Fouille de données et acquisition de connaissances à partir de bases de données de réactions chimiques

Sandra Berasaluce

UFR Sciences et Techniques de la Matière et des Procédés
Département de formation doctorale en Chimie Informatique et Théorique
École doctorale SESAMES

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré — Nancy I

en Chimie Informatique et Théorique

par **Sandra Berasaluce**

Fouille de données et acquisition de connaissances à partir de bases de données de réactions chimiques.

Soutenue publiquement le 20 Décembre 2002

Membres du jury :

Rapporteurs :	Jean-François Boulicaut	Maître de conférences, INSA de Lyon
	Alain Krief	Professeur, Université Notre-Dame de la Paix de Namur, Belgique
Examinateurs :	Roland Ducournau	Professeur, Université Montpellier II
	Claude Laurenço	Directeur de Recherche CNRS, ENSC Montpellier (Co-directeur de thèse)
	Amedeo Napoli	Chargé de Recherche CNRS, LORIA, Nancy (Co-directeur de thèse)
	Jean-Louis Rivail	Professeur, Université Henri Poincaré, Nancy I
Invité :	Geneviève Cassanas	Professeur, Université Montpellier I

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Université de Nancy I (Université Henri Poincaré — Nancy I)
Laboratoire LOrrain de Recherche en Informatique et ses Applications (LORIA)

Référence
BERASALUCE, Sandra *« Fouille de données et acquisition de connaissances à partir de bases de données de réactions chimiques. »* Thèse de doctorat, Université de Nancy I, 20 Décembre 2002

Merci ...

Je ne dérogerai pas à la règle que je suppose exacte (d'après l'examen de ma propre base de données) :

`thèse` $\Rightarrow$ `remerciements`

Un écrivain québécois a dit très justement : « Le meilleur ami de *merci* est *beaucoup* ». Pour moi le mot *beaucoup* a un double sens ici. *Beaucoup* car il y a énormément de personnes que je désire remercier, mais aussi *beaucoup* car toutes ces personnes ont contribué bien plus qu'elles ne le croient à l'accomplissement de ce travail.

Dans ces remerciements, je me permets de faire une analogie avec la chimie. Certaines personnes se retrouveront classées plusieurs fois (peut-être parce qu'ils sont inclassables ...).

Je n'ai pas besoin de faire un grand travail de fouille pour extraire les personnes qui m'ont accompagnée, encouragée et entourée tout au long de ces dernières années. Par contre, il se peut que, vu le grand nombre de données à prendre en compte, j'ai oublié de citer certaines personnes ici. Qu'elles me pardonnent et sachent que, si j'ai omis de les mentionner, c'est à cause de l'intervention humaine de l'analyste.

J'ai passé ces trois dernières années au sein d'un groupe de recherche, assimilable à un *complexe de coordination* dont l'élément central, le *métal de transition*, est Claude Laurenço. À lui seul il arrive à faire l'interface et à maintenir la cohésion d'un groupe de *ligands* relativement hétérogènes mais qui fait la richesse de l'ensemble complexe. J'ai énormément apprécié les discussions, autant avec les *informaticiens chimistes* (Amedeo Napoli, Roland Ducournau, Joel Quinqueton, Philippe Vismara, Yolande Ahronovitz, Pierre Jambau, Yannic Tognetti et plus récemment Philippe Jauffret), qu'avec les *chimistes informaticiens* (Claude Laurenço, Jacques Coste, Gilles Niel, Régine Sola).

C'est dans le cadre de mon DEA que j'ai pour la première fois entendu parler des travaux effectués dans le cadre du GDR. Ceci à Nancy où j'ai été accueillie, ainsi que mes compagnons de la promotion de l'Université Paul Sabatier, pour suivre les cours d'algorithmique prodigués par Jean Lieber. Je classerai les personnes que j'ai rencontrées au cours de cette période et de ma thèse d'*éléments rares*. Rares d'une part car je ne les ai malheureusement vues que de façon très épisodique, mais également rares du fait de leurs qualités d'écoute et de réponse ainsi que de gentillesse. Parmi celles-ci, je désire signaler les membres de l'équipe Orpailleur (Amedeo Napoli, Jean Lieber, Hacène Cherfi, Florence Le Ber, Yannick Toussaint, Sandy Maumus, Rin Al-Hulou, Emmanuel Nauer,

Mathieu D'Aquin, Jean-Luc Metzger, Rafik Taouil, Antoinette Courrier ...), ainsi que les membres du LCTN (Jean-Louis Rivail qui m'a fait l'honneur de présider mon jury de thèse, Christophe Chipot, Walid Harb ...).

En parlant de ce jury de thèse, je ne pourrai omettre de citer ceux qui ont accepté de juger mon travail en tant que rapporteurs de ma thèse. Il n'a pas été facile de trouver des personnes susceptibles de le faire. C'est pourquoi je suis reconnaissante envers Alain Krief et Jean-françois Boulicaut de m'avoir fait bénéficier de leur expertise dans leurs domaines respectifs, la synthèse organique pour l'un, la fouille de données pour l'autre.

Durant la période de mon DEA passée à Toulouse, j'ai également rencontré des personnes d'exception. Je pense en particulier à Romuald Poteau, mon responsable de stage de DEA, qui a su laisser l'électron libre que j'étais choisir sa propre orbitale.

Depuis maintenant quelques mois, j'ai intégré en tant qu'ATER un nouveau laboratoire à la faculté de pharmacie de Montpellier. Dans cette unité d'enseignement et de recherche dirigée de main de maître par Geneviève Cassanas, j'ai été accueillie de façon très spontanée par une équipe dynamique. Je voudrais donc remercier ici Geneviève, Nadia, Abdel, Patrice, Guilhem, Christelle, Robert et Sylvie pour toute l'attention qu'ils me portent et tout ce que j'apprends à leur contact.

Ces trois dernières années, j'ai passé la majeure partie de mon temps au LIRMM, laboratoire dont Michel Habib est devenu le directeur. Dans ce laboratoire, il y a des personnes en tous points exceptionnelles. Parmi celles-ci, je voudrais, en particulier, citer Thérèse Libourel et Isabelle Mougenot à qui je dois, en autres, mes premiers pas dans l'enseignement supérieur. Bien entendu, je désire également saluer mes valeureux camarades de thèse. Notamment ceux en compagnie de qui j'ai rédigé : Yannic que j'ai harcelé (et que j'harcèle toujours) de mes questions incessantes ainsi que Gilles et Jérome, jeunes docteurs et futurs maîtres de conférence (j'espère). Je n'oublierai pas non plus les autres docteurs, thésards et ingénieurs informaticiens qui ont su accepter parmi eux une "chimiste" : Patrice (Pat le chat), Sèverine (Boulette), LeUrEnT (à ne pas confondre avec mon Laurent !), Stéphane (Bouley ! mais futur docteur), David (Davy), Gabriel (Le Gab), Vincent (le barbu), Hervé (Groovy forever), Fabien (le p'tit jeune qui s'la pèt' mais qui va l'avoir sa thèse s'il continue comme ça !), Christophe (Chris), Alexis(s), Momo, Toufik, Fabien(s), Sylvain, Rémi, Florent, Jean-François, Anne-Lise ...

Enfin, dans la classe des *éléments* qui me sont le plus *précieux*, je placerai ma famille, mes amis et celui qui m'attend depuis tant d'années. Qu'ils soient remerciés de leurs encouragements constants et de leur patience inaltérable.

... beaucoup

Table des matières

Table des figures

Liste des tableaux

Première partie

Les systèmes d'information chimique

Chapitre 1

Pourquoi ?

The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.

Vannevar Bush, 1945

Dès les années 30, V. Bush, un conseiller scientifique de Roosvelt, se rend compte de l'importance du traitement et de la gestion de l'information. Nommé directeur de l'Office du Développement et de la Recherche Scientifique des États Unis en 1941, il est conscient de la difficulté du suivi des avancées scientifiques même pour un spécialiste du domaine. En 1945, il expose dans un article désormais célèbre, intitulé « As we may think » [Bush, 1945], le concept du Memex[1]. Alors que l'informatique n'en est qu'à ses balbutiements, l'idée de ce visionnaire est celle d'un dispositif automatique permettant un accès rapide et souple à des données stockés au sein d'une grande masse de documents. Outre l'indexation classique par codes et l'organisation hiérarchique, V. Bush préconise la création et l'utilisation d'associations entre les documents. Les liens établis entre deux parcelles d'information permettent de naviguer entre les documents par des cheminements dirigés par l'analogie. Bien que le Memex n'ait jamais été implanté, il présente des concepts, toujours valides, qui sont les ancêtres de la « navigation hypertextuelle », notion gouvernant l'utilisation du World Wide Web. Dans son article, V. Bush donne des exemples de domaine dans lesquels ce type de système serait particulièrement utile. La chimie est un de ces domaines et la citation donnée en début de ce chapitre reproduit ses propos.

V. Bush pose les bases d'un système d'information idéal à partir duquel on aurait accès à toutes les informations connues et répertoriées de par le monde ayant trait à un domaine. Des dizaines d'années se sont écoulées depuis, l'informatique a fait d'immenses progrès mais le problème des systèmes d'information est toujours actuel.

[1]Memex : contraction de memory extender

Qu'est-ce qu'un système d'information ? C'est essentiellement un système qui doit rendre des services répondant aux besoins de ses utilisateurs. Les services escomptés peuvent être de nature très variée, allant de la recherche documentaire à la résolution de problèmes complexes. Un système d'information doit forcément disposer de bases de données pour accéder à des faits. Cependant pour résoudre des problèmes, il faut des connaissances et des mécanismes de raisonnement pour les exploiter. Ces connaissances permettent de faire les généralisations, abstractions et analogies nécessaires à la compréhension et à la résolution du problème. Un système d'information doit donc combiner les fonctionnalités des systèmes de gestion de bases de données à celles des systèmes basés sur la connaissance. Les systèmes d'information chimique sont des systèmes particuliers dans le sens où ils doivent prendre en compte le langage spécifique au domaine de la chimie : celui des formules structurales. Les données à gérer ne sont donc pas uniquement de type textuel, il faut être capable de manipuler des données de type structural c'est-à-dire généralement représentées par des graphes.

Les besoins en information chimique sont aussi divers que la variété de données produites dans ce domaine largement expérimental. Depuis 1907, les Chemical Abstracts Services répertorient les résultats publiés dans le domaine de la chimie. De nos jours[2], plus de 20 millions de substances dont 80% sont de nature organique ont été indexées dans le CAS Registry Number. Un très grand nombre de réactions a également été décrit. Toutes ces données relèvent du domaine public.

Bien que différents, les besoins des industriels sont également multiples et soulèvent plusieurs difficultés dont celle de la confidentialité. Comme dans de nombreuses sociétés, se pose le problème de la mémoire d'entreprise. Il est nécessaire de conserver la trace des molécules qui ont été préparées mais aussi de celles qui n'ont pas pu l'être. Une grande quantité de données décrivant les propriétés des molécules et les caractéristiques des réactions sont récoltées quotidiennement et doivent être sauvegardées. Enfin, il existe un volume important de brevets et documents administratifs ou juridiques à gérer.

Dans les laboratoires de recherche, aussi bien industriels qu'universitaires, la tenue de cahiers de laboratoire est d'une importance capitale pour le suivi des expérimentations. Pour faciliter l'élaboration et la mise à jour de tels recueils de données, des solutions informatisées appelées « cahiers de laboratoires électroniques » doivent être envisagées sérieusement, si cela n'est pas déjà fait.

Parmi les initiatives visant la gestion de l'information chimique, nous pouvons citer celle du CNRS qui s'est engagé dans la création d'une « chimiothèque nationale » dans le but de répertorier et de faire connaître à des personnes autorisées les disponibilités pour des tests biologiques des diverses molécules obtenues par les chercheurs français. La mise en place d'un tel système centralisateur des données récoltées au cours des travaux des équipes de recherche n'est pas simple et nécessite le choix d'un système informatique adapté aux besoins du stockage d'une grande quantité de molécules et des données associées.

Dans le cadre de l'éducation et de la formation en chimie, l'informatique présente des outils didactiques de transmission de la connaissance qui devraient prendre une place non

[2]source datant d'octobre 2002

négligeable à côté des enseignements classiques [Lehn, 1996]. Des projets tels qu'ENCORE visent de tels objectifs. ENCORE est un projet d'encyclopédie multimédia orientée vers l'enseignement de la chimie organique initié par le Professeur A. Krief. Le but d'un tel système est de proposer une navigation souple à travers un ensemble de documents structurés écrits par des spécialistes d'un domaine particulier de la chimie (par exemple : réactivité des composés organo-séléniés, influence du solvant sur la réaction de substitution, etc.).

Pour notre part, notre but n'est pas de concevoir un système d'information sur tout le domaine de la synthèse organique. Nous nous intéressons particulièrement aux problèmes d'information que rencontrent les industriels. Au sein des sociétés pharmaceutiques, les responsables du développement chimique ont à produire des molécules données dans des délais de temps très courts et avec de fortes contraintes d'efficacité et de sécurité. On comprend dans cette perspective toute l'importance de l'accès à l'information sur le problème à traiter, c'est-à-dire sur la molécule à synthétiser. Notre objectif est donc d'aider les chimistes à utiliser au mieux la connaissance existante.

Contributions et présentation du mémoire

C'est à partir de l'expression de tels besoins qu'a été créé le GDR 1093 du CNRS au sein duquel j'ai effectué ma thèse. Ce groupement de recherche intitulé « Traitement Informatique de la Connaissance en Chimie Organique » a été fondé dans le but de concevoir des systèmes d'information pouvant être intégrés dans l'environnement de travail des chimistes pratiquant la synthèse multiétapes ciblée, en s'appuyant sur l'expertise de spécialistes reconnus du domaine et sur des techniques informatiques avancées. De nature interdisciplinaire, les motivations de mon projet de recherche étaient d'étendre le modèle de connaissances élaboré au cours de travaux effectués précédemment dans le cadre du GDR TICCO. Ceci en vue de prendre en compte les outils indispensables de la synthèse organique, à savoir les méthodes de synthèse. L'objectif visé était de profiter de l'énorme quantité d'informations présentes dans les bases de données de réactions pour enrichir les bases de connaissances d'un prototype d'aide à la compréhension de problèmes de synthèse appelé RESYN ASSISTANT. Ma contribution a donc consisté à modéliser les objets complexes que sont les méthodes de synthèse, tant du point de vue conceptuel qu'informatique, et à utiliser des techniques d'extraction automatique de connaissances à partir de grands volumes de données.

Comme tous les travaux effectués dans le cadre du GDR TICCO, ce mémoire s'adresse au moins à deux communautés scientifiques. Pour celle des chimistes organiciens, il présente une modélisation possible du domaine de la synthèse avec pour leitmotiv la résolution de problèmes de synthèse et la mise en évidence de l'utilité de l'outil informatique. Pour celle des informaticiens, l'application de méthodes éprouvées à un nouveau domaine d'application ouvre des perspectives intéressantes et offre de nouveaux sujets de réflexion.

Pour expliquer la teneur de mes travaux de thèse, j'ai articulé mon mémoire en trois parties. La première a pour but de présenter le problème de la synthèse organique ciblée et des systèmes d'information existants. Dans la deuxième partie, j'ai essayé de faire un rapide exposé de l'acquisition de connaissances préalable à la construction de tout système basé sur

la connaissance. Dans la troisième partie, sont développés les détails de la modélisation du domaine ainsi que l'expérimentation d'extraction de connaissances effectuée dans les bases de données de réactions. Les concepts modélisés ont été implémentés dans le prototype ReSyn Assistant et les nombreux tests ont pu être menés à partir de cette plate-forme.

Le but recherché dans ce mémoire est que tout lecteur saisisse les notions du domaine dont il n'est pas spécialiste. En conséquence, les parties traitant plus particulièrement d'un des deux domaines n'ont pas la prétention de présenter un état de l'art exhaustif de la discipline considérée mais une vue de ce qu'il est nécessaire de connaître pour comprendre la teneur de mes travaux.

Chapitre 2

La résolution de problèmes de synthèse chimique

If there is any key in planning a synthesis, it is to work the problem backwards.

Robert E. Ireland,1969

Sommaire

2.1 Problème de synthèse chimique

2.1.1 Introduction

Synthétiser des molécules est une des principales tâches des chimistes organiciens. La synthèse d'une molécule est l'obtention de la molécule à partir d'autres molécules disponibles (commercialement ou facilement synthétisables), qui sont combinées entre elles

au moyen de réactions. L'enchaînement des réactions employées constitue une voie ou un chemin de synthèse. Une molécule peut être obtenue selon différents chemins de synthèse.

La conception d'une nouvelle synthèse peut être comparée à celle de la construction d'un édifice [Lehn, 1996]. Les briques sont les molécules et les outils sont des réactions. Cependant, il y a entre ces deux problèmes des différences fondamentales telles que le fait que les briques moléculaires sont transformées lors l'application des réactions. Le chimiste est donc un architecte [Hanessian, 1983], un architecte de l'infiniment petit, qui se doit de posséder un esprit très créatif [Lehn, 1998]. En cela la synthèse est souvent comparée à un art.

Durant les deux derniers siècles, les progrès continuels ont fait de la synthèse organique un domaine de recherche riche en développements aussi bien théoriques que pratiques. Les chimistes ont entrepris la synthèse de molécules de plus en plus complexes et doivent développer des trésors d'ingéniosité pour résoudre les nouveaux problèmes auxquels ils sont confrontés. L'essor de l'industrie chimique a profité mais aussi participé à cet effort. Il ne s'agit plus de déterminer un chemin de synthèse pour une molécule donnée mais plutôt la meilleure voie qui permet d'atteindre le but fixé. Selon de nombreux auteurs, « la synthèse idéale » est la voie de synthèse au cours de laquelle la molécule serait préparée à partir des produits de départ disponibles par un procédé simple, en une seule étape, non dangereux, respectueux de l'environnement, rapide et d'un bon rendement. Dans [Wender and Miller, 1993], Wender et Miller mettent en évidence les besoins de notre société en terme d'efficacité, de sécurité et d'écologie devant des problèmes réels de production. Bien entendu cette vision de la synthèse est utopiste mais met en avant les critères qui vont orienter la conception. Comme Corey le signale dans [Corey and Cheng, 1989], les problèmes de synthèse sont devenus d'une telle complexité que l'élaboration de synthèses multiétapes est devenue incontournable. Pour lui, l'efficacité d'une synthèse réside dans la spécificité de chacune des étapes, c'est-à-dire l'obtention d'un seul produit par étape.

Il n'existe pas d'algorithme permettant de résoudre un problème de synthèse. L'alternative est de s'appuyer sur des faits et des modèles développés : la résolution de problèmes fait largement appel à l'analogie. Du point de vue de la chimie théorique, les calculs demandent encore beaucoup trop de temps par rapport à la puissance actuelle des ordinateurs. Même si cette dernière continue de s'accroître en suivant la loi de Moore[1], il faudra attendre deux siècles avant que l'analyse d'une synthèse soit réalisée dans un temps correct (celui d'une pause café !) d'après [Goodman, 2000]. Cependant, la chimie théorique peut être employée pour valider ou réfuter certaines hypothèses grâce à l'étude d'étapes individuelles, évitant ainsi le passage par des expérimentations coûteuses. Les théories de la réactivité chimique conduisent à la prévision de la réactivité à partir de propriétés moléculaires statiques ou dynamiques ou à l'analyse de la réactivité des systèmes π à partir de la détermination des interactions entre orbitales moléculaires π ou par la théorie des orbitales frontières [Rivail, 1994].

[1] La loi de Moore dit que la puissance des ordinateurs double tous les deux ans.

2.1.2 Les différents énoncés

Un problème de synthèse peut s'énoncer de différentes façons selon le nombre d'inconnues du problème. Si nous posons que R est un ensemble de molécules réagissant, P est un ensemble de molécules produites et C les conditions dans lesquelles se font les réactions, nous pouvons formuler les trois questions assez générales suivantes :

1. $? \xrightarrow{?} P$
2. $R \xrightarrow{?} P$
3. $R \xrightarrow{C} P$

On peut envisager encore d'autres formulations de problèmes de synthèse. Le premier cas est le plus général. On ne connaît a priori que la molécule à synthétiser (P) et il faut déterminer à la fois les produits de départ et l'enchaînement des méthodes de synthèse[2] permettant de construire pas à pas la molécule à synthétiser dite « molécule cible » ou « cible ». Dans le deuxième cas, on connaît déjà les molécules de départ (R) et on veut connaître la séquence de méthodes qui permettront d'obtenir le produit souhaité. Pour le dernier problème, les données sont plus complètes que dans les deux cas précédents. Il s'agit de problèmes d'ordonnancement des étapes de la synthèse ou d'optimisation des conditions.

Nous nous situons dans le cas le plus général où seule la molécule à synthétiser est connue. Une voie ou chemin de synthèse, rendant compte du passage de molécules disponibles, « les produits de départ », à la molécule cible, est une solution effective du problème, c'est-à-dire validée par l'expérimentation. Si au départ le problème se réduit à cette formulation, au cours de la résolution d'un problème de synthèse, nous serons confrontés à l'ensemble des différentes classes de problèmes énoncés, et ceci de façon itérative. En effet, si l'on définit les molécules à employer pour synthétiser la cible, le problème se ramènera au deuxième cas de problème évoqué. Une fois les méthodes de synthèse déterminées, il sera nécessaire d'ordonner les étapes et de préciser les conditions réactionnelles ce qui correspond au troisième type de problème.

2.1.3 Les buts et le contexte de la synthèse

Comprendre un problème de synthèse c'est tout d'abord en comprendre les motivations et le cadre. La résolution d'un problème de synthèse n'est pas uniquement dirigée par le but de préparer à tout prix un produit chimique. De nombreuses contraintes, de différents ordres, viennent s'ajouter et influencer la conception de la synthèse. Il s'agit d'un niveau politique qui impose implicitement ou explicitement des règles et des critères de choix. Ainsi, l'enjeu d'une synthèse développée dans le cadre académique ne sera pas le même que celui d'une synthèse menée dans le milieu industriel. Si l'innovation et le développement de nouvelles méthodes sont privilégiés dans le premier cas, des critères de coût, de temps, de facilité de mise en place, de respect de l'environnement et de la sécurité seront primordiaux

[2]Nous reviendrons sur la notion de méthode de synthèse de nombreuses fois, notamment dans la section 6.2.2 où nous en donnerons la définition que nous avons adoptée en la distinguant de la notion de réaction.

dans le second. Bien entendu, les choses ne sont pas figées et, en milieu industriel, une nouvelle méthode ne sera pas forcément rejetée si elle se révèle plus efficace que d'anciennes bien maîtrisées et qu'elle ne coûte pas trop cher à mettre en place. De plus, l'intérêt des industriels est de breveter de nouveaux chemins de synthèse utilisant le plus souvent des méthodes connues : la nouveauté réside dans le nouveau chemin mis en évidence. De même, les universitaires doivent prendre en compte des données telles que le rendement, une méthode de synthèse ne pouvant être retenue en tant que telle uniquement à cause de son originalité, son efficacité est importante. Depuis les premières synthèses, les problèmes auxquels sont confrontés les chimistes ont changé. Les méthodes de résolution de problèmes employées ont évolué en conséquence.

2.2 La planification de synthèse : une étape cruciale

De façon générale, un problème se résout en franchissant les étapes successives [Polya, 1965] que l'on peut voir sur la figure 2.1. Bien entendu, dans le cas d'un problème difficile, cette séquence n'est pas simplement linéaire mais, la plupart du temps, des allers et retours sont nécessaires pour parvenir à une solution meilleure. Il est nécessaire de revoir le problème à chaque étape de la résolution.

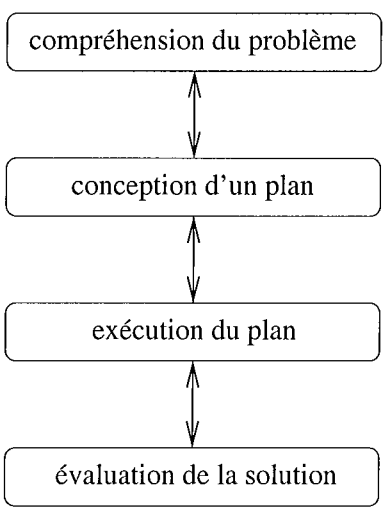


FIG. 2.1 – Les étapes de la résolution d'un problème.

