Prétraitement des bases de données de réactions chimiques pour la fouille de schémas de réactions

Frédéric Pennerath[*,***], Géraldine Polaillon[**], Amedeo Napoli[***]

[*]Supélec, campus de Metz
2 rue Edouard Belin 57070 Metz
frederic.pennerath@supelec.fr
[**]Supélec, campus de Gif-sur-Yvette
3 rue Joliot-Curie 91192 Gif-sur-Yvette
geraldine.polaillon@supelec.fr
[***]Equipe Orpailleur, Loria
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex
amedeo.napoli@loria.fr

Résumé. Un grand nombre de réactions chimiques sont aujourd'hui répertoriées dans des bases de données. Les chimistes aimeraient pouvoir fouiller les graphes moléculaires contenus dans ces données pour en extraire des schémas de réactions fréquents. Deux obstacles s'opposent à cela : d'une part la manière dont les chimistes représentent les réactions par des graphes ne permet pas aux techniques de fouille de graphes d'extraire les schémas de réactions fréquents. D'autre part les bases de données contiennent des descriptions de réactions souvent incomplètes, ambiguës ou erronées. Le présent article décrit un processus de prétraitement opérationnel qui permet de filtrer, compléter puis transformer le contenu d'une base de réactions en des données fiables constituées de graphes abstraits répondant au problème de la fouille de schémas de réactions. Le processus place ainsi les bases de réactions à portée des techniques de fouille de graphes comme en attestent les résultats expérimentaux.

1 Introduction

Les chimistes mettent au point de nouveaux procédés de synthèse de molécules en consultant de très grandes bases de données recensant les réactions chimiques disponibles. Les chimistes aimeraient pouvoir fouiller les graphes moléculaires contenus dans ces données pour en extraire des schémas de réactions fréquents qui serviront de candidats privilégiés lors de nouveaux problèmes de synthèse. Deux obstacles s'opposent à cela. D'une part la manière dont les chimistes représentent les réactions par des graphes ne permet pas aux techniques de fouille de graphes d'extraire les schémas de réactions fréquents. Il existe des algorithmes efficaces (Yan et Han, 2002, 2003; Nijssen et Kok, 2004) pour extraire d'un ensemble E de graphes étiquetés l'ensemble des sous-graphes G connexes fréquents dont le support, défini comme le nombre de graphes de E qui contiennent au moins un sous-graphe isomorphe à G, est supérieur à un certain seuil. Si ces méthodes peuvent s'appliquer avec succès à la fouille de graphes

moléculaires (Fischer et Meinl, 2004), leur application directe aux graphes d'une base de réactions ne conduirait à aucun résultat pertinent : tout au plus pourrait-on mettre en évidence les fragments de graphes moléculaires qui sont fréquemment détruits ou au contraire fréquemment créés lors des réactions sans qu'aucun schéma de réaction, c'est à dire, aucun schéma de transformation entre graphes moléculaires, ne puisse s'en déduire. D'autre part les bases de données contiennent des descriptions de réactions souvent incomplètes, ambiguës ou erronées. Leur fouille sans autre filtrage conduirait à des résultats trop bruités donc inexploitables.

Une étape de prétraitement s'avère donc indispensable pour améliorer la qualité des données fouillées et pour exprimer les données au sein d'un modèle répondant au problème posé. Considérant qu'un problème d'extraction de connaissance ne peut être résolu efficacement que si les connaissances du domaine d'application sont prises en compte à tous les niveaux, que ce soit lors du prétraitement des données, de leur fouille proprement dite ou de l'analyse des résultats, le présent article décrit comment les connaissances du domaine, c'est à dire certains principes établis de chimie organique, ont aidé à la conception d'un prétraitement original des bases de réactions, conçu spécifiquement pour extraire les schémas de réactions fréquents à l'aide des algorithmes existants de recherche de sous-graphes fréquents. L'article se restreint essentiellement à présenter les détails de ce prétraitement, qui n'est qu'une étape d'un processus d'extraction de connaissances plus vaste, dont les principes généraux ont été introduits dans Pennerath et Napoli (2006) et dont les premiers résultats sont exposés dans Pennerath et Napoli (2008).

2 Définitions formelles des données et du problème

Les molécules sont des groupes d'atomes maintenus solidaires par des forces de covalence. Toute molécule se modélise donc naturellement par un *graphe moléculaire* g étiqueté et connexe dont les ensembles de sommets $S(g)$ et d'arêtes $A(g)$ représentent respectivement les atomes et les liaisons de covalence de la molécule. Chaque sommet $s \in S(g)$ est étiqueté par l'élément chimique $e(s)$ de son atome (H pour l'hydrogène, ..., et par défaut C pour le carbone) et chaque arête $a \in A(g)$ est étiquetée par la multiplicité $m(a)$ de la liaison associée (simple, double, etc). Une réaction chimique modifie la structure de certaines molécules et se modélise en première approximation par une transformation de graphes moléculaires. Les chimistes représentent cette transformation par une *équation chimique* (cf figure 1) dont les membres de gauche et de droite représentent respectivement l'ensemble des graphes moléculaires des molécules de départ, appelées *réactants*, et des molécules d'arrivée appelées *produits*. Les bases

