2 - Data selection

The results reported herein were obtained from commercial reaction databases supplied by MDL®.[15] From the database members of the *Synthetic Methodology* set that may be queried with the ISIS system and that contain information on over one million of reactions altogether, we selected only two of them: the *Organic Syntheses* database (ORGSYN2000) and the *Derwent Journal of Synthetic Methods* database (JSM2002). Our criteria of choice were based on their size and quality, both topics will be discussed in section 5. For developing our methodology and programs and for performing these first data mining experiments, we started with ORGSYN2000 which stored 5486 records. The greater size of JSM2002 (75291 records) enabled us afterwards to process a larger volume of information as required by data mining techniques. Selection of the data related to functional interconversions was achieved during the data preprocessing step. It should be noticed that only structural information was considered. We left out data supplying little or no information on the molecular functionality. In this connection we observed that JSM database's fields KEYPHRASES and COMMENTS did not relate to the electronic card but to the original document. Since this one may encompass different cards no specific keywords can be deduced for a given card.

4.3 – Data preprocessing

The purpose of the preprocessing step of data mining is to improve the quality of selected data by cleaning and normalizing them. The recorded information in databases such as ORGSYN2000 or JSM2002 is said to be about particular chemical reactions, and it can be visualized with ISIS/Base as a collection of electronic cards. But what is called a reaction in this context is not obvious. Examining a sample of these electronic cards soon shows that this sample is heterogeneous due to the looseness of the reaction definition in the database conceptual scheme. According to the reaction model, the structural information is mainly involved in an initial state - the reactant set - and a final state - the product set - along with an atom-to-atom mapping between them, but not in any explicit intermediate state. So a card always contains only one chemical equation, which may describe (i) a one-step reaction that leads to a single product, (ii) a set of parallel reactions in which the same reactants lead to several different products (Figure 5), (iii) a reaction sequence involving either single or concurrent reactions but in an overall scheme without details about the intermediates. The corresponding equation is seldom stoichiometric. Moreover, in some cases the mapping relation may be left out, incomplete or chemically irrelevant.[4]

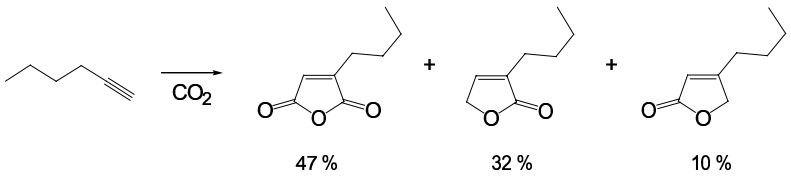

Figure 5. Example of a concurrent reaction from JSM2002-record-5542[45]

In the present work, data preprocessing had consisted in exporting and analyzing the structural information recorded in the databases in order to extract all the functional interconversions and to represent them in a unique format based on our model. During the extraction process, (i) not only pure functional interconversions were recovered but also all transformations adding a function on a non-functionalized carbon atom as well as the function cleavages leading to a new C-C bond,[42] (ii) no distinction was made between one-step reactions and reaction sequences described as single steps. For a given record describing the formation of n

different products, n different representations were generated, each of them describing the formation of one product. Stoichiometry was partially restored by duplicating the representation of some reactants. Faulty atom-to-atom mapping within functions had not to be corrected for it had little harmful effect on the block mapping calculated during the abstraction process when passing from level 0 to level 1 (§ 4.1).

In practice, the ISIS system was used to export the information from databases to Reaction Data Files (RDFs).[46] The subsequent parts of data preprocessing were automatically carried out by RESYN_Assistant.[47] This computer program has been designed to assist the synthetic chemist in the understanding of organic synthesis problems. That means that RESYN_Assistant is able to perceive a target molecule, i.e. to recognize various features from its molecular graph and categorize it according to these features, and to represent several viewpoints of it. The new hierarchical and multi-viewpoint representation should correspond to possible actions that could be performed to get this molecule, e.g. a given carbocyclic molecule containing a 6-membered ring could be obtained via a Diels-Alder cycloaddition. The perception of a molecule involves identifying it as a member of one or more chemical classes. In the case of functionality-based classes, this is done by RESYN_Assistant in three successive steps. First, the system detects the aromatic bonds of the molecule. To do that it computes the set of relevant cycles[3] to which it applies an effective aromaticity-detection algorithm[48, 49] and it gives the aromatic type to the aromatic ring bonds. Second, the functions present in the molecule are all recognized by using an algorithm that produces results consistent with our definition (§ 4.1). Third, each function is identified by a name via a classification process. To this purpose RESYN_Assistant uses a knowledge base in which a function is defined by a graph associated with a term that gives a chemical sense to it, e.g. primary amine, alkynyl ether or phosphinate. Each graph describes the specific structural features of a function and the set of all the graphs is partially ordered by a substructure relation. This relation exists when a graph is a subgraph of one or more others. The Hasse graph of this poset defines the function hierarchy which is searched through by the classification process. A function within the molecule is identified during this process when an isomorphism of its graph with one of the hierarchy is found. Then it takes the corresponding name. This one refers to the function as well as to the chemical class whose the molecule is a member. The knowledge base currently allows to name about five hundred functions, so a molecule may embody functions that cannot be identified by the classification process. Such functions are either named by the user and added as new ones to the knowledge base, or let as “unnamed".

The perception module of RESYN_Assistant has been extended to perceive reactions. From an imported RDF the program creates reaction files in its own internal representation format, each of them describing a single product reaction in an object representation. For each new file RESYN_Assistant perceives the reactant(s) and product. From the given atom-to-atom mapping, it establishes the correspondence between the recognized blocks of the same nature and determines their involvement in the reaction. As seen previously, a functional block can be present in a reactant or in the product, or in both molecules. In this last case it remains unchanged. In other cases the reactant block is destroyed and the product block is created. This corresponds either to an interconversion, or an elimination, or an addition of function. During an interconversion (i) one or more reactant blocks can contribute to create the product block, (ii) a product can possess several formed functions. At the end of the preprocessing step all this information is incorporated into the representation of the functional viewpoint of the reaction that the RESYN_Assistant file describes.

In practical terms, data preprocessing from ORGSYN2000 database used hierarchies of 482 named functions and 1213 unnamed functions. We observed that 5476 records could be strictly converted from the 5486 records present in the initial RDF. The reasons why are : i) the program does not yet recognize some atoms such as deuterium or polymer-labelled atoms, ii) the calculation of the aromaticity for fullerene-like compounds fails. These 5476 records were broken down into

5700 basic reactions containing a single product and corresponding to 6275 functional interconversions. These interconversions could be splitted as follows:

named function	→	named function	3931
unnamed function	→	named function	561
named + unnamed function	→	named function	198
named + unnamed function	→	unnamed function	1585
Total			6275

The above data show the significant role of the unnamed functions during the preprocessing step. It is worth mentioning that half of these unnamed functions includes aromatic derivatives with at least one heteroatom attached to the arene ring. Data preprocessing from JSM2002 was performed on 75291 records that could be converted in 74295 records then in 75516 basic reactions. The data preprocessing resulting files are stored in a folder tree containing (i) the data related to the data mining experiment (see § 4.4), (ii) the structural data related to the functional interconversions (see § 4.6).

4.4 - Data transformation

As seen in section 3, data processed by data mining algorithms we used must conform to a *boolean table* format. Thus we could not directly use abstract representations like block graphs obtained after the preprocessing step but we had to find a means to express the preprocessed data with a minimal lost of information. The level 1, as shown in Figure 4, describes the whole functionality transformations related to a particular reaction.

	destroyed blocks			created blocks			unchanged blocks		
functions / objects	anhydride		hemiacetal	carbonyle		carboxylate	alkene		phenyl
T[1]	X		X	X		X	X		X

	destroyed blocks			created blocks			unchanged blocks		
functions / objects	anhydride		hemiacetal	carbonyle		carboxylate	alkene		phenyl
T1[2]	X		X			X	X		X
T2[2]			X	X			X		X

[1] with a global consideration of the block correspondence. [2] with a specific consideration of the block correspondence.