La première étape est celle par laquelle toute résolution de problèmes devrait débuter [Kahney, 1986]. Comprendre un problème, c'est, comme nous l'avons décrit pour un problème de synthèse dans la section précédente, en connaître les diverses motivations et l'environnement, c'est-à-dire les conditions. Mais c'est aussi en étudier les caractéristiques de façon plus formelle en terme d'espace d'états[3] par exemple, savoir si le problème auquel

[3]Cette notion sera définie dans la section suivante 2.2. Bien qu'il existe d'autres dénominations telles que *espace de recherche* ou *graphe d'états* pour cette notion, nous emploierons dans ce mémoire uniquement le terme d'espace d'états.

on a affaire est un problème bien défini ou non en identifiant quelles sont les données et quelles sont les inconnues, de quel type de problème il s'agit (conception[4], diagnostic[5], etc.). Tous ces éléments concourent à une meilleure appréhension du problème qui doit se traduire par une représentation adéquate du problème. Dans un problème de synthèse tel que nous l'avons énoncé section 2.1.2, la seule véritable donnée est la molécule à synthétiser tandis que les inconnues sont les ensembles de molécules de départ permettant d'obtenir la cible. Bien entendu, le chimiste organicien choisit les produits de départ parmi la grande quantité de molécules chimiques disponibles (commercialement) et peut soit sélectionner des méthodes de synthèse connues, soit en inventer de nouvelles.

Concevoir un plan, c'est tout d'abord avoir une idée du plan. Cette idée de plan peut être très vague dans les premiers temps et s'affine petit à petit au cours de la conception. La planification est une activité qui nécessite des connaissances, de l'expérience mais aussi de la créativité. Finalement, un plan détaille les étapes permettant de construire la molécule cible et représente donc la stratégie de la synthèse élaborée [Deslongchamps, 1984].

Exécuter le plan, c'est expérimenter la solution hypothétique élaborée pour la confronter à la réalité. Ceci est d'autant plus vrai dans un domaine tel que celui de la chimie où l'expérimentation est une phase essentielle mais redoutable. En effet, celle-ci peut valider les hypothèses émises, dans le meilleur des cas, ou réfuter le plan dans sa totalité. C'est une vérification sans appel qui peut impliquer une révision de tout ou partie du plan. L'exécution du plan permet également de préciser des parties du plan en ce sens qu'elle est l'occasion en particulier de spécifier ou de préciser les conditions réactionnelles.

Enfin, évaluer la solution, c'est critiquer les résultats obtenus par le plan qui a été conçu et expérimenté. Cette phase peut être à l'origine d'une amélioration de la solution si par hasard une nouvelle idée survient (par exemple pour raccourcir une des étapes de la synthèse). Elle est également l'occasion de faire une certaine acquisition d'expérience de ce qui a été fait et comment cela a été fait, c'est à dire les stratégies et les tactiques de résolution employées. Ceci permet d'établir des liens avec des problèmes déjà rencontrés et aussi d'envisager la résolution de nouveaux problèmes analogues en appliquant les principes développés.

La résolution d'un problème de synthèse n'échappe pas à ce schéma de résolution de problèmes : l'importance de l'élaboration du plan est primordiale car l'expérimentation est une opération coûteuse en temps aussi bien que sur le plan économique. De plus, il est entendu que de la qualité du plan élaboré découle celle du chemin de synthèse obtenu. Couplée à l'étape de compréhension du problème, elle correspond à la phase conceptuelle de la résolution du problème. C'est à cette phase que nous nous intéressons.

[4]Un problème de conception est un problème pour lequel on doit déterminer les objets et les outils à employer pour construire un objet, ceci de façon conceptuelle. La planification de synthèse est un problème de conception.

[5]Un problème de diagnostic implique de trouver une pathologie à partir d'une liste de symptômes. Un médecin détermine la maladie dont est atteint un client par une tâche de diagnostic. De même, pour un mécanicien face à la panne d'une machine.

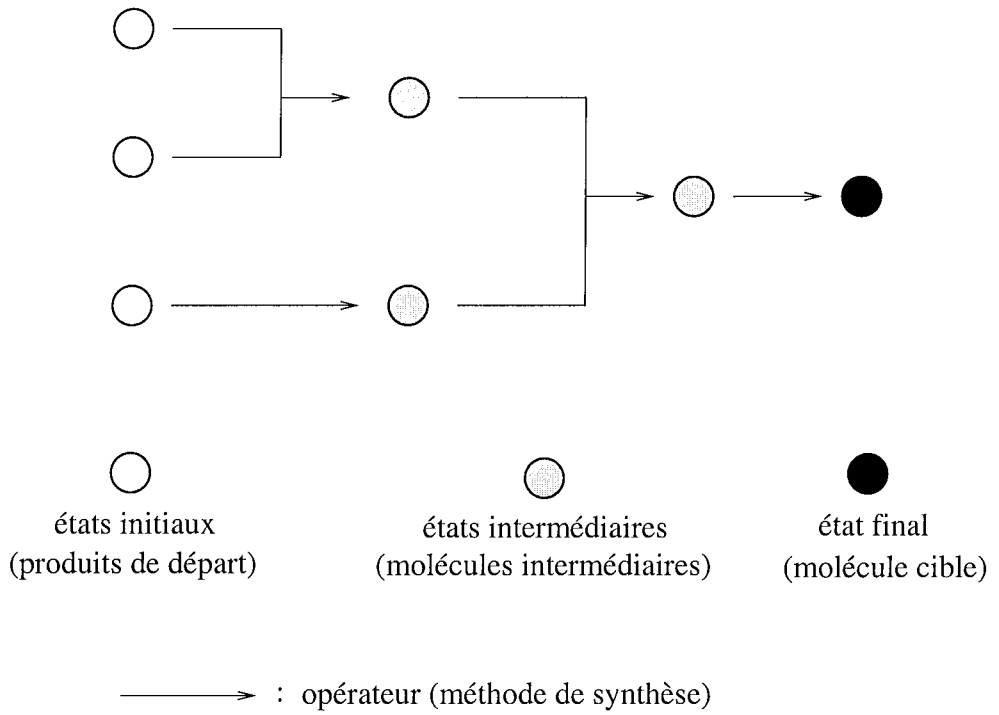


FIG. 2.2 – La représentation générale d'un plan de synthèse.

Plan de synthèse

Dans la typologie de Greeno [Greeno, 1978], un problème de synthèse est un problème de transformation puisque le but est de trouver la ou les séquences de réactions susceptibles de combiner des produits de départ afin d'obtenir la molécule cible.

De façon formelle, ce type de problème est défini par un triplet {I,O,B} où I est l'ensemble des états initiaux, O l'ensemble des opérateurs permettant le passage entre deux états différents et B l'ensemble des états buts. Dans le cas d'un problème de synthèse, I est l'ensemble des ensembles de molécules de départ pertinents, B se résume à la molécule à synthétiser et O est l'ensemble des méthodes de synthèse décrivant le passage d'un ensemble de molécules à un autre ensemble de molécules. L'espace d'états d'un problème se définit comme l'ensemble de tous les états possibles par application systématique d'opérateurs permettant le passage d'un état à un autre état. Dans le cas d'un problème de synthèse, l'espace d'états représente l'infinité des combinaisons possibles entre des millions de molécules par des milliers de méthodes génériques. On ne possède pas de règles permettant de déterminer I a priori : la combinatoire de ce type de problème est très grande. Les solutions du problème sont trouvées par un parcours de l'espace d'états et peuvent être représentées par des plans de synthèse (voir figure 2.2).

Les qualités d'un plan de synthèse

Il est désormais convenu que les principales caractéristiques d'un « bon » plan de synthèse sont :

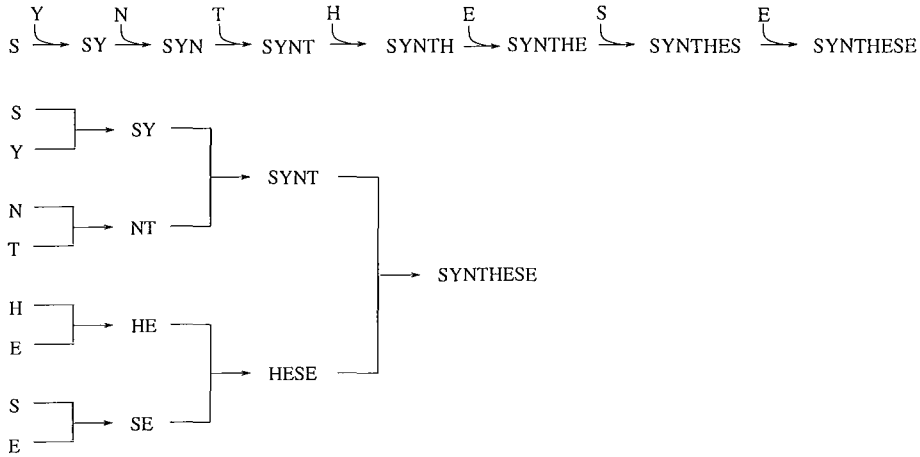


FIG. 2.3 – Comparaison synthèse linéaire/synthèse convergente. Le plan de synthèse linéaire a une profondeur de 7 alors que la solution convergente présente une structure d'arbre bien équilibré d'une profondeur de 3. Supposons que chacune des étapes se fasse avec un rendement de 90%. Pour la synthèse linéaire, le rendement global sera d'au plus 47% tandis que pour la synthèse convergente, il sera de 72%.

- la simplicité,
- la minimisation du nombre de réactions de protection/déprotection[6] grâce au contrôle de la compatibilité des réactifs,
- la flexibilité en indiquant les possibilités de solutions alternatives par rapport à des étapes considérées comme incertaines

Il est nécessaire de préciser ici qu'un plan de synthèse se doit d'être aussi détaillé que possible, en indiquant les conditions opératoires lorsqu'elles peuvent être prévues.

Nous pouvons ajouter à ces critères celui de la convergence. En effet, une synthèse convergente sera préférable à une version linéaire (figure 2.3). Ceci peut être expliqué par un simple calcul du rendement global de la synthèse connaissant les rendements partiels de chacune des étapes [Velluz et al., 1971]. C'est l'application du principe général : « les solutions les plus courtes sont les meilleures ».

Le plan de synthèse de la molécule de 5-keto-9-methoxy-I :2 :3 :4 :4a :5 :IIa :IIb-octahydrochrysofluorene [Chatterjee et al., 1957] est un exemple de plan efficace (figure 2.4). La profondeur est de trois et la structure du plan est bien équilibrée ce qui est la caractéristique d'une synthèse convergente. La représentation de ce plan est simplifiée et ne fait pas apparaître les conditions expérimentales nécessaires.

D'après la terminologie établie par Simon et Newell [Newell and Simon, 1972], un problème de synthèse tel que nous l'avons posé (section 2.1.2) est un problème mal défini car on n'en connaît que l'état final. De plus, l'espace d'états de ce problème tend vers une taille

[6]Les réactions de protection servent à protéger des parties de la molécule qui seraient attaquées dans les conditions réactionnelles envisagées. On protège ces parties en les remplaçant par des structures moins réactives. Une réaction de déprotection permet de récupérer la partie telle qu'elle était avant d'être protégée.

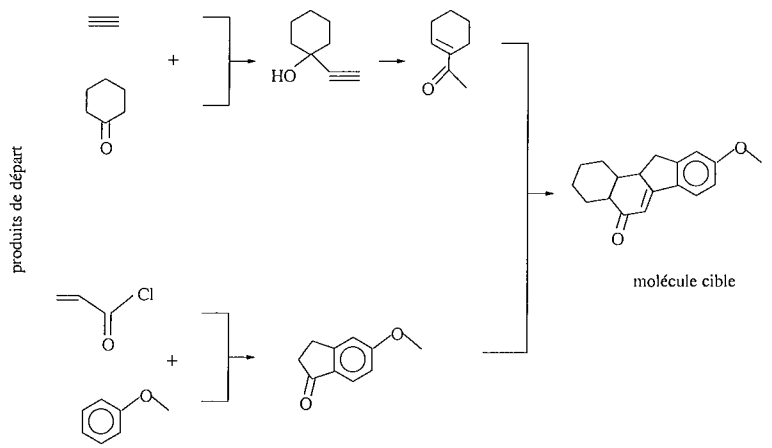


FIG. 2.4 – Un exemple de plan de synthèse convergent.

exponentielle étant donné l'énorme quantité de produits de départ disponibles ainsi que le nombre de méthodes de synthèse potentiellement applicables. En effet, le chimiste a le choix parmi les méthodes connues mais peut aussi développer pour l'occasion une nouvelle méthode.

2.3 Les raisonnements face à un problème de synthèse chimique

Nous venons de voir l'importance de la conception d'un plan de synthèse dans la résolution d'un problème de synthèse. Voyons maintenant par quels processus le chimiste résout un problème de synthèse. Comment parvient-il à faire un choix, le bon choix, parmi toutes les possibilités qui s'offrent à lui ? Quels sont les raisonnements qui dirigent la conception du plan de synthèse de la molécule cible ? Voici ce que nous nous proposons de traiter dans cette section. Ce ne sera qu'un aperçu des principaux raisonnements employés par les chimistes, et nous reviendrons tout au long de ce mémoire sur ces modèles de résolution de problèmes. Bien qu'il soit difficile d'identifier la démarche suivie par les chimistes lors de la résolution de problèmes de synthèse, nous pouvons mettre en évidence deux types de raisonnement très généraux : l'analogie et la classification. Le raisonnement rétrosynthétique, quant à lui, propose un cadre formel dans lequel le problème est résolu de manière analytique.

2.3.1 L'analogie

Qui n'a jamais résolu un problème en s'appuyant sur la solution connue d'un autre problème jugé similaire ? Nous faisons constamment des analogies face aux diverses situations auxquelles nous sommes confrontés, de la plus banale à la plus exigeante en terme de réflexion [Polya, 1965].

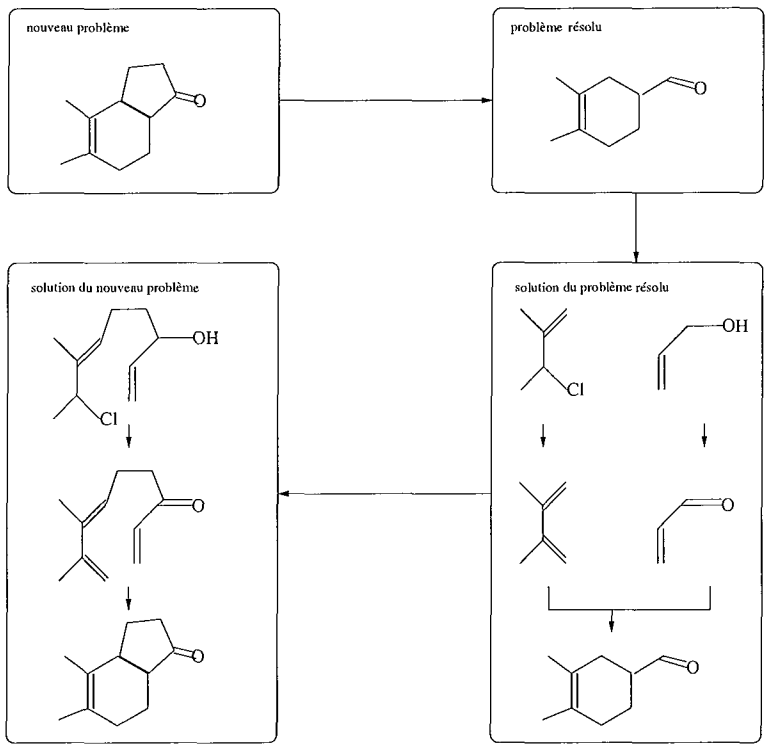


FIG. 2.5 – Le raisonnement par analogie en synthèse.

Les chimistes emploient intensivement ce type de raisonnement et sont en permanence à la recherche d'un problème analogue au nouveau problème qui leur est posé. Ils espèrent ainsi pouvoir réutiliser ou adapter une stratégie qui a déjà prouvé son efficacité. La figure 2.5 donne un exemple de résolution d'un problème de synthèse par analogie. Une fois un problème analogue connu trouvé, la remémoration puis l'adaptation de la solution du problème déjà résolu conduisent à la construction d'une solution pour le problème courant.

Pour pouvoir retrouver les informations sur les problèmes précédents, les chimistes développent des représentations mentales des problèmes résolus dans leur mémoire à long terme. Généralement [Kahney, 1986], de telles représentations se font à divers niveaux d'abstraction pour permettre la recherche, ceci grâce à des processus mentaux tels que la suppression, l'agrégation, la généralisation ou la construction.

Le schéma de la figure 2.6 montre quelques unes des différentes représentations d'une méthode de synthèse, l'acylation d'un cycle benzénique par la méthode de Friedels et Crafts. Le niveau d'abstraction est croissant du niveau 1, où les instances de la méthode sont représentées, au niveau 4 qui donne le bilan général de la réaction de substitution. Le niveau 2 correspond au schéma général de l'acylation d'un cycle benzénique alors que le niveau 3, directement supérieur, décrit la substitution électrophile sur un noyau aromatique. D'autres niveaux intermédiaires peuvent être envisagés. Au premier niveau, une instance de la méthode de synthèse, c'est-à-dire une réaction, est décrite tandis que les autres niveaux correspondent à la description de classes de réactions.

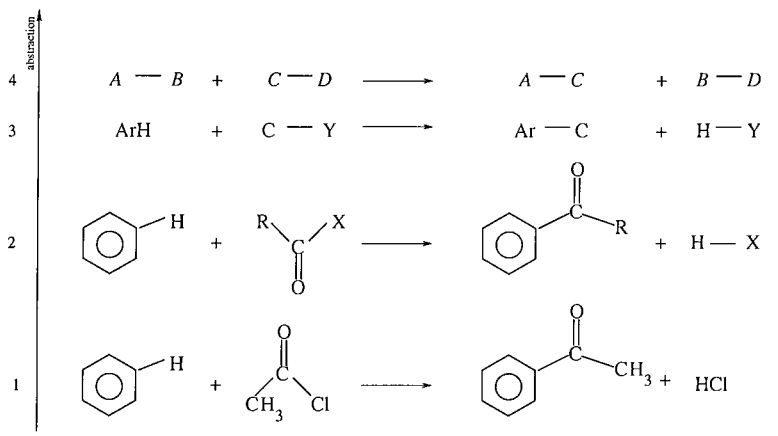


FIG. 2.6 – La représentation d'une réaction à divers niveaux d'abstraction.

On remarque que les différentes descriptions pour une même méthode dépendent aussi des points de vue, c'est-à-dire de l'angle selon lequel on décrit la méthode. En chimie, il y a trois types de points de vue qui sont généralement très utilisés : la topologie (structure planaire), la fonctionnalité (réactivité) et la stéréochimie (chiralité).

2.3.2 La classification

La classification intervient dans de nombreux mécanismes de résolution de problèmes [Rich and Knight, 1991]. Le procédé de classification, que nous employons dans nos tâches quotidiennes, nous permet d'adapter notre comportement vis à vis des catégories d'objets que nous savons identifier. Ce comportement varie pour une même catégorie d'objets selon l'environnement, c'est-à-dire le contexte dans lequel l'entité est observée [Napoli, 1992].

La classification est présente partout en chimie. De très nombreuses classifications de tous types existent et traitent les éléments les plus simples, les atomes avec, par exemple, le tableau périodique de Mendeleev, et les objets les plus complexes, les méthodes de synthèse, en passant par les molécules. La classification des molécules consiste à les regrouper selon certaines caractéristiques principalement structurales. Ensuite, il est ainsi possible de déterminer à quelles familles de composés la molécule étudiée appartient. Déterminer la catégorie d'une molécule permet d'en connaître les principales propriétés, d'avoir une idée de la réactivité ainsi que les grandes lignes des modes de préparation de celle-ci. Par exemple (figure 2.7), le Cholestérol est un stéroïde qui fait également partie de la famille des alcools, et plus particulièrement des stérols, et de celle des alcènes, et plus précisément des cycloalcènes. Cette reconnaissance conduit à la mise en évidence de méthodes connues pour préparer chacune des classes de composés. Ce procédé permet donc de savoir comment l'on pourrait construire certaines parties caractéristiques de la molécule. On comprend donc l'intérêt particulier que revêt la classification dans le contexte de la synthèse. Les familles de composés sont organisées en hiérarchie (par exemple, les stérols sont des stéroïdes et des alcools). Plus on reconnaît l'appartenance à une famille spécifique, plus on a de chance

de retrouver une méthode de synthèse adaptée à la molécule considérée.

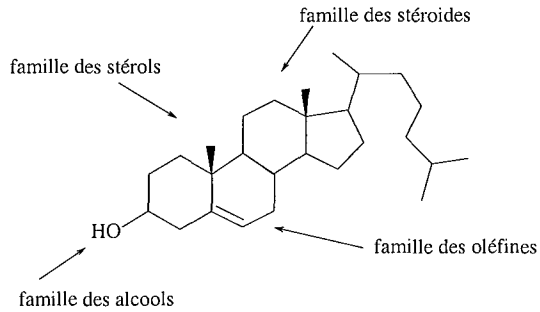


FIG. 2.7 – Quelques unes des grandes familles de composés chimiques auxquelles la molécule de cholestérol appartient.

2.3.3 La rétrosynthèse

E. J. Corey est le premier spécialiste de la synthèse a avoir posé les bases du raisonnement rétrosynthétique. On parle également d'*analyse antithétique* ou d'*analyse rétrosynthétique* ou de *rétrosynthèse*. Nous essaierons dans ce qui suit d'employer ce dernier terme car les deux autres expressions nous paraissent redondantes. Corey a développé toute une théorie basée sur ce paradigme et a mis en évidence une série de stratégies de résolution utilisables.

2.3.3.1 Un raisonnement analytique face à un problème difficile

Comme nous l'avons vu dans la section 2.2, la planification de synthèse est un problème difficile et mal défini. Trouver la synthèse d'une molécule complexe peut être rarement effectué en déterminant en un seul coup d'œil les produits de départ [Warren, 1986]. Il est nécessaire de décomposer le problème en sous-problèmes plus simples susceptibles d'être eux-mêmes décomposés. La méthode de décomposition permet de résoudre plus facilement des problèmes très compliqués [Rich and Knight, 1991]. Sur la figure 2.8, on peut voir une première décomposition du problème lié à la synthèse du Taxol. En analysant la structure de cette molécule, on remarque que deux motifs structuraux très différents sont reliés par une fonction ester : une partie polycyclique ayant un squelette carboné diterpénique et une partie azotée donc un alcaloïde, chacune constituant un sous-problème. Si l'on détermine comment synthétiser séparément ces deux sous-structures, la dernière étape de la synthèse consistera à les connecter en créant la fonction ester.