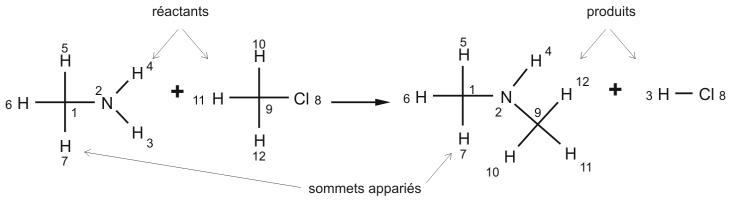

FIG. 1 – *Exemple d'équation de réaction à deux réactants et deux produits.*

de réactions décrivent principalement chaque réaction par son équation chimique, c'est à dire

par la donnée de deux graphes moléculaires $\mathcal{R}$ et $\mathcal{P}$ représentant respectivement les membres gauche et droit de l'équation. Les composantes connexes de $\mathcal{R}$ (resp. $\mathcal{P}$) correspondent aux graphes moléculaires des différents réactants (resp. produits). Deux fonctions d'annotation $\lambda_{\mathcal{R}} : \mathcal{D}_{\lambda_{\mathcal{R}}} \subseteq S(\mathcal{R}) \rightarrow \mathbb{N}$ et $\lambda_{\mathcal{P}} : \mathcal{D}_{\lambda_{\mathcal{P}}} \subseteq S(\mathcal{P}) \rightarrow \mathbb{N}$ des sommets de $\mathcal{R}$ et de $\mathcal{P}$ complètent les données des graphes $\mathcal{R}$ et $\mathcal{P}$. Un sommet s est *apparié* si $\lambda_{\mathcal{R}}$ ou $\lambda_{\mathcal{P}}$ lui associe un indice d'appariement (i.e. $s \in \mathcal{D}_{\lambda_{\mathcal{R}}} \cup \mathcal{D}_{\lambda_{\mathcal{P}}}$). Une équation (cf figure 1) annote chaque sommet apparié par son indice. Deux sommets $s_1 \in \mathcal{R}$ et $s_2 \in \mathcal{P}$ sont *appariés* l'un à l'autre s'ils sont annotés par le même entier. Ils sont alors censés représenter un et un seul même atome. Une équation $(\mathcal{R}, \mathcal{P})$ est dite totalement appariée si tout sommet de $\mathcal{P}$ est apparié. Une base de réactions est donc équivalente à un ensemble de 4-uplets $\{(\mathcal{R}_i, \mathcal{P}_i, \lambda_{\mathcal{R}_i}, \lambda_{\mathcal{P}_i}\})_{1 \leq i \leq n}$.

Un *schéma de réaction* est une équation de réaction incomplète qui permet de représenter un schéma de transformation commun à plusieurs réactions. Le schéma de la figure 2 est un des nombreux schémas généralisant l'équation de la figure 1. Formellement le schéma

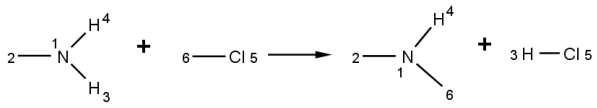

FIG. 2 – *Exemple d'un schéma de réaction généralisant l'équation de la figure 1.*

de réaction $S_1 = (\mathcal{R}_1, \mathcal{P}_1, \lambda_{\mathcal{R}_1}, \lambda_{\mathcal{P}_1})$ généralise l'équation ou le schéma de réaction $S_2 = (\mathcal{R}_2, \mathcal{P}_2, \lambda_{\mathcal{R}_2}, \lambda_{\mathcal{P}_2})$ (et on note $S_1 \subseteq_S S_2$) si les graphes $\mathcal{R}_1$ et $\mathcal{P}_1$ sont isomorphes à des sous-graphes de $\mathcal{R}_2$ et $\mathcal{P}_2$ et si les injections correspondantes de sommets $\theta_R : S(\mathcal{R}_1) \rightarrow S(\mathcal{R}_2)$ et $\theta_P : S(\mathcal{P}_1) \rightarrow S(\mathcal{P}_2)$ sont compatibles avec les relations d'appariement (i.e. $\lambda_{\mathcal{R}_1}(s_1) = \lambda_{\mathcal{P}_1}(s_2) \Leftrightarrow \lambda_{\mathcal{R}_2}(\theta_R(s_1)) = \lambda_{\mathcal{P}_2}(\theta_P(s_2))$). Le *support d'un schéma de réaction* S relativement à un ensemble E de réactions est le nombre de réactions de E généralisées par S. Le *problème de la recherche des schémas de réactions fréquents* dans une base de réactions consiste à déterminer le support de tous les schémas de réactions fréquents dont le support est supérieur ou égal à un seuil f_{min}.

3 L'axiomatisation des connaissances du domaine

Les équations présentes dans les bases de réactions comportent souvent des erreurs. Les chimistes prennent par exemple rarement la peine de décrire tous les appariements entre sommets et certains produits jugés inintéressants sont tout simplement omis de l'équation. Ces négligences sont tolérées dans la mesure où les chimistes n'ont aucune difficulté à réinterpréter correctement les données l'aide des connaissances qu'ils ont du domaine. Dans le cadre d'un processus automatisé d'extraction de connaissance, il devient nécessaire d'identifier les propriétés particulières que présentent les graphes $(\mathcal{R}, \mathcal{P})$ et que le chimiste utilise implicitement. La démarche adoptée consiste à reformuler ces propriétés en axiomes exprimés exclusivement à partir de concepts propres à l'informatique et à la théorie des graphes de manière à ce que

les algorithmes de prétraitement puissent s'en déduire indépendamment des connaissances du domaine :

Conservation des sommets Tout atome se conservant au cours d'une réaction, il existe une bijection ν des sommets du graphe $\mathcal{P}$ vers ceux de $\mathcal{R}$. Cela implique en particulier que les fonctions $\lambda_\mathcal{R}$ et $\lambda_\mathcal{R}$ soient injectives et que $\lambda_\mathcal{R}(s_1) = \lambda_\mathcal{P}(s_2) \Rightarrow \nu(s_2) = s_1$.