Figure 6. Boolean tables obtained by data transformation.

In order to describe the functional interconversions, we considered two distinct modes (Figure 6) : (i) a global consideration of the functionality transformations leads to create a single object T, per analyzed reaction, to which is associated a list of properties, i.e. created and/or destroyed and/or unchanged functions, (ii) a specific consideration of the functionality

transformations is based on the creation of as many different objects Tn as different functions were formed.

The first mode is suitable to study the chemoselectivity of interconversions and was used in the case of the acetal function (§ 4.6); the second mode is more suitable to compare the relative reactivities of the studied functions. Whatever the selected mode spatial information linking any nearby functions is lost.

4.5 – Data mining

The 'Close' and 'Pascal' algorithms, applied to the above boolean tables, generate a large number of k-itemsets, i.e. itemsets of size k, then association rules that are stored in separate ASCII files. Each itemset within an itemset file states a list of properties and its calculated support. Afterwards we considered the support sup(P) of an itemset equal to card (f(P)) in order to handle integer values rather than real ones (see Eq. 1). For example the 3-itemset "carboxylic-acid$_d$ AND primary-amine$_d$ AND secondary-amide$_f$" has a support equal to 121 and this means that the occurrence of the destroyed carboxylic-acid and primary-amine functions and a formed secondary-amide function stands at 121. Since the support of the 2-itemset "carboxylic-acid$_d$ AND primary-amine$_d$" is 154, the confidence of the derived association rule "carboxylic-acid$_d$ AND secondary-amide$_f$ => primary-amine$_d$" is equal to 78.6. This conditional implication means that if the premise "a carboxylic-acid function is destroyed and a secondary-amide function is formed" is true, then the conclusion "a primary-amine function is destroyed" is also true in 78.6 % of the cases present in the database.

The number of itemsets depends (i) on the way of considering the block correspondence, either global or specific, (ii) on the minsup value relative to the frequent itemsets, (iii) on the confidence level selected for producing the association rules. Corresponding data are shown in Table 1.

Table 1. Number of itemsets and association rules resulting from the data mining.

		ORGSYN2000		JSM2002	
		global[1]	specific[2]	global[1]	specific[2]
Itemsets	minsup > 1	26.053	9.707	504.316	139.159
	minsup > 10	659	543	12.834	7.326
	minsup > 100	41	41	1.089	763
Association rules	minsup > 10 and confidence > 0	1.366	1.048	Nd	39.496
	minsup > 10 and confidence > 50	78	140	Nd	2.687
	minsup > 1 and confidence > 0	427.908	72.882	Nd	nd
	minsup > 1 and confidence > 50	225.800	23.801	Nd	1.326.268

[1)] by global consideration of the block correspondence; [2)] by specific consideration of the block correspondence.

4.6 – Interpretation - Evaluation

4.6.1 – From data preprocessing

The RESYN_Assistant data preprocessing (cf. section 4.3) of both ORGSYN and JSM2002 databases led to the results expressed in Tables 2 and 3; only the first 30 most frequent functions are displayed herein. Our results show that both reaction databases share common points though they differ in terms of size and data coverage. Thus among the 482 functions included in our function knowledge base, only 170 are recovered from ORGSYN while 297 functions were recovered from JSM2002.

Table 2 : Most frequent functions found in ORGSYN

Entry	function[1]	n_{react}[2]	n_{prod}[3]	n_{des}[4]	n_{form}[5]	n_{unch}[6]	n_{tot}[7]
1	carbonyl	1545	1259	970	684	575	2229
2	alkene	1378	1234	847	703	531	2081
3	phenyl	1724	1649	233	158	1491	1882
4	carboxylate	1344	1102	656	414	688	1758
5	alcohol	1095	741	856	502	239	1597
6	carboxylic-acid	724	718	465	459	259	1183
7	alkyl-bromide	443	213	405	175	38	618
8	alkyl-chloride	385	174	320	109	65	494
9	nitrile	322	290	154	122	168	444
10	ether	258	312	113	167	145	425
11	N-alkyl-amine	228	183	181	136	47	364
12	acyl-chloride	239	54	237	52	2	291
13	acetal	159	162	94	97	65	256
14	alkyne	171	106	128	63	43	234
15	anhydride	210	35	198	23	12	233
16	N,N-dialkyl-amine	116	105	95	84	21	200
17	carbonyl-ligand	162	91	79	8	83	170
18	aniline	143	21	136	14	7	157
19	carboxylate-ion	87	51	83	47	4	134
20	N,N,N-trialkyl-amine	65	82	40	57	25	122
21	phenol	94	51	66	23	28	117
22	N-alkyl-amide	65	60	54	49	11	114
23	nitrobenzene	82	74	33	25	49	107
24	enol-ether	67	40	57	30	10	97
25	pyridine	73	65	31	23	42	96
26	phenate	54	77	16	39	38	93
27	N,N-dialkyl-amide	62	47	40	25	22	87
28	gem-trifluoride	79	69	17	7	62	86
29	nitro	53	44	32	23	21	76
30	alkyl-iodide	39	38	38	37	1	76

[1)] ranked by decreasing order on the column n_{tot}; [2)] number of occurrences for which the function is present within the reactants; [3)] number of occurrences for which the function is present within the products; [4)] number of occurrences for which the function, present in the reactants, is destroyed; [5)] number of occurrences for which the function, present in the products, is formed; [6)] number of occurrences for which the function, present in both reactants and products, remains unchanged; [7)] number of occurrences for which the function is present as a destroyed, formed and unchanged function.

Table 3 : Most frequent functions found in JSM2002

Entry	function[1]	n_{react}[2]	n_{prod}[3]	n_{des}[4]	n_{form}[5]	n_{unch}[6]	n_{tot}[7]
1	phenyl	38372	36942	3801	2371	34571	40743
2	alkene	26796	21956	17748	12908	9048	39704
3	carbonyl	22717	16878	14202	8363	8515	31080
4	carboxylate	20928	20370	5352	4794	15576	25722
5	alcohol	11013	12044	7470	8501	3543	19514
6	ether	8372	8033	3128	2789	5244	11161
7	nitrile	5161	3855	2285	979	2876	6140
8	carboxylic-acid	3662	2911	2744	1993	918	5655
9	acetal	4193	3556	1876	1239	2317	5432
10	alkyne	4596	1403	3746	553	850	5149
11	carbonyl-ligand	4458	1605	2975	122	1483	4580
12	alkyl-bromide	3594	720	3428	554	166	4148
13	phenate	3017	2983	553	519	2464	3536
14	N-alkyl-amine	2460	1287	2179	1006	281	3466
15	alkyl-chloride	2490	1258	2025	793	465	3283
16	N-alkyl-amide	1889	2458	747	1316	1142	3205
17	N,N,N-trialkyl-amine	1380	1921	620	1161	760	2541
18	N,N-dialkyl-amine	1599	1123	1406	930	193	2529
19	enol-ether	1598	1075	1350	827	248	2425
20	gem-trifluoride	2203	1838	507	142	1696	2345
21	N,N-dialkyl-amide	1390	1541	713	864	677	2254
22	trialkyl-silylether	1705	1479	733	507	972	2212
23	acyl-chloride	1832	97	1813	78	19	1910
24	gem-difluoride	1361	1342	289	270	1072	1631
25	tetraalkyl-silane	1166	685	874	393	292	1559
26	N,O-carbamate	1078	1148	369	439	709	1517
27	phenol	1154	789	725	360	429	1514
28	alkyl-iodide	909	456	887	434	22	1343
29	chlorobenzene	1250	1095	243	88	1007	1338
30	N-alkyl-imine	1004	469	866	331	138	1335

[1)] ranked by decreasing order on the column n_{tot}; [2)] number of occurrences for which the function is present within the reactants; [3)] number of occurrences for which the function is present within the products; [4)] number of occurrences for which the function, present in the reactants, is destroyed; [5)] number of occurrences for which the function, present in the products, is formed; [6)] number of occurrences for which the function, present in both reactants and products, remains unchanged; [7)] total number of occurrences for which the function is present as a destroyed, formed and unchanged function.

