Extraction de connaissances à partir de bases de données de réactions en chimie organique

Sandra Berasaluce, Claude Laurenço, Amedeo Napoli

Extraction de connaissances à partir de bases de données de réactions en chimie organique

Sandra Berasaluce*†, Claude Laurenço†‡, Amedeo Napoli*

* Equipe Orpailleur,
Laboratoire LOrrain de Recherche en Informatique et ses Applications, UMR7503 du CNRS,
Campus Scientifique, BP 239, 54506 - Vandœuvre-lès-Nancy
Amedeo.Napoli@loria.fr
† Equipe Informatique et Chimie,
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR5506 du CNRS,
161, rue Ada, 34392 - Montpellier Cedex 5,
berasalu@lirmm.fr
‡ Laboratoire des Systèmes d'Information Chimique, Hétérochimie moléculaire et macromoléculaire, UMR 5076, ENSCM,
8, rue de l'Ecole Normale, 34296 cedex 5 - Montpellier
cl@lirmm.fr

Résumé

Dans cet article, nous présentons un aspect de l'extraction de connaissances dans des bases de données de réactions chimiques. Ces bases de données sont de première importance, mais leur exploitation actuelle reste limitée à des interrogations classiques. Nous avons fait l'hypothèse que l'application de techniques de fouille de données à de telles bases peut faire émerger des éléments de connaissance sur les réactions, qui peuvent alors être réutilisés pour résoudre des problèmes de synthèse chimique. Pour mener à bien fouille de données et résolution de problème, la représentation et l'exploitation de connaissances du domaine est un préalable obligé. Les premiers résultats d'une expérience de fouille de données dans des bases de réactions sont présentés et analysés ici.

Mots clef : représentation des connaissances ; fouille de données ; recherche d'information ; interprétation de données.

1 Introduction

Notre travail s'inscrit dans un projet à long terme de conception de systèmes d'information chimique visant à aider les chimistes dans la synthèse de molécules organiques complexes [29].

Le but principal de la chimie est d'étudier les transformations des substances. Les réactions chimiques qui sont mises en œuvre dans de telles transformations peuvent être vues à la fois comme des propriétés chimiques des *réactants* — les molécules qui réagissent entre elles — et comme le moyen de préparer les *produits* — les molécules engendrées. Ce dernier aspect est celui qu'exploite l'industrie chimique pour satisfaire les besoins de notre société en produits aux applications les plus diverses.

Faire la synthèse d'une molécule complexe consiste donc à combiner un ensemble de molécules disponibles plus simples — les produits de départ — par un ensemble de réactions afin de construire progressivement, étape par étape, la structure attendue. Le problème n'est cependant pas trivial car, pour une molécule cible donnée, de très nombreux ensembles de produits de départ peuvent être constitués à partir des centaines de milliers de composés chimiques disponibles commercialement et le choix des réactions à employer, pouvant rarement s'appuyer sur des calculs de chimie théorique, s'effectue le plus souvent par analogie avec des problèmes résolus précédemment [18][19][24]. De plus, un chemin de synthèse conduisant des produits de départ à la molécule cible peut avoir plusieurs dizaines d'étapes. L'espace d'états d'un tel problème est donc de taille exponentielle.

La résolution d'un problème de synthèse comporte deux phases principales : l'élaboration d'un plan de synthèse puis l'expérimentation de ce plan. Le chimiste effectue une partie de la conception du plan par *rétrosynthèse*, c'est-à-dire par un raisonnement analytique selon lequel il part de la structure cible, recherche des réactions pouvant y conduire, engendre les réactants correspondants - les précurseurs - et, si ceux-ci ne sont pas des produits disponibles, réitère le processus jusqu'à obtenir des chemins convenables. Cette démarche fait appel à de nombreuses connaissances, notamment sur les réactions : leurs différentes catégories, leur intérêt synthétique et leurs limites, leurs exemples connus, etc. Le plan n'est qu'une hypothèse qui sera confirmée ou infirmée par l'expérimentation «à la paillasse» [16] .

Un certain nombre de systèmes ont été développés depuis les années 70 pour aider la rétrosynthèse, la prédiction de réactions ou la recherche de produits de départ [22][11][16]. Cependant, basés sur la connaissance d'experts, ces systèmes sont difficiles à réaliser et restent encore à l'état de prototypes plus ou moins élaborés. De ce fait, ils sont peu utilisés. L'un des problèmes majeurs rencontrés dans leur réalisation est la constitution et la mise à jour de leurs bases de connaissances. Diverses approches mettant en oeuvre des techniques d'apprentissage automatique ont été suivies pour

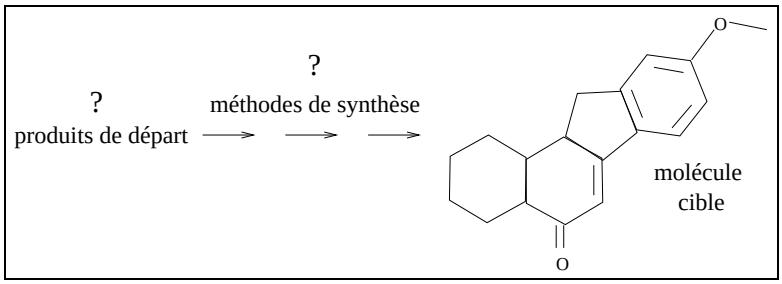

FIG. 1 – Représentation d'un problème de synthèse.

tenter de constituer ces bases de connaissances à partir de bases de données de réactions [10][4][14], mais cette question est encore loin d'être réglée.

Pour notre part, nous travaillons à la conception de nouveaux systèmes d'information chimique dédiés à la résolution des problèmes de synthèse en chimie organique. Ils devraient être susceptibles de combiner des fonctionnalités des systèmes de gestion de bases de données et des systèmes basés sur la connaissance, en disposant de bases qui incluent à la fois des faits et des concepts organisés hiérarchiquement. L'un des aspects de notre recherche est d'étudier ce que peuvent apporter les techniques de fouille de données dans l'extraction de connaissance à partir de bases de données de réactions et, au-delà, dans la structuration de ces bases et à leur interrogation.

Dans cet article, nous présentons les premiers résultats d'une expérience originale : l'application d'algorithmes de recherche de motifs fréquents dans deux bases de réactions chimiques. Les algorithmes utilisés sont des algorithmes par niveaux définis et détaillés dans [1] et [3] (en particulier l'algorithme *Close*). Cette expérience est originale au moins à deux titres : le premier est qu'il s'agit à notre connaissance du seul travail de ce genre réalisé sur des bases de données de réactions, et le second est qu'il s'agit d'un des rares travaux traitant des données dynamiques avec une technique symbolique de fouille (il existe quelques travaux de même nature en analyse de concepts formels, par exemple [9]).

Nous traitons ici les points suivants. Avant de mettre au point ou d'utiliser des méthodes de fouille de données, il est nécessaire de faire une étude préalable des besoins en information auxquels le système devrait pouvoir répondre (§ 2) ainsi que des données que l'on possède et que l'on peut manipuler suite à notre modélisation (§ 3). Nous voyons ensuite (§ 4) ce que proposent les outils de fouille de données qui ont été mis à notre disposition ainsi que les transformations (§ 5.1) des données qui ont été faites pour permettre l'analyse. Puis nous exposons quelques exemples des premiers résultats (§ 5.2) qui ont été obtenus sur l'examen de deux des bases de données de réactions auxquelles nous avons accès. Ensuite (§ 6), nous énumérons les différents problèmes que nous avons rencontrés au cours de ce processus d'extraction de connaissances sur des données issues des bases de données de réactions et nous comparons notre démarche avec celle d'autres systèmes analogues et discutons des apports. Enfin (§ 7), nous concluons et présentons quelques unes des perspectives de ce travail.