Sur ce principe, Corey préconise une analyse logique et systématique de la molécule à synthétiser. Ceci est possible en imaginant des déconnexions au niveau de la structure moléculaire qui découpent celle-ci en structures plus petites, et ainsi de suite. Les sous-structures obtenues deviennent alors des sous-problèmes, censés être plus faciles à résoudre. On construit ainsi *un arbre de rétrosynthèse*. Corey n'a pas été le premier à émettre l'idée

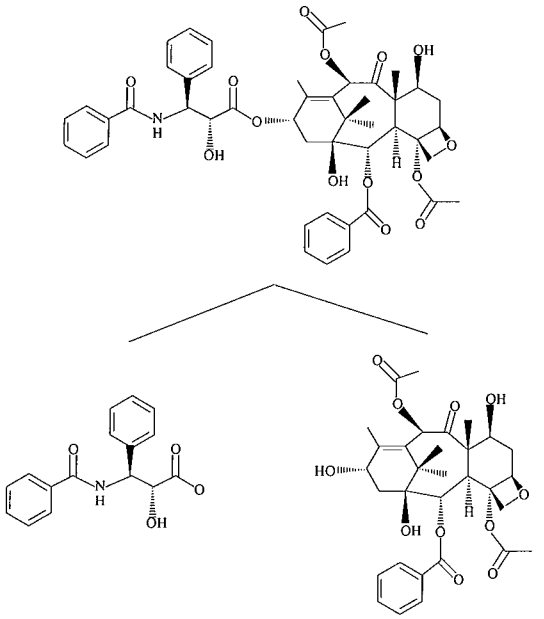


FIG. 2.8 – La décomposition du problème de synthèse du Taxol en deux sous-problèmes. À gauche, un alcaloïde. À droite, un diterpène.

de la rétrosynthèse. Robinson en avait introduit le principe puis Vleduts avait décrit l'utilisation du point de vue des systèmes informatiques d'aide à la synthèse organique [Vleduts, 1963]. Les travaux de Corey, qui lui ont valu l'obtention du prix Nobel de chimie en 1990, ont permis de codifier et d'offrir un cadre méthodologique à la synthèse organique qui jusque là avait été peu formalisée. Lorsque l'on décompose une structure en d'autres sous-structures, on effectue une transformation qui permet de passer de la structure originale aux structures résultats. Nous allons préciser dans la section suivante la signification du terme de transformation dans le domaine de la chimie organique.

2.3.3.2 Notion de transformation

Dans le domaine de la synthèse, le concept de « transformation » peut être vu selon deux angles diamétralement opposés, le sens synthétique et le sens rétrosynthétique.

Dans le langage de la rétrosynthèse, la molécule à laquelle est appliquée la transformation est appelée « cible » et les molécules hypothétiques obtenues par cette opération virtuelle sont nommées « précurseurs ». Par convention, le passage entre la cible et les précurseurs est indiqué par une double flèche, ce qui permet de distinguer une transformation d'une réaction au niveau de laquelle le passage des réactants aux produits est symbolisé par une flèche simple.

sens synthétique	:	réactants	$\rightarrow$	produits
sens rétrosynthétique	:	cible	$\Rightarrow$	précurseurs

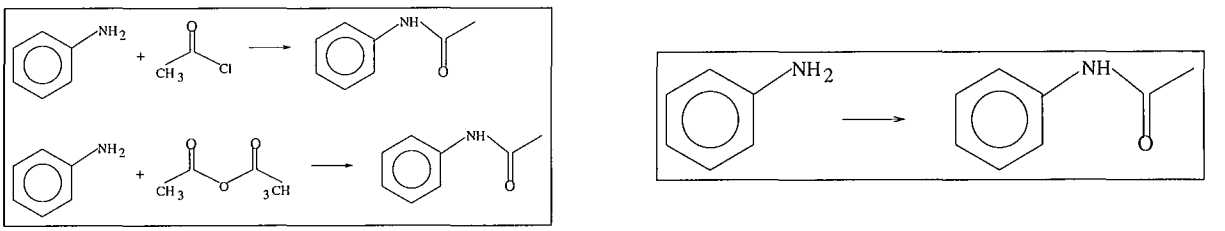


FIG. 2.9 – Deux réactions possibles permettant de synthétiser une phénylacétamide à partir d'une aniline et la transformation correspondant selon la définition de l'IUPAC.

Une transformation selon le sens synthétique

L'IUPAC [7] donne une signification « orientée synthèse » au terme de transformation dans le « Glossary of terms used in physical organic chemistry » [IUPAC, 1994]. En effet, dans la recommandation de 1994 de l'IUPAC, une transformation est définie comme étant la conversion d'un substrat en un produit particulier, sans rendre compte des réactifs ni du mécanisme. Il est précisé qu'une transformation est donc différente d'une réaction qui implique, elle, la description complète des réactants et des produits. D'après cette définition, une transformation consiste en une description partielle d'une réaction. De plus, elle est subjective car la détermination de la molécule qui a le rôle du substrat dans une réaction dépend du point de vue de l'observateur, même si le substrat est censé être la molécule la plus grosse qui réagit. On peut en déduire que deux réactions différentes peuvent être associées à la même transformation (comme sur la figure 2.9) mais aussi que plusieurs transformations peuvent correspondre à la même réaction selon le choix du substrat qui a été fait. Il est à noter qu'une nomenclature des transformations a été développée dans le cadre de l'IUPAC [IUPAC, 1989]. Dans ce cadre, plusieurs types de transformations sont décrits dont les additions, éliminations, substitutions, cyclisations, ouvertures de cycle etc.

Une transformation selon le sens rétrosynthétique

Corey a défini le concept de transformation en rétrosynthèse dans le cadre du projet LHASA [Corey, 1971]. Une transformation correspond exactement à l'inverse d'une réaction. On a une et une seule transformation par réaction saisie de manière stœchiométrique[8] (figure 2.10).

Corey classe les transformations selon le résultat de leur action (sens rétrosynthétique). Il a ainsi distingué les transformations simplifiantes des transformations de modification et de celles de complexification.

[7] IUPAC : International Union of Pure and Applied Chemistry.

[8] Une réaction est représentée par un bilan qui rend compte de la nature des corps réagissants aussi bien que de ceux qui sont formés. L'équation bilan peut également indiquer la quantité des réactifs consommés et celle des produits formés. Dans ce cas là, si le principe de conservation de la matière de Lavoisier est respecté, on dit que l'on a représenté la réaction de façon stœchiométrique. La réaction décrite est équilibrée.

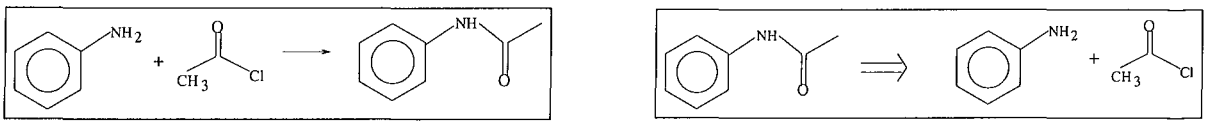


FIG. 2.10 – Une des réactions possibles permettant de transformer une aniline en une phénylacétamide et la transformation correspondant selon la définition révisée de Corey.

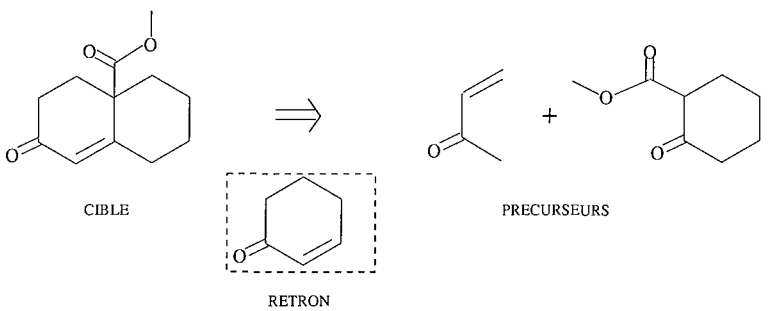
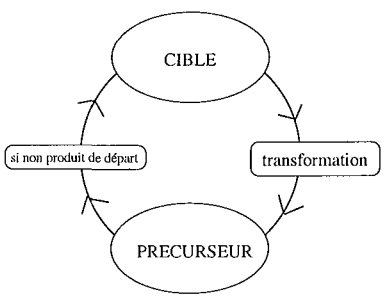


FIG. 2.11 – Le rétron (encadré) associé à la transformation correspondant à un exemple d'utilisation de la méthode d'annellation de Robinson selon Corey.

La démarche rétrosynthétique a pour principe d'appliquer des transformations à la molécule cible pour obtenir des précurseurs généralement plus simples. Si les précurseurs obtenus ne correspondent pas à des produits de départ disponibles, ils sont alors pris pour cible de l'analyse et ainsi de suite.



Divers critères conditionnent l'application d'une transformation à une molécule donnée. La condition minimale est la présence dans la cible du « rétron » de la transformation. Le rétron associé à une transformation est la sous-structure minimale qui doit être présente dans la molécule cible pour pouvoir générer les précurseurs [Corey et al., 1985]. Le schéma de la figure 2.11 représente la transformation et le rétron correspondant à un exemple d'utilisation de la méthode d'annellation de Robinson selon Corey.

2.3.3.3 Arbre de rétrosynthèse

La conception d'un plan de synthèse par rétrosynthèse a pour résultat la construction d'un *arbre de rétrosynthèse*, également appelé *arbre de synthèse*. Cette deuxième expression ne nous paraît pas juste et nous ne l'emploierons pas dans ce qui suit. Un arbre de rétrosynthèse est une représentation qui permet de décrire, selon certaines conventions, l'ensemble des hypothèses faites pendant la planification de la synthèse. La racine d'un tel arbre est la molécule cible, les feuilles correspondant aux produits de départ. Dans [Tonnelier, 1990], un arbre de rétrosynthèse est un arbre ET/OU[9] sur le principe énoncé dans [Bersohn and

[9] c'est-à-dire la représentation de la solution d'un problème obtenue par « chaînage arrière ».

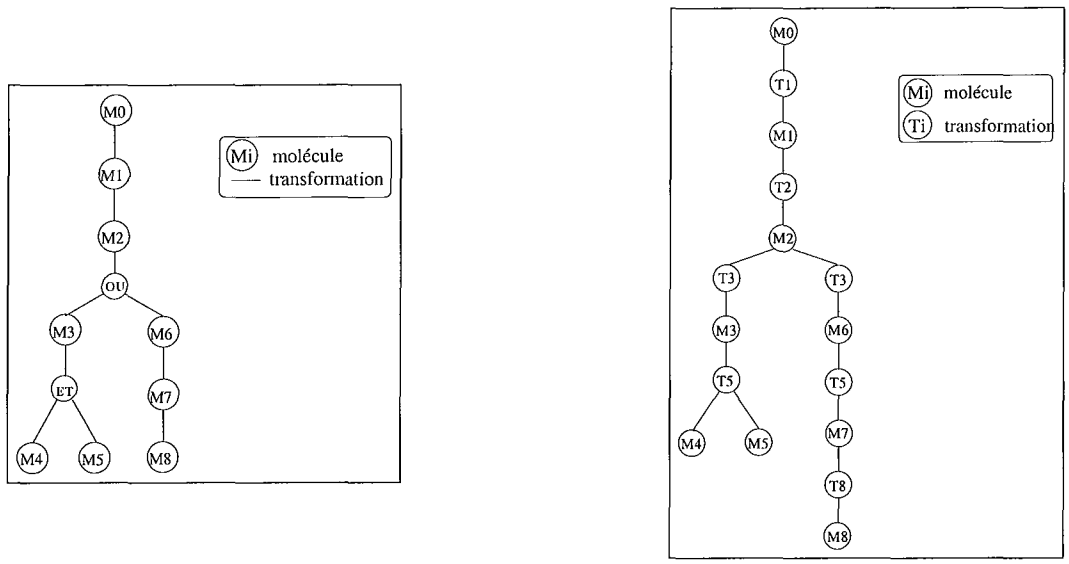


FIG. 2.12 – Un arbre de rétrosynthèse à gauche tel que le définit Tonnelier, le même arbre de rétrosynthèse à droite tel que le définit Gien.

Esack, 1976]. Les noeuds OU donnent les alternatives à la synthèse de la molécule décrite par le noeud parent du noeud OU alors que les noeuds ET donnent les molécules (noeuds filles du noeud ET) censées réagir ensemble pour produire la molécule décrite par le noeud parent du noeud ET. Dans [Gien, 1998], le plan est modélisé par un arbre de rétrosynthèse bipartite. Les deux types de sommets de cet arbre sont les structures, d'une part, et les transformations, d'autre part. La figure 2.12 montre un schéma de chacune de ces deux représentations pour un même plan de synthèse.

2.3.3.4 Stratégies de résolution de rétrosynthèse

La rétrosynthèse, si elle est menée de manière systématique, conduit à une explosion combinatoire étant donné le nombre de transformations et de possibilités d'application de celles-ci à une même molécule. Corey a cherché à exprimer des règles et des méthodes contrôlant le raisonnement rétrosynthétique. Il a notamment mis en évidence cinq grands types de stratégie de résolution.

- stratégie d'application des transformations les plus simplifiantes (méthode de Michael, annellation de Robinson, méthode de Mannich, réarrangement de Claisen, etc.),
- stratégie de reconnaissance de produits de départ ou d'intermédiaires clés à partir de la structure moléculaire de la cible,
- stratégie basée sur la topologie (reconnaissance des liaisons stratégiques d'un point de vue topologique, c'est-à-dire des liaisons dont la coupure permettra de réduire la complexité moléculaire),
- stratégie fondée sur la stéréochimie (reconnaissance des stéréocentres susceptibles d'être établis et ceux à préserver),

– stratégie s'appuyant sur la fonctionnalité présente dans la molécule.

Pour un même problème, Corey conseille d'appliquer plusieurs stratégies de résolution en parallèle pour explorer le maximum d'alternatives possibles.

2.3.4 Les approches alternatives

2.3.4.1 L'approche des chirons

Hanessian a développé une stratégie de résolution de problèmes de synthèse de molécules naturelles [Hanessian, 1983]. Devant la difficulté de synthétiser des molécules chirales de façon sélective, il préconise une stratégie défensive qui consiste à sélectionner des produits de départ pour lesquels un minimum de modifications du point de vue stéréochimique seront nécessaires pour obtenir la molécule cible. Son analyse revient à rechercher des *chirons*, c'est-à-dire des intermédiaires dans lesquels un maximum de centres stéréochimiques sont préformés. Les chirons doivent, de plus, pouvoir être facilement obtenus à partir de molécules naturelles chirales telles que les sucres, les acides aminés, les terpènes, etc. Cette approche est basée sur une analyse de la molécule cible. On peut remarquer que cette approche est un cas particulier de l'approche dirigée par les produits de départ et conduit au développement d'hémisynthèses.

2.3.4.2 L'approche d'Hendrickson

Hendrickson a proposé une approche semi-formelle, dite « logique », qui implique une déconnexion systématique des liaisons de la cible [Hendrickson and Parks, 1992]. La seule condition qui est posée au niveau stratégique est que les déconnexions choisies respectent le *principe de convergence*, donc d'économie. Ce sont les liaisons du squelette qui sont déconnectées, les fragments générés doivent comporter aux moins trois atomes. Un plan d'assemblage est accepté si au bout de deux étapes de déconnexion les fragments correspondent à des squelettes de produits de départ disponibles. Au niveau tactique, les transformations sont divisées en deux catégories : les transformations permettant de construire le squelette de la molécule et celles qui modifient la fonctionnalité. La sélection des transformations à appliquer se fait en fonction de la réduction de la complexité moléculaire engendrée. Étant donné que cette dernière est fortement liée à la notion de squelette, c'est le premier type de transformation qui est préférentiellement choisi

2.4 Conclusion

Comme nous l'avons vu, résoudre un problème de synthèse n'est pas simple (ça ne serait pas un problème si c'était le cas). Comme il n'existe pas de solution algorithmique à ce type de problèmes, le travail des experts de la synthèse est très souvent considéré comme un art. Cependant des outils théoriques peuvent être tout de même utilisés pour vérifier les hypothèses émises à partir des analogies faites par les chimistes. Au niveau de notre projet

de système d'aide à la planification de synthèse, nous n'avons pas l'ambition d'inventer de nouvelles méthodes de synthèse mais d'utiliser au mieux celles qui existent.

Chapitre 3

L'aide à la résolution de problèmes de synthèse

Sommaire

3.1 Introduction

L'informatique a révolutionné les pratiques dans de nombreux domaines. Que ce soit pour le stockage et la recherche d'information ou la résolution de problèmes, les systèmes informatiques proposent des solutions relativement efficaces et conviviales.

La possibilité de représenter les objets chimiques — les molécules — par des graphes moléculaires[1] a rendu l'exploitation informatique encore plus performante. La théorie des graphes, qui doit ses premiers développements à la chimie [Cayley, 1874], offre un cadre

[1]Un graphe moléculaire est un graphe dont les sommets correspondent aux atomes et les arêtes aux liaisons entre les atomes. Nous donnerons une définition formelle de ces graphes dans la partie III de ce mémoire, chapitre 7.

formel au traitement informatique des formules structurales. De nombreux algorithmes permettent de manipuler les graphes moléculaires de façon originale et plus ou moins rapide. Les modes de communication (interrogation et visualisation) possibles avec la machine ont l'avantage d'utiliser le langage graphique des chimistes, qui consiste à représenter les molécules par des formules structurales plus ou moins développées et dont nous avons déjà donné quelques exemples.

Deux types principaux de systèmes dont la finalité est d'aider les chimistes dans la résolution de problèmes de synthèse ont vu le jour à partir des années 60. Ott et Noordik [Ott and Noordik, 1992] classent ces systèmes en deux catégories : les systèmes de bases de données et les programmes d'aide à la synthèse. Les premiers sont des systèmes de gestion de bases de données. Les seconds sont des systèmes d'aide à la résolution de problèmes s'apparentant à des systèmes experts dans le sens où ils utilisent des connaissances pour résoudre des problèmes. Ils sont divisés en programmes de planification de synthèse d'une part, et programmes de prédiction de réactions d'autre part.

Dès 1963, Vleduts pose les bases de ces deux catégories de systèmes [Vleduts, 1963]. Il expose les principes de la représentation et de la manipulation des données intéressantes du point de vue de la synthèse, c'est-à-dire les réactions ou, leurs homologues dans le sens rétrosynthétique, les transformations.

3.2 Les bases de données de réactions

3.2.1 Généralités

Devant la quantité croissante des données collectées chaque jour, la nécessité de disposer de moyens de stockage et d'exploitation de l'information est devenue évidente. La plupart sinon tous les domaines scientifiques profitent des capacités des systèmes informatiques pour gérer les données issues des recherches et des expérimentations effectuées de par le monde. Le domaine de la chimie, et notamment celui de la chimie organique, est particulièrement concerné vu le volume et la diversité des données disponibles.

La première préoccupation du chimiste face à un nouveau problème de synthèse est de faire une étude bibliographique sur la cible. Ceci afin d'essayer de tirer profit de tous les travaux susceptibles d'être en relation avec son problème et donc de l'aider à la compréhension de celui-ci mais aussi d'éviter de ré-expérimenter ce qui a déjà été réalisé. Pour cela, outre ses documents personnels, le chimiste dispose de plusieurs sources d'information [Coste, 2002] : littérature primaire (publications originales), littérature secondaire (index des sources primaires), littérature tertiaire (compilation d'un ensemble de résultats ayant trait à un sujet donné tels que les encyclopédies, les mises au point etc.).

Avant les années 80, la recherche bibliographique se faisait sur des versions papier des documents précédemment cités ou grâce à des systèmes de cartes encodées, précurseurs des bases de données informatiques. La recherche pouvait être améliorée ou facilitée à partir d'index tel que celui des CHEMICAL ABSTRACTS SERVICES [Beach et al., 1979] ou d'encyclopédies telles que le BEILSTEIN HANDBOOK [Luckenbach, 1998]. Il fallait tout

de même être capable d'exprimer la requête en termes textuels, le recours à des requêtes de type graphique étant alors impossible. La recherche d'information était d'autant plus difficile que la question ne pouvait être présentée qu'au niveau structural, c'est-à-dire par un schéma réactionnel.

À partir des années 80, différents systèmes de bases de données sont devenus opérationnels pour répondre aux besoins en documentation des chimistes. Ces bases de données ont pour but de répertorier et de permettre un accès convivial et plus facile qu'une recherche bibliographique manuelle aux données publiées dans la littérature du domaine. Il existe des bases de données de molécules et de réactions. Les bases de données de réactions ont été plus difficiles à mettre en place que celles sur les molécules [Willett, 1986] car la recherche de réactions fait appel à des mécanismes plus complexes que celle de molécules. En effet, contrairement aux molécules pour lesquelles on dispose de formules ou de nomenclatures et de systèmes d'indexation établis depuis longtemps, les réactions, elles, ne bénéficient pas de systèmes de classification universellement reconnus. Avec les systèmes de bases de données de réactions, cette difficulté est contournée par l'emploi de schémas réactionnels. Cependant, l'implémentation du système de requête reste plus complexe que pour la recherche structurale.

Les premiers systèmes de bases de données de réactions ont été développés à la suite des travaux faits sur les systèmes d'aide à la résolution de problèmes de synthèse [Ott and Noordik, 1992]. On peut citer, par exemple, la base ORAC [2] qui a été créée à partir des travaux sur LHASA et REACCS [3] à partir de ceux sur SECS [Johnson, 1988]. En effet, grâce aux résultats de ces recherches, la représentation, la manipulation et la visualisation des structures moléculaires étaient maîtrisées par des techniques informatiques.

Les bases de données et leurs systèmes de gestion sont devenus des outils indispensables pour les chimistes organiciens. En 1999, une enquête initiée dans le cadre du GDR 1093 du CNRS a mis en évidence l'importance des bases de données chimiques dans l'activité quotidienne des chercheurs français en synthèse [Baldy, 1999]. Dans le même temps, des études [Coste et al., 1999] [Parkar and Parkin, 1999] discutent des problèmes rencontrés lors de l'utilisation de ces bases sans toutefois nier le progrès que celles-ci constituent. Ils mettent en évidence les défauts liés en grande partie à la conception des bases.

Il existe de nombreuses bases commercialisées ou en accès libre à partir d'Internet (par exemple WebReactions [Hendrickson and Sander, url]) ou bien encore propres aux entreprises ou aux laboratoires de recherche. Ces bases répondent à des besoins très variés et couvrent des domaines aussi divers que la chimie fine, la biochimie (PDB[4]), la médecine (MedLine). Elles contiennent des données ayant trait à des références bibliographiques pour les systèmes documentaires tels que Current Contents ou les CHEMICAL ABSTRACTS SERVICES, toutes sortes d'informations sur les données physiques ou chimiques pour les systèmes de type encyclopédique tel que BEILSTEIN. Devant le besoin accru en systèmes de gestion de l'information chimique, de nombreuses sociétés de service informatiques pro-

[2] ORAC : ORGANIC REACTION ACCESS BY COMPUTER.
[3] REACCS : REaction ACcess System.
[4] PDB : Protein Data Bank.

posent des solutions personnalisables, sous forme de bases de données capables de stocker des molécules ou des réactions (Accelrys [Accelrys, url], Tripos [Tripos, url], Daylight [Daylight, url], MDL [MDL, url], etc).

Notre projet étant tourné vers l'aide à la planification de synthèse, nous nous intéressons plus particulièrement aux bases décrivant des réactions. Dans ce qui suit, nous parlerons des bases de données de réactions les plus usitées, c'est-à-dire les bases distribuées par la société MOLECULAR DESIGN LIMITED (MDL) ainsi que celles fournies par CHEMICAL ABSTRACTS SERVICES (CAS).

3.2.2 Les principes et les fonctionnalités

Les bases dont nous allons détailler brièvement les principes et les fonctionnalités sont au nombre de trois : CASREACT, BEILSTEIN ainsi que celles accessibles via le système de gestion de bases de données ISIS [5].

La base CASREACT

CASREACT[6] [CASREACT, url] est une base de données de réactions développée par les CHEMICAL ABSTRACTS SERVICES [CAS, url][Blake and Dana, 1990] depuis 1988. Aujourd'hui[7], elle contient 6 millions de réactions extraites de plus de 339000 articles et brevets. La base est mise à jour toutes les semaines à raison d'environ 600 à 1300 réactions. CASREACT est interrogeable via plusieurs systèmes, à savoir STN International, STN on the Web, Sci Finder et Sci Finder Scholar [Ridley, 2002].

La base BEILSTEIN

La base BEILSTEIN [Heller, 1998] a été construite par la société BEILSTEIN [Beilstein, url] et est désormais distribuée par la société MDL. Elle est interrogeable à partir du système CROSSFIRE [Lawson, 1998]. Beilstein est une base importante en taille : elle couvre la documentation de pratiquement 9 millions de réactions dont un peu plus de 6 millions sont saisies de façon complète[8].