The five first functions are present in both databases with high occurrence frequencies. The following functions differ noticeably in their respective ranking; we observe that some functions, though recovered from ORGSYN database (Entries 15, 18 and 19 – Table 2) are not included in the most frequent functions recovered from JSM database. The reverse is also true (Entries 25, 29 and 30 – Table 3). For example, the carboxylate-ion function (Entry 19 – Table 2) is highly frequent since ORGSYN database contains many preparations of carboxylic acid salts.

The occurrence frequencies observed between both databases show significant differences; the carboxylic acid function is really more frequent in ORGSYN database than in JSM database. On the opposite the carbonyl-ligand function (C≡O) is twice more present in JSM database than in ORGSYN database. These discrepancies could be explained by different data selection criteria and editor motivations.

Table 4 : Reactivity and/or stability of the most frequent functions found in JSM

Entry	functions[1]	p_r[2]	p_p[2]	p_{dr}[2]	p_{cp}[2]	p_{ur}[2]	p_{ud}[2]
1	phenyl	94.2	90.7	9.9	6.4	90.1	93.6
2	alkene	67.5	55.3	66.2	58.8	33.8	41.2
3	carbonyl	73.1	54.3	62.5	49.5	37.5	50.5
4	carboxylate	81.4	79.2	25.6	23.5	74.4	76.5
5	alcohol	56.4	61.7	67.8	70.6	32.2	29.4
6	ether	75.0	72.0	37.4	34.7	62.6	65.3
7	nitrile	84.1	62.8	44.3	25.4	55.7	74.6
8	carboxylic-acid	64.8	51.5	74.9	68.5	25.1	31.5
9	acetal	77.2	65.5	44.7	34.8	55.3	65.2
10	alkyne	89.3	27.2	81.5	39.4	18.5	60.6
11	carbonyl-ligand	97.3	35.0	66.7	7.6	33.3	92.4
12	alkyl-bromide	86.6	17.4	95.4	76.9	4.6	23.1
13	phenate	85.3	84.4	18.3	17.4	81.7	82.6
14	N-alkyl-amine	71.0	37.1	88.6	78.2	11.4	21.8
15	alkyl-chloride	75.8	38.3	81.3	63.0	18.7	37.0
16	N-alkyl-amide	58.9	76.7	39.5	53.5	60.5	46.5
17	N,N,N-trialkyl-amine	54.3	75.6	44.9	60.4	55.1	39.6
18	N,N-dialkyl-amine	63.2	44.4	87.9	82.8	12.1	17.2
19	enol-ether	65.9	44.3	84.5	76.9	15.5	23.1
20	gem-trifluoride	93.9	78.4	23.0	7.7	77.0	92.3
21	N,N-dialkyl-amide	61.7	68.4	51.3	56.1	48.7	43.9
22	trialkyl-silylether	77.1	66.9	43.0	34.3	57.0	65.7
23	acyl-chloride	95.9	5.1	99.0	80.4	1.0	19.6
24	gem-difluoride	83.4	82.3	21.2	20.1	78.8	79.9
25	tetraalkyl-silane	74.8	43.9	75.0	57.4	25.0	42.6
26	N,O-carbamate	71.1	75.7	34.2	38.2	65.8	61.8
27	phenol	76.2	52.1	62.8	45.6	37.2	54.4
28	alkyl-iodide	67.7	34.0	97.6	95.2	2.4	4.8
29	chlorobenzene	93.4	81.8	19.4	8.0	80.6	92.0
30	N-alkyl-imine	75.2	35.1	86.3	70.6	13.7	29.4

[1] ranked by decreasing order of frequency of occurrences Fo; [2] p_r, p_p, p_{dr}, p_{cp}, p_{ur} and p_{ud} are expressed as a percentage of respectively n_{react}/n_{tot}, n_{prod}/n_{tot}, n_{des}/n_{react}, n_{form}/n_{prod}, n_{unch}/n_{react} and n_{unch}/n_{prod}.

To get a deeper insight into function stability and reactivity, we converted the results expressed in Tables 2 and 3 and calculated the following percentages (Table 4) :

- $p_r = n_{react}/n_{tot}$ and $p_p = n_{prod}/n_{tot}$ indicate how frequent is a given function in reactants and products respectively.

- $p_{dr} = n_{des}/n_{react}$ and $p_{ur} = n_{unch}/n_{react}$ specify the relative proportions of destroyed and unchanged functions respectively over the total number of functions recovered during the data preprocessing. These percentages give some information on function reactivity, i.e. a given function is more reactive as p_{dr} increases.

- $p_{cp} = n_{form}/n_{prod}$ and $p_{ud} = n_{unch}/n_{prod}$ indicate rather a need for a function to be formed. We may say that for high p_{cp} values this function is more accessible in terms of preparation, i.e. a significant number of related synthetic methods or at least efficient and proven methods can create such a function.

As a consequence of the preceding remarks, analysis of the JSM database results shows a high stability for some functions – p_{ur} and/or $p_{ud} > 80\%$ – a high reactivity for some others – $p_{dr} >$

80%. In the case of phenyl, phenate, gem-trifluoride, chloro- and nitrobenzene functions we observe a large stability, a well-known behaviour for most chemists. On the other hand it is worth commenting on the carbonyl-ligand function which is more present in reactants than in products with a quite high p_{dr} value. In fact this function is often used to introduce a carbonyl group for ester formation or cyclic ketones. This latter use is illustrated during the Pauson-Khand reaction. Since the carbonyl-ligand function is widely used as a ligand with transition metals, it shows a weak frequency in the products and a very high p_{ur}.

Among the reactive functions we find the acyl-chloride and alkyl-bromide functions (Entries 23, 12), the *N*-alkyl-amine, *N*,*N*-dialkyl-amine, enol-ether, alkyne and alkyl-chloride functions (Entries 14, 18, 19, 10 and 15). Somewhat less reactive are the tetraalkyl-silane, carboxylic acid, alcohol, alkene, phenol and carbonyl functions ($60 < p_{dr} < 80\%$). All these reactive functions are more present in reactants than in products ($p_r > p_p$).

Some functions are either often or quite often formed ($p_{cp} > 60\%$ – Entries 5, 8, 12, 14, 15, 17, 18, 19, 23 – Table 3). Other functions have a dual behaviour (p_{dr} and $p_{cp} > 60\%$); this is the case for alcohol, carboxylic acid, alkyl-bromide, *N*-alkyl-amine, alkyl-chloride, *N*,*N*-dialkyl-amine, enol-ether and acyl-chloride functions. Since all these ones are the most widely functions used in organic synthesis their high p_{dr} and p_{cp} values are not surprising.

These preliminary results give an overview on the function stability and reactivity but they do not enable the organic chemist to draw conclusions about important questions when he is confronted to a synthesis problem, for example:

- which are the the most frequent functional interchanges ?
- when a function is to be formed, from which function(s) must he perform the desired interchange?

We wondered afterwards how effective was the data mining contribution by comparison with the data preprocessing contribution. Therefore frequent itemsets and corresponding association rules were used as defined in section 4. With respect to the functionality we only studied functional interchanges to focus on the retrosynthetic point of view, thus we left out function formations from a nonfunctional bond and function eliminations. Consequently sets of destroyed or formed or unchanged functions, respectively called $\{F_d\}$, $\{F_f\}$ and $\{F_u\}$ contain a lesser object amount than in preceding results.

4.6.2 – From data transformation and data mining

Taking into account the above-mentioned questions, we mainly studied the JSM database in the following paragraphs and fixed the minsup threshold equal to 10. This value was chosen after examining the total number of k-itemsets in order to reduce the number of related association rules and thus the analysis complexity.

