La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique

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

[*]Supélec,
Campus de Metz, 2 rue Edouard Belin 57070 Metz
frederic.pennnerath@supelec.fr
[**]Equipe Orpailleur, Loria
Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex
amedeo.napoli@loria.fr

Résumé. La synthèse en chimie organique consiste à concevoir de nouvelles molécules à partir de réactifs et de réactions. Les experts de la synthèse s'appuient sur de très grandes bases de données de réactions qu'ils consultent à travers des procédures d'interrogation standard. Un processus de découverte de nouvelles réactions leur permettrait de mettre au point de nouveaux procédés de synthèse. Cet article présente une modélisation des réactions par des graphes et introduit une méthode de fouille de ces graphes de réaction qui permet de faire émerger des motifs génériques utiles à la prédiction de nouvelles réactions. Enfin l'article fait le point sur l'état actuel de ce travail de recherche en présentant le modèle général dans lequel s'intégrera un nouvel algorithme de fouille de réactions chimiques.

1 Introduction

Le problème auquel s'intéresse cet article est la découverte de nouvelles familles de réactions chimiques à partir de *bases de données de réactions*. Cet article montre en quoi ce problème peut se reformuler en un problème particulier de *fouille de graphes*. La découverte de nouvelles réactions présente un grand intérêt pour la *synthèse* en chimie organique, discipline dont le but est la conception de molécules complexes à partir de composants chimiques usuels et de réactions. En effet, plus un expert de la synthèse a de réactions à sa disposition, plus il peut créer de nouveaux produits à partir d'un ensemble donné de molécules et plus il peut optimiser le plan de synthèse d'une molécule cible donnée. Par ailleurs, la découverte de dizaines de millions de réactions a vite rendu leur recensement nécessaire à travers la constitution de très grandes bases de données de réactions. Ces *bases de données réactionnelles* sont plus particulièrement exploitées par les experts de la *rétrosynthèse*. Cette méthode consiste à inférer le plan de synthèse d'une molécule cible en recherchant les réactions qui permettent d'aboutir à la cible, puis à réitérer récursivement le processus en prenant pour cibles les réactifs des réactions ainsi trouvées et ce jusqu'à l'obtention de réactifs de départ jugés ordinaires. La rétrosynthèse peut donc tirer un excellent parti de tout modèle prédictif capable de propo-

ser des réactions qui n'ont jamais été testées mais qui ont de forte chance d'être réalisables expérimentalement.

Pour établir un tel modèle prédictif qui soit suffisamment fiable, certaines méthodes d'apprentissage automatique ont été appliquées aux bases de données réactionnelles, notamment des méthodes de voisinage symbolique (Régin, 1995). Mais leurs résultats restent limités tant leurs calculs de généralisation (appliqués de surcroît à des graphes) s'avèrent longs , et tant leurs procédures d'induction se révèlent sensibles aux inexactitudes engendrées par la pauvreté des graphes moléculaires en tant que modèle de représentation des réactions. A ce titre l'emploi par Berasaluce et al. (2004) d'une méthode de *recherche de motifs fréquents* (Agrawal et Srikant, 1994) s'est révélé judicieux dans la mesure où de telles méthodes sont à la fois adaptées à de grands volumes de données et robustes aux incohérences partielles des données puisque basées sur des probabilités. La principale faiblesse d'une telle approche est de travailler sur des données booléennes et donc de ne pouvoir réellement prendre en compte la topologie des atomes dans une molécule, pourtant essentielle à la compréhension des réactions.

Les travaux présentés dans cet article se situent dans le prolongement de ceux de Berasaluce et al. (2004). La différence majeure réside dans la prise en compte de la topologie des molécules par le recours à des techniques de *fouille de graphes*, c'est-à-dire de généralisation des méthodes de fouille de données booléennes à des graphes. L'apport principal de cet article est de montrer comment la recherche de nouveaux schémas de réactions à partir de bases de réactions peut se reformuler en un problème particulier de fouille de graphes. Pour se faire, les notions de chimie organique utiles à la compréhension du problème sont introduites (section 2). Les réactions décrites dans les bases de données réactionnelles sont ensuite modélisées sous la forme de graphes de réaction, à partir desquels des schémas de réactions particuliers appelés graphes de réaction partiels sont dérivés (section 3). Après un bref état de l'art des algorithmes de fouille de graphes, le problème d'apprentissage des mécanismes réactionnels est reformulé comme un problème particulier de fouille de graphes de réaction partiels (section 4).