Les bases accessibles via ISIS

Ces bases sont diffusées par la société MDL [MDL, url]. Il s'agit de[9] :

- ChemInform RX (CIRX) : versions datant de 1992 à 2001
- ORGSYN : version 2000.1
- REACCS-JSM : version 2001.1
- REFLIB qui contient Thelheimer (version 90.1.5) et Comprehensive Heterocyclic Chemistry (CHC) (version 92.2.3)

[5] ISIS : Integrated Scientific Information System.

[6] CASREACT : CHEMICAL ABSTRACTS REACTION SEARCH SERVICE.

[7] sources datant de juillet 2002.

[8] les réactions que nous considérons comme incomplètes sont saisies sous la forme de réactions n'indiquant que le produit qui est formé (*half-reaction*).

[9] sources datant d'août 2002.

- Solid-Phase Organic REactions (SPORE) : version 2001.2

Ces bases sont interrogeables via le système de gestion de base de données ISIS. Ces bases sont orientées « méthodologie de synthèse » et sont, en conséquence, plutôt de nature sélective, c'est-à-dire que les méthodes de synthèse ne sont représentées dans les bases que par un nombre réduit d'exemples chacune. Au total un peu plus d'un million de réactions sont stockées. Certaines de ces bases sont mises à jour périodiquement alors que d'autres n'évoluent plus.

3.2.2.1 Les données

Tout d'abord, donnons quelques définitions employées dans les différentes bases et qui reprennent les recommandations de l'IUPAC.

Une réaction implique le passage d'un ensemble de molécules R (cet ensemble pouvant être réduit à un seul élément) à un autre ensemble de molécules P dans certaines conditions C. Les molécules formées sont nommées produits de la réaction. Selon leur rôle au cours de la réaction, les molécules réagissant sont appelées : réactant ou réactif/catalyseur. La distinction entre réactant et réactif/catalyseur vient du fait qu'un réactant doit participer pour au moins un atome de carbone dans la formation des produits contrairement aux réactifs/catalyseurs qui, eux, n'interviennent qu'au plus pour des hétéroatomes[10].

Les bases de données de réactions sont des collections d'exemples particuliers de réactions. Ceux-ci y sont décrits de la façon suivante :

- des données structurales illustrant le schéma réactionnel,
- des données α-numériques détaillant les résultats expérimentaux, les conditions réactionnelles ainsi que les références bibliographiques.

Dans les bases de données, le schéma réactionnel englobe la description des réactants et des produits. Les structures moléculaires sont codées sous forme de tables de connectivité décrivant principalement la nature des atomes et des liaisons entre les atomes. Étant donné que les réactifs et les catalyseurs ne sont pas présents au niveau du schéma réactionnel, les réactions saisies sont très rarement équilibrées[11]. Les réactions telles qu'elles sont présentées dans les bases de données de réactions sont à rapprocher des transformations selon l'IUPAC. Une correspondance, également appelée appariement ou *mapping*, entre les atomes des réactants et ceux des produits est donnée pour rendre compte de ce qu'il s'est passé au cours de la réaction. Sur le schéma (a) de la figure 3.1, nous pouvons voir quelques unes des conventions de dessin notamment en ce qui concerne les indications de nature stéréochimique[12]. On remarque que les atomes de carbone ne sont pas représentés par leur symbole et que la plupart des atomes d'hydrogène sont omis. Seuls les hydrogènes « significatifs » sont explicitement indiqués. La correspondance entre les atomes des réactants et ceux des produits permet de constater les changements qui ont eu lieu au cours de la réaction en

[10] hétéroatome : atome autre qu'un atome de carbone ou d'hydrogène.

[11] Une réaction est équilibrée lorsque tous les atomes présents d'un côté de la flèche se retrouvent de l'autre côté.

[12] c'est-à-dire de type tridimensionnel.

terme de modifications des liaisons entre les atomes (schéma (b) de la figure 3.1). On peut noter que les hydrogènes « significatifs » ne portent pas de numéros de correspondance (ceci est le cas dans les bases distribuées par la société MDL). L'ensemble des atomes et des liaisons modifiés constitue ce que l'on appelle le *centre de réaction* ou *cœur de réaction* ou encore *site de réaction.* Sur le schéma (c) de la figure 3.1, cet ensemble de liaisons a été mis en évidence. L'appariement des atomes conduit aussi à la détermination des changements observés du point de vue stéréochimique. Ces diverses modifications sont codées dans les bases grâce à des indications sur les liaisons (cassée/créée/modifiée/inchangée, en arrière/en avant/dans le plan pour les liaisons simples, cis/trans pour les liaisons doubles, stéréocentre ou non pour les atomes).

Initialement, la détermination du site réactionnel était laissée à l'appréciation de l'opérateur qui saisissait la réaction dans la base de données. Devant le manque d'homogénéité due à cette approche, des travaux ont permis de développer des techniques d'appariement automatique des atomes des réactants et des produits d'un schéma réactionnel. Ces méthodes sont basées sur le principe de déformation minimale qui impose que le nombre de liaisons modifiées (créées, cassées ou changeant de type) soit le plus petit possible. Les algorithmes consistent à déterminer en priorité la plus grande sous-structure commune (ou MCS pour *Maximal Common Subgraph*) entre les graphes des réactants et des produits [Moock et al., 1988b][Lynch and Willett, 1978][McGregor and Willett, 1981][Funatsu et al., 1988].

Dans les bases de données de réactions accessibles vie ISIS, aucune différence de représentation ni de traitement n'est faite entre une réaction dite « monoétape » et les réactions bilan « multiétapes ». Il existe, cependant, un lien entre entre les réactions bilan et chacune de leurs étapes lorsque celles-ci ont été saisies[13]. Une fonction du système de gestion des bases permet de retrouver toutes les réactions faisant partie du même « schéma ». Dans CASREACT, quand une réaction bilan est répertoriée, les sous-bilans correspondant à toutes ses sous-séquences le sont eux aussi.

Généralement, les bases de données de réactions sont couplées à des bases de données de molécules et/ou de documents. Cela permet une navigation dans un très vaste volume de données hétérogènes.

La figure 3.2, inspirée de [Laurenço, 1998] et [Coste, 2002], résume ce que nous venons de dire sur les données présentes dans les bases de données de réactions. Dans le cas de BEILSTEIN, on a également accès aux propriétés des réactants et des produits. Dans certaines systèmes comme ISIS, il est aussi possible d'accéder à la description structurale des réactifs/catalyseurs.

3.2.2.2 L'interrogation des bases de données de réactions

Les bases de données de réactions, doivent, comme tous les systèmes d'information, remplir certaines conditions pour répondre de façon convenable aux besoins des utilisateurs. Ces caractéristiques sont principalement au nombre de trois :

[13]Toutes les étapes correspondant à une réaction bilan ne sont pas forcément saisies.

(a)

(b)

(c)

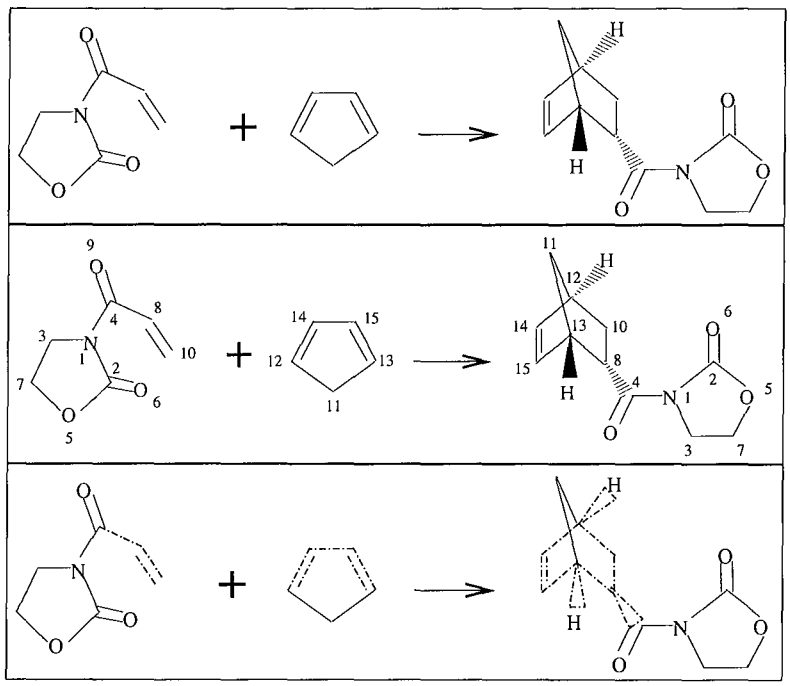


FIG. 3.1 – (a) la réaction 499 de la base ORGSYN version 98 ; (b) la correspondance entre les atomes des réactants et des produits telle qu'elle est indiquée dans la base de données ; (c) le centre de réaction correspondant à la correspondance indiquée.

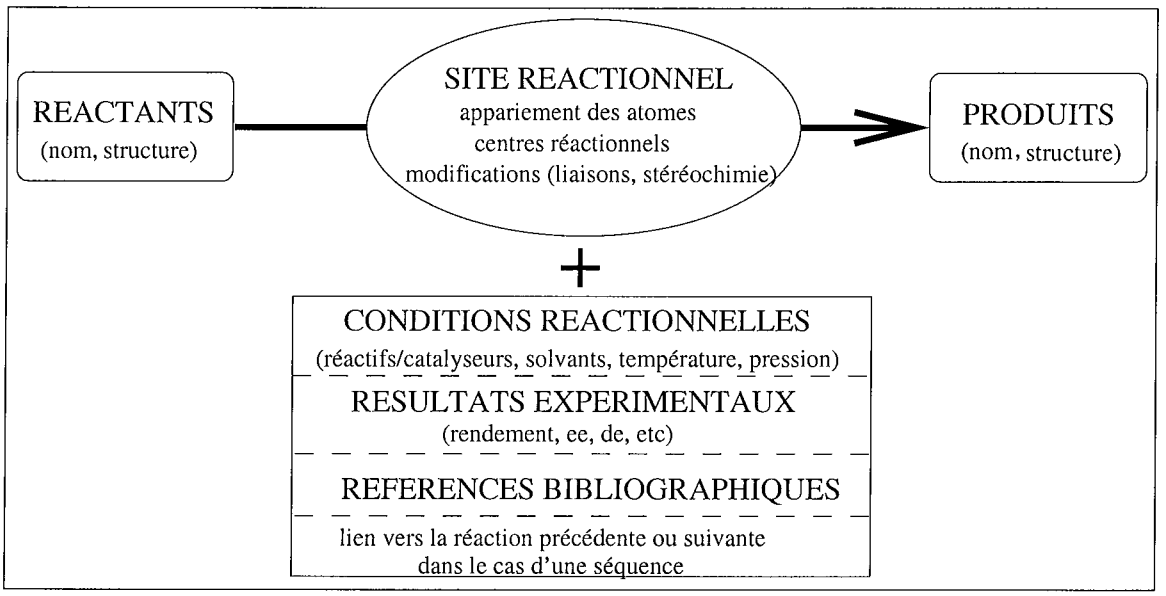


FIG. 3.2 – Les données dans les bases de données de réactions.

- exhaustivité (sur l'ensemble du domaine ou sur un thème particulier du domaine) ou sélectivité,
- modules de recherche rapides et puissants,
- interface machine/utilisateur adaptée.

Complétude

Dans [Coste, 2002], Coste fait la différence entre les bases exhaustives, sélectives et thématiques. Les bases à but exhaustif ont pour ambition de couvrir une grande partie, voir la totalité, des résultats publiés dans la littérature du domaine [Beilstein, url][CASREACT, url]. Les bases sélectives sont supposées illustrer toute la chimie organique, mais en ne présentant qu'un nombre limité de réactions pour chaque classe de méthode. Les bases thématiques sont construites pour rapporter l'ensemble des recherches d'un domaine particulier (chimie des organométalliques avec les Cahiers Bibliographiques de Chimie Organométallique, chimie des hétérocycles avec CHC) ou sélectionner des exemples représentatifs de méthodes de synthèse (par exemple, REACCS-JSM). Il faut donc bien faire la différence entre les bases à but exhaustif et celles dites sélectives ou thématiques. De plus, il est légitime de se poser des questions sur le contenu réel des bases.

Modes de recherche dans les bases de données de réactions

Dans le tableau 3.1, nous présentons le type de requêtes orientées « synthèse » qu'il est possible de poser à un système d'information tel qu'un système de gestion de bases de données de réactions [Laurenço, 1998][Coste, 2002].

Pour répondre à ces requêtes, les systèmes de gestion de bases de données de réactions proposent divers modes de recherche. Outre la possibilité de faire des interrogations par

? $\rightarrow$ P	quels ensembles de réactants R permettent d'obtenir P ?
R $\rightarrow$?	quels ensembles de produits P sont créés à partir de R ?
R $\rightarrow$ P	quelles sont les réactions qui transforment R en P ?
R $\overset{r/c}{\rightarrow}$ P	quelles sont les réactions qui transforment R en P dans les conditions r/c ?
R ?	dans quelles réactions l'ensemble R réagit ?
P ?	dans quelles réactions l'ensemble P est produit ?
? $\overset{r/c}{\rightarrow}$?	dans quelles réactions les conditions r/c sont utilisées ?

TAB. 3.1 – Les requêtes que l'on peut poser à une base de données de réactions où R signifie réactants, P produits, r/c réactifs ou catalyseurs. R et P peuvent être des ensembles de structures exactes ou de sous-structures.

mots clés ou par termes bibliographiques, beaucoup de bases permettent d'effectuer des recherches structurales et sous-structurales performantes dans des temps très corrects. Certains systèmes disposent de modules annexes de recherche par similitude ou basée sur le schéma réactionnel mais aussi de fonctions de tri selon des critères variés.

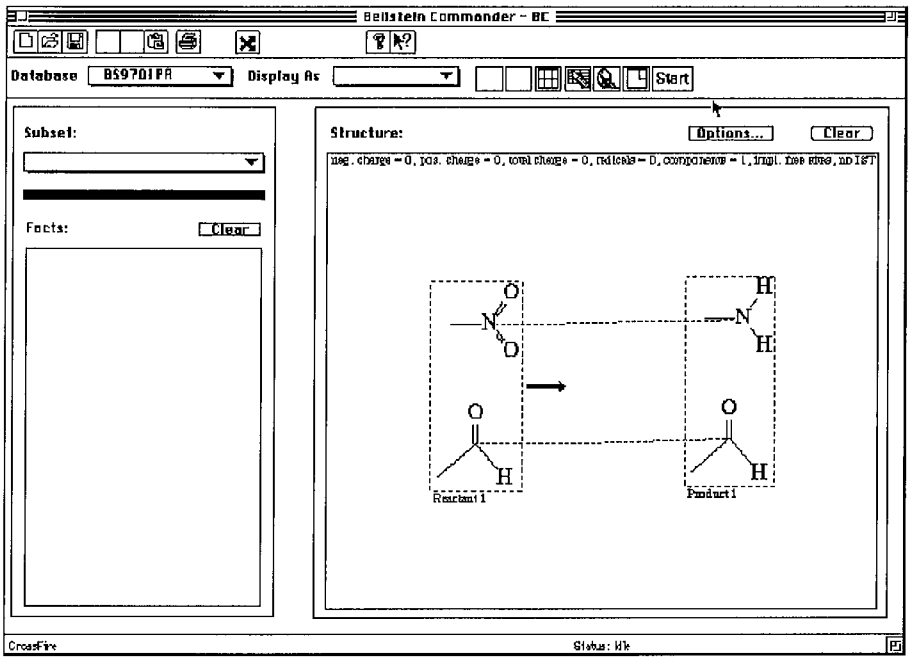
▷ *recherche par mots clés*

On peut interroger les bases sur tous les champs textuels et numériques (rendement, température, auteur, ...). Les mots clés se rapportant au type de réaction ne sont pas renseignés dans toutes les bases. Si CASREACT est bien documentée de ce point de vue, on peut déplorer que, parmi les bases accessibles via le système ISIS, une seule, SPORE, présente cette option de recherche par mots clés.

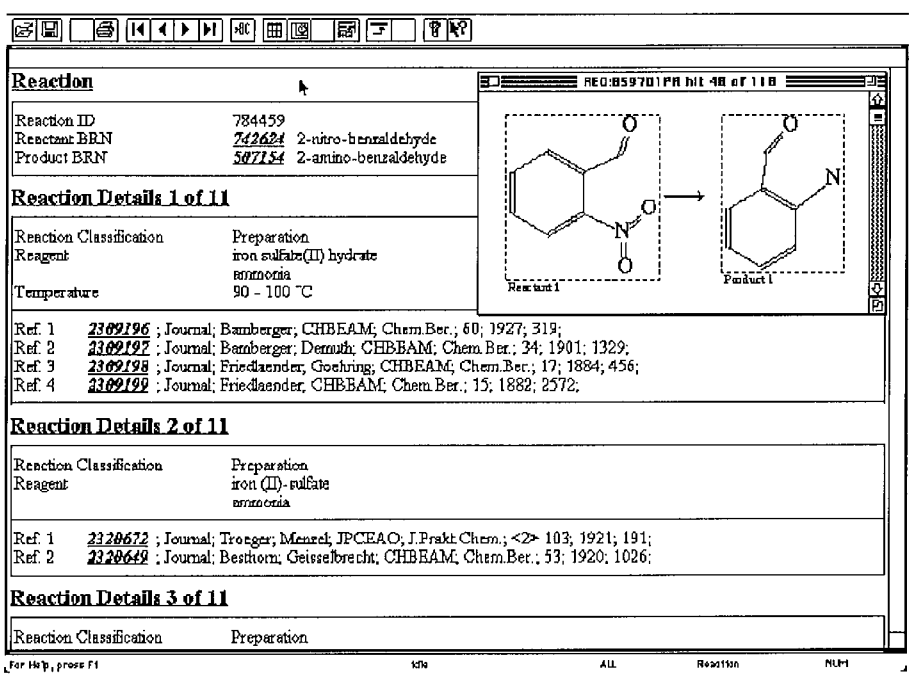
▷ *recherche structurale et sous-structurale*

Les représentations graphiques engendrent un pouvoir d'expression bien supérieur au langage naturel et permettent donc de poser des requêtes qui ne pourraient l'être de manière textuelle. Sur la figure 3.3, nous montrons une recherche sous-structurale dont le but est de récupérer des réactions permettant de réduire une fonction nitro sans modifier un aldéhyde. Le trait tiré entre les atomes du réactant et du produit indique la correspondance que l'on impose entre les atomes. La capture d'écran de droite montre un des résultats de la recherche. On remarque qu'effectivement au cours de cette réaction il y a eu réduction sélective d'une fonction nitro.

Dans le cadre de la théorie des graphes, une recherche structurale exacte consiste à tes-



une requête sous-structurale



une réaction donnée en réponse

FIG. 3.3 – Recherche par sous-structure dans une base de données de réactions.

ter l'isomorphisme entre deux graphes alors que, pour une recherche sous-structurale, c'est l'inclusion de graphes qui est vérifiée. En terme de complexité, la recherche de l'isomorphisme de graphes est un problème non classé tandis que l'isomorphisme de sous-graphes est NP-Complet [Garey and Johnson, 1979]. Les graphes moléculaires étant étiquetés[14], les problèmes d'isomorphismes dans le cadre des graphes moléculaires sont de complexité un peu moindre que dans le cas général. La résolution de ces deux problèmes est par conséquent coûteuse en temps de calcul et différentes solutions ont été proposées dans le cadre des systèmes d'information chimique [Barnard and Walkowiak, 1998]. Les procédés de recherche structurale qui ont été implémentés dans la plupart des systèmes de gestion de bases de données chimiques contournent le problème par un filtrage intensif des données (ou « screening »). Généralement, le principe est de décrire et d'indexer les molécules en termes de motifs reconnus dans la structure et de manipuler des fichiers inverses issus de l'indexation. Parmi les types de motifs assez discriminants pour permettre le filtrage, les sous-structures cycliques occupent une place importante [Dury et al., 2001]. Celui-ci permet de réduire considérablement l'ensemble des structures sur lesquelles il faudra ensuite faire une recherche atome par atome (ABAS[15]) [Moock et al., 1988a]. Si ce procédé ne permet pas forcément de donner toutes les réponses, il ne fournira pas de réponses fausses. Certains systèmes tels que BEILSTEIN ne procèdent pas de cette façon. Les concepteurs de BEILSTEIN ont développé un moteur de recherche puissant et rapide grâce à une indexation des molécules originale. En effet, dans ce système, chaque molécule est décrite par autant de notations linéaires qu'elle comporte d'atomes. Une notation linéaire est établie par un parcours en largeur du graphe moléculaire à partir d'un des atomes de la molécule. La phase de filtrage est très efficace car elle consiste à comparer des chaînes de bits représentant les notations linéaires obtenues. Après cette phase de filtrage, le système procède à l'étape de recherche atome par atome. Dans le cas des deux méthodes, le système présente des résultats très rapidement (quelques secondes pour une recherche pour plusieurs centaines de milliers de molécules).

▷ *recherche par similitude*

Il existe plusieurs manières de faire des recherches par similitude dans les bases de données de réactions [Willett, 1998]. La grande majorité des recherches par similitude est fondée sur le calcul de recouvrement entre descripteurs. La similitude entre deux schémas réactionnels est donnée par la distance calculée à partir du recouvrement, c'est-à-dire une valeur quantitative souvent exprimée en pourcentage. Pour calculer le degré de similitude entre deux objets a et b, l'une des formules les plus connues est le coefficient de Tanimoto :

$$T = \frac{C}{A + B - C}$$

avec C le nombre de descripteurs communs à a et b, A le nombre de descripteurs trouvés uniquement dans a, B le nombre de descripteurs rencontrés exclusivement dans b. On pourra

[14] voir section 7.1.1 pour une définition formelle de ces graphes.

[15] ABAS : Atom By Atom Search.

trouver dans [Willett, 1998] une collection des métriques usitées provenant du domaine de l'analyse de données.

Dans les bases interrogeables par ISIS, on peut demander une similitude structurelle (portant sur la « ressemblance » entre les molécules) ou réactionnelle (« ressemblance » entre les centres de réaction), ou bien les deux en même temps. Les mesures de similitude dans ces bases reposent sur un ensemble conséquent de descripteurs (933 pour les molécules et 230 sur les réactions) [Grethe and Moock, 1990]. C'est un calcul utilisant le principe du coefficient de Tanimoto qui est utilisé.

▷ *recherche par centre de réaction plus ou moins étendu*

Dans les bases de données interrogeables grâce au système ISIS, un mode de recherche permet de récupérer les réactions qui correspondent à la « même transformation » que celle indiquée dans la requête. Dans la base, à chaque réaction sont attribués trois codes de hachage affectés respectivement au centre de réaction, au centre de réaction augmenté de la sphère des premiers voisins[16], au centre de réaction augmenté des sphères des premiers et seconds voisins. Les codes de hachage sont calculés par l'algorithme CLASSIFY [Grethe, 1995].

▷ *module de partitionnement des réponses*

Une requête posée dans une base de données est susceptible de ramener un grand nombre de réponses. L'examen manuel d'une longue liste d'enregistrements est fastidieux, voire irréalisable. Cet état de fait se vérifie souvent au cours de l'interrogation des bases de données de réactions. Pour aider l'utilisateur, les concepteus des bases de données de réactions proposent des fonctions de regroupement (ou « clustering ») des réponses en catégories selon des critères que l'on doit fixer. L'analyse des différents clusters sera plus facile. Dans les bases accessibles via le système ISIS, il est possible de faire un tel partitionnement à partir des champs textuels de la base (journal, auteur, année, conditions réactionnelles, formules des solvants et réactifs, etc.) ainsi que du type de la réaction d'un point de vue structural (cette fonction étant basée sur les codes de hachage présentés dans le paragraphe précédent). Il est possible de conduire un tel processus sur un ensemble d'au plus 6000 enregistrements. Certains systèmes documentaires sur les réactions disposent d'autres systèmes de classification et d'indexation structurale des réactions [Ziegler, 1979][Rose and Gasteiger, 1994]. Il n'existe pas de module de regroupement de réponses dans la base BEILSTEIN.