Valence des sommets Le nombre de liaisons simples auxquelles un atome stable participe étant défini par l'élément chimique de l'atome, tout sommet s de $\mathcal{R}$ ou $\mathcal{P}$ et d'étiquette $l(s)$ a un degré pondéré $deg(s)$ (i.e. la somme des multiplicités $m(a)$ des arêtes a incidentes à s) égal à l'image de $l(s)$ par une fonction appelée *valence* :

$$deg(s) = valence(l(s)) \tag{1}$$

Réagencement des arêtes Du fait de la propriété de *conservation des sommets*, les réactions consistent uniquement à briser, créer ou changer la multiplicité des liaisons. Le graphe produit $\mathcal{P}$ s'obtient donc du graphe de départ $\mathcal{R}$ en ajoutant à la multiplicité $m(a)$ de chaque arête $a = \{s_1; s_2\}$ un entier $r(a)$ vérifiant (en supposant $m(a) = 0$ si $a \notin A(\mathcal{R})$) :

$$r(a) = m(\{\nu^{-1}(s_1); \nu^{-1}(s_2)\}) - m(a) \tag{2}$$

La propriété de *valence des sommets* implique alors :

$$\forall s \in S(\mathcal{R}), \sum_{s' \in S(\mathcal{R}), s' \neq s} r(\{s; s'\}) = 0 \tag{3}$$

Minimalité de la distance d'édition Une réaction transforme ses réactants en ses produits en suivant statistiquement la séquence (t_j) de transformations élémentaires qui minimise l'énergie thermodynamique nécessaire. Cette énergie est proportionnelle en première approximation à la distance d'édition $d(\mathcal{R}, \mathcal{P}) = \sum c(t_j)$ pour passer de $\mathcal{R}$ à $\mathcal{P}$ en supposant que le coût $c(t_j)$ soit l'énergie nécessaire à la transformation t_j. D'après l'axiome de *réagencement des arêtes*, les transformations élémentaires consistent uniquement à diminuer ou augmenter la multiplicité des arêtes. La distance peut donc se réécrire comme $d(\mathcal{R}, \mathcal{P}) = \sum_{A(\mathcal{R}) \cup A(\mathcal{P})} c(a)$ où le coût $c(a)$ correspond à l'énergie nécessaire pour modifier de $r(a)$ unités, la multiplicité $m(a)$ de l'arête a. Ce coût est nul si $r(a) \geq 0$ puisque la formation d'une liaison libère de l'énergie, et peut être supposé proportionnel au nombre d'arêtes élémentaires brisées lorsque $r(a) < 0$. Finalement :

$$d(\mathcal{R}, \mathcal{P}) = \min_r \left(\sum_{a \in A(\mathcal{R}), r(a) < 0} |r(a)| \right) \tag{4}$$

4 Le processus de fouille des schémas de réactions

La figure 3 présente les différentes étapes du processus de fouille de schémas de réactions à partir de bases de réactions. Ce processus entièrement opérationnel comprend toutes les étapes du processus d'extraction de connaissances tel que décrit par Fayyad et al. (1996) : l'expert commence tout d'abord par sélectionner les réactions qu'il souhaite fouiller dans une base de