2 Les besoins en information

On ne sait pas avec précision combien de réactions particulières de la chimie organique ont été décrites jusqu'à présent dans la littérature, mais il est certain que leur nombre est supérieur à 10 millions. Par ailleurs, aucune nomenclature n'a été développée pour les nommer avec précision. Si de multiples systèmes de classification ont été proposés, il n'en existe pas de parfaitement adapté à la catégorisation des réactions en tant que méthodes de synthèse. Ceci ne facilite ni le stockage ni la recherche d'information sur ces réactions. Néanmoins, depuis les années 80, un certain nombre de bases de données de réactions ont été créées et sont mises à jour régulièrement. Certaines ont la vocation d'être générales et (relativement) exhaustives, comme la base Beilstein, tandis que la plupart sont spécialisées dans un domaine particulier [6]. Toutes sont diffusées par des sociétés commerciales. Leur originalité est de tirer parti de la représentation en termes de graphes qui peut être faite des molécules et qui autorise des recherches par structure et sous-structures (sous-graphes) très performantes sur ces bases. Bien qu'étant des outils très utilisés par les spécialistes de la synthèse, les systèmes de gestion de ces bases ne sont pas exempts de défauts. Hormis la présence de données erronées, on peut constater que les informations, souvent hétérogènes, ne sont pas structurées, les bases étant de simples collections de réactions particulières, que les langages de requête sont peu précis et que les procédures de regroupement des résultats sont peu efficaces. Il en résulte que les interrogations fournissent de grands ensembles de réponses dont il est fastidieux d'extraire l'information pertinente.

Avant d'étudier les données contenues dans les bases de données, les diverses informations que l'on peut en déduire et celles que l'on peut calculer, il est nécessaire de déterminer les services dont ont besoin les utilisateurs d'un système d'information chimique dédié à l'aide à la synthèse organique. Pour cela, nous avons envisagé le type de requêtes qu'un utilisateur du système voudrait pouvoir poser. Nous nous sommes interrogés sur le genre d'information, n'étant pas déjà contenue explicitement ou calculée, qui serait inté-

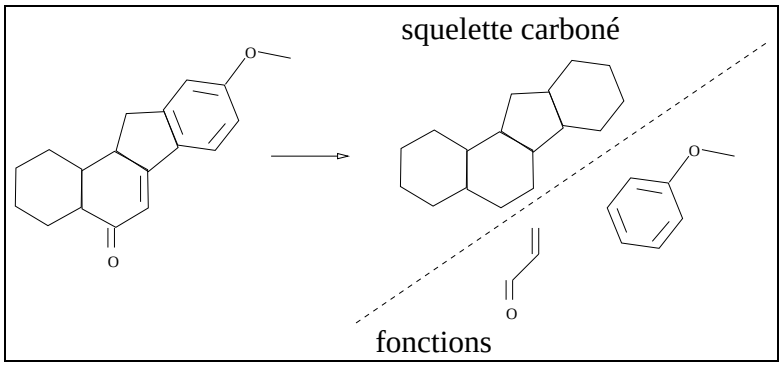

FIG. 2 – Une première analyse du problème.

ressant de connaître sur les réactions. Cependant, au vu de la quantité de questions que l'on voudrait poser à un système d'information chimique, nous avons décidé, dans un premier temps, de nous restreindre à l'un des aspects fondamentaux des réactions : le point de vue de la fonctionnalité.

Ainsi qu'a pu le dire Ireland, la synthèse de la plupart des molécules organiques peut être décomposée en deux problèmes : la préparation du squelette carboné et l'introduction, la modification et/ou l'élimination des divers groupements fonctionnels [12]. Les fonctions sont des groupements d'atomes liés de certaine façon entre eux dont la présence dans les molécules confèrent à celles-ci la majeure partie de leur réactivité et donc de leurs propriétés. Si l'on prend l'exemple de la molécule cible de la figure 1, on obtient l'analyse représentée sur le schéma de la figure 2.

Ceci amène à une catégorisation élémentaire des réactions : celles qui permettent de construire/modifier le squelette des molécules et celles qui agissent sur la fonctionnalité des molécules. Parmi celles-ci on distingue les interchanges fonctionnels[1], les fonctionnalisations[2] et les éliminations de fonctions[3]. Nous nous intéressons ici plus particulièrement aux réactions d'interchange fonctionnel. qui sont des méthodes très utilisées en synthèse.

Dans ce cadre, on peut énumérer quelques unes des requêtes que l'on aimerait poser à un système d'information chimique orienté vers l'aide à la planification de synthèse :

- `Q1` : Quelles sont les fonctions à partir desquelles peut être obtenue telle autre fonction ?
- `Q2` : Quelles sont les réactions qui permettent de passer de telle fonction à telle autre fonction ?
- `Q3` : Quelles sont les fonctions qui restent inchangées lorsque l'on passe de telle fonction à telle autre fonction ?

Il serait aussi souhaitable, si l'on présente une liste de réponses, que les solutions proposées soient classées selon leur «pertinence». En effet, les fonctions sont plus ou moins réactives et il est toujours préférable d'employer, si c'est possible, les méthodes de synthèse dont le rendement sera probablement le meilleur.

3 Les objets du domaine : molécules et réactions

3.1 Les réactions

Une réaction chimique permet de passer d'un ensemble de molécules (*les réactants*) à un autre ensemble de molécules (*les produits*) dans certaines circonstances (*les conditions réactionnelles*).

Dans les bases de réactions auxquelles nous avons accès, les données sont essentiellement de deux types :

- Des données structurelles qui décrivent les molécules participant à la réaction. On dispose aussi de la correspondance entre les atomes des réactants et des produits, permettant de déterminer ce qu'il s'est passé au cours de la réaction (voir Figure 3 la réaction 13426 de la base REACCS-JSM[4] version 2000). Ces données peuvent aussi décrire des molécules qui ne réagissent pas directement tels que les catalyseurs et les solvants.
- Des données de type textuel sur le rendement et les conditions physiques (température, pression ...) dans lesquelles la réaction se réalise ainsi que les références bibliographiques relatives à la réaction.

Nous nous intéressons dans notre travail aux données structurelles. Les molécules pouvant se représenter sous forme de graphes étiquetés sur les nœuds (type des atomes) et les arêtes (type des liaisons), nous pouvons les considérer au niveau informatique dans le cadre formel de la théorie des graphes [28].

Les graphes moléculaires des réactants et des produits sont représentés par une succession de tables de connectivité, une par molécule participant à la réaction. Sur le plan de l'effet

[1] Un interchange fonctionnel est une réaction qui permet de passer d'une ou plusieurs fonctions à une autre fonction.
[2] Une fonctionnalisation est une réaction qui crée une fonction à partir de parties non fonctionnelles de la molécule.
[3] Une élimination de fonction, comme son nom l'indique, conduit à l'obtention d'une partie non fonctionnelle à partir d'une fonction.
[4] La base REACCS-JSM est la version électronique du Journal of Synthetic Methods.