Use of frequent itemsets may be done stepwise. At first studying frequent 2-itemsets enables the analyst to determine some basic relations between functions. For example if we want to know from which destroyed functions F_d another function F_f is formed we have to look at all itemsets of type : $F_d \wedge F_f$. If we are interested in a formed function from two destroyed functions we have to study all itemsets of type : $F_{d1} \wedge F_{d2} \wedge F_f$. In many cases the choice of a synthetic method to form a new function depends on other functions which are present in both reactants and products but should be kept unchanged along the transformation. The corresponding frequent itemsets are of type: $F_f \wedge F_i \wedge F_d$. This latter approach can be applied to search for a protective group supposed to be stable under given experimental conditions.

Use of association rules bring further information by comparison with the use of frequent itemsets. If we want to get a molecule containing the function F_f and to know the more frequent way how to form it, we must retrieve association rules corresponding to : "if the function F_f is formed then it is formed from function(s) F_d". In such cases the rules contain "F_f" as a premice and

destroyed function item(s) as a conclusion. These retrieved rules are of type : $F_f \Rightarrow \{F_d\}$ where $\{F_d\}$ represents the destroyed functions set. Comparing the values of corresponding rules confidences enables a classification of function combinations answering the starting question. If we want to determine how a function F_f is formed from two destroyed functions F_1 and F_2 and if one of these two, F_1 for example, is still known, the studied association rules are of type : $F_f \wedge F_{d1} \Rightarrow \{F_{d2}\}$ where $\{F_{d2}\}$ represents the destroyed functions set F_2.

In order to highlight both use and relevance of our approach, we closely examined carboxylate, alkene and acetal functions. These functions show a high occurrence frequency enabling us to work with a significant enough occurrence number. Unless otherwise noted, the following studies were performed using the specific consideration of the block correspondence (Figure 6).

Ester function

The studied frequent 2-itemsets are of type : $F_d \wedge \text{carboxylate}_f$ and the targeted rules are of type : $\text{carboxylate}_f \Rightarrow \{F_d\}$. The destroyed functions used to form a carboxylate function are ranked by decreasing confidence (Table 5). Over the 101 data mining retrieved functions are only displayed the 10 more frequent destroyed functions. Among these we retrieved expected functions involved in esterification and transesterification transformations. Generally these functions have a high occurrence number n_{tot}; anhydride and vinyloxycarbonyl functions, which are not shown in Table 2, occur respectively 1209 and 890 times.

Table 5 : Most frequent functions involved in the carboxylate function formation

Entry	destroyed function[1]	sup(P)[2]	conf[3]
1	alcohol	1030	21.5
2	carboxylic-acid	660	13.8
3	carboxylate	651	13.6
4	carbonyl	567	11.8
5	anhydride	419	8.7
6	alkene	334	7.0
7	ether	206	4.3
8	acetal	181	3.8
9	acyl-chloride	175	3.7
10	vinyloxycarbonyl	140	2.9

[1)] ranked by decreasing order on the column sup(P); [2)] indicates the number of objects containing this item; [3)] confidence.

The functional interchanges between an alcohol function or a carboxylic acid function to a carboxylate function are the most frequent interchanges, in the range of 2-3 %. This result was still pointed out in a previous study on CASREACT database.[22] It is well-known that this carboxylate function mainly results from a two function combination. We therefore used frequent 3-itemsets of type : $F_{d1} \wedge F_{d2} \wedge \text{carboxylate}_f$ and targeted rules of type : $\text{carboxylate}_f \Rightarrow \{F_{d1}\} + \{F_{d2}\}$ then we compared confidence values for each rule (Table 6). The most frequent combination is "alcohol + anhydride" which is significantly more frequent than the "alcohol + acyl-chloride" combination though the acyl-chloride function is often considered as more reactive than the anhydride function. This may be explained since function relative reactivities are not the only parameters being considered when a chemist selects reactants to be used in a carboxylate function formation. In fact other criteria dramatically affect the choice of reactants such as easy handling, stability on storage,

cost, environmental or safety considerations. Function combinations whose confidence is in the range of 2.5-3.1 are still of some usefulness. Among these ones the "alcohol + carboxylate" combination (Entry 1) describes logically the transesterification reaction.

Table 6 : Most frequent function combinations involved in the carboxylate function formation

Entry	destroyed functions[1]	sup(P)[2]	conf[3]
1	alcohol + anhydride	204	4.3
2	alcohol + carboxylic-acid	149	3.1
3	alkene + carboxylic-acid	143	3.0
4	alcohol + carboxylate	140	2.9
5	carbonyl + carboxylate	122	2.5
6	alcohol + acyl-chloride	63	1.3
7	alkene + carboxylate	47	1.0

[1)] ranked by decreasing order on the column sup(P); [2)] indicates the number of objects containing this item; [3)] confidence.

More surprising are combinations involving the functions "alkene + carboxylic-acid" or "carbonyl + carboxylate" (Entries 3 and 5). In both cases, we searched which synthetic methods could correspond to the related transformations present in JSM database. About the former one, we observed these transformations were instances of the following synthetic methods: (i) carboxylic acid addition on alkenes or dienes through its hydroxyl group, (ii) lactonization reaction, (iii) cyclization of functionalized dienes possessing a carboxylic acid function, (iv) formation of branched esters of type *tert*-butyl ester, (v) a reaction sequence – epoxidation/addition of the carboxylic acid group. About the latter case, the transformations involved the following synthetic methods: (i) ring opening of hindered cycloalkanones, (ii) a reaction sequence – carbonyl group reduction/transesterification or /spirolactonization, (iii) a five-membered ring expansion to seven-membered rings, (iv) a reaction sequence – carbanion addition on the carbonyl group/lactonization.

Association rules enable the analyst to go deeper into the analysis of obtained results. For example, if one searches for which transformations form a carboxylate function while keeping an ether function unchanged, he has to examine rules of type : $\text{carboxylate}_f \wedge \text{ether}_i \Rightarrow F_d$ or of type : $\text{carboxylate}_f \wedge \text{ether}_i \Rightarrow F_{d1} \wedge F_{d2}$ to obtain destroyed function combinations. According to the latter query, the following rule has been retrieved : $\text{carboxylate}_f \wedge \text{ether}_i \Rightarrow \text{alcohol}_d \wedge \text{anhydride}_d$ with a confidence value of 10.6 %. Moreover results of the preceding query can suggest to the analyst a more precise question : which destroyed functions could be involved if he uses an alcohol function to form a carboxylate function while keeping an ether function unchanged ? Among the retrieved association rules, the following rule "$\text{carboxylate}_f \wedge \text{ether}_i \wedge \text{alcohol}_d \Rightarrow \text{anhydride}_d$" indicates that in 40,1 % of cases an anhydride function was used.

Alkene function

We studied this function in a similar way as the carboxylate function. Examining association rules of type : $\text{alkene}_f \Rightarrow \{F_d\}$ enables us to retrieve the whole functions involved in the alkene function formation. Over the 64 data mining retrieved functions are only displayed the 27 more frequent destroyed functions corresponding to a confidence threshold ≥ 0.5 (Table 7). For each entry we paid a special attention to the synthetic methods related to the corresponding frequent 2-itemsets by browsing JSM database records. This reasoning was done in order to confirm the relevance of data mining results. Comments on the retrieved synthetic methods are shown in the right column.