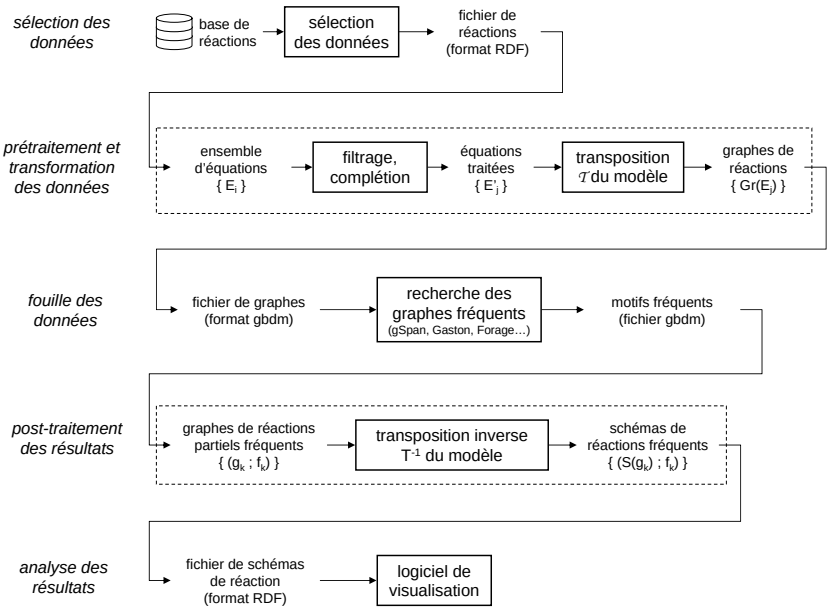

FIG. 3 – *Étapes du processus de fouille des schémas de réactions.*

réactions à l'aide d'un langage de requêtes spécifique puis sauvegarde la réponse de sa requête dans un fichier *Reaction Data File* ou RDF[1]. L'étape de prétraitement développée en section 6, filtre et corrige l'ensemble $\{E_i\}$ des équations de départ en un ensemble $\{E'_j\}$ d'équations totalement appariées. Cet ensemble peut alors être transformé en l'ensemble des graphes de réactions équivalents $\{G_r(E'_j)\}$ dont le modèle est développé en section 5. Les descriptions de ces graphes étiquetés sont sauvegardées dans un fichier au format gbdm afin d'être exploités par un algorithme de fouille de graphes, comme `Gaston` (Nijssen et Kok, 2004) ou `gSpan` (Yan et Han, 2002) pour la recherche des sous-graphes fréquents ou `Forage` (Pennerath et Napoli, 2007) pour l'extraction des schémas de réactions les plus informatifs. Dans tous les cas, l'algorithme produit un fichier gbdm contenant un ensemble $\{(g_k, f_k, ...)\}$ de motifs g_k associés à leur fréquence f_k plus éventuellement d'autres propriétés (score, information, etc). L'étape de post-traitement permet de convertir l'ensemble des graphes de réactions partiels g_k en l'ensemble $\{S(g_k)\}$ des schémas de réactions équivalents, qui sont ensuite triés selon les valeurs décroissantes d'une des propriétés spécifiée par l'expert (par exemple le score ou la fréquence), avant d'être sauvegardés dans un fichier RDF. L'expert peut alors analyser les schémas obtenus et leurs propriétés à l'aide d'un logiciel de visualisation d'équations de réactions.

[1]Ce format de fichier est un des formats les plus répandus pour l'échange de descriptions de réactions. Il n'est aucunement relié au langage *Resource Description Framework* du Web sémantique.

5 La transformation des données : les graphes de réactions

Tout algorithme de recherche de motifs fréquents exploite la propriété de monotonicité de la relation de subsomption entre motifs qui correspond en l'occurrence à la relation d'inclusion $\subseteq_S$. Les méthodes existantes de fouille de graphes sont incapables de s'adapter à cette relation d'inclusion pour deux raisons essentielles : d'une part ces méthodes ne génèrent, pour des raisons de réduction combinatoire, que des motifs de graphes connexes, ce qui n'est pas le cas des schémas de réactions. D'autre part ces méthodes n'intègrent pas naturellement la relation d'ordre $\subseteq_S$ entre schéma de réactions puisque cette relation repose sur la notion étrangère d'appariement entre sommets (i.e. les fonctions $\lambda_{\mathcal{R}}$ et $\lambda_{\mathcal{P}}$). Le modèle des graphes de réactions introduit initialement dans Vladutz (1986) pour produire une représentation connexe d'une réaction, a également l'avantage de résoudre le second problème Pennerath et Napoli (2006) : tirant parti des axiomes de *conservation des sommets* et de *réagencement des arêtes*, le graphe de réaction $G_r(S)$ associé à un schéma de réaction $S = (\mathcal{R}, \mathcal{P}, \lambda_{\mathcal{R}}, \lambda_{\mathcal{P}})$ totalement apparié revient à confondre le graphe des produits $\mathcal{P}$ avec celui des réactants $\mathcal{R}$ en fusionnant les sommets appariés afin d'identifier les arêtes inchangées, détruites et créées lors de la réaction (par des étiquettes d'arêtes 0, $-$ et $+$). Le graphe de réaction de l'équation de la figure 1 est représenté sur la figure 4. Cette transformation $S \mapsto G_r(S)$ est bijective : les

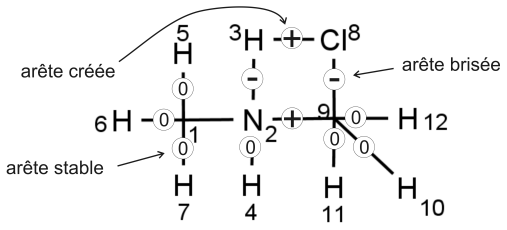

FIG. 4 – *Graphe de réaction équivalent à l'équation totalement appariée de la figure 1.*

graphes $\mathcal{R}$ et $\mathcal{P}$ de S s'obtiennent de $G_r(S)$ en supprimant respectivement les arêtes créées (marquées $+$) et brisées (marquées $-$) puis en remplaçant dans $\mathcal{R}$ et $\mathcal{P}$ tout ensemble A d'arêtes multiples par une seule arête a de multiplicité $m(a) = |A|$. Un graphe de réaction $G_r(S)$ est donc un graphe connexe rigoureusement équivalent à un schéma de réaction S totalement apparié. On démontre que la relation d'ordre $\subseteq_g$ de sous-graphe isomorphe définie sur l'ensemble des graphes de réactions est isomorphe à la relation d'inclusion entre schémas de réactions : $S_1 \subseteq_S S_2 \Leftrightarrow G_r(S_1) \subseteq_g G_r(S_2)$. La figure 5 illustre l'équivalence des deux relations d'ordre sur l'exemple du schéma de réaction de la figure 2 inclus dans l'équation de la figure 1. Le problème de recherche des schémas de réactions fréquents dans un ensemble $(E_i)_{1 \leq i \leq n}$ d'équations de réactions est donc équivalent à celui de la recherche des graphes fréquents dans l'ensemble des graphes de réactions équivalents $(G_r(E_i))_{1 \leq i \leq n}$. Les algorithmes existants de fouille de graphes peuvent donc résoudre le problème de la recherche de schémas de réactions fréquents (du moins si on se restreint aux motifs contenant au moins une arête de type $-$ ou $+$).