2 Les graphes moléculaires et les schémas de réactions

Une *molécule* est un assemblage géométrique d'atomes solidaires liés par des liaisons de covalence. La *forme développée* ou *graphe moléculaire* est une représentation de la topologie des liaisons d'une molécule sous forme d'un graphe étiqueté $g(V, E, \lambda)$ où les sommets $V(g) = V$ et les arêtes $E(g) = E$ représentent respectivement les atomes et les liaisons de covalence de la molécule et où la fonction d'étiquetage $\lambda : V \rightarrow \mathcal{L}$ associe (au minimum) à un sommet l'élément chimique, tel le carbone (C) ou l'hydrogène (H), de l'atome représenté par ce sommet. Un graphe moléculaire est un graphe particulier en ce sens que les sommets

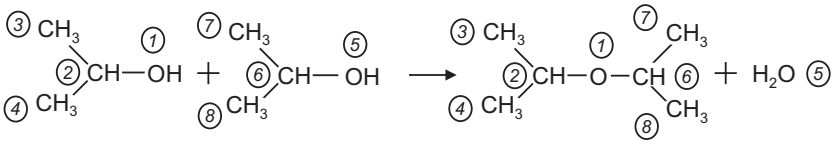


FIG. 1 – *Equation de la réaction de déshydratation du propan-2-ol avec appariement partiel des atomes (numéros encerclés)*

représentant un même élément chimique ont des *degrés* (i.e. le nombre d'arêtes incidentes à un sommet) tous égaux à la *valence* de cet élément (4 pour C, 1 pour H). Une *réaction chimique* est quant à elle un processus physique qui, en chimie organique, brise certaines liaisons de covalence pour en créer de nouvelles, transformant ainsi un ensemble de molécules appelées *réactifs* en un ensemble de nouvelles molécules appelées *produits*. Elle se représente par une *équation chimique*, comme illustrée sur la figure 1, mettant en rapport les formes développées des réactifs et des produits.

Les *schémas de molécules* (resp. les *schémas de réactions*) sont des graphes moléculaires (resp. des équations chimiques) dont certains sommets représentent des variables, remplaçant par une opération dite de *contraction*, un *radical*, c'est-à-dire un groupe d'atomes connectés. De telles variables peuvent être typées auxquels cas leurs ensembles de définition se restreignent à des radicaux d'un type particulier. Certains types de schémas qualifiés dans cet article de *partiels* peuvent de plus, autoriser à ce que certains sommets ne soient pas *saturés*, c'est-à-dire que leur degré puisse être strictement inférieur à la valence de leur élément chimique. Un graphe moléculaire satisfait un schéma partiel s'il contient un sous-graphe qui par une série de contractions (compatibles avec les types des variables du schéma) est isomorphe au schéma. Ainsi le schéma de réactions représenté sur la figure 2 est satisfait par la réac-

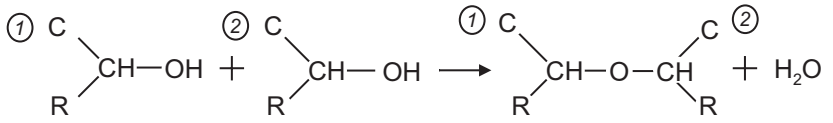


FIG. 2 – *Un schéma de la déshydratation d'un alcool secondaire*

tion de la figure 1, le groupe méthyl CH_3 étant une valeur acceptable pour une variable R de type alkyle, représentant toute chaîne linéaire d'atomes de carbone saturée en hydrogène. Ce schéma est partiel puisque les atomes de carbone (de valence 4) numérotés 1 et 2 ne sont pas incidents à 4 liaisons et sont donc non saturés. Un schéma de molécules (resp. de réactions) permet de représenter une molécule (resp. une réaction) générique instanciée par une *famille de molécules* (resp. une famille de réactions) de la même manière qu'un concept d'un langage de représentation des connaissances est satisfait par ses instances.

3 Une modélisation des réactions chimiques adaptée à l'apprentissage

3.1 Les graphes de réaction

Les bases de données réactionnelles décrivent chaque réaction au minimum par son équation chimique ainsi que par les *conditions réactionnelles* nécessaires à son déclenchement. En général les atomes des produits sont appariés avec ceux des réactifs pour traduire le principe de conservation des atomes au cours d'une réaction chimique. L'appariement de la réaction de la figure 1 est ainsi représenté par des numéros encerclés identifiant chaque atome. Sous cette forme la description des réactions est difficilement exploitable puisque les descriptions associées à la cause, c'est-à-dire aux réactifs, et ceux associées à la transformation, c'est-à-dire aux

produits, pourtant indissociables, ne peuvent être mis en corrélation que par une information totalement étrangère (les appariements) à leur mode de description (les graphes moléculaires). L'introduction d'un *graphe de réaction*, illustré sur la figure 3 (a), permet de rattacher cause et effet de la réaction en un seul objet. Ce graphe de réaction résulte de la superposition des atomes appariés entre les graphes moléculaires des réactifs et des produits. Formellement ce graphe se construit à partir des graphes moléculaires des réactifs auxquels on ajoute les arêtes élémentaires nouvellement crées par la réaction (une liaison multiple étant dissociée en un nombre de liaisons élémentaires égal à sa multiplicité). De plus chaque arête est marquée d'une étiquette précisant s'il s'agit d'une arête inchangée, brisée ou créée. La figure 3 (a) représente le graphe de réaction associé à l'équation chimique de la figure 1. Les arêtes stables, brisées et créées y sont représentées respectivement en trait continu, en pointillés et en en trait épais.

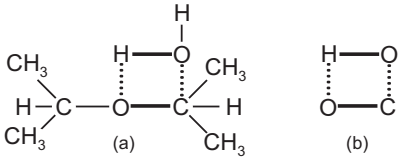


FIG. 3 – *Graphe de réaction (a) et graphe de cœur (b) de l'équation de la figure 1*

A partir d'un graphe de réaction R et des ensembles de ses arêtes inchangées, brisées et créées respectivement notés $E^0(R)$, $E^-(R)$ et $E^+(R)$, peuvent être définis trois sous-graphes présentant un intérêt particulier.