Table 7 : Functions involved in the alkene function formation

Entry	destroyed function[1]	sup(P)[2]	conf[3]	Corresponding synthetic methods[4]
1	alkene	3800	29.4	metatheses (RCM[5] and ROM-CM[5]); ozonolysis/intramolecular aldolization; ozonolysis/Wittig reaction; Diels-Alder reaction; ene-reaction
2	carbonyl	1343	10.4	Wittig olefination
3	alkyne	1253	9.7	reduction
4	alcohol	575	4.5	oxidation/olefination
5	phenyl	532	4.1	Birch reduction (dearomatization)
6	carboxylate	373	2.9	reduction/olefination; ester allyl elimination
7	ether	354	2.7	alcohol protection ;
8	allene	315	2.4	Claisen-, oxy-Cope rearrangements reduction; heterocyclizations; exocyclic double bonds formation
9	tetraalkyl-silane	278	2.2	Peterson olefination
10	alkyl-bromide	276	2.1	elimination; Wittig olefination; dimerization
11	enol-ether	257	2.0	Claisen-, oxy-Cope rearrangements
12	vinyl-bromide	204	1.6	Heck coupling (diene formation); organometallic compounds coupling; reduction
13	vinyl-iodide	177	1.4	Heck coupling (diene formation); organometallic compounds coupling; reduction
14	furan	166	1.3	Diels-Alder reaction/bicycle opening reaction
15	alkyl-chloride	161	1.2	elimination; Wittig olefination
16	phenol	135	1.0	Claisen rearrangement
17	phosphonate	109	0.8	Horner-Wadsworths-Emmons olefination
18	acetal	86	0.7	deprotection/olefination or Ferrier rearrangement
19	N,N-dialkyl-enamine	82	0.6	enamine alkylation
20	trialkyl-vinylsilane	81	0.6	addition electrophilic compounds
21	phenate	80	0.6	Claisen rearrangement
22	dialkyl-carbonate	79	0.6	elimination
23	vinyloxycarbonyl	79	0.6	[8+2] or [2+2] cycloaddition; Diels-Alder reaction or CO_2 extrusion; allylic oxydation
24	trialkyl-silylenol-ether	78	0.6	enol ether addition on various nucleophilic compounds and Michael acceptors
25	sulfone	75	0.6	Julia olefination
26	pyrrole	72	0.6	porphyrine derivatives formation
27	vinyl-chloride	71	0.6	enyne formation; reduction

[1)] ranked by decreasing order on the column sup(P); [2)] indicates the number of objects containing this item; [3)] confidence ≥ 0.5; [4)] the different synthetic methods and reaction sequences are delimited by a semicolon respectively by a /; [5)] RCM = Ring-Closing Metathesis; ROM-CM = Ring-Opening Metathesis–Cross-Metathesis.

With respect to knowledge to be deduced from the preceding association rules, we noticed first that the most frequent used function to form an alkene function is another alkene function (Entry 1); related synthetic methods are olefinic metatheses (RCM, ROM-CM), allylic rearrangements, pericyclic transformations (Diels-Alder, ene-transformation), multi-step sequences (ozonolysis followed by a second transformation). We observed also the presence of functions such as carbonyl, tetraalkyl-silane, phosphonate and sulfone functions (Entries 2, 9, 17 and 25) are necessary for olefination sequences. Nevertheless the phosphonium function is absent from this list;

indeed phosphonium salts are not explicitly mentioned as reactants but tertiary phosphines are rather indexed as reagents, i.e. as textual data and therefore do not appear as molecular graphs; as a consequence the occurrence frequency of the phosphonium function is lower than the data mining fixed threshold (minsup = 10) ruling out this function. On the other hand alkyl halides (Entries 10 and 15) are present as precursors of phosphonium salts or of various carbanions. These alkyl halides are involved also in elimination reactions. Alcohol and carboxylate functions (Entries 4 and 6) play an important role in olefination reactions but after respectively oxidation to an aldehyde or reduction to a ketone. Other examples of reduction reactions involve alkyne, phenyl and allene functions (Entries 3, 5 and 8). We noticed that ether, enol-ether, phenol, acetal and phenate functions (Entries 7, 11, 16, 18 and 21) are used in famous rearrangements such as Claisen, oxy-Cope and Ferrier rearrangements.

Basic knowledge about alkene formation is contained in teaching manuals but synthetic methods are classified according to mechanistic points of view such as nucleophilic and electrophilic substitution, free-radical substitution, addition to Carbone-Carbone and Carbone-Heteroatome multiple bonds, eliminations, rearrangements, oxidations and reductions. If we compare our results with the domain knowledge, the data mining process, though performed on a selective database such as JSM, enabled us to retrieve most information related to this knowledge. A further interest of the data mining lies in its ability to propose a classification of synthetic methods according to their occurrence frequency thus giving some insights about the usefulness of requested synthetic methods.

The preceding remarks were illustrated by examining association rules of type : $alkene_f \Rightarrow \{F_{d1}\} + \{F_{d2}\}$, since this brought some useful answers about a wide question such as “which are the destroyed function combinations necessary to form an alkene function ?” Since the carbonyl function was one of the most frequent functions detected in Table 5 and also generally used for olefination reactions, the above-mentioned question was limited to the study of association rules of type : $carbonyl_d \wedge alkene_f \Rightarrow F_d$. The results summed up in Table 8 showed that the phosphonate function was mostly employed with a carbonyl group to form an alkene function. Many addition/elimination sequences were retrieved as instances of these synthetic methods.

Table 8 : Alkene function formation from the carbonyl function and a second function

Entry	destroyed functions[1]	sup(P)[2]	conf[3]	Corresponding synthetic methods[4]
1	carbonyl + phosphonate	82	6.1	Horner-Wadsworths-Emmons olefination
2	carbonyl + alkene	67	5.0	Michael addition/elimination or condensation
3	carbonyl + alkyl-bromide	31	2.3	Wittig olefination; carbanion addition
4	carbonyl + tetraalkyl-silane	27	2.0	Peterson olefination; carbanion addition/elimination
5	carbonyl + alkyne	16	1.2	propargylsilane addition
6	carbonyl + alkyl-chloride	15	1.1	Wittig olefination; carbanion addition
7	carbonyl + N,N-dialkyl-enamine	12	0.9	enamine addition/elimination

[1)] ranked by decreasing order on the column sup(P); [2)] indicates the number of objects containing this item; [3)] confidence ≥ 0.5; [4)] the different synthetic methods and reaction sequences are delimited by a semicolon respectively by a slash.

Acetal function

The acetal function belongs to the most widely used protective groups for aldehydes and ketones.[50, 51] Acetal cleavage is generally performed under aqueous acid or Lewis acid conditions. This function is known to be stable under basic conditions, towards nonacidic oxidative reagents or hydride reductions. On the other hand it reacts in the presence of oxophilic Lewis acids

since one of the C–O bonds can be substituted by organometallic compounds derived from Mg, Cu, Zn, Al, Si.

Table 9 : Most frequent function combinations involved in the acetal function formation

Entry	destroyed function[1]	Conf[2]
1	alcohol + enol-ether	8.2
2	alcohol + carbonyl	7.7
3	acetal + alcohol	5.6
4	alcohol + hemiacetal	2.4
5	alcohol + thioacetal	1.9
6	alcohol + ether	1.8
7	acetal + carbonyl	1.7
8	alkene + alcohol	1.6
9	alkene + carbonyl	1.6
10	alcohol + gem-fluoro-ether	1.5
11	carbonyl + trialkyl-silylether	1.5
12	enol-ether + carbonyl	1.4
13	alcohol + gem-bromo-ether	1.2
14	alcohol + gem-chloro-ether	1.0
15	ether + carbonyl	1.0

[1)] ranked by decreasing order on the column sup(P); [2)] confidence ≥ 1.

It seemed to us interesting to study the acetal function from both points of view related to its formation then to its stability. As regards the former point, studying rules of type : $acetal_f \Rightarrow \{F_{d1}\} + \{F_{d2}\}$ gave the results summed up in Table 10. Whereas the alcohol, enol-ether and carbonyl functions (Entries 1-2) are the most frequently used to create the acetal function, entries 3-5, 7 emphasize the importance of transacetalization reactions and similar synthetic methods. Carbohydrate chemistry is also well represented by the presence of hemiacetal (Entry 4) and gem-haloether functions (Entries 10, 13 and 14). In this specific domain, glycosylation reactions are known for their wide usefulness. Epoxide-containing heterocycles are represented by the ether function (Entry 15) though they are more seldom employed as starting materials to form an acetal.