▷ *mode logique*

La plupart des systèmes permettent aussi de croiser les résultats des requêtes grâce à des instructions logiques (OR, AND, NOT) combinant les listes de réponses comme cela est possible avec les langages de requêtes standard.

[16] en terme de proximité.

3.2.3 Les défauts et les problèmes

Les bases de données et leurs systèmes de gestion tels qu'ils ont été conçus répondent plus à des besoins encyclopédiques ou documentaires que d'aide à la résolution de problèmes de synthèse. On peut relever de nombreux défauts dont certains sont communs aux bases de données en général alors que d'autres sont liés au modèle conceptuel, ou au manque de modèle, dans la conception des bases de données de réactions.

Contenu des bases de données de réactions
Tout d'abord, on peut s'interroger sur les prétentions à l'exhaustivité ou contrairement à la sélectivité affichées par les sociétés qui développent les bases de données de réactions. Des études ont été menées à ce sujet. Dans [Parkar and Parkin, 1999] et [Coste et al., 1999], les auteurs tentent de comparer la base BEILSTEIN, à but exhaustif, aux bases sélectives fournies par MDL. Coste et col. remarquent notamment que le contenu de la base BEILSTEIN est très inégal, le nombre de journaux sources variant selon les années. Les deux analyses en viennent à la même conclusion : les deux types de bases doivent être interrogés conjointement car ces bases sont complémentaires.

Regroupement des réponses
Dans ISIS, la méthode de regroupement selon le site réactionnel pose problème. L'algorithme CLASSIFY n'étant pas biunivoque[17], il est fréquent de retrouver dans une même catégorie deux réactions n'ayant pas de rapport entre elles ou bien deux réactions censées appartenir au même ensemble dans deux partitions différentes.

Typologie des réactions
L'un des principaux problèmes est lié à la typologie des réactions [Coste et al., 1999]. Comme nous l'avons dit précédemment, les bases de données répertorient des exemples particuliers de réactions. Les requêtes posées aux systèmes de gestion de bases de données de réactions donnent de grandes listes de réactions que les méthodes de clustering sont incapables de regrouper en méthodes de synthèse données. Or, c'est cette information qui intéresserait en priorité l'expert de la synthèse. En effet, les chimistes procèdent par allers-retours entre diverses représentations mentales, c'est-à-dire de concepts plus ou moins abstraits aux faits et des faits aux concepts. Ces différents niveaux d'abstraction des données ne sont pas présents dans les bases de données de réactions qui ne s'attachent qu'à la documentation du niveau le plus bas. Ceci pose donc le problème de l'organisation des bases de données de réactions qui ne travaillent pas sur des données suffisamment structurées pour répondre réellement efficacement aux questions d'intérêt synthétique.

Erreurs sur les données
De nombreuses erreurs peuvent être constatées dans les données. Les définitions qui sont données sur les différents objets chimiques (réactant, produit, réactif, catalyseur ...) ne sont

[17] au sens de bijectif.

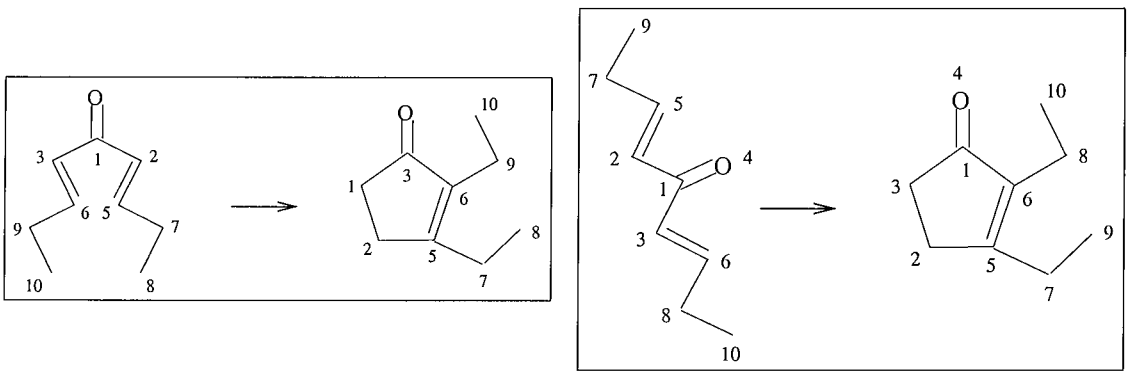


FIG. 3.4 – Deux appariements différents des atomes proposés pour une même réaction dans les bases de MDL. À gauche, la réaction 64555 de la base REACCS-JSM version 2000.1. À droite, la réaction 118967 de la base REFLIB version 95.1.2.

pas ou peu respectées dans la réalité. Ainsi nous avons pu voir des molécules classées en tant que réactif/catalyseur participer à la construction du produit pour au moins un atome de carbone ce qui est contraire aux conventions données dans la documentation des bases. Au niveau des conditions réactionnelles, il y a un manque flagrant de standardisation. Cet aspect des réactions a été étudié en détail en particulier dans la thèse de Hugues Vogel [Vogel, 2000].

Les problèmes d'appariement des atomes

Une autre source de problèmes réside dans l'établissement de l'appariement des atomes des réactants avec ceux des produits. Dans [Parkar and Parkin, 1999], les auteurs citent une étude établissant que seulement la moitié des réactions ayant les mêmes références bibliographiques ont le même centre de réaction. Ceci sous-entend que les algorithmes de détermination du site réactionnel ne sont pas univoques. Le résultat obtenu dépend de l'ordre dans lequel les atomes des molécules ont été saisis. À titre d'exemple, la figure 3.4 présente deux exemplaires d'une même réaction trouvés dans deux bases différentes, REACCS-JSM version 2000.1 et REFLIB 95.1.2. L'appariement est différent d'une base à l'autre. Après vérification, dans ce cas, c'est effectivement le module de *mapping* automatique du système ISIS qui est à l'origine de cette différence.

Sur le dessin de la réaction 64555 de la base REACCS-JSM version 2000.1 (figure 3.4), apparaît un deuxième défaut de la correspondance établie de façon automatique. En effet, devant certaines situations, l'algorithme se trouve dans une impasse et est incapable de faire un choix. Il en résulte une absence d'appariement de certains atomes.

Dans la documentation du système ISIS, il est signalé que les appariements incorrects peuvent être dus à trois types de conditions : (1) soit les structures des réactants et des produits sont symétriques de façon non standard, (2) soit le centre réactionnel implique la création ou la destruction de liaisons dans lesquelles intervient un atome d'hydrogène, (3) soit des substituants migrent d'un emplacement à un autre au cours de la réaction. Si nous comprenons tout à fait qu'il est difficile de ne pas se tromper dans le troisième cas qui

est celui des réarrangements, les deux autres situations semblent moins incontournables. Dans le deuxième cas, c'est le principe de changement minimal de la structure carbonée qui est certainement en cause. En effet, lors de la recherche de la sous-structure commune maximale, l'algorithme commence par apparier les liaisons de même type alors que l'essence même d'une réaction est de modifier le type des liaisons. En particulier les liaisons carbone-carbone de type multiple, les liaisons carbone-hétéroatome et les liaisons hétéroatome-hétéroatome. Les liaisons entre atomes de carbone et d'hydrogène sont plus difficiles à rompre ou à créer. Nous avons constaté au cours des travaux effectués dans le cadre de cette thèse une grande proportion d'appariements incomplets ou erronés.

Les erreurs de mapping ont pour conséquence la génération de centres de réaction qui produisent des résultats chimiques non pertinents lors de la recherche par schéma de réaction ainsi que dans le regroupement des réponses. Elles se rajoutent aux faiblesses de l'algorithme CLASSIFY précédemment identifiées.

Le dernier problème concernant la correspondance des atomes est l'absence d'appariement. L'étude des bases de données faite dans [Coste et al., 1999] a montré que, dans BEILSTEIN, seulement la moitié des réactions est représentée sous forme graphique, lesquelles ne sont pas du tout appariées dans 30% des cas, et dans CIRX, 9% de réactions n'ont pas été mises en correspondance. Le pourcentage de réactions non analysées du point de vue de l'appariement des atomes est donc non négligeable.

Au vu des problèmes (erreur et absence) de l'appariement des atomes dans les bases, Coste et col. conseillent de faire plusieurs requêtes pour un même problème — des requêtes sans mapping et d'autres testant différents mappings. Ceci tant que la « qualité » de l'appariement restera tel qu'il l'est à ce jour.

Exemples négatifs

Enfin, il existe une lacune dans les bases de données de réactions : celle de l'absence des réactions qui ne conduisent pas au produit escompté. Ce travail serait pourtant bien utile mais difficile à réaliser car peu d'auteurs expliquent en détail les échecs rencontrés dans leurs synthèses. Ce type de résultats, pourtant très riches d'enseignements, n'est pas reporté dans les articles. CASREACT a tout de même commencé à introduire de tels contre-exemples depuis sa création [Blake and Dana, 1990]. Il est possible d'accéder à ces réactions particulières par mot clé (« failed reaction »). La société Accelrys s'est aussi engagée dans la construction d'une base de données de ces réactions [Failed Reactions, url]. Dans ces bases, pour chaque réaction, le schéma réactionnel attendu ainsi que celui du produit réellement obtenu sont répertoriés.

3.2.4 Conclusion

Les bases de données de réactions, malgré leurs défauts, constituent une mine d'informations. Les chimistes, même s'ils n'en sont pas toujours conscients, trouvent en ces bases de données des outils très appréciables qu'ils utilisent dans leur travail quotidien.

Les exemples de réactions se rapportant de près ou de loin aux problèmes qui sont retrouvés aident au raisonnement par analogie. Avertis des problèmes des bases de données de réactions, nous montrerons dans ce mémoire (partie III, chapitre 8) ce que nous pouvons extraire par rapport à la problématique de la synthèse multiétapes ciblée.

3.3 Les systèmes d'aide à la résolution de problèmes de synthèse

3.3.1 Généralités

Les systèmes d'aide à la résolution de problèmes de synthèse ayant vus le jour depuis la fin des années 60 sont assez diversifiés. Le but de cette section n'est pas de décrire en détail chacun de ceux-ci mais d'en dégager les principales caractéristiques. Des descriptions plus complètes des principaux systèmes sont données dans [Laurenço, 1985] puis dans [Gien, 1998] et [Schildknecht, 2001].

La première différenciation porte sur l'approche adoptée, empirique ou formelle, dont nous allons parler dans ce qui suit.

On peut distinguer dans un deuxième temps les systèmes travaillant dans le sens synthétique ou dans le sens rétrosynthétique. Le but des premiers est la prédiction des réactions, voir l'élucidation de mécanisme, tandis que les seconds ont pour objectif de construire pas à pas un arbre de rétrosynthèse. Quelques systèmes proposent une analyse combinant ces deux modes.

Les autres critères de catégorisation des outils réalisés sont :
- interactivité ou fonctionnement automatique,
- domaine d'application général ou spécialisé,
- niveau de connaissance manipulée,
- volume de connaissances utilisées,
- traitement de la stéréochimie ou non,
- interface homme/machine textuel ou graphique,
- représentation en machine des données sur les molécules et les réactions (notation linéaire, table de connexion, matrices d'adjacence etc.),
- langage de programmation (langage de bas niveau, langage spécialisé, langage évolué).

3.3.2 Les différentes approches

De manière générale, les systèmes d'aide à la synthèse sont divisés en deux catégories selon l'approche employée. Ainsi, deux courants de pensée coexistent : l'approche formelle, d'une part, et l'approche empirique, d'autre part. Il existe aussi d'autres démarches intermédiaires telles que celle adoptée par Hendrickson.

3.3.2.1 L'approche empirique

Partant de l'idée que l'expérience des chimistes dirige la résolution des problèmes de synthèse, Corey base son approche sur la connaissance. Les connaissances intéressantes du point de vue de la rétrosynthèse sont celles relatives aux transformations dont il a été question dans la section 2.3.3.2. Les informations à posséder sur les transformations sont de natures diverses. On doit savoir, par exemple, quel est le contexte minimal d'application, c'est-à-dire décrire le rétron (section 2.3.3.2), les limites d'applicabilité ainsi que l'opérateur qui permettra d'obtenir les précurseurs. Il est donc nécessaire d'avoir une base de transformations. Une analyse systématique n'étant pas réalisable compte tenu de la taille de l'espace de recherche, il faut établir des heuristiques pour diriger la rétrosynthèse. Corey a mis en évidence des stratégies de résolution proches de celles utilisées par les chimistes (section 2.3.3.4). Les stratégies de résolution sont directement liées aux propriétés topologiques, fonctionnelles et stéréochimiques de la molécule cible. Le choix de la stratégie de résolution dépend donc des caractéristiques de la molécule cible. Celles-ci sont déterminées par un processus de perception. Ainsi, dans les systèmes empiriques, ce sont des considérations topologiques et/ou fonctionnelles et/ou stéréochimiques qui orienteront la sélection du type de transformation à appliquer [Jambaud, 1996]. Les stratégies de rétrosynthèse sont à la disposition de l'utilisateur pour contrôler et guider la construction de l'arbre de rétrosynthèse, le plus souvent pas à pas. Les transformations sont stockées individuellement sous forme d'un programme. Dans LHASA, elles comportent :

- un nom,
- un rétron,
- un score initial,
- des indications à caractère stratégique ou tactique,
- des enquêtes sur l'environnement du rétron lorsqu'il est présent dans la structure cible (cas favorables et cas défavorables),
- des indications sur les conditions réactionnelles,
- des instructions de manipulation de graphe permettant d'appliquer la transformation au graphe cible et d'engendrer ceux des précurseurs.

Les indications ainsi que les tests d'applicabilité des transformations sont codés dans un langage proche de celui des chimistes (par exemple, CHMTRN [18] conçu pour LHASA). Le rétron est stocké dans une table de connexion, une matrice ou un code linéaire et les actions à mener pour générer les précurseurs sont décrites dans le langage de programmation spécialisé avec des instructions du type `break bond`, `make bond` etc.

De manière générale, les systèmes basés sur l'approche empirique obéissent à l'organisation représentée sur la figure 3.5. Selon les cas, l'utilisateur a la possibilité de communiquer avec le système et d'interagir dans la résolution du problème.

Plusieurs systèmes ont été construits selon cette approche. Parmi les plus remarquables, le premier, OCCS[19] [Corey, 1967], a été rapidement suivi de LHASA [Corey et al., 1985] et de SECS [Wipke et al., 1977]. LHASA est très certainement le plus évolué des systèmes

[18] CHMTRN : CHeMical TRaNslation.

[19] OCCS : ORGANIC CHEMICAL SIMULATION OF SYNTHESIS.

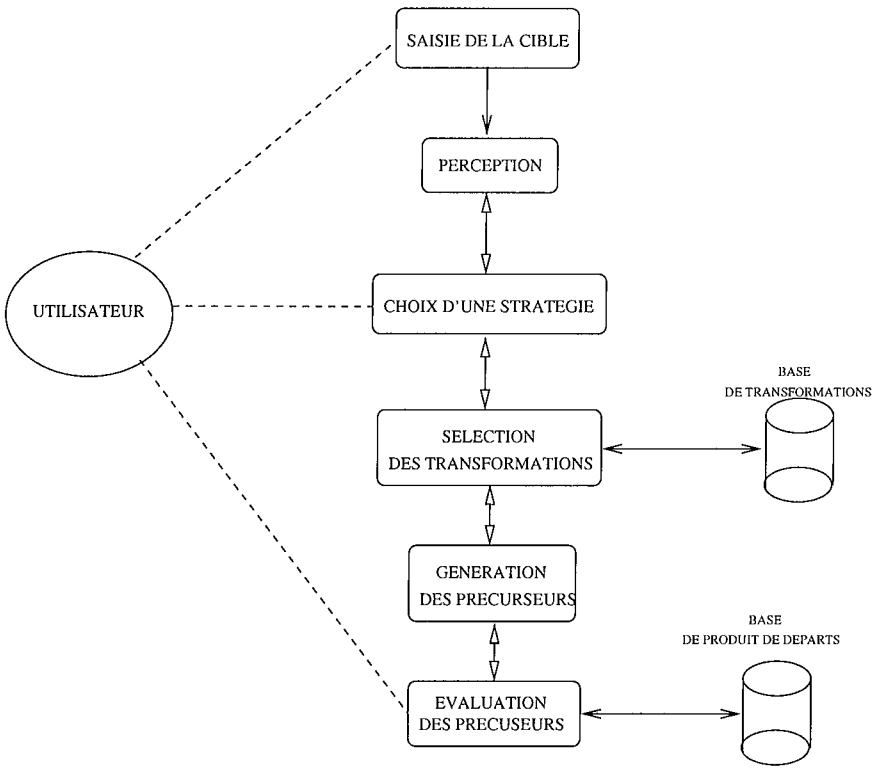


FIG. 3.5 – L'organisation des systèmes basés sur l'approche empirique. Les doubles flèches entre les modules du système indiquent l'enchaînement des modules en précisant les allers et retours possibles. Les traits en pointillés signalent les modules à partir desquels l'utilisateur peut éventuellement interagir avec le système.

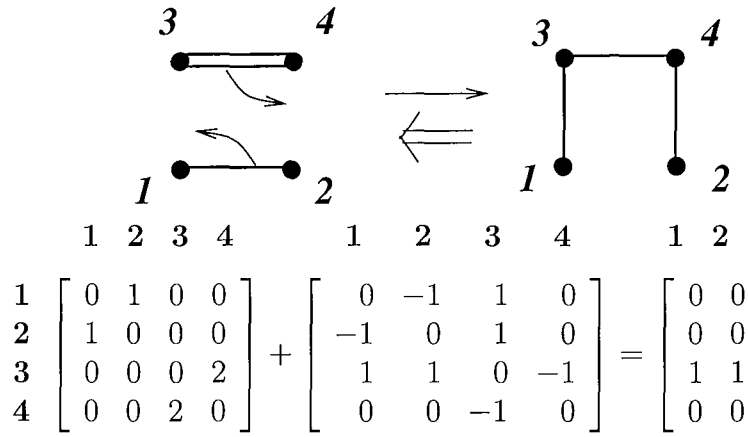
réalisés jusqu'à présent. Dans LHASA, le chimiste intervient dans le choix de la stratégie de résolution en sélectionnant parmi celles proposées par le système celle qu'il veut voir appliquer. SECS est basé sur le même principe que LHASA mais tenait à l'origine en plus compte de la stéréochimie. Le langage ALCHEM[20] a été développé à cette occasion. Depuis cet aspect a été également intégré dans LHASA. Dans ces logiciels, une interface graphique permet de saisir les molécules et de visualiser les résultats.

SYNCHEM[21] et SYNCHEM2 sont les résultats des travaux de l'équipe de Gelernter [Gelernter et al., 1990]. C'est un système uniquement automatique réduisant l'espace de recherche au moyen d'heuristiques basées sur la perception de « synthèmes » (groupements structuraux ou fonctionnels d'intérêt synthétique) et le calcul du « mérite »[22] des transformations. La recherche heuristique est effectuée grâce à un algorithme de type AO*.

[20] ALCHEM : A Language for CHEMistry.

[21] SYNCHEM : SYNthetic CHEMistry.

[22] Le mérite d'une transformation est un score qui est attribué à une transformation en fonction de critères tels que le rendement, la commodité et la fiabilité.



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix} \end{array} + \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

FIG. 3.6 – Un schéma de redistribution électronique ainsi que sa représentation selon l'équation fondamentale de Ugi.

3.3.2.2 L'approche formelle

Cette approche se base sur un formalisme mathématique des réactions développé par Ugi et Dugundji au début des années 70 [Ugi et al., 1994]. Le concept à l'origine de ce modèle est celui de l'extension de l'isomérie structurale à celle d'isomérie d'ensembles de molécules : « deux ensembles de molécules (EM) sont dits isomères s'ils ont la même formule brute et s'ils contiennent le même nombre total d'électrons de valence mais des distributions différentes de liaisons entre les atomes ». Dans le modèle d'Ugi-Dugundji, une réaction correspond à la conversion d'un EM en un EM isomère. Chaque EM étant représenté par une matrice symétrique, appelée *be-matrice*[23], la matrice représentative de la réaction, ou *r-matrice*, est obtenue par résolution de l'équation fondamentale du modèle d'Ugi-Dugundji :
B + R = E avec B : matrice de l'EM des réactants, E : matrice de l'EM des produits et R la r-matrice (figure 3.6). Dans le sens rétrosynthétique, il faut simplement soustraire la matrice de réaction choisie à la matrice de la cible pour obtenir la matrice de l'ensemble des précurseurs.

Les analyses synthétiques et rétrosynthétiques sont dirigées par le principe de changement de structure minimum. La règle heuristique employée impose donc une redistribution électronique, dite *distance chimique*[24], minimale. Reposant sur des fondements logiques et mathématiques, les schémas réactionnels obtenus ne correspondent pas forcément à des exemples réels de réaction mais uniquement à des schémas de mouvements électroniques possibles. Ne se limitant pas aux réactions connues, un système basé sur cette approche est donc susceptible de découvrir des plans de synthèse ou des réactions originaux.

L'équipe de Ugi a mis en application cette démarche formelle dans les systèmes RAIN[25]

[23]be pour bond/electron.

[24]distance chimique entre les EM B et E : $d(B,E) = \sum |b_{ij} - e_{ij}| = \sum |r_{ij}|$ avec b_{ij}, e_{ij} et r_{ij} éléments des matrices B,E et R respectivement.

[25]RAIN : Reactions And Intermediates Networks.

FIG. 3.7 – Deux molécules qui peuvent être analysées par SYNGEN. À gauche, la molécule d'acide lysergique. À droite, la molécule de vétivone.

[Ugi et al., 1994], IGOR[26] [Herges, 1988] et CICLOPS[27]. RAIN est capable d'effectuer une analyse dans les deux sens, synthétique et rétrosynthétique, ce qui lui permet de construire les réseaux de réactions conduisant d'un EM B à un EM E. RAIN ne prend pas en compte la stéréochimie. Créé à l'origine pour générer des réactions péricycliques, IGOR se révèle être plus polyvalent que ce pour quoi il était prévu car il est susceptible de prédire des réactions de divers types. D'autres systèmes tels que EROS[28], ASSOR[29], AHMOS[30] ont été aussi réalisés à partir de cette approche et génèrent des chemins de synthèse possibles entre deux ensembles de molécules.

Cependant, cette approche n'est pas satisfaisante du point de vue de la rétrosynthèse.

3.3.2.3 L'approche d'Hendrickson

L'approche systématique d'Hendrickson s'est matérialisée avec le système SYNGEN. Hendrickson emploie le terme de réaction à la place de celui de transformation et définit une réaction comme la combinaison de réactions unitaires ou *demi-réactions*. Une demi-réaction décrit les modifications affectant un atome de carbone à la fois. Au total, il suffit de 25 demi-réactions pour générer toutes les réactions de SYNGEN. Le problème de cette démarche est que le nombre de réponses fournies est très grand du fait de la simplicité de la stratégie de résolution employée et du fonctionnement automatique du système. Par exemple, pour des molécules telles que l'Acide lysergique ou le Vétivone (figure 3.7), SYNGEN génère respectivement 386 et 2068 possibilités de plan [SYNGEN, url].

3.3.2.4 L'approche des chirons

Hanessian a mis en œuvre son approche des chirons au niveau d'un système interactif appelé CHIRON[31]. Ce logiciel a pour objectif d'assister le chimiste en l'aidant à identifier des chirons dans la structure de la molécule cible. La recherche de chirons se fait dans une base

[26] IGOR : Interactive Generation of Organic Reactions.
[27] CICLOPS : Computer In Chemistry, Logic Oriented Planning Synthesis.
[28] EROS : Elaboration of Reactions for Organic Synthesis.
[29] ASSOR : Allgemeines SimulationsSystem Organischer Reaktionen.
[30] AHMOS : Automatic and Heuristic Modeling for Organic Chemistry.
[31] CHIRON : CHIRal synthON

de plus de 2000 composés et est dirigée par des paramètres donnés par l'utilisateur ainsi que des critères de recouvrement en terme de squelette, de fonctionnalité et de stéréochimie mais aussi de réactivité. En effet, le système est capable d'établir des scores à partir de règles chimiques telle que la proximité d'un groupement activant pour la création d'une liaison. S'il donne des résultats intéressants en mettant en évidence des relations entre une molécule cible et des produits de départ potentiels, CHIRON fournit souvent beaucoup trop de réponses qu'il est nécessaire de filtrer.