- Le *graphe du cœur* $\mathcal{C}(R) = R \cdot (E^-(R) \cup E^+(R))$ est le sous-graphe de R réduit [1] par l'ensemble de ses arêtes brisées ou créées. Ce graphe représente le *cœur de la réaction*, c'est-à-dire l'ensemble des atomes dont les liaisons de covalence sont modifiées lors de la réaction. Le graphe de cœur du graphe de réaction de la figure 3 (a) est représenté sur la figure 3 (b).
- Le *graphe des réactifs* $\mathcal{R}(R) = R \cdot (E^-(R) \cup E^0(R))$ est le sous-graphe de R réduit par l'ensemble de ses arêtes brisées ou inchangées. Ce graphe représente l'union des graphes moléculaires des réactifs. Le graphe des réactifs associé au graphe de réaction de la figure 3 est identique à la partie gauche de l'équation chimique de la figure 1.
- Le *graphe des produits* $\mathcal{P}(R) = R \cdot (E^0(R) \cup E^+(R))$ est le sous-graphe de R réduit par l'ensemble de ses arêtes inchangées ou créées. Ce graphe représente l'union des graphes moléculaires des produits. Le graphe des produits associé au graphe de réaction de la figure 3 est identique à la partie droite de l'équation chimique de la figure 1.

On démontre que le graphe de cœur est un ensemble connexe de cycles de longueurs paires disjoints par leurs arêtes. Chaque cycle est une suite alternée de liaisons brisées et de liaisons créées. La preuve repose sur une démonstration similaire à celle, demeurée célèbre, qu'Euler apporta au problème des ponts de Konigsberg (Pour plus de détails, on se référera à un manuel de théorie des graphes comme par exemple Gondran et Minoux (1995)). Vu que dans un graphe de cœur un cycle de longueur 0 n'a pas de sens et qu'un cycle de longueur 2 peut être vu comme

[1] La réduction notée $G \cdot E$ d'un graphe G par un ensemble E d'arêtes est le plus petit sous-graphe de G tel que $E(G \cdot E) = E$.

une liaison stable, on peut conclure que les graphes de cœur de réaction sont des systèmes de cycles alternés de longueurs paires supérieures ou égales à 4.

D'un point de vue purement informationnel, le graphe de réaction est rigoureusement équivalent à une équation chimique appariée puisqu'il est possible de passer indifféremment d'un formalisme de représentation à l'autre. Mais les avantages du graphe de réaction sont multiples : outre certains avantages en terme de complexité algorithmique qui sont ici hors sujet, le graphe de réaction permet de représenter naturellement le lien entre la cause et l'effet d'une réaction et ce en un seul objet, via un graphe connexe. Cette association est indispensable pour généraliser les réactions et approcher l'expression des mécanismes réactionnels sous-jacents qui rattache nécessairement les effets à leurs causes. Enfin le graphe de cœur peut servir à réaliser une partition et donc une indexation efficace des réactions d'une base de données réactionnelles.

A ce titre, on introduit ici la notion de *réaction nulle* qui désigne une absence de toute réaction lors de la mise en présence d'un ensemble donné de réactifs dans des conditions expérimentales données. Le graphe de cœur d'une réaction nulle est évidemment le graphe vide (sans sommets). Les graphes de réaction des réactions nulles sont les seuls à ne pas être connexes et dans ce cas uniquement se confondent avec les graphes des réactifs (ou indifféremment les graphes des produits). Les réactions nulles servent d'exemples négatifs supplémentaires utiles pour interdire certaines généralisations irréalistes de réactions. Malheureusement les bases de données réactionnelles ne contiennent pas la description de réactions nulles, ce qui est pour le moins normal vu l'intérêt tout aussi nul qu'elles présentent en synthèse organique. On peut cependant construire des réactions nulles en émettant l'hypothèse que la plupart des réactions d'une base de données réactionnelles forment des produits stables. Le graphe des produits de cette réaction peut alors servir de graphe de réaction pour une nouvelle réaction nulle. C'est pourquoi on suppose désormais qu'une base de données réactionnelle est un ensemble de graphes de réaction indexés par leur graphe de cœur éventuellement nul.

3.2 Les graphes de réaction partiels

L'introduction des graphes de réaction permet de définir un type particulier de schémas de réactions (au sens de celui défini dans la section 2) tenant compte de l'appariement des atomes des réactifs et des produits. Ces schémas appelés *graphes de réaction partiels*, tiendront lieu de motifs dans nos algorithmes de fouille de graphes.