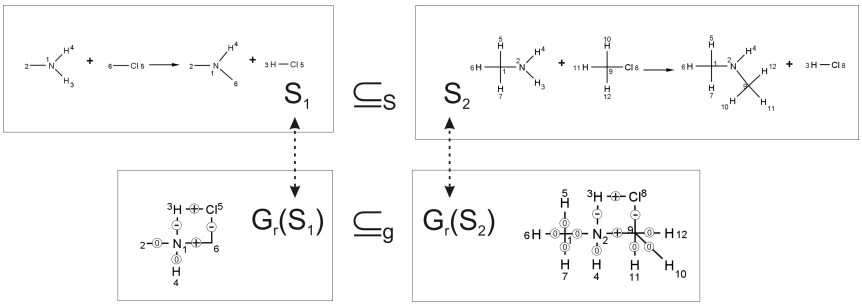

FIG. 5 – *Equivalence des relations d'ordre $\subseteq_S$ et $\subseteq_g$.*

6 Le prétraitement des données

6.1 Les données du domaine et leurs imperfections

Une base de réactions décrit une équation chimique par un 4-uplet $(\mathcal{R}, \mathcal{P}, \lambda_\mathcal{R}, \lambda_\mathcal{P})$. Ce 4-uplet respecte très rarement tous les axiomes de la section 3. Les équations telles que celles des figures 6 et 7 peuvent être qualifiées différemment selon la nature de leurs non conformités (Berasaluce, 2002). Une équation est ainsi :

Non saturée quand les atomes d'hydrogène ne sont pas explicités et que le principe de la valence n'est pas respecté (i.e. quand $\{s \in S(\mathcal{R}) \cup S(\mathcal{P}) | deg(s) < valence(l(s))\} \neq \emptyset$, cf figure 6).

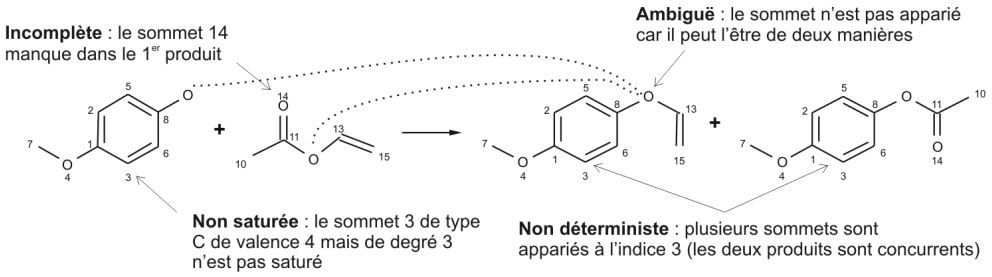

FIG. 6 – *Exemple d'équation non saturée, non déterministe, incomplète et ambiguë.*

Non déterministe quand les produits d'une équation ne sont pas produits simultanément mais concurremment selon des rendements statistiques respectifs (i.e. quand il existe au moins deux sommets s_1 et s_2 de deux composantes connexes distinctes de $\mathcal{P}$ portant le même indice d'appariement $\lambda_\mathcal{P}(s_1) = \lambda_\mathcal{P}(s_2)$, cf figure 6).

Non équilibrée quand certains réactants ou produits doivent être dupliqués pour que le principe de conservation des atomes soit respecté (i.e quand il existe une combinaison linéaire $H_\mathcal{R} \times C_\mathcal{R} = H_\mathcal{P} \times C_\mathcal{P}$ telle que les vecteurs de pondération $C_\mathcal{R}$ et $C_\mathcal{P}$ aient des coefficients strictement positifs non tous égaux et que la matrice $H_\mathcal{R}$ (resp. $H_\mathcal{P}$) ait pour

coefficients h_{ij} les nombres de sommets d'étiquette i dans le j$^{\text{ème}}$ réactant $\mathcal{R}_j$ (resp. j$^{\text{ème}}$ produit $\mathcal{P}_j$)).

Erronée quand l'équation ne peut être équilibrée et que certains éléments chimiques sont plus présents dans les produits que dans les réactants (i.e. quand il existe une étiquette de sommet d'avantage présente dans $\mathcal{P}$ que dans $\mathcal{R}$, cf figure 7).

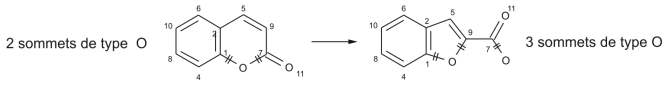

FIG. 7 – *Exemple d'équation erronée.*

Incomplète quand l'équation n'est ni équilibrable ni erronée parce que certains produits secondaires sont omis de l'équation (i.e. quand il existe une étiquette de sommet plus présente dans $\mathcal{R}$ que dans $\mathcal{P}$, cf figure 6)

Ambiguë quand l'équation n'est pas totalement appariée parce que plusieurs appariements sont envisageables (i.e. quand $\mathcal{D}_{\lambda_\mathcal{P}} \neq \mathcal{P}$, cf figure 6).