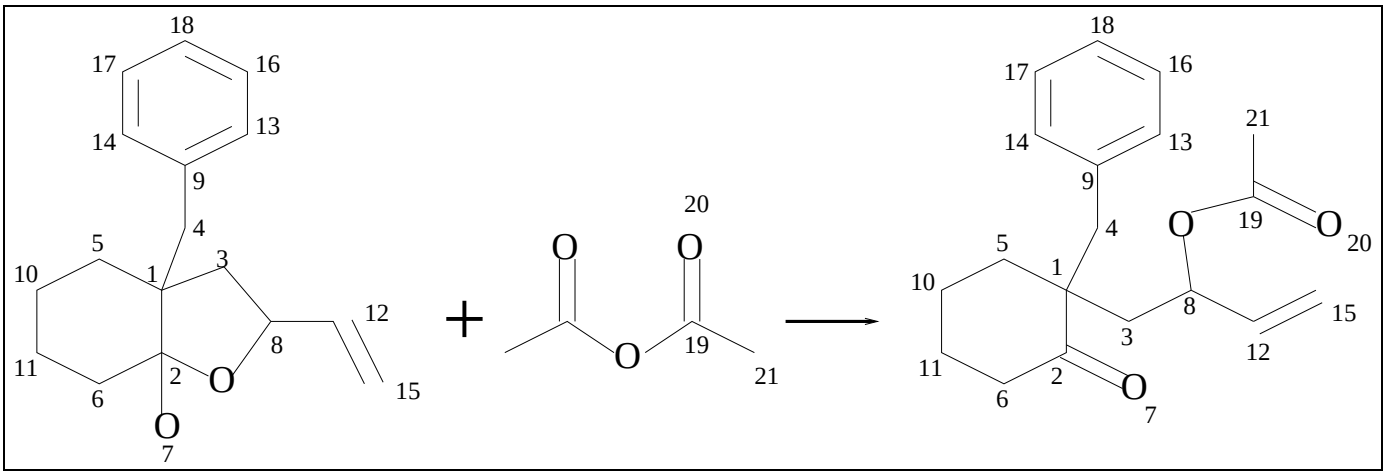

FIG. 3 – Représentation d'une réaction par un schéma de réaction ainsi de la correspondance donnée entre les atomes (réaction 13426 de la base REACCS-JSM version 2000).

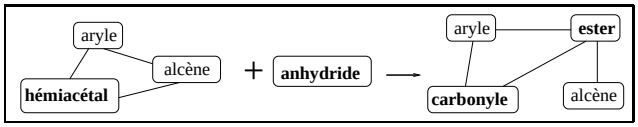

FIG. 4 – Analyse fonctionnelle de la réaction. Les blocs fonctionnels qui ont été modifiés sont en gras, les autres sont inchangés au cours de la réaction.

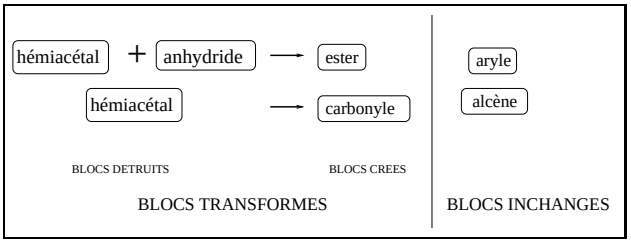

FIG. 5 – Les données sur les réactions d'après l'analyse fonctionnelle de la réaction

de la réaction, cette description n'indique que les modifications subies par les liaisons (création, destruction, modification de type) des molécules. Or l'intérêt d'une réaction en tant que méthode de synthèse réside dans les objectifs structuraux (par exemple un cycle, une fonction alcool ...) qu'elle permet d'atteindre, le type de molécule qu'elle permet de préparer et à partir de quelles autres sortes de composés chimiques.

Pour arriver à exprimer de telles connaissances sur les réactions, il est nécessaire d'enrichir et de structurer les données présentes dans les bases. Nous nous sommes pour cela appuyés sur la modélisation des molécules mise en place lors de travaux antérieurs [29][17][21].

3.2 La perception des molécules

Les méthodes élaborées au cours de ces travaux permettent de détecter dans les molécules la présence de certains motifs structuraux remarquables selon les points de vue de la *topologie* (la topologie décrit les relations de voisinage), de la *fonctionnalité* ou de la *stéréochimie* (la stéréochimie décrit l'arrangement relatif spatial des atomes dans les molécules qui sont en fait des objets tridimensionnels). La reconnaissance de ces sous-structures particulières conduisent les chimistes à classer les molécules et à connaître une grande partie de leurs propriétés. Nous appelons «blocs» ces motifs structuraux remarquables. Les interactions entre ces trois points de vue ont été étudiées sous l'angle de la négociation dans la résolution de problèmes tels que la détermination de stratégies de rétrosynthèse [13].

Pour ce qui est du point de vue de la fonctionnalité, outre une définition d'un bloc fonctionnel à partir de laquelle nous détectons automatiquement tous les blocs fonctionnels présents dans une molécule, nous disposons d'une base de connaissances sur les fonctions qui nous permet de nommer un certain nombre des fonctions perçues. Cette base de connaissances est organisée en une hiérarchie de graphes selon une relation d'inclusion de graphes.

Pour l'instant, notre base de connaissances comprend 304 fonctions nommées. Une fois qu'un bloc fonctionnel a été détecté dans une molécule, on classe ce bloc dans la hiérarchie. Si le bloc est isomorphe à l'un des graphes de la hiérarchie, alors on le nomme avec le nom donné à ce graphe. Sinon, le bloc est nommé "sous-structure inconnue".

On peut représenter la molécule de façon plus abstraite par un graphe des blocs décrivant les relations de voisinage des blocs perçus dans la molécule [30]. Dans le cas du graphe des blocs fonctionnels (voir figure 4), les nœuds sont les blocs fonctionnels tandis que les arêtes représentent les chemins[5] minimaux entre les blocs.

[5]En réalité, ce sont des chemins non fonctionnels, c'est-à-dire des chemins composés d'atomes de carbone simplement liés les uns aux autres.

3.3 Le monde des blocs

La perception des molécules et la correspondance des atomes permettent d'établir une correspondance entre blocs (deux blocs ayant en commun au moins un atome se correspondent) présents dans les réactants et dans les produits. Considérons, par exemple, la réaction de la figure 3. Nous y avons indiqué la correspondance des atomes telle qu'elle est donnée dans la base de données de réactions. La réaction n'est pas équilibrée (c'est-à-dire que l'on ne retrouve pas tous les mêmes atomes de part et de d'autre de la flèche), et donc la correspondance des atomes est incomplète. De plus, il y a un conflit sur un atome d'oxygène que l'algorithme de détermination de la correspondance n'a pas su régler.

Après perception des différentes molécules et mise en correspondance des fonctions perçues, on peut représenter la réaction, de façon abstraite, comme sur la figure 4. On en déduit qu'au cours de la réaction une fonction carbonyle et une fonction ester ont été créées (en gras à droite de la flèche), qu'une fonction anhydride et une fonction hémiacétal ont été détruites (en gras à gauche de la flèche) et que deux fonctions sont restées inchangées (fonctions alcène et aryle).

La correspondance obtenue entre les différents blocs nous permet, de plus, de préciser que :

- la fonction ester créée a été construite à partir des fonctions anhydride et hémiacétal détruites,
- la fonction carbonyle créée a été construite à partir de la fonction hémiacétal détruite.