Un graphe de réaction partiel est défini comme un schéma de molécule contenant un et un seul graphe de cœur et dont les variables sont toutes mono-atomiques, c'est-à-dire qu'elles ne peuvent représenter qu'un seul atome. Les types des variables mono-atomiques se confondent alors avec un ensemble $(\mathcal{L}, \sqsubseteq)$ d'*étiquettes généralisées* ordonné selon un ordre de subsomption $\sqsubseteq$: les étiquettes les plus spécifiques sont les éléments chimiques et l'étiquette la plus générale, notée $\top$, remplace tout atome quel qu'il soit. Les étiquettes intermédiaires dans l'ordre induit par $\sqsubseteq$ permettent de regrouper les éléments dont les propriétés chimiques sont similaires, comme par exemple la famille des halogènes dont font notamment partie le chlore et le brome. La figure 4 (a) présente un graphe de réaction partiel qui est satisfait par la réaction de la figure 1.

Ces schémas de réactions sont introduits ici car ils revêtent d'excellentes propriétés vis à vis des algorithmes de fouille de graphes. Ainsi chaque sommet ne peut être apparié qu'à un seul atome ce qui facilite la gestion des appariements entre les graphes partiels et les graphes

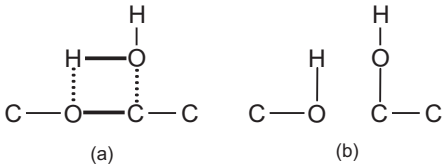


FIG. 4 – *Un graphe de réaction partiel (a) et son réacton (b)*

de réaction de la base de données (et ce contrairement aux radicaux en général, comme les groupes alkyle R). Mais surtout il est facile d'énumérer tous les graphes de réaction partiels possibles. Si on suppose dans un premier temps que l'on sache générer l'ensemble des graphes de cœur possibles, alors l'ensemble des spécialisations d'un graphe partiel de réaction qui permettent d'aboutir aux graphes partiels directement plus spécifiques est rapidement identifiable à l'aide de structures de données appropriées et optimisées qui ne sont pas détaillées ici. En effet seulement trois types de spécialisations sont envisageables :

- L'adjonction d'une nouvelle arête stable entre deux sommets non saturés.
- L'adjonction d'une nouvelle arête stable entre un sommet non saturé et un nouveau sommet étiqueté par $\top$.
- La spécialisation d'une étiquette généralisée en une étiquette immédiatement plus spécifique.

Il est donc possible par une suite de spécialisations de générer tout graphe de réaction partiel à partir de son graphe du cœur. Par ailleurs et compte tenu de la connaissance des mécanismes réactionnels dont les chimistes disposent, la grande majorité des réactions sont le résultat d'une succession de réactions élémentaires dont le graphe de cœur se réduit à un seul cycle. Notre étude peut donc se restreindre à l'étude des réactions élémentaires dont les graphes partiels peuvent être générés à partir de l'ensemble des cycles alternés de longueur paire supérieure ou égale à 4 et dont tous les sommets sont étiquettés par $\top$. Ce dernier ensemble est lui même trivial à générer. La figure 5 représente la suite (incomplète) de spécialisations passant du

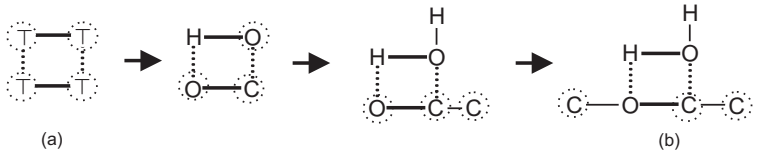


FIG. 5 – *Une suite incomplète de spécialisations passant du graphe de cœur (a) au graphe de réaction partiel (b)*

graphe de cœur (a) au graphe de réaction partiel (b) de la figure 4. Les atomes encerclés de pointillés correspondent aux sommets non saturés capables d'accepter une nouvelle arête.

4 La recherche de réactions génériques fiables vue comme un problème de fouille de graphes

4.1 L'apprentissage des mécanismes réactionnels

Au cœur de chaque réaction élémentaire se trouve un *mécanisme réactionnel*, c'est-à-dire un processus temporel et déterministe de transformation qui brise certaines liaisons de covalence pour en créer d'autres. Les graphes de réaction partiels peuvent servir de modèles de représentation des mécanismes réactionnels. Ce modèle n'est pas exact et induit des erreurs de prédiction lorsqu'il est confronté à une base d'exemples de réactions. Ces erreurs se manifestent par des exemples qui satisfont la cause de la réaction générique sans en satisfaire l'effet. L'effet d'un mécanisme réactionnel est modélisé par le graphe de cœur du graphe partiel, traduisant la redistribution des liaisons de covalence, alors que la cause du mécanisme est modélisée par un graphe que l'on décide d'appeler *réacton*. Le réacton d'un schéma désigne l'ensemble des schémas des réactifs, c'est-à-dire le graphe des réactifs déduit du graphe de réaction partiel. Sur la figure 4 est illustré le réacton (b) associé au graphe partiel (a).