6.2 Les étapes du prétraitement

Le prétraitement des données se décompose en une succession d'étapes. Chaque étape peut être perçue comme une fonction qui transforme tout 4-uplet qu'elle reçoit en entrée en un ensemble éventuellement vide de 4-uplets. Tout 4-uplet en sortie de e_i devient ensuite un 4-uplet en entrée de e_{i+1}. L'ordre des étapes est défini de telle sorte qu'une étape résout une catégorie particulière de défauts sans jamais introduire un type de défaut résolu par une étape précédente. La succession des étapes garantit que les graphes de réactions obtenus en sortie de la dernière étape correspondent aux équations des réactions les plus plausibles. Les étapes sont dans l'ordre :

La saturation des molécules qui consiste simplement à connecter à chaque sommet s de $\mathcal{R}$ et de $\mathcal{P}$, $valence(l(s)) - deg(s)$ sommets d'hydrogène par des arêtes simples pour satisfaire l'axiome de *valence des sommets*.

La scission des équations non déterministes : si l'équation $(\mathcal{R}, \mathcal{P}, \lambda_\mathcal{R}, \lambda_\mathcal{P})$ est non déterministe, toute composante connexe $\mathcal{P}_k$ de $\mathcal{P}$ dont le rendement associé dépasse un seuil configurable donne lieu à une équation déterministe $(\mathcal{R}, \mathcal{P}_k, \lambda_{\mathcal{R}_k}, \lambda_{\mathcal{P}_k})$ où $\lambda_{\mathcal{R}_k}$ et $\lambda_{\mathcal{P}_k}$ sont les restrictions de $\lambda_\mathcal{R}$ et $\lambda_\mathcal{P}$ sur $\mathcal{P}_k$.

L'élimination des équations erronées : si il existe une étiquette de sommet (i.e. un élément chimique) qui est présente dans $\mathcal{P}$ sans l'être dans $\mathcal{R}$, l'équation candidate est erronée et est éliminée.

La pondération des équations non équilibrées est un problème théorique complexe de programmation linéaire en nombres entiers (Sena et al., 2006). En pratique les équations comptent rarement plus de trois réactants et trois produits. L'étape de pondération se contente donc pour toute équation non équilibrée de tester toutes les duplications évidentes de réactants et/ou de produits jusqu'à obtenir une équation équilibrée (en testant

la condition $H_{\mathcal{R}} \times C_{\mathcal{R}} = H_{\mathcal{P}} \times C_{\mathcal{P}}$). Si toutes les tentatives de pondération échouent, on distingue deux cas : si il existe une étiquette de sommet plus représentée dans $\mathcal{P}$ que dans $\mathcal{R}$, l'équation candidate est considérée erronée et est éliminée. Dans le cas contraire l'équation est incomplète mais peut passer à l'étape suivante.

La complétion des appariements : A ce stade du prétraitement, l'équation obtenue est presque toujours ambiguë. Un rejet systématique des équations ambiguës est donc impossible. La solution adoptée consiste à produire l'ensemble des équations totalement appariées les plus plausibles qui se déduisent de l'équation ambiguë. Cet ensemble contient vraisemblablement l'unique réaction se produisant réellement en plus d'éventuelles équations factices. L'effet néfaste introduit par la présence de ces artefacts est toutefois limité par le filtrage statistique qu'opère la recherche des motifs fréquents en ignorant les schémas de réactions non représentatifs.

Pour désambiguïser une équation, il est nécessaire d'apparier tous les sommets non appariés de $\mathcal{P}$ à des sommets non appariés de $\mathcal{R}$. Compte tenu du nombre exponentiel d'appariements possibles, il est nécessaire de restreindre la procédure aux appariements les plus plausibles. On introduit à cette fin la notion de compatibilité entre sommets : deux sommets $s_1 \in S(\mathcal{P})$ et $s_2 \in S(\mathcal{R})$ sont *compatibles* si aucun des deux n'est déjà apparié, si ils ont la même étiquette, si ils sont tous deux adjacents à un sommet apparié et si ils partagent le même ensemble maximal de sommets voisins appariés : s_1 est incompatible avec s_2 si il existe un sommet $s_3 \in S(\mathcal{R})$ compatible avec s_1 qui a plus de voisins appariés avec des voisins de s_1 que s_2 n'en a avec s_1. En notant $\mathcal{V}(s)$ l'ensemble des sommets voisins du sommet s, cette dernière condition équivaut à :

$$|\lambda_{\mathcal{R}}(\mathcal{V}(s_3) \cap \mathcal{D}_{\lambda_{\mathcal{R}}}) \cap \lambda_{\mathcal{P}}(\mathcal{V}(s_1) \cap \mathcal{D}_{\lambda_{\mathcal{P}}})| > |\lambda_{\mathcal{R}}(\mathcal{V}(s_2) \cap \mathcal{D}_{\lambda_{\mathcal{R}}}) \cap \lambda_{\mathcal{P}}(\mathcal{V}(s_1) \cap \mathcal{D}_{\lambda_{\mathcal{P}}})|$$

L'appariement de deux sommets n'est plausible que si ces derniers sont compatibles, selon le principe de *minimalité de la distance d'édition*. L'élimination des appariements incompatibles permet d'élaguer l'arbre de recherche dans l'espace d'état des appariements possibles. La procédure consiste alors en un algorithme de backtracking qui effectue à chaque étape de sa progression l'appariement d'un sommet de $\mathcal{P}$ non encore apparié avec un sommet de $\mathcal{R}$ qui lui est compatible puis met à jour les fonctions $\lambda_{\mathcal{R}}$ et $\lambda_{\mathcal{P}}$. Un retour arrière se produit soit dès qu'un sommet non apparié de $\mathcal{P}$ n'est plus compatible avec aucun sommet de $\mathcal{R}$ soit dès que tout sommet de $\mathcal{P}$ est apparié, l'équation associée étant alors passée à l'étape de traitement suivante.

