

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.pennerath@supélec.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-