Then we compared our results with basic knowledge that can be deduced from a standard teaching manual[19] about acetal formation. To this purpose the manual proposes the following series of either synthetic methods or starting materials: (i) acetalization of carbonyl compounds in the presence of alcohols and diols, (ii) transacetalization to cyclic acetals, (iii) insertion of alcohols from a diazoalkane, (iv) Williamson reaction from gem-dihalides or gem-haloethers, (v) orthoester reduction, (vi) addition of Grignard compounds on orthoesters, (vii) cyclization of β-hydroxy-ethers, (viii) oxymercuration-demercuration of alkynes with alcohols, (ix) Prins reaction (alkene + formaldehyde), (x) epoxide addition on carbonyl derivatives.

Comparing data mining results with the preceding synthetic methods shows a good parallel between both approaches. Data mining techniques applied on the JSM database describe most taught synthetic methods and moreover enable the user to sort them on their occurrence frequency.

Table 10 : Most frequent destroyed or formed functions while an acetal function is kept unchanged

Entry	destroyed function[1]	conf[2]	Entry	formed function[1]	conf[2]
1	alcohol	16.8	18	alcohol	21.2
2	alkene	11.9	19	alkene	17.4
3	carbonyl	11.3	20	carbonyl	7.9
4	ether	7.3	21	carboxylate	7.8
5	carboxylate	5.9	22	acetal	7.6
6	acetal	5.1	23	ether	6.1
7	alkyne	3.9	24	N-alkyl-amide	1.7
8	anhydride	2.9	25	phenyl	1.6
9	enol-ether	2.9	26	N-alkyl-amine	1.4
10	alcohol + anhydride	2.0	27	enol-ether	1.3
11	N-alkyl-amine	1.9	28	N,N,N-trialkyl-amine	1.1
12	hemiacetal	1.9	29	N,N-dialkyl-amine	1.1
13	alkyl-bromide	1.5	30	carboxylic-acid	1.0
14	acyl-chloride	1.4	31	hemiacetal	1.0
15	carboxylic-acid	1.1			
16	thioacetal	1.0			
17	vinyl-bromide	1.0			

[1)] ranked by decreasing order on the column confidence. [2)] confidence ≥ 1.

Since the main interest of acetal function lies in its ability to protect efficiently the carbonyl group, we were interested in determining which functions, either destroyed or formed, were used while keeping unchanged an acetal function. Intended association rules are of type: $acetal_i \Rightarrow \{F_d\}$ or $acetal_i \Rightarrow \{F_f\}$ and derive from the global consideration of the block correspondence (see Fig. 6) that is more appropriate for studying the chemoselectivity of functional transformations.

Our results are displayed in Table 11 where the left column depicts 17 destroyed functions among the 42 retrieved destroyed functions and the right column depicts 14 formed functions over the 26 formed functions. We observe first that main transformations keeping an acetal unchanged involve functions used to form an acetal function (see Table 10). We assume that different experimental conditions may differentiate the transformations leading to the acetal formation from those keeping it unchanged. It is also worth noting that very reactive destroyed functions such as alkyne (Entry 7), anhydride (Entry 8, 10), *N*-alkyl-amine (Entry 11), carboxylic-acid (Entry 15) and vinyl-bromide (Entry 17) function may react without interfering with acetal function. This observation still holds for formed functions such as amine functions (Entries 24, 26, 28). Other information about carboxylate, *N*-alkyl-amide, phenyl functions is less interesting because of their known stability. Nevertheless we note that the thioacetal function, used as an alcohol or ketone protective group, may react selectively.

4.7 - Visualization

The display of our results was performed by generating html formatted files from the directory tree containing the structural data. A html file lists all files contained in a selected directory and its subdirectories. The generation of these files is possible after the complete perception process is finished, i.e. after the preprocessing step (§ 4.3). A partial display of the functional interconversion directory is given in Figure 7. The first line states the selected directory and the whole number of contained files into brackets then the formed functions are listed alphabetically below. Activating a given function, e.g. the acetal function, enables the user to

browse through a new list of destroyed functions involved in the selected function formation. The diols are represented in our example in the middle of the Figure 7 by the item *alcohol_alcohol (11)* which means that 11 functional interconversions from a diol to an alcohol function were analyzed during the preprocessing step.

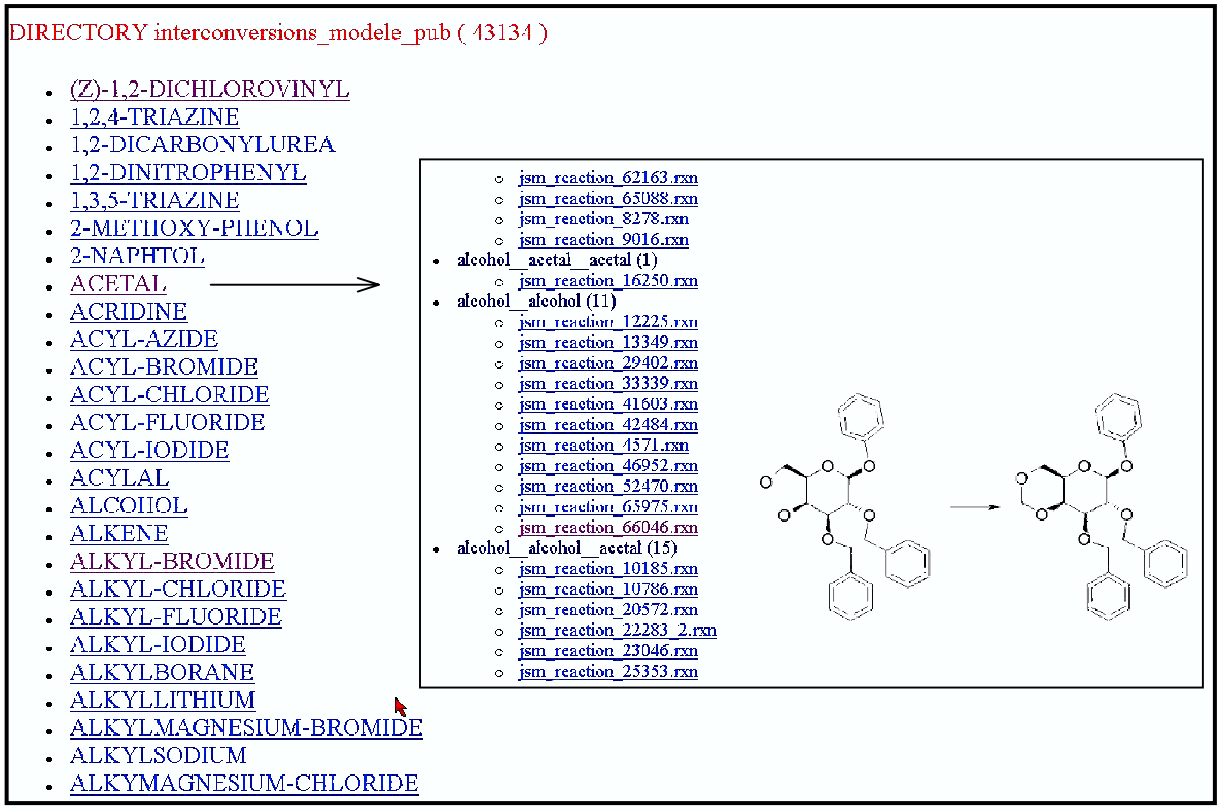

Figure 7 : Partial display of the functional interconversion html file. Example of the JSM_reaction_66046.

5 – DISCUSSION

Tackling the functionality problem is quite complex and the presented data mining experiments offer the reader new insights on how to manage it. If we try to make an appraisal of the present study, three main types of problems were encountered depending on the modelling step, the selected databases and their content and the data mining techniques respectively.

With respect to the former point, our functionality model has to be refined since some functions cover functional groups which are generally used by organic chemists in a different manner. Since no single carbon-carbon bond is included in any functions, some functional groups commonly used by organic chemists are not specifically recognized by our approach. For example the carbonyl function represents both aldehydes and ketones, the alcohol function refers to alcohols as well as 1,2- to 1,n-diols. It is also worth mentioning that some functions represent any acyclic or cyclic atom sequences since the topological point of view was not differentiated from the functional one during this study. That is the case for the alkene, ether, carboxylate, acetal, anhydride, amine, amide functions among others. We deal with this point of view in another paper by studying the ring construction methods.[21] The concomitant consideration of several points of view should improve both preprocessing and data mining results.