La construction du graphe de réaction : A ce stade du prétraitement, toutes les équations sont totalement appariées. Chaque équation E est donc remplacée par son graphe de réaction $G_r(E)$, conformément à la section 5.

L'élimination des graphes de réactions non réalistes : L'étape de complétion des appariements transforme une équation de départ en un nombre limité d'équations E_i totalement appariées qui n'ont pas nécessairement le même degré de plausibilité. De ce fait, le nombre n_i d'arêtes brisées (étiquetées $-$) de chaque graphe de réaction $G_r(E_i)$ est calculé. Tout graphe $G_r(E_i)$ dont le nombre n_i n'est pas égal à $n_{min} = \min(n_i)$ peut alors être éliminé au nom du principe de *minimalité de la distance d'édition*. Après élimination, toute équation initiale E aboutit à un ensemble d'équations totalement appariées de même n_i minimal. Le nombre de ces équations est le plus souvent égal à 1, parfois

nul lorsque E s'avère erronée, parfois égal à 2 ou 3 (mais rarement plus) lorsque E peut valablement être interprétée selon différents appariements de sommets. Lorsque le nombre d'appariements possibles est supérieur à un seuil (fixé à 4), on estime que l'appariement de l'équation initiale est insuffisant et que les équations qui en découlent sont trop incertaines et doivent être éliminées.

La complétion des arêtes créées : Dans le cas d'une réaction incomplète, les sous-graphes présents initialement dans $\mathcal{R}$ mais pas dans $\mathcal{P}$ impliquent la présence de sommets s de $G_r(E)$ tels que le nombre $deg^+(s)$ d'arêtes créées incidentes à s soit strictement inférieur au nombre $deg^-(s)$ d'arêtes brisées incidentes à s alors que ces deux nombres devraient être égaux d'après l'axiome de *valence des sommets*. Dans le cas particulier où seuls deux sommets s_1 et s_2 sont dans ce cas, l'équation 3 implique nécessairement l'égalité des lacunes (i.e. $deg^-(s_1) - deg^+(s_1) = deg^-(s_2) - deg^+(s_2)$). L'ajout de $deg^-(s_1) - deg^+(s_1)$ arêtes de type $+$ entre s_1 et s_2 permet alors de reconstruire a posteriori les produits manquants.

L'équation de la figure 6 est ainsi scindée en deux équations déterministes qui présentent chacune deux appariements plausibles. La figure 8 présente une des équations résultantes.

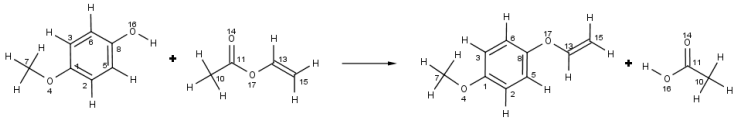

FIG. 8 – *Équation en sortie du prétraitement dérivant de l'équation initiale de la figure 6.*

7 Résultats des tests

L'algorithme de prétraitement a été implémenté et testé sur plus de 95000 réactions monoétapes issues de l'intégralité des deux bases de réactions Orgsyn et JSM[2] particulièrement importantes pour la variété des méthodes de synthèse qu'elles proposent. Les résultats des tests résumés dans le tableau de la figure 9 reposent sur la définition de différents taux indicateurs. Soit ainsi $\mathcal{I}$ l'ensemble initial des équations de la base de données. Soit $\mathcal{E} \subseteq \mathcal{I}$ l'ensemble des

base	taille	taux t_e d'erreur	taux t_c de complétude	taux t_a d'ambiguïté	taux t_r de réussite
Orgsyn	4856	25 %	11 %	10 %	63 %
JSM	91468	11 %	19 %	9 %	72 %
total pondéré	96324	12 %	19 %	9 %	71 %

FIG. 9 – *Résultats du prétraitement sur les bases de réactions Orgsyn et JSM.*

équations erronées. Le *taux d'erreur* des données est défini par $t_e = \frac{|\mathcal{E}|}{|\mathcal{I}|}$. L'ensemble $\mathcal{I} \setminus \mathcal{E}$

[2]Ces bases de réactions sont des produits commerciaux de la société MDL (www.mdli.com) accessibles par les universitaires via le portail titanesciences.inist.fr .