3.3.2.5 CAMEO

CAMEO[32] est un système de prédiction de réaction conçu par l'équipe de Jorgensen [Métivier, 1987]. Pour évaluer la faisabilité d'une réaction, le système, à la manière du chimiste, utilise de nombreux modèles et connaissances. Ainsi huit modules indépendants de traitement mécanistique sont implantés (réactions nucléophiles et baso-catalysées, réactions électrophiles et acido-catalysées, réactions impliquant des carbènes, réactions impliquant des radicaux libres, etc.). Après une perception (connectivité, type d'éléments, charge, aromaticité, fonctionnalité, symétrie) et la prédiction de certaines données physiques (pK_a, chaleur de formation, énergie de dissociation de liaison) des produits de départ, le système invite l'utilisateur à choisir un module parmi les huit disponibles. Bien que le système ne soit pas capable de choisir le module de départ de façon automatique, les modules sont susceptibles d'en utiliser d'autres au cours de la résolution du problème. Les résultats donnés par CAMEO sont intéressants et originaux car ils s'appuient sur des modèles mécanistiques. Cependant certains mécanismes restants peu connus, tous les modules ne sont pas performants.

3.3.2.6 WODCA

WODCA [33], qui a été élaboré à partir des années 90 par Ihlenfeldt et Gasteiger[Gasteiger et al., 1992], présente une philosophie un peu différente des systèmes précédents. En effet, Gasteiger a voulu rendre le système plus interactif et plus souple que ceux existant. Ceci pour rendre compte du cheminement du chimiste qui effectue des raisonnements parallèles et des révisions constantes au cours de la résolution d'un problème de synthèse. Le but principal du système est d'arriver à trouver des produits de départ acceptables. Pour cela plusieurs méthodes de recherche et fonctions d'évaluation sont proposées. Il est aussi possible de décomposer la cible en fragments en déconnectant des liaisons stratégiques, déterminées par l'utilisateur ou le système. Les différents fragments peuvent alors être individuellement pris pour cible. Ne possédant pas de base de transformations, WODCA n'indique pas à partir de quelles transformations les fragments peuvent être obtenus du point de vue tactique.

[32] CAMEO : Computer Assisted Mechanistic Evaluation of Organic reaction.
[33] WODCA : WODCA

3.3.3 Discussion

L'élaboration des systèmes que nous venons de présenter a permis de formaliser la synthèse organique ciblée. L'effort de conceptualisation d'un domaine aussi empirique est en lui même remarquable.

Cependant, contrairement aux bases de données de réactions qui ont été utilisables dès les années 80, les systèmes d'aide à la synthèse, quant à eux, n'ont jamais vraiment atteint un stade opérationnel. Ils sont restés à l'état de prototypes bien que quelques extensions industrielles aient été réalisées par exemple pour LHASA, SECS, etc.

Comme nous l'avons dit en introduction de ce chapitre, les divers systèmes d'aide à la résolution de synthèse qui ont été développés dès la fin des années 60 s'apparentent à des systèmes experts. Ce qui les rapproche des systèmes experts est le fait que les systèmes empiriques doivent disposer d'une base de connaissances. La construction et la mise à jour de cette base pose le même problème que pour les systèmes experts. Quant au point de vue informatique, ils en sont éloignés car ils reposent sur des méthodes procédurales et non des règles de production et un moteur d'inférence. Les implantations en terme de systèmes experts ont donné de mauvais résultats.

Quelle que soit l'approche suivie, les systèmes d'aide à la résolution que nous avons présentés ont été pour la plupart conçus par des chimistes selon des principes de conception aujourd'hui dépassés. Les méthodes informatiques et les langages procéduraux employés en rendent la maintenance lourde et difficile.

Dans les systèmes empiriques, les bases de transformations doivent être mises à jour régulièrement pour tenir compte de l'évolution des connaissances. Mais cette opération est fastidieuse. Enfin, toute modification d'ordre conceptuel des bases ou des programmes est trop coûteuse, voir irréalisable.

Les prototypes basés sur l'approche formelle d'Ugi ou semi-formelle d'Hendrickson ne rencontrent pas le problème de mise à jour des connaissances car ils ne nécessitent que peu de schémas pour décrire la totalité de ce qu'il est possible de faire. Cependant, pour les systèmes basés sur le modèle d'Ugi, la démarche adoptée fait qu'ils sont plus aptes à inventer de nouvelles réactions ou à expliciter des mécanismes qu'à aider à la rétrosynthèse.

De manière générale, les modules stratégiques sont insuffisamment développés pour permettre d'élaguer la recherche. L'arbre de rétrosynthèse obtenu est, dans la majeure partie des cas, beaucoup trop grand pour être exploré par l'utilisateur.

Pour finir, ces systèmes ne présentent pas de module d'explication des résultats. Il n'est pas possible de savoir comment la solution a été obtenue, c'est-à-dire quels cheminements ont permis d'y parvenir.

Les recherches dans le domaine de l'« intelligence artificielle » offrent de nos jours des méthodologies et des outils permettant la construction de systèmes évolués. À une conception monolithique ont succédé des approches dans lesquelles chaque connaissance est représentée de façon adéquate par rapport à l'usage que l'on veut en faire. Ces systèmes appelés *systèmes basés sur la connaissance* manipulent des connaissances représentées de diverses façons et disposent de différentes méthodes de raisonnement pour résoudre les problèmes. Nous allons présenter dans la section suivante le système basé sur la connaissance d'aide à

la compréhension de problèmes de synthèse multiétapes ciblée sur lequel nous travaillons.

3.3.4 Resyn Assistant : un système basé sur la connaissance d'aide à la compréhension de problèmes de synthèse

ReSyn Assistant est un prototype d'aide à la résolution de problèmes de synthèse développé à partir de 1997 dans le cadre du GdR 1093 du CNRS.

Créé en 1993 par le département des Sciences Chimiques du CNRS et dirigé par Claude Laurenço, le GdR 1093 Traitement Informatique de la Connaissance en Chimie Organique (Ticco) s'est terminé fin 2000. Le GdR Ticco a constitué un groupe de recherche interdisciplinaire permettant un dialogue entre spécialistes de la synthèse, informaticiens et modélisateurs de la connaissance en chimie, industriels aussi bien qu'universitaires. Cette remarque est importante à plus d'un titre car jusque là aucune équipe de recherche dans le domaine n'avait mis en jeu des compétences aussi diverses. Ce type de structure paraît pourtant incontournable dans la conception de systèmes d'information appliqués à un domaine aussi complexe que celui de la synthèse multiétapes ciblée. Dès la création du groupement de recherche, les travaux du GdR 1093 Ticco ont été soutenus par les industries pharmaceutiques telles que Roussel Uclaf, Sanofi-Synthelabo, Servier. Ils ont impliqué plusieurs laboratoires de recherche publics en chimie (Ccipe[34]) et en informatique (Lirmm[35], Loria[36]) ainsi que des sociétés de service informatique telles que Sema Group et Framentech-Cognitech. Précédent la création du GdR 1093 Ticco, une collaboration scientifique a débuté dès 1990 entre une société de l'industrie pharmaceutique, Roussel Uclaf, une société de service informatique spécialisée en intelligence artificielle, Framentech-Cognitech, et deux laboratoires de recherche publics, le Lirmm et le Ccipe. Elle a eu pour projet de construire un prototype d'aide à l'élaboration de plans de synthèse tenant compte de contraintes industrielles et de connaissances expertes. Les motivations premières de la société Roussel Uclaf, à l'origine de ce projet, était la récupération, la transcription et la conservation des connaissances d'un expert reconnu de la synthèse, Edmond Toromanoff. Le prototype, ReSyn[37], fruit de cette collaboration, a ensuite servi de plate-forme de développement pour les travaux du GdR 1093 Ticco [Vismara et al., 1998]. Ce prototype, basé sur la rétrosynthèse, a permis, à terme, de construire de façon interactive ou automatique des plans de synthèse. La construction et la mise à jour des bases de connaissances nécessaires est facilitée par la modélisation des objets chimiques sous forme de graphes. Dans le système, la connaissance chimique est représentée à l'aide d'un langage de frames très évolué (Y3 [Ducournau, 1989]). Dans cette conception orientée objet [Napoli and Laurenço, 1993; Laurenço et al., 1990], l'utilisation du raisonnement par classification permet à la fois la mise à jour des connaissances et la détermination de l'applicabilité d'une transformation. L'évolution et l'ajout des connais-

[34]Ccipe : Centre CNRS-INSERM de Pharmacologie et d'Endocrinologie.

[35]Lirmm : Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier.

[36]Loria : Laboratoire LOrrain d'Informatique et ses Applications.

[37]ReSyn pour **Re**tro**Syn**thèse.

sances sont facilités par une représentation éclatée des transformations.

En 1997, un nouveau projet de système d'aide à la compréhension de problèmes de synthèse a été initié par la société SANOFI-SYNTHELABO. Pour des raisons diverses parmi lesquelles on peut citer la portabilité et la maintenance, ce deuxième prototype RESYN ASSISTANT[38] a été développé dans le langage JAVA. La conception de RESYN ASSISTANT a bénéficié des travaux effectués dans le cadre du GDR. Parmi ceux-ci, l'étude des modes de raisonnement employés par les spécialistes de la synthèse a conduit au développement de systèmes basés sur les principes dégagés : l'analogie [Py, 1992], le raisonnement distribué [Jambaud, 1996], le raisonnement à partir de cas [Lieber, 1997; Lieber and Napoli, 2000]. Des développements à partir d'algorithmique de graphes [Vismara, 1995] ou de satisfaction de contraintes [Régin, 1995; Tognetti and Vismara, 2000] facilitent et améliorent la manipulation des objets chimiques. La modélisation des connaissances dans le cadre de la synthèse multiétapes ciblée est l'objectif principal de ces travaux [Gien, 1998].

L'ambition du GDR 1093 TICCO est la conception de nouveaux systèmes d'information chimique intégrant à la fois des fonctionnalités de bases de données et celle des systèmes basés sur la connaissance. Par définition, le système basé sur la connaissance RESYN ASSISTANT manipule différents types de connaissances (connaissances des objets du domaine, connaissances stratégiques etc.), de représentations et de méthodes de résolution de problèmes. Les travaux du GDR TICCO profitent d'un environnement favorable à l'échange scientifique et à l'exploitation des techniques avancées de l'informatique actuelle. Les connaissances et les graphes moléculaires sont représentés à divers niveaux d'abstraction [Vismara and Laurenço, 2000].

Nécessairement, et comme son nom l'indique, RESYN ASSISTANT a pour finalité d'aider le chimiste dans sa tâche. Le système est donc fortement interactif et n'impose aucun raisonnement ni cheminement de résolution prédéfini mais doit proposer les résultats de son analyse et expliciter les solutions obtenues grâce à ses connaissances. Ceci pallie la relative « mécanisation du raisonnement » qui est le propre des systèmes existants.

Le tableau 3.2 résume les différences entre notre approche et celle des systèmes présentés section 3.3. Le choix des critères de comparaison présentés ne prétend pas à l'exhaustivité. Cependant, il représente les points importants de différence qui existent généralement entre les systèmes procéduraux et les systèmes basés sur la connaissance.

[38] RESYN ASSISTANT a été l'objet de trois années de travail de Pierre Jambaud dans le cadre d'un contrat postdoctoral financé par SANOFI-SYNTHELABO ainsi que la région Languedoc-Rousillon.

	SYSTÈMES EXISTANTS	RESYN ASSISTANT
type de système	système procédural	système basé sur la connaissance
connaissances	transformations programmées	très variées : objets, stratégies de résolution, heuristiques, raisonnements
organisation des connaissances	un seul niveau	modèles, hiérarchies, abstractions
raisonnements	déduction	très variés : classification, analogie, etc.
langage de programmation	langage procédural	langage orienté objet
interface avec l'utilisateur	faible interactivité ou automatisation	forte interactivité

TAB. 3.2 – Comparaison de RESYN ASSISTANT avec les systèmes existants.

Deuxième partie

L'acquisition des connaissances

Chapitre 4

La modélisation et la conception d'un système basé sur la connaissance

One of the few hard and fast results to come out of the first three decades of AI research is that intelligence requires knowledge.

E. Rich and K. Knight

Sommaire

4.1 Introduction

Dès le premier chapitre de [Rich and Knight, 1991], il est dit que pour raisonner il faut des connaissances. Ceci peut paraître évident mais n'a pas été pris en compte dès les débuts de l'*Intelligence Artificielle*. Il y a cinquante ans, les problèmes qui étaient résolus par les systèmes informatiques ne requéraient pas réellement d'intelligence mais uniquement la

puissance de calcul que les calculateurs étaient capables de fournir. À cette époque, on ne s'était pas encore vraiment préoccupé de la simulation de « comportement intelligent » et du raisonnement proprement dit pour résoudre un problème. « L'intelligence » des ordinateurs résidait alors dans le volume de calculs effectués et la rapidité avec laquelle ils étaient capables de donner des réponses par rapport aux performances humaines. Depuis, il est reconnu que la capacité de calcul ne fait pas tout et qu'elle n'est qu'une « illusion de l'intelligence ». Les connaissances jouent un rôle majeur dans la résolution d'un problème [David, 1995].

À partir du milieu des années 60, le projet de construire des systèmes dont l'ambition était d'atteindre les performances d'un expert face à un problème donné a pris forme. Devant l'accroissement de la difficulté des problèmes abordés, la philosophie de l'intelligence artificielle a alors évolué. Elle ne correspondait plus à trouver une technique générale voir universelle de résolution de problèmes (comme avec le GPS[1] de Simon et Newell) mais plutôt à se concentrer sur le domaine d'application du problème pour en découvrir les mécanismes propres [Kayser, 1997]. Pour cela, le programme doit compiler les connaissances du domaine nécessaires à la résolution du problème. La problématique de l'acquisition des connaissances est donc apparue avec celle de la conception des « systèmes experts de première génération » (SE1G). Ces systèmes ont été et seront appelés par la suite « systèmes à base de règles » (SBR) car leur fondement repose sur la manipulation d'une base de règles de production, c'est-à-dire dire des règles de la forme « si *condition* alors *conclusion* ou *action* ». Au vu des problèmes dus aux principes sur lesquels ceux-ci reposaient, une nouvelle approche a abouti à l'émergence de « systèmes à base de connaissances » (SBC) appelés aussi « systèmes experts de seconde génération » (SE2G) [David, 1995]. Le terme de « système à base de connaissances » (SBC) est souvent employé dans la littérature pour désigner les systèmes experts quelque soit leur type [David, 1995]. Tout comme [Charlet et al., 2000], nous réservons, pour notre part, ce terme à celui des systèmes experts de seconde génération. Dans cette seconde approche, l'objectif des systèmes intelligents n'est plus tant de simuler un comportement humain mais plutôt celui de l'assister dans ses tâches les plus complexes en suivant un raisonnement. D'un aspect compétitif et reproductif, on en vient à des notions d'aide et d'adaptation. Ainsi les systèmes deviennent coopératifs et interactifs.

Un des premiers systèmes experts fut le célèbre DENDRAL dont le but était d'aider les chimistes dans l'élucidation de structures en chimie organique à partir de spectres de masse et de résonance magnétique. Aujourd'hui, les problèmes traités par les systèmes experts couvrent de nombreux domaines d'application (aéronautique, agriculture, milieux financier et administratif, biologie, chimie, électronique, industrie, etc.), ceci pour divers types de tâches (diagnostic, conseil, planification, conception) [Stefik, 1995]. Certains systèmes sont devenus des produits opérationnels, en particulier dans les tâches de diagnostic, d'interprétation et de surveillance.

Depuis les premiers travaux sur les systèmes experts, de nombreux progrès ont été faits dans la construction de tels logiciels. Des méthodologies ont été développées et per-

[1] GPS : General Problem Solver

mettent de guider la conception des systèmes à bases de connaissances. Devant l'évolution des techniques développées et des courants de pensée, le terme même d'« acquisition de connaissance » s'est fait supplanté par celui d'« ingénierie des connaissances » (*knowledge engineering* [Charlet et al., 2000]).

Nous allons voir dans ce qui suit les problèmes soulevés par l'acquisition de connaissance ainsi que les solutions qui y ont été apportées.

Nous nous intéressons dans ce chapitre à la conception des SBC. Après avoir vu la définition de ceux-ci et le déroulement général du processus, nous nous focaliserons sur une étape primordiale : la modélisation. Nous parlerons ensuite du problème de l'alimentation de la base de connaissances.

Dire que l'acquisition des connaissances est le goulet d'étranglement de la construction des systèmes à base de connaissances est devenu un « cliché » [Musen, 1993]. Cependant, il n'en est pas moins vrai que cette étape est une phase déterminante de la conception des bases de connaissances sur lesquelles reposent les SBC. La qualité du système dépendra essentiellement de celle des connaissances possédées et manipulées.

4.2 L'architecture d'un système à base de connaissances

Les SBC trouvent leur utilité au niveau de la résolution de problèmes qui n'ont pas de solution algorithmique simple connue. Pour résoudre ce type de problèmes, un système de résolution de problèmes doit s'appuyer sur un vaste ensemble de connaissances. Celles-ci sont manipulées à l'aide de méthodes de raisonnements pour produire de nouvelles connaissances. Pour être efficace, il est souhaitable que le processus de résolution de problème soit guidé par l'utilisation de stratégies.

4.2.1 L'approche au niveau connaissance

Le coeur d'un SBR ou d'un SBC est la base de connaissances qu'il possède et l'usage qu'il en fait pour produire une solution. Cependant, les SBC sont basés sur une toute autre hypothèse que les SBR. La multiplicité et l'hétérogénéité des connaissances et représentations des SBC se substituent à l'unicité du formalisme des SBR. Considérer les connaissances par niveaux de généralité, comme le prône Newell avec l'approche au *knowledge level*, s'oppose à la démarche, adoptée dans la conception des SBR qui consiste à introduire les connaissances de façon non organisée au fur et à mesure qu'elles sont acquises. Les travaux sur les SBR ont eu le mérite de faire émerger le fait que les connaissances doivent être séparées du moteur d'inférence du système. Le « niveau connaissance » implique la description des connaissances à des niveaux d'abstraction adéquats. Ce paradigme encourage une approche descendante dont la première préoccupation est l'élaboration d'un *modèle conceptuel*. Ce modèle doit être réfléchi en toute indépendance par rapport aux contraintes de représentation et d'opérationnalisation. Dans cette approche, la conception d'un système à base de connaissances se déroule en plusieurs temps bien distincts (figure 4.1). Le processus d'acquisition des connaissances couvre les activités de modélisation et d'élicitation des

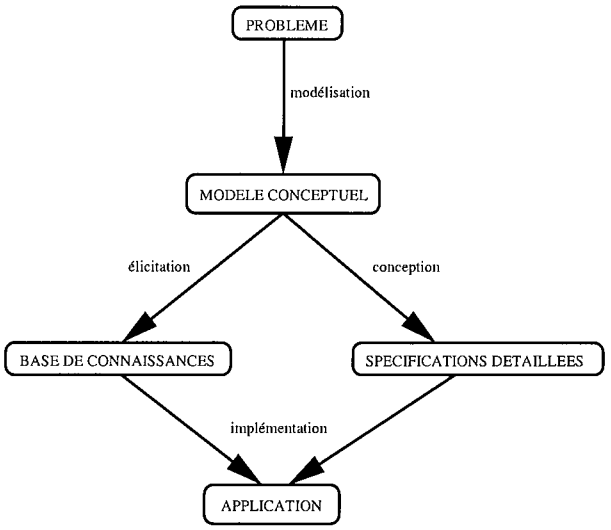


FIG. 4.1 – La conception d'un système à base de connaissances au « niveau connaissance » (d'après Newell).

connaissances. La construction du modèle implique une étude préalable des objectifs du système. Elle permet de conduire d'une part l'élicitation des connaissances qui permettra d'enrichir la base de connaissances et d'autre part de spécifier les caractéristiques et fonctionnalités du système. Le système est mis en œuvre par implémentation des spécifications établies et de la base de connaissances.

4.2.2 Buts et typologie des problèmes

« Un système est conçu pour une pratique »[Charlet et al., 2000]. La construction d'un SBC, comme tout autre système, correspond à l'expression d'un besoin. Cette attente est celle de la résolution de problèmes complexes dans un domaine peu formalisé. La définition des objectifs du système permet de mettre en évidence les caractéristiques générales du problème ainsi que de choisir les méthodes de résolution à employer. On ne modélise jamais un domaine dans l'absolu, sans la définition d'un traitement, d'une utilisation particulière. On modélise en vue de concevoir un système qui doit atteindre des objectifs bien déterminés. Identifier le type de problèmes à résoudre peut être une première avancée. La méthode KADS dont il va être question en section 4.3.2.1 propose une typologie des tâches requérant de la connaissance (figure 4.2). Celle-ci distingue les problèmes d'analyse de ceux de synthèse. Le suivi de la méthodologie KADS implique la description du modèle d'organisation avant tout autre chose [Speel et al., 2001]. Même s'il est très informel, ce type de modèle permet de décrire le contexte général du problème et est utile à la compréhension du problème dont nous en avons déjà parlé section 2.1.3.

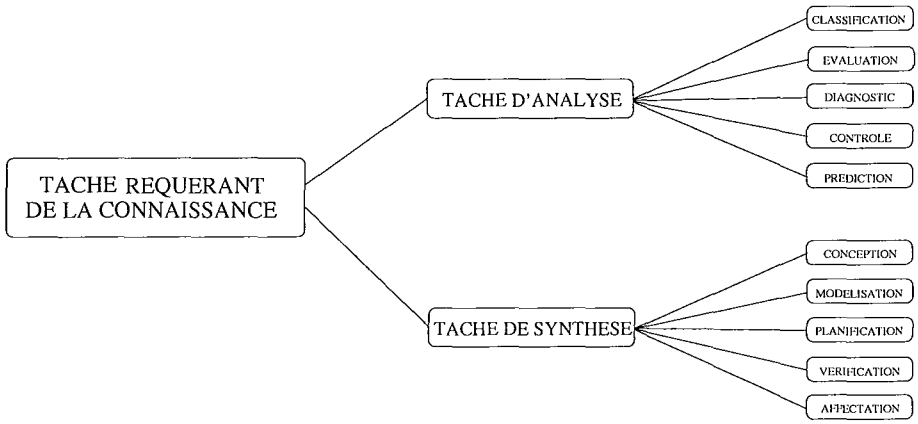


FIG. 4.2 – La typologie des problèmes requérant de la connaissance selon KADS.

4.3 La modélisation

Établir un *modèle* est une étape primordiale dans la construction de systèmes à base de connaissances. La modélisation, contrairement au transfert d'expertise, est une activité essentiellement et nécessairement créative [Musen, 1993]. Dans ce qui suit, nous nous sommes inspirés de la présentation de la problématique de la modélisation qui est donnée dans [Speel et al., 2001].

4.3.1 Le modèle conceptuel

Un *modèle conceptuel* est une vue et une interprétation d'un domaine d'étude donnant une spécification possible des connaissances de ce domaine. Le modèle conceptuel permet la représentation du monde[2] au « niveau connaissance » introduit par Newell en 1982 [Newell, 1982]. Le modèle doit s'abstraire de toute considération de représentation sur l'implémentation ultérieure. Il fait la liaison entre les connaissances expertes et le système de résolution de problèmes à proprement parler.