C'est sur cette représentation des données (ici les réactions), comme indiqué sur le schéma de la figure 5, que vont être appliqués les algorithmes de fouille de données. Cette transformation des réactions en blocs peut être vue comme la phase de préparation des données dans le processus global d'extraction de connaissances dans des bases de données [8].

4 Éléments sur les techniques symboliques de fouille de données

La fouille de données est une des étapes du processus d'extraction de connaissances [8]. En effet avant d'être traitées par les techniques de fouille de données, les données doivent être sélectionnées, pré-traitées et transformées. Après avoir été obtenus, les résultats de la fouille doivent être interprétés et validés par l'analyste, expert du domaine sur lequel portent les données étudiées.

4.1 La place de l'analyste dans le processus d'extraction de connaissances

Pour notre part, nous considérons qu'un expert du domaine relatif aux données, dit l'*analyste*, est au centre du processus d'extraction des connaissances, et qu'il est chargé de contrôler ce processus. L'extraction de connaissances possède un caractère expérimental, voire empirique. En fonction de ses objectifs, l'analyste sélectionne des données (fenêtrage) et utilise des outils de fouille de données pour faire émerger des modèles expliquant les données. Ensuite, l'analyste retient et valide les modèles qui représentent un point de vue satisfaisant pour lui, pour une réutilisation ultérieure.

Pour mener à bien une telle activité, l'analyste met à contribution ses connaissances mais aussi un ensemble d'outils et de fonctionnalités regroupés au sein d'un système d'extraction de connaissances comportant idéalement :

(1) une ou des bases de données avec leurs gestionnaires,
(2) un système de représentation des connaissances du domaine (raisonnements et résolution de problèmes),
(3) le système de fouille de données proprement dit,
(4) une interface se chargeant des interactions et de la visualisation des résultats intermédiaires et finaux.

Ainsi, l'analyste peut s'appuyer non seulement sur ses propres connaissances, mais aussi sur un système de représentation des connaissances du domaine. En particulier, le processus d'extraction des connaissances ne se conçoit pas en l'absence d'un modèle du domaine des données.

4.2 La recherche de motifs fréquents et l'extraction de règles

Le but principal des algorithmes de fouille de données est de rechercher des régularités dans un grand volume de données. Pour cela, différentes approches ont été proposées :

- classification conceptuelle et classification par treillis,
- recherche de motifs fréquents,
- recherche de règles d'association.

Nous disposons de programmes de fouille de données qui permettent d'extraire les motifs fréquents, et parmi ces derniers les motifs fermés, ainsi que d'extraire les règles d'association à partir des motifs fréquents [1][2][3]. Les algorithmes de recherche de motifs fréquents traitent des données qui se présentent de la façon suivante :

- un objet,
- une liste de propriétés associées à cet objet.

Soit un ensemble fini d'objets $\mathcal{O}$, un ensemble fini de propriétés (items) $\mathcal{P}$ et une relation binaire entre ces deux ensembles $\mathcal{R}$. On appelle *base de données* le tableau booléen $\mathcal{O} \times \mathcal{P}$ où x$\mathcal{R}$y = 1 si l'objet x et la propriété y sont en relation par l'intermédiaire de la relation $\mathcal{R}$. Dans ce cas, on dit que l'objet x *possède* la propriété y.

Un *motif* est sous-ensemble de $\mathcal{P}$. Un motif P est *inclus* dans l'objet O si P et O sont en relation.

On définit le support d'un motif comme le rapport entre le nombre d'occurences de ce motif et le nombre d'objets contenus dans la base de données :

$$sup(\mathsf{P}) = \frac{card(f(\mathsf{P}))}{card(\mathcal{O})}$$

avec f(P)= { O $\in \mathcal{O}$ / O contient P }.

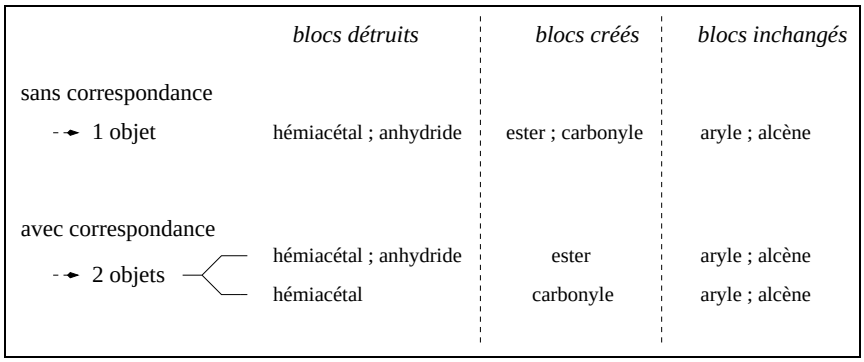

FIG. 6 – Les données transformées pour la fouille de données.

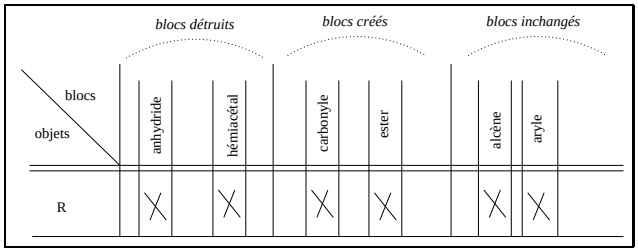

FIG. 7 – Transformation sans prise en compte de la correspondance.

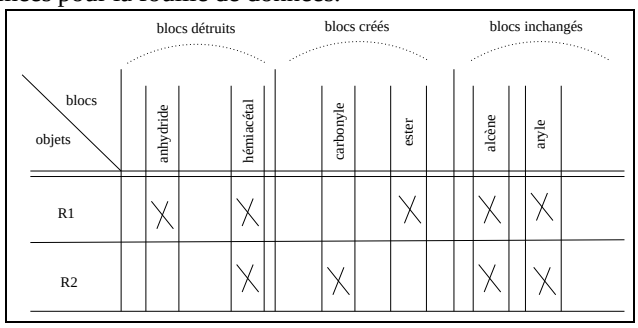

FIG. 8 – Transformation avec prise en compte de la correspondance.

Un motif P est dit *fréquent* si sup(P) $\geq$ minsup avec minsup = seuil au dessous duquel le motif n'est pas considéré comme fréquent.

Á partir des motifs fréquents déterminés, des algorithmes de fouille de données ont pour but de calculer des *règles d'association*. Une règle est constituée d'une prémisse et d'une conclusion et est caractérisée par une *confiance*. La prémisse et la conclusion d'une règle sont des motifs, c'est-à-dire des conjonctions de propriétés. On peut représenter une règle de la façon suivante : R : A $\Rightarrow$ B avec la confiance x.

La confiance (conf) d'une règle (R) est calculée d'après les supports du motif de la prémisse, d'une part, et du motif résultat de la conjonction des motifs de la prémisse et de la conclusion, d'autre part, selon la formule suivante :

$$conf(R) = \frac{sup(A \bigcup B)}{sup(A)}$$

Le support de la règle R est : sup(R) = sup(A $\bigcup$ B).