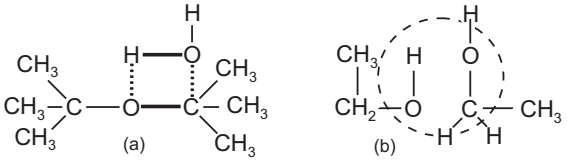


FIG. 6 – *Un exemple positif (a) et un exemple négatif (b) à 110 °C du graphe de réaction partiel de la figure 4*

Plus formellement étant donné un ensemble $\mathcal{R}$ des réactions (y compris des réactions nulles) dans des conditions réactionnelles fixées, on note $g_r(r)$ (resp. $g_{re}(r)$) le graphe de réaction (resp. le graphe des réactifs) d'une réaction r. Etant donné un graphe partiel de réaction s, la réaction r satisfait s si s est un sous-graphe de $g_r(r)$, c'est-à-dire si la cause (l'environnement topologique du cœur) et l'effet (la redistribution des liaisons de covalence) décrits par s se retrouve dans r. A l'inverse une réaction r infirmera s si le réacton $g_{re}(s)$ de s est un sous-graphe du graphe des réactifs $g_{re}(r)$ sans que r satisfasse s, c'est-à-dire si la cause se trouve dans r alors que l'effet ne s'y trouve pas. Un exemple de réaction qui satisfait (resp. qui infirme) un schéma s est dit *positif* (resp. *négatif*) pour s. La figure 6 exhibe un exemple positif (a) et un exemple négatif (b) pour le graphe de réaction partiel de la figure 4. Il est important de noter que ces exemples ne sont valables que dans des conditions réactionnelles bien précises (ici une température supérieure à 110 °C). Si la température devient inférieure à ce seuil de 110 °C, l'expérience montre que la réaction de la figure 1 devient une réaction nulle et passe donc du statut d'exemple positif à celui d'exemple négatif.

Etant donné un ensemble $\{r_i\}_{1\leq i\leq n}$ d'exemples de graphes de réaction éventuellement nulles, il est alors possible de définir la fréquence positive $\mathrm{f}^+(s)$ (resp. la fréquence négative $\mathrm{f}^-(s)$) d'un graphe partiel s comme le nombre d'exemples positifs (resp. négatifs) pour s. Plus la fréquence négative d'un schéma est faible, plus le schéma est *fiable*, plus la fréquence positive d'un schéma est grande, plus le schéma est *général*.

4.2 La fouille de graphes

La majorité des méthodes de fouille de données et d'apprentissage s'applique à des données où chaque objet est décrit par la liste des *attributs* booléens qu'il vérifie. La *recherche des motifs fréquents et d'extraction des règles d'association* (Agrawal et Srikant, 1994) est une de ces méthodes les plus employées. Etant donné une base d'objets décrits par la liste de leurs attributs, choisis dans un ensemble $\mathcal{A}$, l'extraction de motifs fréquents consiste à énumérer dans un premier temps les conjonctions d'attributs, ou *motifs*, qualifiés de *fréquents*, c'est-à-dire dont le nombre d'occurrences (définies par les objets contenant simultanément **tous** les attributs du motif) dans la base de données, encore appelé *fréquence* ou *support*, est supérieur à un seuil fixé arbitrairement. L'extraction des règles d'association consiste dans un deuxième temps à déduire les règles d'association non triviales entre motifs fréquents dont la probabilité conditionnelle, ou *confiance*, est supérieure à un second seuil.

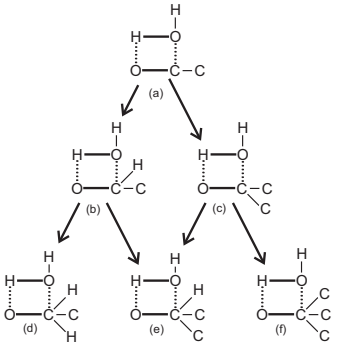
Certains travaux initiés notamment par Inokuchi et al. (2000) et Kuramochi et Karypis (2001) ont depuis permis de transposer le paradigme des motifs fréquents au cas où les données sont des graphes. L'ensemble $(2^{\mathcal{A}}, \subseteq)$ des motifs booléens ordonné par l'inclusion est alors remplacé par l'ensemble $(\mathcal{G}, \leq_{\mathcal{G}})$ des graphes étiquetés ordonné par la relation $\leq_{\mathcal{G}}$ d'inclusion d'un sous-graphe isomorphe[2]. Ce problème d'isomorphisme, inexistant dans le cas des données booléennes, complique sensiblement les algorithmes et substitue à certaines primitives algorithmiques de complexité linéaire d'autres de complexité exponentielle. Les algorithmes de fouille de graphes les plus performants comme *gSpan* (Yan et Han, 2002) ou Gaston (Nijssen et Kok, 2004) sont de types verticaux, c'est-à-dire qu'ils parcourent l'espace de recherche (i.e l'ordre partiel des motifs) en profondeur, en effectuant un retour arrière lorsqu'un motif n'est pas fréquent. Outre sa simplicité, cette recherche en profondeur permet une économie de mémoire suffisante pour stocker les appariements entre le motif courant et les graphes de la base, restreignant ainsi les possibilités d'appariements des motifs plus spécifiques. Les algorithmes de fouille de graphes ont ouvert la voie à la *fouille de molécules* résumée dans Fischer et Meinl (2004) et ont notamment permis d'isoler au sein d'un corpus de molécules des sous-structures fréquentes, responsables d'une activité chimique particulière. Mais ces méthodes n'ont jamais été utilisées pour fouiller des bases de réactions chimiques.