Concerning the second point, the choice of the ORGSYN and JSM databases was guided by their coverage relevance. Indeed the ORGSYN database provides an electronic version of the entire series of Organic Syntheses and offers an access to new general synthetic methods. The principle

followed by the publishers of Organic Syntheses is particularly interesting since each synthetic method has been checked by expert laboratories. This practice confers to these data a high value : they are proven compound preparations. As mentioned by Ireland, ORGSYN is a gold mine of molecules and techniques.[20] On the other hand, the JSM database is a document-based organic reaction database presenting a high coverage in synthetic organic chemistry from 1975 to date. To qualify for selection a reaction must be novel or have a particular advantage over an existing method. In addition the reaction must have a clear experimental method, be repeatable and proceed in good yield. For these reasons, ORGSYN and JSM databases seemed to us quite suitable for a study of chemical functionality. Both databases contain very selected data unlike ChemInform databases or larger databases such as CASREACT or Beilstein databases. Thus the information to be retrieved is necessarily more focused. We considered ORGSYN and JSM databases as more useful sources in order to explore a knowledge domain rather than to get exhaustive information about particular reactions. As a consequence, the itemset supports may seem low by comparison with those observed during other data mining experiments such as marketing studies. We thus studied frequent 1-, 2- and 3-itemsets in order to work with acceptable itemset supports and to extract a quite general information. Use of association rules derived from higher size itemsets should enable the user to get a more specific information but with a very low confidence.

As seen in section 4.3, a weak amount of database records could not be preprocessed leading to an unavoidable data loss, i.e. 0.2 and 1.3 % in the case of ORGSYN and JSM databases respectively. More dramatically a bias is introduced in this study since a non-negligible amount of transformations has been assigned an incorrect or incomplete atom-atom mapping.[4] This results from an imperfect reaction substructure search algorithm.[9] Since false atom-atom mapping occurs randomly, the concerned transformations could not be counted exactly but false mapping containing records were estimated at 6% after a crafted sampling of the JSM database. Finally a great number of functionality transformations correspond to reaction sequences. Even if interesting answers may result from such outcomes, the subsequent information lacks details and subtleties about the omitted single-step reactions and may introduce exotic functional transformations.

Finally with respect to the data mining techniques, it is worth mentioning that very few data mining studies deal with data possessing dynamic features such as chemical reactions.[52-55] These studies have different objectives but are mainly concerned by molecular graph manipulation rather than reaction database mining. Another study on the lattice-based classification of dynamic knowledge units helped us in representing knowledge units through conceptual graphs.[56] This work is more focused on formal concept analysis and lattice construction rather than on data mining concerns. An extension of these techniques to more complex objects has been performed but could not be directly used in our study.[57] Thus we had to transform complex objects such as molecules and reactions; while molecules could be easily represented as graphs, the reaction representation required graph rewrite rules. This required transformation of graphical data into a boolean table led to some losses of information, more particularly the block neighbourhood information. Moreover in the case of the specific consideration of the block correspondence, each transformation was associated to different objects, T1 to Tn, but no link was created to specify that these objects refer to the same transformation; as regards the example of Figure 4, we cannot know that the hemiacetal function destroyed in T1 to give a carboxylate function is the same that the one giving the carbonyl function in T2. This consideration of the block correspondence introduces a bias towards the function occurrence frequency since unchanged functions are counted as many times as new objects are created.

6 – CONCLUSION

During this knowledge discovery process, we benefited from large chemical reaction databases, a source of numerous and very varied functional transformations. Such a diversity

involves the presence of scarce but meaningful instances of functional interchanges leading to somewhat low itemset supports as well as representative instances of well-known generic synthetic methods. This pleads for the essential role of the analyst all along the process since he is better placed to compare the result validity with his starting objectives.

Handling of the dynamic functional changes into three kinds of blocks, i.e. formed, destroyed and unchanged blocks, enabled the introduction of a further abstraction level. That proved very suitable to obtain a general but quite detailed overview within the databases content. Moreover this approach might be reused in other contexts involving dynamic data.

Frequent itemsets or association rules are generic elements that can be used either to index or to retrieve reactions. Indeed comparing values of rule confidences enables a classification of function combinations answering the starting question. This classification provides new insights about the usefulness of related synthetic methods. Frequent itemsets and association rules can be considered as meta-information on the studied databases highlighting new knowledge elements. We plan to use these structured metadata to build a knowledge database that could be queried through questions mentioned in section § 2. Other research perspectives following this work include the adaptation of sequential pattern algorithms to chemical reactions and introducing the concept analysis for lattice-based classification of the data.

ACKNOWLEDGEMENT

We gratefully acknowledge the Molecular Design Ltd company for allowing us to use extracted data from their reaction databases within an academic research.

REFERENCES

[1] Laurenço, C.; Py, M.; Napoli, A.; Quinqueton, J.; Castro, B. Representation of organic synthesis knowledge using an object-oriented language. *New J. Chem.* **1990**, *14*, 921-931.

[2] Gien, O., PhD Thesis, Université Montpellier II, Montpellier, 1998.

[3] Vismara, P.; Laurenco, C. An abstract representation for molecular graphs. *DIMACS Series In Discrete Mathematics and Theoretical Computer Science* **2000**, *51*, 343-356.

[4] Coste, J.; Gien, O.; Dietz, A.; Laurenco, C. Use of reaction databases. *Actual. Chim.* **1999**, *7*, 27-32.

[5] Ott, M.A.; Noordik, J.H. Computer tools for reaction retrieval and synthesis planning in organic chemistry. A brief review of their history, methods, and programs. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 239-246.

[6] Chen, L.; Gasteiger, J.; Rose, J.R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon-Hydrogen Bonds to Carbon-Carbon Double Bonds. *J. Org. Chem.* **1995**, *60*, 8002-8014.

[7] Regin, J.C.; Gascuel, O.; Laurenco, C. Machine learning of strategic knowledge in organic synthesis from reaction databases. *AIP Conf. Proc.* **1995**, *330*, 618-623.

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

[9] Chen, L.; Nourse, J.G.; Christie, B.D.; Leland, B.A.; Grier, D.L. Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1296-1310.

[10] Gasteiger, J., Pfortner, M., Sitzmann, M., Hollering, R., Sacher, O., Kostka, T.;Karg, N. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspect. Drug Discovery Des.* **2000**, *20*, 245-264.

[11] Funatsu, K. (Eds.), Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, American Chemical Society, 2001, CINF-068.

[12] 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.* **1990**, *30*, 492-504.

[13] Blurock, E.S. Automatic extraction of reaction information from databases using classification and learning techniques. *Chem. Inf.* **1990**, *2*, 25-35.

[14] Jauffret, P.; Vogel, H.; Schildknecht, S.; Kaufmann, G. Learning synthetic knowledge from reaction databases: Dealing with experimental conditions. *Proc. Int. Chem. Inf. Conf.* **2000**, 137-163.

[15] Mezey, P.G. Functional Groups in Quantum Chemistry. In *Advances in Quantum Chemistry*; Eds P O Lowden Eds., Academic Press, 1996, pp. 163-222.

[16] Pearson, D.P.J. Reaction information needs of the synthetic chemist. *Mod. Approaches Chem. React. Searching, Proc. Conf.* **1986**, 18-27.

[17] Schier, O.; Nuebling, W.; Steidle, W.; Valls, J. System for the documentation of chemical reactions. *Angewandte Chemie, International Edition in English* **1970**, *9*, 599-604.

[18] Fugmann, R.; Kusemann, G.; Winter, J.H. The supply of information on chemical reactions in the IDC system. *Information Processing & Management* **1979**, *15*, 303-323.

[19] Smith, M.B.; March, J. March's Advanced Organic Chemistry : Reactions, Mechanisms and Structure. In John Wileys & Sons Eds., Wiley-Interscience, New-York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 2001.