Une règle dérive forcément d'un motif fréquent : $sup(A \Longrightarrow B) = sup(A \cup B) \geq$ minsup, d'où $A \cup B$ est un motif fréquent. Tous les détails concernant les algorithmes de recherche de motifs fréquents et d'extraction de règles sont donnés dans les références [1], [2] et [3]

5 Extraction de connaissances dans les bases de données de réactions

5.1 Le traitement et les transformations des données

Pour permettre d'appliquer les algorithmes de fouille de données à la description que nous possédons des réactions, il est nécessaire de mettre au point un traitement et un format appropriés. Comme on l'a vu dans la section 4, les données exploitées par les algorithmes de fouille de données que nous utilisons doivent être mises sous la forme d'un tableau de booléens. On ne peut donc pas utiliser directement les représentations abstraites sous forme de graphes de blocs. Il faut trouver le moyen d'exprimer les données de façon adéquate et en perdant le moins d'information possible.

Différentes transformations des données ont été envisagées. En effet, plusieurs choix sont possibles :

- prise en compte de la correspondance des blocs ou non,
- prise en compte des fonctions créées et/ou des fonctions détruites et/ou des fonctions inchangées.

Dans la base de données sur laquelle nous voulons appliquer des outils de fouille, nous avons considéré que les objets sont des réactions et les propriétés les fonctions (fonctions créées et/ou fonctions détruites et/ou fonctions inchangées) présentes dans les molécules participant à la réaction. Si

l'on reprend l'exemple de réaction de la section 3, on peut le décrire de deux façons selon que l'on prenne en compte la correspondance entre les blocs ou non. Les deux possibilités de description sont détaillées dans le schéma de la figure 6.

Si l'on considère le cas où l'on ne tient pas compte de la correspondance entre les blocs, on n'a qu'un objet (R) associé à la réaction considérée. La figure 7 montre ce que sera la ligne correspondant à cet objet dans le tableau booléen de la base de données.

Dans le cas où l'on prend en compte la correspondance des blocs, on associe à la réaction deux objets différents (R1 et R2). On aura donc deux lignes dans le tableau de booléens de la base de données comme sur la figure 8. Dans la base de données, il n'y a aucune indication sur le fait que les 2 objets R1 et R2 font référence à la même réaction (on ne sait donc pas en examinant ces données que "l'hémiacétal" qui est détruit dans R1 pour donner un "ester" est le même que celui qui a donné un "carbonyle" dans R2).

Bien entendu, en passant des représentations abstraites à cette description booléenne, on perd aussi les indications de voisinage des différents blocs dans les molécules.

5.2 Les premiers résultats

Les résultats que nous présentons ici concernent deux des bases de données de réactions commercialisées[6]. Comme dans toute fouille de données, l'application du processus doit se faire sur une quantité importante de données. C'est pourquoi nous avons choisi la base ORGSYN version 2000 (environ 5500 réactions) pour faire les tous premiers tests puis la base JSM version 1999 (plus de 60000 réactions) pour passer à une échelle plus appropriée à la validation de résultats de fouille de données.

5.2.1 Statistiques sur la réactivité des fonctions dans les bases de données de réactions

Nous nous interrogeons ici sur le contenu des bases. On désire répondre à des questions du type : «quelles sont les fonctions qui sont le plus souvent présentes dans les molécules des bases de données de réactions ?», «est-ce que ces fonctions sont souvent créées/détruites/inchangées ?», «quels sont les principaux interchanges de fonctions observés ?». Les motifs fréquents et leur support peuvent donner des éléments de réponse à ces interrogations. Le problème que nous souhaitons résoudre peut s'énoncer de la façon suivante : nous voulons construire une fonction (fonction créée) et savoir comment le faire (fonctions détruites) en prenant en compte certains impératifs (fonctions inchangées).

Exploitation des motifs de taille 1

Les motifs fréquents permettent dans un premier temps d'étudier, de façon générale, le degré de réactivité de chaque fonction. En effet, en comparant les supports relatifs des motifs de taille 1, on peut examiner statistiquement dans quelles proportions les fonctions réagissent.

Si l'on pose pour la fonction *f*,

- f_c = item représentant la fonction *f* créée
- f_d = item représentant la fonction *f* détruite
- f_i = item représentant la fonction *f* inchangée

Soit sup(f_c), sup(f_d), sup(f_i) les supports respectifs de f_c, f_d et f_i sur la base de données. Ici, le support d'un motif représente le nombre d'occurences du motif dans la base et non plus le rapport présenté section 4. Le nombre total de fois où la fonction *f* apparaît dans la base de données est égal à : n_{tot} = sup(f_c) + sup(f_d) + 2×sup(f_i). En effet, la fonction est comptée à chaque fois qu'elle est créée (présente dans les produits et pas dans les réactants), détruite (présente dans les réactants et pas dans les produits) mais aussi inchangée (présente dans les réactants et dans les produits). De même, le nombre de fois où la fonction est présente dans les réactifs s'écrit : n_{react} = sup(f_d) + sup(f_i), ainsi que le nombre de fois où on la détecte dans les produits est égal à : n_{prod} = sup(f_c) + sup(f_i). Bien entendu, $n_{tot} = n_{react} + n_{prod}$.

On peut donc calculer les différents pourcentages suivants :

1. $p_r = n_{react}/n_{tot}$ et $p_p = n_{prod}/n_{tot}$ indiquent la présence relative de la fonction dans les réactants et dans les produits,
2. $p_{dr} = sup(f_d)/n_{react}$ et $p_{ir} = sup(f_i)/n_{react}$ donnent des indications sur la réactivité de la fonction,
3. $p_{cp} = sup(f_c)/n_{prod}$ et $p_{ip} = sup(f_i)/n_{prod}$ indiquent si la fonction est facile ou non à faire.

Quelques résultats de l'expérimentation sur les motifs de taille 1

Tout d'abord, nous constatons que l'on rencontre plus de fonctions différentes dans la base JSM 1999 que dans la base ORGSYN 2000. En effet, parmi les 304 fonctions présentes dans notre base de connaissances, 176 fonctions ont été identifiées dans au moins une des molécules de la base JSM 1999 tandis que seulement 129 sont trouvées dans ORGSYN 2000. Ceci est normal car la base JSM 1999 compte 10 fois plus d'exemples de réactions qu'ORGSYN 2000.

D'autre part, on peut observer les fonctions qui sont les plus réactives ainsi que celles qui sont la plupart du temps inactives. Au vu des pourcentages définis précédemment et des connaissances du domaine, nous avons établi les règles suivantes :

- Fonctions très stables : les fonctions très stables sont celles qui apparaissent relativement souvent dans la base et qui sont la plupart du temps inchangées c'est-à-dire des pourcentages p_{ir} et p_{ip} d'environ 90%. On remarque, par exemple, que la fonction «aryle» se retrouve dans de nombreuses molécules, comme cela a déjà été constaté dans [15]. De plus, grâce à notre représentation, nous pouvons ajouter que, dans la plupart

[6] http ://www.mdli.com

des cas, la fonction «aryle»[7] détectée reste inchangée, ce qui correspond à la stabilité chimique bien connue des fonctions aromatiques. Nous retrouvons là une propriété en accord avec le domaine de connaissances.

– Fonctions très réactives : ce sont des fonctions qui, quand elles sont présentes, que ce soit dans les réactants ou dans les produits, sont transformées. Les valeurs des pourcentages p_{dr} et p_{cp} doivent être élevées (plus de 90%) tandis que celles des pourcentages p_{ir} et p_{ip} doivent être faibles (moins de 10%). Le fait que ces fonctions très réactives soient majoritairement créées quand elles sont présentes dans les produits peut paraître paradoxal. Ceci s'explique simplement : ne trouvant pas ces fonctions dans les produits de départ, il est nécessaire de les créer quand on en a besoin.