4.3 La fouille de graphes de réaction partiels

Les modèles retenus pour la représentation des mécanismes réactionnels sont les graphes de réaction partiels. Etant donné une base de données réactionnelles, chaque graphe partiel g_r peut être associé à une fréquence positive $\mathrm{f}^+(g_r)$ et une fréquence négative $\mathrm{f}^-(g_r)$. Ces deux fréquences sont des fonctions décroissantes par rapport à l'ordre partiel $(\mathcal{G}_r, \leq_{\mathcal{G}})$ des graphes de réaction partiels (i.e. $g_1 \leq_{\mathcal{G}} g_2 \Rightarrow \mathrm{f}^{+/-}(g_1) \geq \mathrm{f}^{+/-}(g_2)$). Un graphe partiel est d'autant plus pertinent que sa fréquence positive est élevée et que sa fréquence négative est faible, comme l'illustre la figure 7 élaborée dans le prolongement des exemples précédents. Dans cet exemple, la base de réactions est censée contenir N_1, N_2 et N_3 réactions mettant respectivement en jeu des alcools primaires, secondaires et tertiaires (i.e des molécules dont l'atome de carbone C portant le groupe alcool OH est relié respectivement à un, deux et trois autres atomes de carbone). Les réactions mettant en jeu des alcools primaires sont nulles alors

[2]Deux graphes sont isomorphes s'il existe une bijection entre les ensembles de leurs sommets et une bijection entre les ensembles de leurs arêtes qui toutes deux préservent les relations d'incidence arête-sommet ainsi que les étiquettes.

que les alcools secondaires et tertiaires réagissent selon le coeur de la figure 4 (b). En outre la base de réaction est supposée représenter de manière homogène les différents types d'alcool. N_1, N_2 et N_3 sont donc supposés tous égaux à N, de manière à pouvoir calculer simplement les fréquences des graphes de réaction partiels de la figure 7 (a), regroupées dans le tableau (b). Il apparaît ainsi que le graphe (c) est plus pertinent que le graphe (a) du fait d'une fréquence négative plus grande pour une même fréquence positive et que le graphe (e) du fait d'une fréquence positive plus faible pour une même fréquence négative.



(a) Graphes de réaction partiels

Graphe partiel	Fréquence positive	Fréquence négative	Fiabilité	Différence
(a)	$N_2 + N_3 = 2 \cdot N$	$N_1 = N$	$2/3$	N
(b)	$N_2 = N$	$N_1 = N$	$1/2$	0
(c)	$N_2 + N_3 = 2 \cdot N$	0	1	$2 \cdot N$
(d)	0	$N_1 = N$	0	$-N$
(e)	$N_2 = N$	0	1	N
(f)	$N_3 = N$	0	1	N

(b) Fréquences associées

FIG. 7 – *Graphes de réaction partiels ordonnés et leurs fréquences*

Différentes familles de graphes partiels peuvent être étudiées, chacune accordant une importance différente aux fréquences positives et négatives. En raison de sa simplicité, on s'intéresse ici à la famille $\mathcal{G}(s^+)$ des graphes partiels dont la fréquence négative est minimale tout en ayant une fréquence positive supérieure ou égale à un seuil s^+ fixé arbitrairement. Ainsi l'exemple de la figure 7 permet d'établir que $\mathcal{G}(2 \cdot N) = \{c\}$ et $\mathcal{G}(N) = \{c, e, f\}$ en supposant par ailleurs que tous les graphes non représentés contenant les graphes (c), (e) et (f) ont une fréquence positive qui leur est strictement inférieure (et donc, pour être totalement exact, que les graphes partiels représentés regroupent l'ensemble des graphes de leur fermé).

Pour calculer facilement une approximation $\widehat{\mathcal{G}}(s^+)$ de cet ensemble, on choisit dans le cas tangent où un graphe g_1 est inclus dans un autre graphe g_2 et que ces deux graphes ont les mêmes fréquences négatives, de ne conserver dans $\widehat{\mathcal{G}}(s^+)$ que le graphe le plus spécifique g_2. Ce choix certes arbitraire, parfois même contraire au choix du graphe le plus pertinent, permet lorsqu'il est combiné au caractère décroissant de la fréquence négative, d'approximer l'ensemble $\mathcal{G}(s^+)$ par la frontière positive de s^+ c'est-à-dire par l'ensemble des graphes partiels fréquents (positivement par rapport à s^+) maximaux, soit l'ensemble $\widehat{\mathcal{G}}(s^+) = \{g \,/\, \mathrm{f}^+(g) \geq s^+ \textit{ et } \forall g' g' \geq_{\mathcal{G}} g \Rightarrow \mathrm{f}^+(g') < s^+\}$. Dans le cas de l'exemple, $\widehat{\mathcal{G}}(2 \cdot N)$ égale bien $\mathcal{G}(2 \cdot N)$ mais $\widehat{\mathcal{G}}(N) = \{e, f\}$ omet l'élément c pourtant le plus pertinent. Les inexactitudes introduites par cette approximation sont toutefois compensées par la possibilité de traiter le problème comme celui d'une recherche de graphes fréquents maximaux et de le résoudre en tant que tel grâce à certains algorithmes adaptés comme Spin de Huan et al. (2004). Deux modifications doivent malgré tout être apportées à un algorithme vertical de fouille de graphes pour qu'il soit adapté à notre formulation du problème :