[20] Ireland, R.E. Organic Synthesis. In *Foundations of Modern Organic Chemistry*; K L Jr Rinehart Eds., Prentice-Hall, Inc., Englewood Cliff, NJ, 1969.

[21] Berasaluce, S.; Niel, G.; Napoli, A.; Laurenço, C. Data mining in reaction databases : Extraction of knowledge on ring construction methods. *J. Chem. Inf. Comput. Sci.* **2004**, submitted.

[22] Blower, P.E., Jr.; Myatt, G.J.; Petras, M.W. Exploring Functional Group Transformations on CASREACT. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 54-58.

[23] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smith, P. From data mining to knowledge discovery : An overview. In *Advances in knowledge discovery and data mining*, AAAI Press / MIT Press, 1996, pp. 37-57.

[24] Agrawal, R.; Imielinski, T.; Swami, A. (Eds.), Proceedings of the 1993 (ACM SIGMOD) Conference, 1993.

[25] Agrawal, R.; Swami, A. (Eds.), Proceedings of the 20th {VLDB} Conference, 1994.

[26] Simon, A.; Napoli, A. Algorithme de fouille de données dans une représentation des données par objets : une application au domaine médical. In *Ingénierie des connaissances. Evolutions récentes et nouveaux défis*; M. Zaclad J. Charlet, G. Kassel and D. Bourigault Eds., Eyrolles, 2000, pp. 197-207.

[27] Brachman, R.J.; Anand, T. The Process of Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining*; U.M. Fayyad and G. Piatetsky-Shapiro and P. Smyth and R. Uthurusamy Eds., AAAI Press / MIT Press, Menlo Park, California, 1996, pp. 37-57.

[28] Bastide, Y.; Taouil, R.; Pasquier, N.; Stumme, G.; Lakhal, L. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques* **2002**, *21*, 65-95.
[29] Jambaud, P., PhD Thesis, Université de Montpellier II, Montpellier, 1996.
[30] Zaki, M.J.; Ogihara, M. Theoretical Foundations of Association Rules. *Proc. of 3rd SIGMOD' 98 Workshop on Research Issues in Datamining and Knowledge Discovery (DMKD' 98)* **1998**, *Seattle, USA*, 85-93.
[31] Willett, P. The reaction indexing problem: a historical viewpoint. *Mod. Approaches Chem. React. Searching, Proc. Conf.* **1986**, 1-17.
[32] Bawden, D. Classification of chemical reactions: potential, possibilities and continuing relevance. *Journal of Chemical Information and Computer Sciences* **1991**, *31*, 212-216.
[33] Corey, E.J. Computer-assisted analysis of complex synthetic problems. *Quart. Rev., Chem. Soc.* **1971**, *25*, 455-482.
[34] Corey, E.J.; Cramer, R.D., III; Howe, W.J. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. *J. Amer. Chem. Soc.* **1972**, *94*, 440-459.
[35] Gotor, V.; Pulido, R. An improved procedure for regioselective acylation of carbohydrates: novel enzymatic acylation of a-D-glucopyranose and methyl a-D-glucopyranoside. *Journal of the Chemical Society, Perkin Transactions 1: Organic and Bio-Organic Chemistry* **1991**, 491-492.
[36] Tsukamoto, T.; Nomura, H.; Morita, S.; Okada, J. Kinetic studies on the oxidation of glucose by immobilized glucose oxidase. *Chemical & Pharmaceutical Bulletin* **1983**, *31*, 3377-3384.
[37] Dixon, D.J.; Ley, S.V.; Tate, E.W. Highly cis- or trans-selective oxygen to carbon rearrangements of anomerically linked 6-substituted tetrahydropyranyl enol ethers. *Journal of the Chemical Society, Perkin Transactions 1: Organic and Bio-Organic Chemistry* **1999**, 2665-2667.
[38] Lee, J.; Barchi, J.J., Jr.; Marquez, V.E. Synthesis of a rigid diacylglycerol analog having a bis-g-butyrolactone skeleton separated by a cyclopentane ring. *Chemistry Letters* **1995**, 299-300.
[39] Boto, A.; Hernandez, R.; Suarez, E. Synthesis of acyclic nucleosides and other C-1 substituted alditols from carbohydrates using a tandem alkoxyl radical b-fragmentation-nucleophilic addition. *Tetrahedron Letters* **2001**, *42*, 9167-9170.
[40] As defined by IUPAC the functional group is an atom or a group of atoms that have similar chemical properties whenever it occurs in different compounds. It defines the characteristic physical and chemical properties of families of organic compounds.
[41] Sello, G. A new definition of functional groups and a general procedure for their identification in organic structures. *J. Am. Chem. Soc.* **1992**, *114*, 3306-3311.
[42] Corey, E.J.; Wipke, W.T.; Cramer, R.D., III; Howe, W.J. Techniques for perception by a computer of synthetically significant structural features in complex molecules. *J. Amer. Chem. Soc.* **1972**, *94*, 431-439.
[43] Dietz, A.; Fiorio, C.; Habib, M.; Laurenço, C. Representation of stereochemistry using combinatorial maps. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **2000**, *51*, 117-128.
[44] Berasaluce, S.; Niel, G.; Napoli, A.; Laurenço, C. Data mining in reaction databases : Extraction of knowledge on ring construction methods. **2004**.
[45] Gabriele, B.; Salerno, G.; Costa, M.; Chiusoli, G.P. Palladium-catalyzed formation of maleic anhydrides from CO, CO2 and alk-1-ynes. *Chemical Communications (Cambridge)* **1999**, 1381-1382.
[46] Hendrickson, J.B. Systematic characterization of structures and reactions for use in organic synthesis. *J. Amer. Chem. Soc.* **1971**, *93*, 6847-6854.
[47] RESYN_Assistant is a software developed in JAVA language within the framework of the CNRS-GDR 1093 "Traitement Informatique de la Connaissance en Chimie Organique" (director : C. Laurenço); see ref [x] for more details on this software.
[48] Metivier, P., PhD Thesis, Université Louis Pasteur - Strasbourg I, Strasbourg, 1987.
[49] Laurenco, A., *Resyn-Assistant : modélisation et programmation du calcul de l'aromaticité des cycles d'une molécule*. 1998, Technical Report, LIRMM: Montpellier.
[50] Kocienski, P.J. Protecting groups. In *Protecting groups*; R. Noyori and B. M. Trost D. Enders Eds., Georg Thieme Verlag, Stuttgart, New-York, 1994.
[51] Greene, T.W.; Wuts, P.G.M. Protective Groups in Organic Synthesis. In *Protective Groups in Organic Synthesis. 2nd*, John Wiley and Sons, Inc., New York, 1991.
[52] Chittimoori; Holder, L.; Cook, D. Applying the Subdue Substructure Discovery System to the Chemical Toxicity Domain. In *Proceedings of the Twelfth International Florida AI Research Society Conference*, 1998, pp. 90-94.
[53] Inokuchi, A.; Washio, T.; Motoda, H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *Principles of Data Mining and Knowledge Discovery*, 2000, pp. 13-23.
[54] Dehaspe, L.; Toivonen, H.; King, R.D. Finding frequent substructures in chemical compounds. In *4th International Conference on Knowledge Discovery and Data Mining*; R. Agrawal and P. Stolorz and G. Piatetsky-Shapiro Eds., AAAI Press., 1998, pp. 30-36.

[55] Kuramochi, M.; Karypis, G., *An efficient algorithm for discovering frequent subgraphs*. 2002, Department of computer science/Army HPC Research Center, University of Minnesota.
[56] Ganter, B.; Rudolph, S. (Eds.), Conceptual Structures: Broadening the Base : 9th International Conference on Conceptual Structures, Stanford, Springer-Verlag Heidelberg, Lecture Notes in Computer Science, 2001, 143-156.
[57] Polaillon, G., PhD Thesis, Université Paris IX-Dauphine, Paris, 1998.