Cette interprétation des résultats a été faite par les analystes — en l'occurrence les deux premiers auteurs — qui sont des spécialistes de la synthèse en chimie organique, ce qui confirme le rôle majeur de l'analyste dans le processus d'extraction des connaissances dans les bases de données [27][5].

Exploitation des motifs de taille supérieure

Nous nous intéressons ici aux motifs de taille supérieure ou égale à 2 impliquant les combinaisons de fonctions détruites et les fonctions créées par cette combinaison. Lorsque l'on tient compte de la correspondance entre les blocs, la recherche de motifs fréquents sur chacune des 2 bases nous indique que les réactions transformant une fonction alcool ou une fonction acide en une fonction ester ou bien, inversement, permettant le passage d'une fonction ester en acide carboxylique ou en alcool (voir figure 9 le schéma de ces deux réactions), sont celles qui ont pratiquement la plus grande fréquence d'apparition (environ 2%). Ce résultat a déjà été remarqué dans une étude faite en 1996 sur la base CASREACT[8] [15].

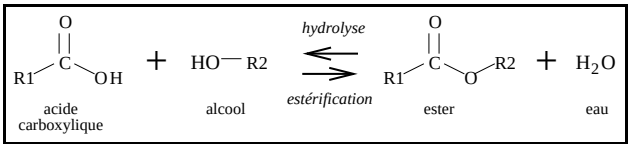

FIG. 9 – La réaction d'estérification et la réaction inverse d'hydrolyse des esters.

Les fréquences d'apparition peuvent apparaître faibles, voire très faibles, par rapport à celles observées habituellement dans les applications de fouille de données. Mais étant donné que nous avons étudié des bases de données sélectives, le résultat n'est pas surprenant vu que le but de ce type de base de données est de couvrir le maximum de catégories de réactions et non de présenter énormément d'exemples de réactions pour une même catégorie de réaction.

5.2.2 Interprétation et formulation de requêtes

Nous nous intéressons ici aux résultats de la fouille par rapport à des problèmes de synthèse donnés. Comme nous l'avons dit précédemment, l'énoncé général du problème est le suivant : on veut construire une fonction (fonction créée) et savoir comment le faire (fonctions détruites) en prenant en compte certains impératifs (fonctions inchangées). L'exploitation des motifs fréquents qui nous préoccupe alors est celle des motifs de taille supérieure ou égale à 2 associant des items de natures différentes (fonctions créées et/ou détruites et/ou inchangées). Vu le type de questions que l'on se pose dans ce cadre, il faut regarder les motifs obtenus par exploitation des bases de données dans lesquelles on tient compte de la correspondance des blocs. Les règles d'association permettent de classer les réponses. Le travail consiste à déterminer la forme générale des motifs ou des règles d'association correspondant à la requête que l'on désire présenter au système. Les résultats des interrogations posées ci-dessous ont été validés, ceux qui paraissaient les plus surprenants ont été compris après examen des réactions associées aux motifs.

Les deux premiers types de requête, présentés ci-dessous, s'apparentent aux questions `Q1` et `Q2`, le troisième type de requête répond à la question `Q3`, posées à la section 2.

• **Premier exemple de requête**

Le premier type de requête que l'on peut se poser face à un problème de synthèse d'une molécule avec une certaine fonctionnalité est de savoir s'il existe des exemples de construction de cette fonctionnalité et, si oui, comment cette fonctionnalité est construite.

Exploitation des motifs fréquents

Si l'on veut savoir à partir de quelles fonctions (F_d) on peut obtenir telle autre fonction (F_c), on doit regarder tous les motifs de la forme : $F_d \wedge F_c$.

Supposons que nous voulions connaître les fonctions permettant d'obtenir un ester, nous devons regarder les motifs de taille 2 composés d'un item de fonction détruite F_d et de l'item de fonction créée "$ester_c$"(c'est-à-dire : $F_d \wedge ester_c$). Par exemple, pour chercher à savoir si un ester peut dériver d'un alcool, il faut vérifier que $alcool_d \wedge ester_c$ est fréquent.

Exploitation des règles d'association

De la même manière que précédemment, nous voulons obtenir une molécule appartenant à la classe des esters et nous désirons savoir quelle est la façon la plus courante (au sens de la fréquence) d'obtenir des molécules de cette catégorie. Nous voulons récupérer des règles dont l'interprétation est : "si un ester est créé alors c'est à partir des fonctions F_d". On doit donc s'intéresser aux règles dont la prémisse contient "$ester_c$" et la conclusion est composée d'items associés à des fonctions détruites. Ceci nous amène

[7] La fonction aryle est un cycle de 6 atomes ayant une distribution électronique particulière dont la présence dans une molécule détermine des propriétés typiques du benzène et de ses dérivés.

[8] `http ://www.cas.org/CASFILES/casreact.html`

à connaître les principales fonctions ou ensemble de fonctions à partir desquelles on peut obtenir un ester. Ce sont des règles de la forme : $\text{ester}_c \Rightarrow \{\text{F}_d\}$ avec $\{\text{F}_d\}$ ensemble de fonctions détruites. En comparant les valeurs des confiances de chacune des règles, nous pouvons classer les possibilités de combinaison de fonctions qui permettent d'obtenir une fonction ester.

- **Deuxième exemple de requête**

Nous nous intéressons ici au fait qu'une fonction peut être construite à partir de 2 autres fonctions et donc aux motifs de la forme : $\text{F}_{d1} \wedge \text{F}_{d2} \wedge \text{F}_c$. Parmi les résultats des requêtes précédentes, on remarque, en accord avec la connaissance du domaine, que la formation d'une fonction ester implique des combinaisons de deux fonctions.

Exploitation des motifs fréquents

Si nous désirons connaître toutes les combinaisons de deux fonctions qui forment un ester, les motifs que nous recherchons sont de la forme : $\text{F}_{d1} \wedge \text{F}_{d2} \wedge \text{ester}_c$.

Exploitation des règles d'association

Ayant sélectionné la première fonction réagissant F_{d1}, on veut connaître et classer la liste des fonctions $\{\text{F}_{d2}\}$ avec lesquelles F_{d1} réagit pour donner F_c. On doit comparer les confiances des règles d'association de modèle général : $\text{F}_{d1} \wedge \text{F}_c \Rightarrow \text{F}_{d2}$.

Dans notre cas, on remarque que pour construire un ester, on emploie souvent un alcool associé à une autre fonction. Si l'on choisit donc de partir d'un alcool, on va regarder les règles de la forme : $\text{alcool}_d \wedge \text{ester}_c \Rightarrow \text{F}_d$.

- **Troisième exemple de requête**

Bien souvent le choix de la méthode à employer pour obtenir une fonction dépend des autres fonctions présentes et que l'on désire préserver. Il est donc très utile de savoir si l'on connait des exemples de réactions qui remplissent les conditions désirées. Les motifs fréquents qui répondent à ce type de requête sont de la forme : $\text{F}_c \wedge \text{F}_i \wedge \text{F}_d$.

Exploitation des motifs fréquents

En examinant les motifs fréquents, on peut déjà savoir s'il existe des réactions dans la base de données qui construisent la fonction voulue en laissant inchangées d'autres fonctions et, si oui, lesquelles. Si l'on veut connaître les fonctions donnant un ester, un éther étant inchangé, les motifs à regarder sont de la forme générale : $\text{ester}_c \wedge \text{éther}_i \wedge \text{F}_d$.

