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