- D'une part la génération des motifs par les algorithmes de fouille doit être modifiée afin de construire les motifs, c'est-à-dire les graphes de réaction partiels, à partir d'un graphe de cœur, c'est-à-dire un cycle alterné de liaisons brisées et créées. Ceci est nécessaire afin de garantir la présence unique et entière d'un graphe de cœur dans chaque motif. Cette modification est facile à intégrer au sein des algorithmes de type verticaux dont les motifs croissent à partir d'un motif initial, en général égal au motif vide mais qui peut être initialisé à une autre valeur.
- D'autre part la phase de détermination des fréquences par le balayage de la base de graphes doit être modifiée pour être capable de calculer deux fréquences. En particulier la fréquence négative doit se calculer en testant dans les graphes de la base l'inclusion non du motif mais celle d'un motif secondaire (le réacton) dérivé du motif initial.

Il est donc possible d'adapter certains algorithmes existants de fouille de graphes pour qu'ils extraient l'ensemble des graphes de réaction partiels maximaux fréquents (positivement). Cet ensemble de résultats peut ensuite être soumis à un expert en chimie afin qu'il analyse l'exactitude et l'originalité des schémas de réactions découverts.

4.4 Discussion

Le choix arbitraire de l'ensemble de graphes partiels maximaux fréquents comme ensemble d'étude mérite une analyse critique. On propose ici une formalisation abstraite du problème général de l'apprentissage de mécanismes réactionnels afin de mieux situer la pertinence de la méthode proposée au paragraphe 4.3. Le constat initial pour une telle formalisation est qu'un graphe de réaction partiel représente d'autant mieux un mécanisme réactionnel que sa fréquence positive est grande et que sa fréquence négative est faible. Ces deux fréquences étant toutes deux des fonctions décroissantes par rapport à $\leq_{\mathcal{G}}$, la recherche des graphes partiels les plus représentatifs n'a de sens que si on se donne un critère d'optimalité capable de pondérer l'importance accordée à la fréquence positive relativement à la fréquence négative. Ce critère peut se définir formellement à partir d'un ensemble totalement ordonné $(\mathbb{E}, \leq)$ et d'une fonction $c : \mathbb{R}^2 \rightarrow \mathbb{E}$ qui associe aux couples des fréquences $(\mathrm{f}^+(g), \mathrm{f}^-(g))$ d'un graphe partiel g un élément de $\mathbb{E}$. La seule contrainte imposée est que c soit une fonction croissante (resp. décroissante) de f^+ (resp. de f^-). Entre deux graphes comparables par $\leq_{\mathcal{G}}$, le graphe dont le critère c est le plus petit peut ainsi être éliminé. Les graphes partiels résistant à cette élimination, c'est à dire les maxima locaux de c, sont sans nul doute les plus pertinents au sens de c. Cet ensemble peut alors être mis sous la forme d'une liste $\mathcal{L}$ triée selon les valeurs décroissantes de leur image par c, c'est-à-dire par ordre d'intérêt décroissant. Seuls les motifs dont le critère c est supérieur à un seuil arbitraire c_0 sont conservés dans $\mathcal{L}$. La liste $\mathcal{L}$ constitue l'ensemble des graphes de réaction partiels les plus pertinents relativement à c et à la base de données réactionnelles considérée.

Le choix de c est arbitraire et ouvre de nombreuses possibilités pour qualifier différemment l'ensemble $\mathcal{L}$ des résultats. L'ensemble $\mathcal{G}(s^+)$ présenté au paragraphe 4.3 peut ainsi être formalisé par l'ensemble $\mathcal{L}$ associé à un ensemble $\mathbb{E} = \mathbb{R}^+ \times \mathbb{R}^-$ muni de l'ordre lexicographique et par le critère $c : (f^+, f^-) \mapsto (min(f^+, s^+), -f^-)$. Le caractère alambiqué de l'expression de c paramétré de surcroît par un seuil arbitraire s^+ tend à prouver le manque de pertinence du choix de $\mathcal{G}(s^+)$. Cet ensemble a été présenté dans cet article uniquement parce que certains algorithmes existants de fouille de graphes fréquents permettent de l'estimer directement. Nous envisageons plutôt de travailler avec d'autres critères c, comme celui que nous