Exploitation des règles d'association

Si l'on veut comparer les solutions proposées par (F_d), on pourra étudier les règles de la forme : $\text{F}_c \wedge \text{F}_i \Rightarrow \text{F}_d$.

Par exemple, si l'on veut connaître les exemples de réactions qui laissent inchangée une fonction éther alors qu'un ester est formé, on regarde les règles ayant pour modèle général : $\text{ester}_c \wedge \text{éther}_i \Rightarrow \text{F}_d$.

Ceci ne nous donne pas les *combinaisons* de fonctions à partir desquelles on a observé la formation d'un ester laissant un éther invariant. Pour répondre à cette question, nous avons la règle suivante : $\text{ester}_c \wedge \text{éther}_i \Rightarrow \text{alcool}_d \wedge \text{anhydride}_d$ (10,6%).

De plus, si après examen des résultats de la requête précédente, on décide d'employer un alcool et que l'on veut savoir quelle fonction on va faire réagir avec cet alcool, la règle [$\text{ester}_c \wedge \text{éther}_i \wedge \text{alcool}_d \Rightarrow \text{anhydride}_d$ (40,1%)] nous indique que dans plus de 40% des cas de la base de données un anhydride a été employé.

6 Discussion

Durant cette première expérience de fouille de données, nous avons rencontré différents problèmes. Il ressort au premier abord des contraintes qui sont dues à la transformation obligatoire des données sous forme d'un tableau booléen et aux caractéristiques complexes des données traitées. Parallèlement, l'exploitation de connaissances du domaine est une nécessité : il ne serait pas envisageable de mener à bien une telle expérience sans connaissance du problème visé - l'aide à la planification de synthèse de molécules - et des objets manipulés, molécules et réactions. Par ailleurs, l'interprétation des résultats nécessite également l'intervention d'un analyste expert du domaine, comme discuté dans [27][5].

- Les molécules et les réactions sont des objets très complexes. Elles peuvent se représenter, moyennant une simplification, sous la forme de graphes pour les premières, de règles de réécriture de graphes pour les secondes. Cependant, les programmes de fouille de données manipulent des données de nature textuelle codées dans des tableaux de booléens. Nous avons donc dû traiter et transformer les données de façon à pouvoir appliquer des techniques de fouille, ce qui nous a conduit à nouveau à une perte d'information. Une extension des techniques de fouille à des objets de nature plus complexe, comme cela est étudié dans [23], doit être fortement envisagée.
- Un autre problème réside aussi dans le choix des seuils du support (pour les motifs) et de la confiance (pour les règles). En effet, la fréquence des motifs est bien moindre que celle observée dans les données habituellement traitées par les algorithmes de fouille de données. Toutefois, il faut se rendre compte qu'un seuil de 2% pour 5000 réactions représente une famille de 100 réactions, ce qui est un chiffre tout à fait convenable pour une famille standard de réactions.
- Le traitement des données doit se faire en fonction des questions qui se posent. Pour tester la réactivité comparée des fonctions par exemple (voir 5.2.1), il faut simplement étudier les motifs de taille 1 sur des tableaux où la correspondance des blocs n'est pas exprimée (la prise en compte de la correspondance multiplie les objets pour une même réaction et introduit un biais pour le

type de question posée). Pour les autres questions, par exemple celles qui prennent en compte les réactions, il faut considérer explicitement la correspondance des blocs.
- Introduire la correspondance entre blocs peut provoquer des effets de bord indésirables. Ainsi, le fait de créer plusieurs objets pour une même réaction provoque la duplication en autant d'exemplaires des fonctions inchangées, et donc, un calcul de support erroné pour de telles fonctions. Toutefois, c'est la seule façon qui soit possible pour prendre en compte la correspondance entre blocs. Là encore, l'obligation de travailler avec un tableau booléen introduit une contrainte et peut aussi mener à des erreurs.
- Un nombre très important de règles redondantes peut être engendré à partir des motifs fréquents (comme discuté dans [1]). Toutefois, ce problème peut être minimisé lorsque l'on ne considère que des règles dites informatives [1] : des règles où la prémisse est minimale et la conclusion maximale en terme d'éléments. Ceci s'applique par exemple à la recherche de toutes les combinaisons de fonctions possibles pour créer une fonction donnée. Cependant, quand nous avons des requêtes précises à poser au système, la traduction de ces requêtes sous forme de règles nous conduit à nous intéresser à des règles ayant une prémisse plus importante en taille que la conclusion.
- On peut discuter de l'intérêt de certaines règles. En effet, nous avons remarqué que certaines fonctions étaient très fréquentes (5.2.1). D'une manière générale les motifs et les règles d'association contenant ces fonctions ne nous intéresseront pas ou peu.

Travaux analogues

Il est paradoxal de constater que très peu de travaux existent actuellement sur la fouille de données présentant un caractère dynamique comme des règles - ici des réactions - ou des données temporelles. Un parallèle peut être fait avec le travail présenté dans [9] où l'analyse de concepts formels est utilisée pour étudier la représentation et la classisfication de «connaissances dynamiques», ici un automate et les transitions qui lui sont associées. Les éléments de connaissances sont représentés sous forme de graphes conceptuels, et les transitions sont considérées comme des couples d'états auxquels sont associées des actions : c'est à partir d'un tel tableau, de nature booléenne, comme celui qui décrit nos réactions, qu'est construit un treillis (classification par treillis [1]) et que sont extraites des règles d'association.

Soulignons de nouveau que ce travail de recherche est un des (très) rares à utiliser la classification symbolique - extraction de motifs fréquents et de règles d'associations - pour mener à bien une expérience de fouille de données sur des données présentant un aspect dynamique. Habituellement, ce sont plutôt des techniques de fouille numériques [20] qui sont mises en oeuvre.

Nos présents travaux ont bénéficié de l'acquis des recherches effectuées dans le cadre du GDR 1093 du CNRS (Traitement Informatique de la Connaissance en Chimie Organique, 1993-2000, directeur : C. Laurenço). Ce GDR a eu pour objectif d'étudier des systèmes d'information chimique susceptibles d'aider les chimistes à résoudre des problèmes de synthèse en chimie organique. De nature interdisciplinaire, il a réuni des chercheurs en chimie et en informatique, tant du secteur publique que de l'industrie. Les principaux thèmes de recherche développés ont été la modélisation de la synthèse organique [11], la représentation des connaissances chimiques sous forme de graphes et de hiérarchies de graphes [28][30][7] et différents mécanismes de raisonnement pouvant lui être associés tels que la classification [28][21], le raisonnement par analogie [24] ou à partir de cas [18][19], la négociation [13] ainsi que l'apprentissage de connaissances stratégiques [26][25]. Une large part de ces recherches a été menée sur l'algorithmique des graphes «chimiques» [28] et notamment sur la résolution du problème de l'isomorphisme de sous-graphes partiels à l'aide de CSP [26]. Plusieurs systèmes ont été réalisés [29] dont RESYN_Assistant, le prototype d'un système d'aide à la compréhension des problèmes de synthèse organique que nous utilisons comme plate-forme d'expérimentation.