Dans ce cadre, le système est décrit par un agent de résolution de problèmes dont il faut expliciter le comportement. L'agent a un ou plusieurs buts, des connaissances et des actions pour les atteindre. Il doit choisir parmi celles-ci ce qu'il doit faire pour résoudre son problème. Le modèle à concevoir doit rendre compte du comportement de l'agent. Il n'existe pas de modèle supérieur à d'autres dans l'absolu. Étant donné que la réalité est très souvent trop complexe pour pouvoir être décrite dans ses moindres détails, un modèle est toujours une approximation. Le défi posé par la modélisation est de se rapprocher de la réalité tout en restant le plus simple possible. Le modèle conceptuel correspond à une vue abstraite du domaine mais il est difficile de construire un modèle conceptuel unique prenant en compte toutes les caractéristiques d'un problème. Bien souvent un modèle est établi selon un point de vue particulier. Il sera alors nécessaire de faire coexister différents

[2]restreint au domaine d'application

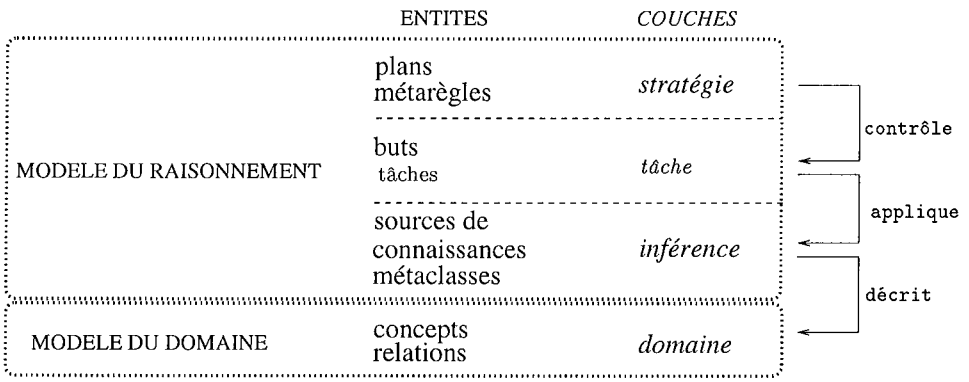


FIG. 4.3 – Le modèle conceptuel : un modèle en couches.

modèles à la fois pour tenir compte de tous les aspects et contextes du problème.

Pour résoudre un problème, il faut différentes connaissances. Toutes les connaissances n'ont pas le même rôle et l'on s'accorde à les séparer en couches [Newell, 1982]. Tout d'abord se trouve la couche des connaissances du domaine, puis celle des connaissances stratégiques pour déterminer ce qu'il faut faire, celle des connaissances tactiques pour choisir comment le faire et enfin celle des connaissances opérationnelles qui permettent d'exécuter les actions sélectionnées par les deux types de connaissance précédents. Ces trois derniers types de connaissances seront présents dans la couche résolution, c'est-à-dire la couche qui englobe les connaissances des méthodes de raisonnements. En conséquence, le modèle conceptuel se décline en quatre couches comme cela est décrit sur le schéma 4.3 inspiré de la méthode KADS.

4.3.1.1 Le modèle du domaine

Le modèle du domaine regroupe les concepts relatifs au domaine ainsi que les relations entre ces concepts. Il s'agit d'identifier la structure des types de connaissances impliquées dans la résolution d'un problème du domaine. Les constructions d'*ontologie* ou de *thésaurus* correspondent à l'activité de conception du modèle du domaine. Dans le domaine de l'acquisition des connaissances, l'acception du terme « ontologie » couvre les schémas qui contiennent et mettent en valeur la compréhension du vocabulaire spécifique au domaine auquel est rattachée l'ontologie [Chandrasekaran et al., 1999]. Une ontologie regroupe donc les spécifications explicites des termes et leur signification dans le domaine d'application. Comme c'est pratiquement toujours le cas, une ontologie se présente sous la forme une hiérarchie de concepts. La construction d'une ontologie est en conséquence à rapprocher et associer à la conceptualisation des connaissances du domaine. Différents outils ont été réalisés pour assister les modélisateurs dans cette tâche particulière. L'utilité des ontologies et thésaurus est multiple. D'une part, leur élaboration rentre dans le cadre de la modélisation du domaine. D'autre part, l'utilisation d'ontologie et de thésaurus permet d'obtenir un consensus entre les personnes participant au projet en leur fournissant les bases d'un vocabulaire commun. Ceci est d'autant plus important dans les domaines d'application

réels et empiriques.

4.3.1.2 Les modèles de raisonnement

Le modèle de raisonnement correspond à l'expression de trois types de connaissances :
- les inférences,
- les tâches,
- les méthodes.

Chacune de ces connaissances présente un niveau de granularité différent.

Les inférences sont les entités élémentaires du raisonnement. Elles sont faites sur les connaissances du domaine.

Les tâches répondent à un but et décrivent les inférences à combiner pour atteindre ce but. On peut avoir différents niveaux de tâches et une tâche est susceptible d'être décomposée en sous-tâches.

Les méthodes, enfin, représentent le niveau qui consiste à expliquer comment contrôler l'enchaînement des tâches ou réagir en cas d'échec pour pouvoir résoudre le problème.

Les tâches et les tâches génériques

Mettant en évidence le fait que des problèmes de nature différente peuvent faire appel à des méthodes de résolution identiques, Chandrasekaran propose d'isoler un certain nombre de primitives de raisonnement ni trop générales, ni trop spécifiques [Chandrasekaran and Johnson, 1993]. Ces tâches génériques pourront être employées en tant que briques de base (*building blocks*) au niveau de plusieurs problèmes n'ayant a priori pas de point commun. Chandrasekaran a sélectionné une dizaine de ces structures récurrentes parmi lesquelles on trouve la classification (plus précisément la classification hiérarchique), l'abduction, l'abstraction de données, la conception de routine, l'appariement d'hypothèse, etc [David, 1995]. Ce découpage est à rapprocher de la typologie des tâches (figure 4.2) donnée dans KADS.

Quelques méthodes de résolution

L'approche des méthodes à rôle déterminé (*role limiting methods*) due à Mc Dermott se base sur l'identification de méthodes de résolution de classes de problèmes [McDermott, 1988]. Au niveau des méthodes mises en évidence, les connaissances doivent remplir des rôles définis. Ainsi les différentes méthodes telles que *cover-and-differenciate*, *propose-and-revise*, *acquire-and-present*, *extrapolate-from-a-similar-case* etc. constituent un ensemble de techniques qui pourront être utilisées dans l'analyse de la résolution du problème. Mc Dermott s'est appliqué à construire des systèmes à base de connaissances mettant en oeuvre une des méthodes abstraites qu'il a définies. Par exemple, dans [Kahn et al., 1985], est décrit un système d'aide à l'acquisition de connaissances dans le cas de tâches de diagnostic : MORE. La stratégie employée est celle de la classification hiérarchique.

4.3.1.3 Les modélisateurs

La plupart du temps dans la littérature, l'activité de modélisation est réservée à l'« ingénieur des connaissances » (appelé aussi « cogniticien »), c'est-à-dire à un expert des sciences cognitives. Dans [Kahn et al., 1985], il est même suggéré que celui-ci doit acquérir une certaine maîtrise du domaine d'application qui lui permette de savoir quel type de connaissances est nécessaire.

Sans nier l'importance de l'intervention de cogniticiens, nous envisageons une modélisation menée par un ensemble d'experts particuliers du domaine, que Vogel appelle « des spécialistes » [Vogel, 1988]. Un spécialiste est un expert qui a une vue théorique de sa compétence et qui a fait une réflexion sur son domaine. Grâce à leur étude des connaissances profondes, les spécialistes sont donc capables de se faire une vue adaptée de la manière d'aborder un problème et d'en rendre compte en construisant un modèle. Les spécialistes sont susceptibles de faire la part entre les concepts (abstraction) et les faits (expérience). Cette approche a largement été employée dans le cadre du GDR TICCO [Laurenço, 1998]. Les réunions d'un groupe de réflexion « Synthèse-rétrosynthèse », rassemblant experts et modélisateurs, ont permis de mettre en évidence des connaissances stratégiques sur la synthèse organique. Au cours de ces rencontres, les démarches des experts se révélaient au travers de mises en situation de résolution de problèmes de synthèse. Une des caractéristiques supplémentaires des modélisateurs est d'avoir une perception du domaine informatique assez précise pour savoir ce qui est « faisable » et permettre une collaboration fructueuse avec les informaticiens [Napoli, 1992].

Qu'ils soient cogniticiens ou spécialistes du domaine, les modélisateurs ont à leur disposition toute une batterie d'outils plus ou moins formels pour les aider dans leur tâche de construction de base de connaissances.

4.3.2 Méthodologies et langages de modélisation

Ces dix dernières années, l'acquisition des connaissances a été un domaine particulièrement fertile en développements d'outils méthodologiques [Schreiber et al., 2000]. Ceux-ci visent à assister les concepteurs de systèmes à base de connaissances dans les tâches ardues de modélisation mais aussi d'élicitation de connaissances. Nous n'allons pas dans ce qui suit faire un état de l'art de ces systèmes mais donner quelques indications sur les plus représentatifs. Les différentes méthodologies de modélisation offrent un cadre systématique et structurant au développement des systèmes à base de connaissances et visent à proposer un support de la phase exploratoire des interviews à la finalisation du système opérationnel. La plupart des méthodologies mettent en avant l'utilisation de méthodes génériques au niveau d'une approche descendante (KADS, etc.) tandis que d'autres comme KOD implique une méthode ascendante.

4.3.2.1 KADS, CommonKADS et PROTÉGÉ

KADS, KADSII et CommonKADS

KADS est très certainement la méthodologie la plus employée. C'est une méthode manuelle et informelle[3]. Elle préconise la succession de construction de modèles : modèle d'organisation, modèle de connaissances, modèle de communication. Le modèle d'expertise [David, 1995] ou modèle de connaissances est la description du processus de résolution d'un agent. Dans KADS, trois niveaux de connaissances sont nécessaires pour décrire le comportement d'un agent de résolution de problèmes :

- Le niveau du « domaine » : c'est celui des concepts du domaine et des relations entre les concepts.
- Le niveau des « structures d'inférence » : une inférence représente le plus bas niveau de connaissance.
- Le niveau des « tâches » : une tâche est définie par un but et l'enchaînement d'inférences permettant d'atteindre ce but.

KADSII et CommonKADS se différencient de KADS par l'ajout d'un quatrième niveau : le niveau stratégique des méthodes qui rendent compte de l'enchaînement des tâches. Avec cette extension, on retrouve les deux modèles de connaissances que sont le modèle du domaine (le premier niveau) et le modèle de raisonnement (les trois autres niveaux) présentés en section 4.3.1. KADS propose de choisir les tâches parmi un ensemble prédéfini.

PROTÉGÉ-2000

PROTÉGÉ [PROTÉGÉ, url] est un outil d'aide à la construction de modèles conceptuels. Ce logiciel permet notamment de construire des ontologies. Pour cela, l'utilisateur a à sa disposition des méta-connaissances qu'il utilise pour construire son propre modèle. Ce modèle peut ensuite être instancié par l'intermédiaire de formulaires qui sont automatiquement générés à partir des définitions des classes [Noy et al., 2000]. Dans [Schreiber et al., 2000], Schreiber et col. présentent une étude de cas d'utilisation de PROTÉGÉ-2000 comme outil d'aide à la construction de modèles conceptuels selon la méthodologie de KADS. Cette expérimentation a montré que PROTÉGÉ-2000 pouvait générer des éditeurs d'interfaces proches de la représentation adoptée dans la méthodologie de CommonKADS. PROTÉGÉ-2000 est un des rares outils de ce genre. Il est très utilisé de par le monde et est en constante évolution. Il suffit de voir le nombre croissant d'utilisateurs du système ainsi que la quantité de messages échangés quotidiennement sur la liste de diffusion pour s'en convaincre.

4.3.2.2 UML

Le modèle conceptuel est le lien entre l'expertise et le système de résolution de problèmes comme nous l'avons dit en section 4.3.1. Pour cela, il faut que le modèle soit compréhensible et que l'expert puisse se l'approprier. Le modèle doit être descriptible dans un langage suffisamment et facilement accessible. Les langages de modélisation apportent une solution à ce problème.

[3] bien que dans la dernière version il y ait une association avec UML, un langage de modélisation que nous présentons brièvement dans la section 4.3.2.2

Vu l'importance de la modélisation, des langages servant à représenter le modèle de façon homogène, standard et facilement compréhensible pour un non informaticien ont été adoptés. Grâce à ces langages, les modèles deviennent accessibles à tous les membres participant au projet. Parmi ceux-ci, le langage de modélisation objet unifié (UML pour Unified Modeling Language) se distingue et est devenu en 1997 une norme OMG[4] [Object Management Group, url; Rational, url]. C'est un langage pour la visualisation, la spécification, la construction et la documentation de modèles. UML permet d'élaborer des modèles objet indépendamment du langage de programmation objet qui sera employé par la suite. C'est un langage formel défini par un méta-modèle. Issu du paradigme objet, il sert de support à l'analyse basée sur les concepts en offrant des moyens de représentation de ceux-ci, de leurs associations, de leurs propriétés, etc. UML définit des vues qui autorisent la description de chacun des aspects d'un système [Booch et al., 1999]. Chaque vue du modèle est représentée sous forme graphique par un diagramme. Il existe neuf types de diagramme. Ceux-ci se divisent en deux sous-ensembles. Les diagrammes statiques sont au nombre de cinq : le diagramme de classes, le diagramme des objets, le diagramme de cas d'utilisation, le diagramme de composants et le diagramme de déploiement. Les représentations dynamiques sont exhibées dans les diagrammes de séquence, de collaboration, d'états ou d'activité. UML permet donc à la fois de représenter les points de vue structuraux et comportementaux relatifs au domaine d'application.

Mettre un expert en situation de résolution de problèmes conduit à dégager la structure du processus de raisonnement. Dans [Vilain et al., 2000a; Vilain et al., 2000b], il est proposé notamment un outil pour représenter l'interaction entre utilisateur et système (appelé UID pour User Interaction Diagram) qui est une extension du diagramme des cas d'usages présents dans UML. La technique des cas d'expérience est très intéressante. L'observation de l'expert dans son travail de résolution de problèmes est pleine d'enseignements.

4.3.2.3 Autres méthodes

KOD

KOD[5] [Vogel, 1988] présente une philosophie différente de celle adoptée par les démarches méthodologiques que l'on vient de présenter. En effet, au suivi de procédés génériques et prédéfinis, KOD oppose une approche ascendante dans laquelle la première phase est une collecte non dirigée d'un maximum de connaissances. Les données verbales récupérées sont par la suite organisées [Duribreux-Cocquebert and Houriez, 2000]. Les modèles obtenus par ce processus « ad hoc » n'ont pas le caractère générique de ceux produits par les méthodologies telles que KADS.

Approche mixte

Dans [Duribreux-Cocquebert and Houriez, 2000] et [Martin-Mattei and Ricard, 1992], les auteurs expérimentent une approche mixte de modélisation de connaissances en combinant

[4]OMG : Object Management Group

[5]KOD : Knowledge Oriented Design.

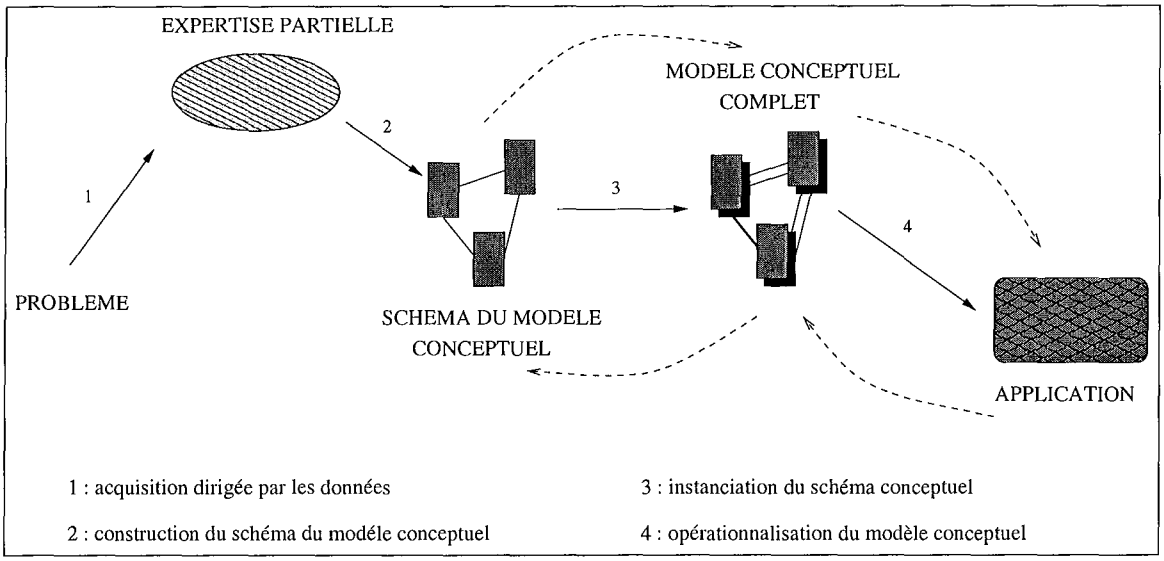


FIG. 4.4 – La construction d'un système à base de connaissances par une approche mixte d'après [Aussenac-Gilles et al., 1992].

une approche descendante de type KADS et une approche ascendante de type KOD. Le choix du modèle dans la phase de démarche descendante est orienté par une première étape de recueil partiel de connaissances selon l'approche ascendante. Cette démarche mixte conduit à un nouveau processus d'acquisition des connaissances décrit sur la figure 4.4. Le processus débute par une phase initiale d'acquisition de connaissances dirigée par les données. Elle consiste à spécifier les données utiles à la résolution de problèmes, soit par analyse du discours libre de l'expert (méthode KOD), soit par exploitation d'un recueil dirigé de l'expertise (méthode KADS). À partir des résultats de cette étape, il est possible de construire un premier schéma du modèle conceptuel. L'instanciation du schéma conceptuel conduit ensuite à la construction du modèle conceptuel complet. Le schéma offre un cadre structurant au recueil d'expertise car il met en relief le type et le rôle des connaissances à acquérir.

L'implémentation, ensuite, a pour but de créer un système opérationnel qui permet d'expérimenter le modèle dans un cadre de résolution effective de problèmes. À tout moment, l'exécution d'une étape est susceptible de révéler des défauts que l'on corrige en revenant à l'étape en cause. La construction d'une base de connaissances est alors vue comme un processus itératif permettant de réviser et raffiner progressivement le modèle. Des allers et retours sont possibles et conseillés entre le schéma conceptuel, le modèle complet et l'expérimentation au travers de l'application réelle.

4.4 Le transfert d'expertise

Les bases de connaissances sont construites à partir d'un dialogue avec les experts du domaine. Mais la récupération des connaissances expertes n'est pas chose aisée. Le problème

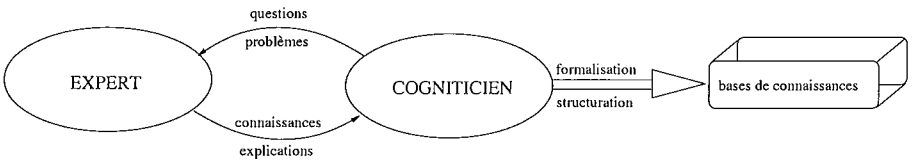


FIG. 4.5 – Le processus de transfert d'expertise.

se pose aussi bien au niveau des SBR que des SBC.

4.4.1 Le transfert d'expertise dans les SBR

À l'origine, l'apport de connaissances dans les SBR se présente sous la forme d'un transfert d'expertise (figure 4.5). Celui-ci consiste à essayer d'extraire et transférer les connaissances de l'expert, connaissances qui seront ensuite organisées de façon à être introduites dans le programme. Le transfert d'expertise implique l'intervention de deux personnes : l'expert, bien entendu, et le cogniticien. La première phase du processus consiste en un dialogue entre l'expert et le cogniticien pendant lequel l'expert est invité à exprimer ses connaissances, à savoir les concepts, le savoir-faire ainsi que les heuristiques de raisonnement. Dans un deuxième temps, le cogniticien a pour tâche de représenter de façon formelle les connaissances exprimées de façon à ce qu'elles soient utilisables par le SBR. Dans certains cas, le cogniticien s'aide de fiches ou de formulaires. Les connaissances formalisées viennent enrichir la base de connaissances, qui est une base de règles dans les SBR. Pour tenir compte des faits ou informations incertains, l'expert a quelquefois la possibilité d'associer à la règle énoncée un facteur de confiance.

Un SBR repose sur un domaine de connaissance et sa qualité dépend de celle des connaissances. Plusieurs critères peuvent aider à mesurer de façon qualitative la performance d'un tel système [David, 1995]. Il s'agit de la facilité d'acquisition de connaissances, de l'existence d'un module d'explication correct, de la robustesse du système face aux changements et enfin de la réutilisabilité de ce qui a été produit dans le but de répondre à de nouveaux besoins. L'évaluation de ces diverses caractéristiques et fonctionnalités fait ressortir certains problèmes.

- Acquisition des connaissances : celle-ci constitue le fameux « goulet d'étranglement » étant données les difficultés liées au transfert d'expertise.
- Explication : cette fonctionnalité, pourtant essentielle, est très limitée. Elle se résume à la retranscription de l'enchaînement de règles qui ont permis d'inférer la solution. L'explication obtenue est très mécanique (chaînage avant/arrière) et non naturelle car on ne peut accéder aux connaissances implicites nécessaires.
- Robustesse : la mise à jour de la base de règles peut avoir des répercussions non souhaitables et imprévisibles, ce qui rend l'évolution ou la correction de celle-ci difficiles. De plus, les SBR sont fragiles quand il s'agit d'aborder des problèmes à la frontière de leur domaine de compétence. Un problème lié est celui de l'arrêt du système par manque d'information.

- Réutilisabilité : les connaissances étant exprimées sans distinction sous forme d'un ensemble non structuré de règles, méthodes et modèles du domaine sont rarement explicitées et donc réutilisables.

Les bases de connaissances telles qu'elles sont présentes dans les SBR sont peu ou pas organisées, les unités de connaissances étant toutes représentées de façon identique et sans relation explicite les unes avec les autres. Dans les années 80, cette méthode de construction d'une base de connaissances a rencontré un vif succès dû à l'apparente facilité et simplicité du procédé qui permettait, a priori, un prototypage rapide [David, 1995; Speel et al., 2001]. Cependant, les défauts engendrés par ce type de conception ont amené à repenser les principes de construction des SBC.

Un des autres points à signaler est le niveau des connaissances exprimées. Dans les SBR, toute connaissance est représentée, indifféremment de sa nature ou de son type, de façon identique aux autres, sous la forme de règle de production. Les connaissances sont donc très dépendantes de leur représentation symbolique, qui ne permet pas de décrire de relations entre les différentes connaissances exprimées. La base de connaissances obtenue consiste en une accumulation de règles. Normalement, les règles sont censées être indépendantes les unes des autres. C'est en particulier de là que vient le problème de « cohérence » des bases de règles. Toute modification d'une partie de la base de règles peut avoir des répercussions inattendues et inévitables.

4.4.2 Les problèmes du transfert d'expertise dans les SBC et SBR

Le principal problème inhérent au transfert d'expertise réside dans le fait qu'« An expert is a man who has stopped thinking - he knows! » comme l'a énoncé l'architecte américain Frank Lloyd Wright. En fait, il serait préférable de nuancer cette affirmation en disant plutôt qu'un expert « pense sans le savoir ». Pour raisonner, l'expert se base sur son expérience et son intuition. Ses réponses face à un problème en deviennent ainsi quasiment automatiques. D'où la difficulté de faire expliciter à un expert des connaissances qui sont pour lui implicites. De plus, même si l'expert entrevoit ce qu'il doit expliquer, il n'arrive pas forcément à bien exprimer ses connaissances. Les experts ont en général du mal à verbaliser et à communiquer leur savoir, à rendre compte de leur comportement dans la résolution d'un problème et à formaliser leurs connaissances de façon à les livrer d'une façon ordonnée et exploitable par une machine (car « nous savons plus que ce que nous pouvons dire » [Ermine, 2000]). Ainsi, la portée des connaissances recueillies lors du transfert d'expertise se limite souvent aux connaissances de surface. On peut, en revanche, profiter de la propension qu'ont les experts à raconter des cas de résolution de problème (les cas peuvent être ensuite exploités par exemple dans un système de RÀPC[6]).