appelons *fiabilité*, défini par $c(f^+, f^-) = \frac{f^+}{f^+ + f^-}$ compris entre 0 et 1. Dans la mesure où la base de réactions couvre de manière homogène l'ensemble des phénomènes physiques qu'elle est censée décrire, la fiabilité constitue une approximation de la *confiance* (c'est-à-dire de la probabilité conditionnelle) associée au schéma de la réaction, interprété dans ce cas comme une règle de transformation reliant une hypothèse à une conclusion. En ce sens la fiabilité est un critère plus pertinent qui par ailleurs croit bien avec f^+ et décroît avec f^-. Nous exigeons de plus que la fréquence f^+ reste supérieure à un seuil s^+ de manière à ne retenir que les schémas suffisamment généraux (en particulier en écartant comme schémas les graphes de réaction de la base dont la fiabilité est évidemment égale à son maximum 1 mais dont la fréquence positive valant aussi 1 est trop faible) et surtout de manière à limiter l'espace de recherche infini à un sous domaine fini (contrainte qui s'avère indispensable pour éviter toute récursion infinie de la part des algorithmes de fouille verticaux). Un autre critère envisageable est la différence $c(f^+, f^-) = f^+ - f^-$. Le tableau (b) de la figure 7 évalue pour chaque graphe les deux critères de fiabilité et de différence. Les maxima locaux de fiabilité sont (c), (e) et (f). Il est à noter que la fiabilité ne peut prendre en compte la généralité d'un schéma et favoriser ainsi le graphe partiel (c) vis à vis des deux autres. Pour pallier ce défaut, il faudrait rajouter la fréquence positive comme critère secondaire à la manière du double critère du paragraphe 4.3. Le critère de différence permet de tenir compte très simplement de la généralité : il donne bien le graphe (c) comme unique maximum. A terme d'autres critères inspirés de la théorie de l'information permettront d'optimiser le compromis fiabilité - généralité.

La méthode d'optimisation de tels critères c, que nous pensons intéressante, est incompatible avec les algorithmes de fouille de graphes actuels. La recherche de maxima locaux de fonctions c non monotones dans $(\mathcal{G}_r, \leq_{\mathcal{G}})$ vont à l'encontre du principe de ces algorithmes qui parcourent l'ensemble des motifs selon un ordre prédéterminé en évitant toutes comparaisons avec les motifs voisins. Nous envisageons donc un nouveau type d'algorithme adapté à la formalisation du critère c qui ne peut être détaillé ici par manque de place.

5 Conclusions

Le problème de la recherche de réactions génériques fiables peut grâce à une modélisation adéquate se reformuler en un problème de fouille de graphes. Ce problème s'apparente à celui de la recherche de motifs qui sont à la fois fréquents dans une base et *non fréquents* dans une autre, avec cette originalité qu'un motif s'exprime différemment dans chacune des deux bases. Toutefois des particularités induites tant par la connaissance des mécanismes réactionnels que par notre propre modélisation du problème nous poussent à entrevoir une méthode générale de fouille de graphes pour laquelle les algorithmes existants sont inadaptés. Un nouvel algorithme inspiré de cette modélisation est en cours de développement et devrait à terme valider les idées introduites dans cet article. De nombreuses expériences sur des bases de données réactionnelles seront alors possibles et permettront notamment d'étudier l'influence du critère d'optimisation c. La méthode introduite n'étant pas spécifique à la chimie, il sera également possible de l'appliquer à d'autres types de données modélisables sous forme de graphes. Cette perspective motivante ne doit cependant pas faire oublier que les performances de calcul et plus encore la pertinence des résultats aux yeux des chimistes restent deux inconnues majeures du problème que seule l'expérimentation pourra lever.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th Conference on Very Large Data Bases (VLDB-94)*, pp. 478–499.

Berasaluce, S., C. Laurenço, A. Napoli, et G. Niel (2004). An experiment on knowledge discovery in chemical databases. In J.-F. Boulicaut, F. Esposito, F. Giannotti, et D. Pedreschi (Eds.), *PKDD*, Volume 3202 of *Lecture Notes in Computer Science*, pp. 39–51. Springer.

Fischer, I. et T. Meinl (2004). Graph based molecular data mining - an overview. In M. P. Wil Thissen, Peter Wieringa et M. Ludema (Eds.), *IEEE Conference on Systems, Man and Cybernetics*.

Gondran, M. et M. Minoux (1995). *Graphes et algorithmes*. Eyrolles.

Huan, J., W. Wang, J. Prins, et J. Yang (2004). Spin : mining maximal frequent subgraphs from graph databases. In W. Kim, R. Kohavi, J. Gehrke, et W. DuMouchel (Eds.), *KDD*, pp. 581–586. ACM.

Inokuchi, A., T. Washio, et H. Motoda (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *PKDD '00 : Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, pp. 13–23. Springer-Verlag.

Kuramochi, M. et G. Karypis (2001). Frequent subgraph discovery. In *ICDM '01 : Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 313–320.

Nijssen, S. et J. N. Kok (2004). A quickstart in frequent structure mining can make a difference. In *KDD '04 : Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 647–652. ACM Press.

Régin, J.-C. (1995). Développement d'outils algorithmiques pour l'intelligence artificielle. Application à la chimie organique. Thèse de l'Université des Sciences et Techniques du Languedoc, Montpellier.

Yan, X. et J. Han (2002). gspan : Graph-based substructure pattern mining. In *ICDM '02 : Proceedings of the 2002 IEEE International Conference on Data Mining*, Washington, DC, USA, pp. 721. IEEE Computer Society.

Summary

Synthesis in organic chemistry consists in designing new molecules from reactants and reactions. Synthesis experts use very large databases of chemical reactions through standard querying procedures. A discovery process of new reactions would be very useful to set up new synthesis processes. This paper presents a model of chemical reactions based on graphs and introduces a method for mining those reaction graphs, that makes generic patterns emerge and can be used for discovering new reactions. Finally this article sums up the progress of this research work by presenting a general model in which a new reaction mining algorithm will fit.