des équations non erronées conduit à un ensemble d'équations déterministes $\mathcal{D}$. Les équations complètes de $\mathcal{D}$ forment un sous-ensemble $\mathcal{C}$. Le *taux de complétude* des données est défini par $t_c = \frac{|\mathcal{C}|}{|\mathcal{D}|}$. La procédure d'appariement appliquée à $\mathcal{D}$ ne réussit que sur un sous-ensemble $\mathcal{A}$ de $\mathcal{D}$. Le *taux de réussite* du prétraitement est défini comme $t_r = \frac{|\mathcal{A}|}{|\mathcal{D}|}$. En fonction de la précision des appariements initiaux dans $\mathcal{I}$, chaque équation de $\mathcal{D}$ alimente l'ensemble résultat $\mathcal{G}$ d'un nombre variable de graphes de réactions. Le *taux d'ambiguïté* des données est défini par $t_a = \frac{|\mathcal{G}|}{|\mathcal{D}|} - 1$. Globalement 92 % des équations correctement converties (i.e. de $\mathcal{A}$) conduisent à un seul graphe de réaction, 7 % à deux, 0,5 % à 3 et 0,3 % à 4. Si le taux de réussite global de 71 % peut paraître relativement faible, un examen approfondi atteste que la plupart des équations rejetées recourent à des représentations non normalisées (i.e. pour mettre en évidence des catalyseurs, des complexes organo-métalliques, etc) présentant des appariements de sommets trop pauvres. L'outil de prétraitement a également permis d'apprécier assez précisément le soin apporté à la spécification des données des bases de réactions. La courbe de la figure 10 représente ainsi l'évolution des indicateurs de qualité des données de la base JSM, divisée arbitrairement en 7 extraits chronologiques de tailles semblables (entre 10 et 20000 échantillons chacun). Si le taux de complétude t_c reste relativement constant, le taux d'erreur t_e observe

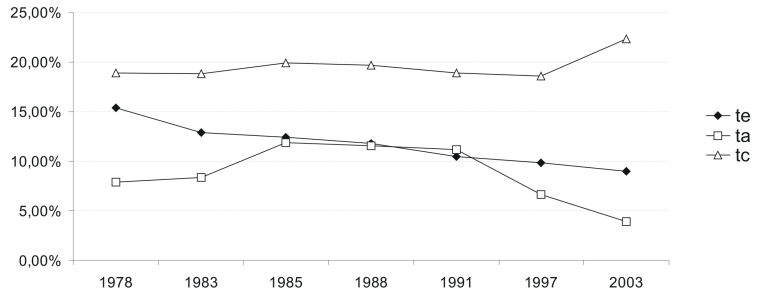

FIG. 10 – *Évolution temporelle des indicateurs de qualité sur la base JSM.*

une diminution régulière, probablement due à une sélection de plus en plus stricte des données. Le taux d'ambiguïté t_a observe une décroissance plus récente, traduisant une qualité croissante des appariements.

8 Conclusions

Les résultats obtenus prouvent qu'il est possible d'extraire les schémas de réactions fréquents des bases de réactions à l'aide des algorithmes existants de fouille de graphes. L'outil `Forage` a ainsi pu extraire de différents ensembles de réactions les schémas de réactions fréquents, fermés fréquents et les plus informatifs fréquents (Pennerath et Napoli, 2007). Si ce résultat devrait à terme permettre aux chimistes d'identifier des schémas de réactions intéressants, il pourrait également inciter les spécialistes de la fouille de données à s'intéresser de plus près à ces objets d'étude riches et complexes que sont les réactions chimiques.

Les auteurs remercient Gilles Niel et Claude Laurenço de l'équipe Système d'Information Chimique de l'ENSC de Montpellier pour l'assistance et l'expertise qu'ils ont apportées.

Références

Berasaluce, S. (2002). *Fouille de données et acquisition de connaissances à partir de bases de données de réactions chimiques*. Thèse de chimie informatique et théorique, Université Henri Poincaré Nancy 1.

Fayyad, U. M., G. Piatetsky-Shapiro, et P. Smyth (1996). From data mining to knowledge discovery : An overview. In *Advances in Knowledge Discovery and Data Mining*, pp. 1–34.

Fischer, I. et T. Meinl (2004). Graph based molecular data mining - an overview. In *IEEE Conference on Systems, Man and Cybernetics*.

Nijssen, S. et J. N. Kok (2004). A quickstart in frequent structure mining can make a difference. In *Proceedings of the tenth ACM SIGKDD conference*, pp. 647–652. ACM Press.

Pennerath, F. et A. Napoli (2006). La fouille de graphes dans les bases de données réactionnelles. In *EGC*, Volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pp. 517–528. Cépaduès-Éditions.

Pennerath, F. et A. Napoli (2007). Mining most informative subgraphs. *Mining and Learning With Graphs 2007 Conference, Firenze*.

Pennerath, F. et A. Napoli (2008). Le problème de l'extraction des graphes d'intérêt maximal. application à la fouille de réactions chimiques. *Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA 2008), Amiens*.

Sena, S., H. Agarwalb, et S. Senc (2006). Chemical equation balancing : An integer programming approach. *Mathematical and Computer Modelling 44*, 678–691.

Vladutz, G. (1986). Do we still need a classification of reactions ? In P. Willet (Ed.), *Modern Approaches to Chemical Reaction Searching*, pp. 202–220. Gower Publishing.

Yan, X. et J. Han (2002). gspan : Graph-based substructure pattern mining. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 721. IEEE Computer Society.

Yan, X. et J. Han (2003). Closegraph : mining closed frequent graph patterns. In *Proceedings of the ninth ACM SIGKDD conference*, New York, NY, USA, pp. 286–295. ACM Press.

Summary

A large number of chemical reactions have been collected in databases. Chemists would like to mine molecular graphs contained in these data in order to extract frequent chemical reaction patterns. Two obstacles then come up: first, the way chemists represent reactions with graphs does not suit the extraction of chemical reaction patterns by graph mining algorithms. Second, databases contain reaction descriptions that are often incomplete, ambiguous or erroneous. The present article describes the preprocessing steps to filter, to complete and then to transform a reaction database into a proper set of abstract graphs fitting the reaction pattern mining problem. The process brings reaction databases into the scope of graph mining methods as shown by experimental results.