7 Conclusions et perspectives

Nous avons montré dans cet article une application des techniques de fouille à des données très particulières, les réactions chimiques. Nous avons vu que les réactions sont des objets complexes que l'on doit décrire à divers niveaux d'abstraction pour pouvoir en extraire des informations pertinentes. La représentation originale par blocs des réactions que nous avons mise au point nous permet d'avoir des informations globales sur les réactions. La fouille de données sur la représentation effectuée du point de vue de la fonctionnalité conduit à obtenir des motifs fréquents et des règles d'association en accord avec les connaissances du domaine. L'inconvénient réside dans la perte d'informations structurelles due au traitement de tableaux uniquement booléens. Les résultats obtenus sont encourageants et donnent une certaine satisfaction aux chimistes, car ils leur montrent que les bases de réactions peuvent non seulement être interrogées de façon classique mais aussi être exploitées plus directement pour traiter un problème de synthèse.

Comme perspectives immédiates, en dehors des extensions dont il est question ci-dessus, il serait intéressant d'appliquer des techniques de fouilles de données sur les conditions réactionnelles, même si ces données sont plus ou moins fiables. Pour pallier les défauts de ces données, on peut sélectionner les bases les mieux contrôlées. Mais il sera aussi nécessaire de mettre en place un pré-traitement particulièrement efficace des données.

Par ailleurs, les techniques de fouille de données peuvent également être utilisées pour faire émerger des méthodes de synthèse générales comme celles de construction de cycles, à sept chaînons par exemple, qui sont des objets particulièrement intéressants du point de vue de la synthèse.

Remerciements

Les auteurs remercient la société Molecular Design Ltd de leur avoir permis d'utiliser les données extraites des bases de données de réactions qu'elle commercialise.

Références

[1] Y. Bastide. *Data mining : algorithmes par niveaux, techniques d'implantation et applications*. Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand, décembre 2000.

[2] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal. Mining frequent patterns with counting inference. *ACM SIGKDD Explorations*, 2(2) :66–75, 2000.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1) :65–95, 2002.

[4] E. S. Blurock. Automatic extraction of reaction information from databases using classification and learning techniques. In *Chemical Information 2*, pages 25–35. Springer-Verlag, Berlin Heidelberg, 1991.

[5] R.J. Brachman et T. Anand. The process of knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et R. Uthurusamy, éditeurs, *Advances in Knowledge Discovery and Data Mining*, pages 37–57. AAAI Press / MIT Press, Menlo Park, California, 1996.

[6] J. Coste, O. Gien, A. Dietz et C. Laurenço. A propos de l'utilisation des bases de données de réactions. *L'actualité chimique*, pages 27–32, Juillet 1999.

[7] A. Dietz, C. Fiorio, M. Habib et C. Laurenço. Representation of stereochemistry using combinatorial maps. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 51 :117–128, 2000.

[8] U. M. Fayyad, G. Piatetsky-Shapiro et P. Smith. From data mining to knowledge discovery : An overview. In *Advances in knowledge discovery and data mining*. AAIT/MIT Press, 1996.

[9] B. Ganter et S. Rudolph. Formal concept analysis methods for dynamic conceptual graphs. In *Proceedings of the 9th International Conference of Conceptual Structures (ICCS'2001)*, pages 143–156, Stanford University, California, USA, August 2001. Springer-Verlag.

[10] H. Gelernter, J. R. Rose et C.Chen. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput. Sci.*, 30 :492–504, 1990.

[11] O. Gien. *Modélisation de la synthèse organique multi-étapes. Développement d'outils informatiques d'aide à la conception de plans de synthèse.* Thèse de doctorat, Université Montpellier II, mars 1998.

[12] R. E. Ireland. *Organic Synthesis*. Prentice-Hall, Inc., N. J., 1969.

[13] P. Jambaud. *Le dialogue comme processus de résolution de problème. Une application en chimie organique.* Thèse de doctorat, Université Montpellier II, décembre 1996.

[14] P. Jauffret, H. Vogel, S. Schildknecht et G. Kaufmann. Learning synthetic knowledge from reaction databases : dealing with experimental conditions. In H. Collier, éditeur, *Proceedings of the 2000 International Chemical Information Conference*, pages 137–163. Infonortics Ltd., Tetbury (England), 2000.

[15] P. E. Blower Jr., G. J. Myatt et M. W. Petras. Exploring functional group transformations on CASREACT. *J. Chem. Inf. Comput. Sci.*, 37 :54–58, 1997.

[16] C. Laurenço. *Synthèse Organique Assistée par Ordinateur*. Thèse de doctorat d'État, Université Louis Pasteur (Strasbourg I), septembre 1985.

[17] C. Laurenço, M. Py, A. Napoli, J. Quinqueton et B. Castro. Représentation de connaissance en synthèse organique à l'aide d'un langage à objets. *New J. Chem.*, 14(12) :921–931, 1990.

[18] J. Lieber. *Raisonnement à partir de cas et classification hiérarchique – Application à la planification de synthèses en chimie organique*. Thèse de doctorat, Université Henri Pointcaré (Nancy I), octobre 1997.

[19] J. Lieber et A. Napoli. Planification à partir de cas et classification. In J. Charlet, M. Zacklad, G. Kassel et D. Bourigault, éditeurs, *Ingénierie des connaissances — Évolutions récentes et nouveaux défis*, pages 357–369. Eyrolles, Paris, 2000.

[20] J.-F. Mari, F. Le Ber et M. Benoît. Fouille de données agricoles par modèles de markov cachés. In *Journées francophones d'ingénierie des connaissances, IC2000, Toulouse, France*, pages 197–205, 2000.

[21] A. Napoli et C. Laurenço. Représentations à objets et classification : Conception d'un système d'aide à la planification de synthèses organiques. *Revue d'Intelligence Artificielle*, 7(2) :175–221, 1993.

[22] M.A. Ott et J.H. Noordik. 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*, 111 :239–246, 1992.

[23] G. Polaillon. *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme*. Thèse de doctorat, Université Paris IX-Dauphine, décembre 1998.

[24] M. Py. *Un agent rationnel pour raisonner par analogie*. Thèse de doctorat, Université Montpellier II, novembre 1992.

[25] J.-C. Régin. *Développement d'outils algorithmiques pour l'Intelligence Artificielle et application à la chimie organique*. Thèse de doctorat, Université Montpellier II, décembre 1995.

[26] J.C. Régin, O. Gascuel et C. Laurenço. Machine learning of strategic knowledge in organic synthesis from databases. In *AIP Conference Proceedings, Computational Chemistry*, pages 618–623, 1995.

[27] A. Simon. *Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données*. Thèse de doctorat, Université Nancy 1, septembre 2000.

[28] P. Vismara. *Reconnaissance et représentation d'éléments structuraux pour la description d'objets complexes. Application à l'élaboration de stratégies de synthèse en chimie organique*. Thèse de doctorat, Université des Sciences et Techniques du Languedoc (Montpellier II), décembre 1995.

[29] P. Vismara, P. Jambaud, C. Laurenço et J. Quinqueton. RESYN : objets, classification et raisonnement distribué en chimie organique. In R. Ducournau, J. Euzenat, G. Masini et A. Napoli, éditeurs, *Langages et modèles à objets : États des recherches et perspectives*, volume 19 of *Collection didactique*, chapitre 14, pages 397–419. INRIA, 1998.

[30] P. Vismara et C. Laurenço. An abstract representation for molecular graphs. *DIMACS Series In Discrete Mathematics and Theoretical Computer Science*, 51 :343–366, 2000.