De par leur origine essentiellement pratique et expérimentale, les connaissances d'un expert lui sont privées. Elles diffèrent d'un expert à l'autre. Il s'ensuit le problème de la collecte des connaissances auprès de plusieurs experts à la fois [McGraw and Harbison-Briggs, 1989]. L'établissement et l'utilisation d'une ontologie peut être une solution à ce

[6]RÀPC : RAISONNEMENT À PARTIR DE CAS.

problème étant donné qu'une ontologie précise le vocabulaire, les concepts et leurs relations. Toute la difficulté réside dans sa construction et son acceptation.

Il faut remarquer que les méthodologies élaborées dans le cadre du développement des SBC dont il est question en section 4.3.2 apportent une aide à la modélisation et l'élicitation des connaissances. Ces techniques ont justement pour but de faciliter la verbalisation des connaissances implicites, de dégager la structure profonde des connaissances contrairement à une simple retranscription d'expertise qui ne récupère souvent que les connaissances macroscopiques (*raw data* [Motta et al., 1990]).

4.5 Conclusion

Comme nous venons de le voir, la construction d'une base de connaissances est un processus complexe et difficile. Elle doit être menée avec le plus grand soin car la qualité finale des résultats du système conçu en dépend. L'acquisition des connaissances repose principalement sur le modèle conceptuel élaboré. À partir de ce travail de modélisation, il sera ensuite possible d'éliciter les connaissances et donc d'alimenter la base de connaissances. Cette opération peut (et doit en partie) être faite manuellement. Cependant, une automatisation ou semi-automatisation de cette activité est également envisageable grâce à un processus d'extraction de connaissances à partir de bases de données existantes. C'est de cette façon de voir l'acquisition des connaissances dont il va être question dans la suite.

Chapitre 5

L'extraction de connaissances dans les bases de données

... knowledge discovery in databases is considered to be the non-trivial extraction of implicit, previously unknown, and potentially useful information from data ...

Frawley et col., 1992

Sommaire

5.1 Introduction

Le développement des techniques et des capacités de stockage a permis la création d'énormes bases de données en constante croissance. Aujourd'hui, des mégabases contenant

des petaoctets[1] de données diverses ont été constituées. Tous les domaines de l'activité humaine sont concernés par cette explosion de la collecte de données : commerce (pour des besoins de marketing, de gestion des stocks), industrie (pour la gestion de production, la prévention de risques, le contrôle de qualité, la maintenance), monde judiciaire (détection de fraudes), monde scientifique (médecine, pharmacologie, génome, chimie), internet (navigation, recherche d'information), etc. La gestion de volumes de données aussi importants pose le problème de l'exploitation de l'information qui y est potentiellement contenue. En effet, les systèmes de gestion de bases de données ont été élaborés pour exécuter des tâches répétitives de consultation et de mise à jour tout en maintenant la cohérence des bases. Les langages de requête des systèmes transactionnels ne sont pas conçus pour répondre aux nouvelles attentes. L'analyse, l'organisation et l'interprétation des données sont devenues des enjeux cruciaux dans une société telle que la notre.

Ces dernières années ont vu l'apparition de l'informatique décisionnelle pour pallier les manques de l'informatique de production. Dès les années 90, des entrepôts de données sont construits à cette fin [Gilleron and Tommasi, 2000; Benitez-Guerrero et al., 2001]. L'architecture de ces nouveaux moyens de stockage des données a permis la mise au point de traitements en ligne des données (OLAP[2]). Cette technique produit des analyses assez fines et multidimensionnelles des données qu'il n'était pas possible de mener avec les langages de requête des bases de données. Le principe est celui du cube de données. Pour analyser les faits suivant plusieurs dimensions, on construit un modèle multidimensionnel dans lequel les données sont décrites selon les dimensions de l'analyse. S'il y a trois dimensions (par exemple, des données sur des ventes peuvent être décrites en terme de clients, produits et dates d'achat), on a effectivement un vrai cube. Bien entendu, la notion de cube de données est étendue à des dimensions supérieures à 3. Cependant même ce procédé a montré des limites et un nouveau domaine de recherche, celui de l'*extraction de connaissances à partir de bases de données*, s'est très vite développé. D'une utilisation traditionnelle pour les besoins tactiques (gestion de la production), les données sont désormais prises en compte dans les entreprises à des fins stratégiques pour augmenter les ventes, réduire les coûts mais aussi gérer les risques et prévenir les fraudes. De nombreuses sociétés de service se sont spécialisées dans le conseil dans ces domaines. Outre l'intérêt grandissant du monde économique, l'extraction de connaissances à partir de bases de données trouve également de très nombreuses applications dans les domaines social, scientifique et industriel. Le développement de techniques efficaces croît parallèlement à l'apparition de nouveaux besoins (prise en compte de nouveaux types de données en particulier).

L'extraction de connaissances à partir de bases de données a deux buts différents : la description et la prédiction. Le premier objectif vise à donner une vue globale des données en les regroupant dans des ensembles le plus possible homogènes. Les données sont de cette façon plus facilement appréhendées. Le deuxième objectif envisagé est l'estimation d'informations a priori inconnues. Plus globalement, l'extraction de connaissances à partir de bases de données a pour but d'extraire des informations potentiellement utiles à partir

[1] Millions de milliards d'octets.

[2] On-Line Analytical Processing.

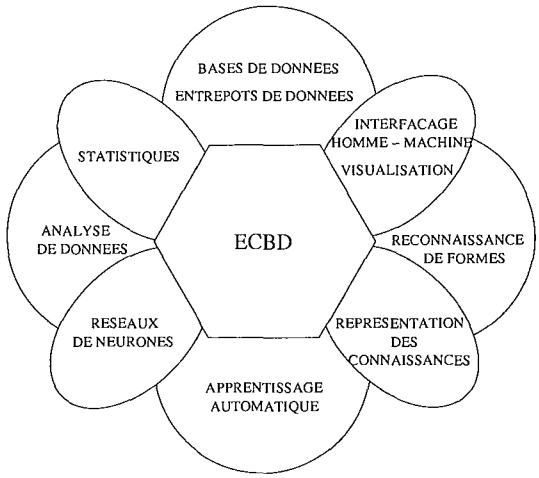


FIG. 5.1 – Les domaines de recherche connexes à l'extraction de connaissances dans les bases de données.

de grandes quantités de données. Il est donc légitime de penser que les connaissances mises en évidence pourraient venir enrichir la base de connaissances d'un SBC. Une partie de l'acquisition des connaissances pourrait alors être alors automatisée ou plutôt semi-automatisée car, comme nous allons le voir, le processus d'extraction de connaissances est avant tout un procédé itératif et interactif.

5.2 Le processus d'extraction de connaissances à partir de bases de données

L'expression « extraction de connaissances à partir de bases de données » (ECBD ou également ECD) vient de la traduction de *knowledge discovery in databases* (KDD). Ce terme a été introduit par Piatetsky et Shapiro lors d'un des workshops de la conférence IJCAI'89. L'ECBD a établi des passerelles avec de nombreux autres domaines (figure 5.1). Elle emprunte à ces différentes disciplines les outils théoriques aussi bien que pratiques dont elle a besoin. L'ECBD travaille à partir de données. Il est naturel que l'étude des bases ainsi que des entrepôts de données constitue une base. Pour analyser et représenter les données, l'ECBD utilise des méthodes développées dans les domaines des statistiques et de l'analyse de données mais aussi de l'intelligence artificielle. L'ECBD s'intéresse à la présentation des résultats tout au long du processus et donc aux techniques de visualisation et à l'interfaçage homme-machine

5.2.1 Les étapes du processus

Par le terme d'extraction de connaissances à partir de bases de données, on désigne tout le cycle de découverte d'informations. Ce processus est complexe et se déroule en plusieurs

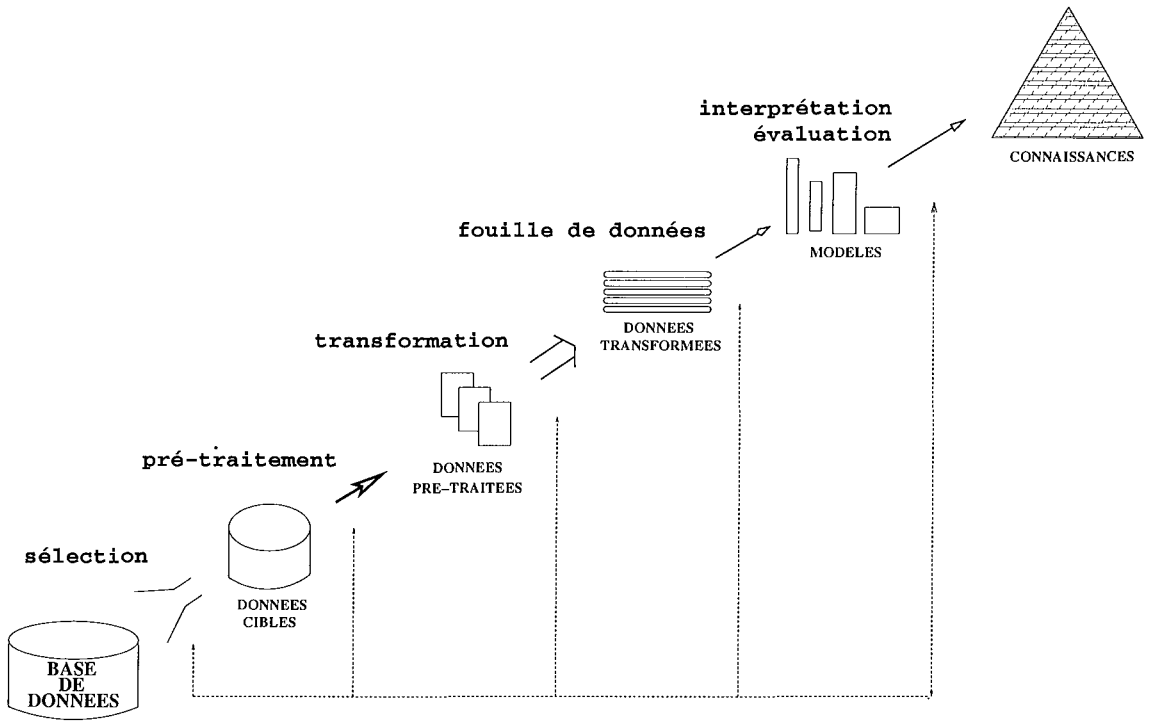


FIG. 5.2 – Les étapes du processus d'extraction de connaissances.

étapes (figure 5.2).

Tout processus d'extraction de connaissances répond à un besoin, tout comme la construction d'un SBC. Il est donc nécessaire de bien identifier, en premier lieu, les objectifs de l'application. Une étude des spécifications des résultats attendus, de leur critère d'évaluation, des connaissances disponibles doit être conduite préalablement. La compréhension du problème est donc la première étape du processus. Les caractéristiques définies permettront de guider la suite du processus, et en particulier la sélection des données.

Sélection des données

Lorsque l'on dispose de très grandes bases de données, il n'est pas obligatoire d'examiner l'ensemble de toutes les données. L'intégralité des données ne correspond pas forcément au type de données intéressant par rapport aux buts fixés. La sélection des données permet de constituer un jeu d'essai adapté au cadre de l'étude. Dans le domaine des statistiques, on constitue un échantillon que l'on espère être représentatif de la population étudiée.

Dans [Simon, 2000], cette étape de sélection est précédée d'une phase de recensement des données car celles-ci ne sont pas forcément disponibles dans des bases préexistantes. Cette étape permet de spécifier les données qu'il faudra récolter (par une enquête ou par l'examen de bases annexes, par exemple) pour les besoins de l'étude en cours.

Pré-traitement

Une fois ce sous-ensemble représentatif récupéré, les données vont devoir être pré-traitées

pour essayer d'augmenter leur qualité (consistance, complétude, précision, homogénéité, etc.). En effet, plus les données sont bruitées, moins le résultat final sera bon. Le pré-traitement est composé de plusieurs phases. L'étape de nettoyage permet de compléter les données manquantes, d'éliminer les valeurs aberrantes et les doublons. L'homogénéisation conduit à l'intégration de données provenant de plusieurs sources différentes tandis que la standardisation conduit à une normalisation et une mise à l'échelle des données. Ces pré-traitements peuvent être très complexes, en particulier dans le cadre de l'intégration de plusieurs bases de données et de données peu structurées comme les données semi-structurées [Al Hulou et al., 1999].

Transformation

Une fois les données préparées, il faut les transformer de façon à les représenter dans un codage adéquat pour l'application de méthodes de fouille. Cette étape est également importante car elle impose de remanier les données de façon à ce qu'elles soient adaptées à la phase de fouille. Ceci implique d'établir une description au niveau d'abstraction adapté aux questions posées à l'origine du processus d'ECBD.

Fouille de données

La ou les méthodes de fouille doivent avoir été choisies en fonction des buts recherchés (classification, association, segmentation) et/ou de leur disponibilité. La fouille de données (*data mining*) est une des étapes du processus d'ECBD. Il y a fréquemment un amalgame entre ces deux termes, au point que certains auteurs parlent indifféremment de fouille de données et d'extraction de connaissances dans les bases de données.

Cette étape consiste à appliquer les techniques aux données préparées et transformées. Pour certaines méthodes, il sera nécessaire d'ajuster des paramètres, de faire plusieurs essais afin d'obtenir les solutions optimales.

Interprétation - Validation

Les éléments issus de la fouille sont ensuite interprétés. Leur interprétation conduit à la validation ou la réfutation qui est susceptible de remettre en cause tout ou partie du processus. Plusieurs méthodes de validation sont envisageables. Les deux modes principaux sont les statistiques et l'expertise. Les modèles issus d'une méthode de classification pourront être vérifiées en premier lieu par un expert puis la validation sera complétée par des tests statistiques sur des bases de cas existantes. Pour les techniques d'apprentissage non supervisées telles que la segmentation et l'association, la détermination de la pertinence des modèles obtenus est essentiellement une affaire d'expertise. Pour ce qui est de la classification supervisée, une validation croisée est faisable à partir de trois ensembles de données : un ensemble d'apprentissage, un ensemble de test et un ensemble de validation. Il est même possible d'effectuer une validation croisée permettant de calculer l'erreur d'un modèle construit sur un ensemble par rapport aux entités de ce même ensemble. Ce problème de validation se retrouve également classiquement en apprentissage automatique.

5.2.2 Un processus interactif et itératif

L'extraction de connaissances dans les bases de données doit être considérée comme un processus. Nous venons d'en présenter le cycle de base. Or une des caractéristiques du déroulement d'un tel processus est de ne pas être (généralement) linéaire. Les résultats obtenus dans une des étapes peuvent remettre en cause les choix faits dans les phases précédentes. Des études ont montré que le temps pris par l'étape de fouille à proprement parlé représente moins de 20% du temps total du procédé [Brachman and Anand, 1996]. Le reste est réservé aux opérations de traitement préalable des données. En effet, l'application à des données réelles présente de nombreuses difficultés [Famili et al., 1997]. Après un cycle complet du processus, l'analyse des résultats obtenus indique s'il est nécessaire de les affiner par un retour-arrière. Il est aussi possible de se rendre compte au cours d'une des étapes quelconques que l'une des précédentes n'a pas été conduite de façon satisfaisante. En conséquence, on reviendra sur l'étape mise en cause avant de continuer plus en avant le processus. L'ECBD procède par une succession de retours-arrières sur les données, le modèle et les résultats jusqu'à l'obtention de « connaissances valides[3] ». Sur la figure 5.2, les flèches en pointillés schématisent l'ensemble des retours possibles entre les différentes étapes du processus d'ECBD. On remarque qu'un retour est possible à toutes les étapes du processus. Le fait que l'ECBD soit un processus itératif en fait naturellement un processus interactif. Il est illusoire de penser que l'on pourrait mettre au point une méthode automatique qui prendrait des données brutes en entrée et donnerait des résultats convenables en sortie sans aucune intervention humaine. Par ailleurs, le système d'ECBD construit doit proposer une interface performante pour permettre à l'utilisateur d'intervenir à chaque fois que cela est nécessaire. Le facteur humain qui rentre en compte dans tout le processus est personnifié par « l'analyste » dont nous allons voir le rôle primordial dans la section suivante.

5.2.3 Une activité centrée sur l'analyste

La place de l'analyste dans le processus d'extraction de connaissances est très importante sinon primordiale. La définition citée en début de ce chapitre est très réductrice et ne rend pas compte de l'influence des données de départ. Reprenant cet énoncé dans [Brachman and Anand, 1996], Brachman et Anand redéfinissent l'extraction de connaissances en mettant en avant l'interaction entre l'analyste, spécialiste du domaine d'application, et les données initiales. L'analyste est à la fois l'expert, le modélisateur et l'utilisateur. L'analyste travaille sur un domaine d'application dont il est expert. En conséquence, c'est lui qui sélectionne les données par rapport aux besoins du problème posé. C'est également lui qui est le plus à même d'apporter les corrections nécessaires aux données. Au niveau de la transformation des données, l'analyste devient le modélisateur qui dirige la modification des données de façon à ce qu'elles soient le mieux adaptées possible au problème traité en tenant compte des exigences des techniques de fouille. Il participe aussi au choix de la technique à employer par rapport aux spécifications du problème, de l'algorithme à utiliser en fonction de la nature des données et du type de connaissances recherchées. Même s'il ne

[3] au sens de plausibles [Valdés-Pérez, 1999]

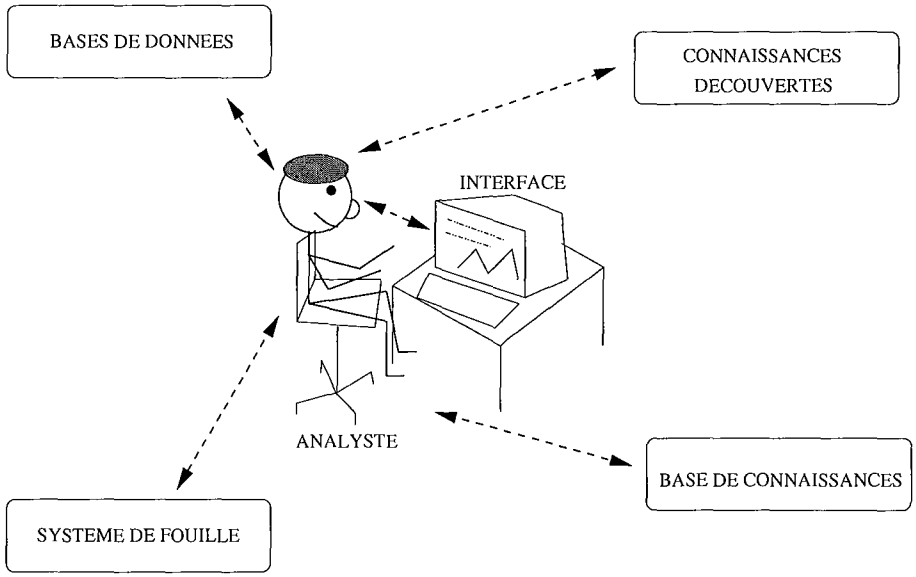


FIG. 5.3 – La place de l'analyste dans un système d'extraction de connaissances à partir de bases de données.

doit pas forcément être un spécialiste des techniques de fouille de données, l'analyste est nécessairement au fait de ce qu'elles proposent. Dans [Simon, 2000], l'analyste est d'abord comparé à un orpailleur dont le métier est de découvrir des « pépites » de connaissances. Tout comme pour d'autres auteurs, devant la faiblesse de la métaphore, l'activité de l'analyste est ensuite rapprochée de celle d'un archéologue qui exhumerait les connaissances des données.

Les fonctionnalités proposées au sein d'un système d'ECBD doivent assister l'analyste dans toutes les étapes du processus. Ainsi, les systèmes d'ECBD comportent :

- une ou des bases de données ainsi que leurs systèmes de gestion,
- un système de gestion et de représentation des connaissances pour la manipulation de données complexes,
- un système de fouille de données,
- une interface permettant à l'utilisateur d'interagir et de visualiser les résultats à chaque étape du processus.

5.3 La fouille de données

5.3.1 Les méthodes de fouille

La fouille de données a pour objectif de faire émerger par des méthodes algorithmiques des tendances ou des schémas (« patterns ») à partir d'un grand volume de données prétraitées. Ceci peut se faire selon deux orientations distinctes [Fayyad et al., 1996] :

- dans un but explicatif de la masse des données soumise à l'analyse,

- dans un but de prise de décision par prévision de valeurs inconnues ou de comportements futurs.

5.3.1.1 Les tâches abordées

Partant d'un ensemble de données insuffisamment décrit pour en avoir une bonne connaissance, la fouille de données a plusieurs tâches : celles de décrire, de prédire et/ou d'analyser les corrélations entre les données.

La description
Il s'agit de résumer les données en en donnant une description abstraite. Les techniques employées sont des méthodes de classification non supervisées dont l'objectif est de construire un ensemble de classes à partir d'un ensemble de données non structuré. Ces méthodes sont aussi appelées méthodes de segmentation ou de regroupement. Une telle classification peut être engendrée par des méthodes d'analyse de données (classifications hiérarchiques ascendantes ou descendantes, classification pyramidale, nuées dynamiques, plus proches voisins), des méthodes neuronales (cartes de Kohonen), des méthodes symboliques (*clustering*). Généralement, la classification n'est pas dirigée mais certaines méthodes permettent de l'orienter. Les hiérarchies conceptuelles sont plus facilement interprétables que celles basées sur des calculs numériques.

La prédiction
Le problème consiste à savoir classer correctement de nouveaux exemples grâce aux règles construites à partir du jeu d'essai. Ces méthodes sont basées sur une classification automatique supervisée. C'est-à-dire que les classes sont déjà prédéfinies par l'utilisateur et décrites par des exemples. Ce sont des techniques d'induction. Parmi celles-ci on peut citer au niveau symbolique les systèmes d'induction de règles et les arbres de décision, au niveau numérique des méthodes basées sur les réseaux de neurones telles que le *perceptron*. Les méthodes symboliques ont l'avantage de produire des modèles facilement compréhensibles pour un utilisateur et intègrent les connaissances du domaine. On peut envisager une méthode de partitionnement préalable pour obtenir les classes initiales.

L'analyse de liens
Ces techniques permettent de mettre en évidence les liens existant entre des objets donnés, les propriétés ou les objets et les propriétés. Elles déterminent des associations de la forme : « Si on a cet ensemble de faits alors, dans x% de cas, on a également cet autre ensemble de faits ». Les méthodes développées sont du point de vue numérique les régressions, et du point de vue symbolique la recherche de règles d'association ou de motifs fréquents ou de motifs séquentiels. Elles ont principalement un but descriptif, l'utilisation pour la prédiction est généralement très contestée.

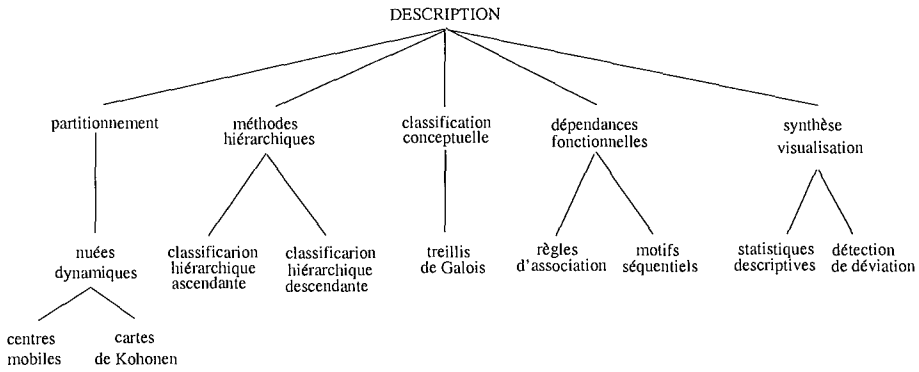


FIG. 5.4 – Quelques unes des méthodes dont le but est la description d'un ensemnble de données.

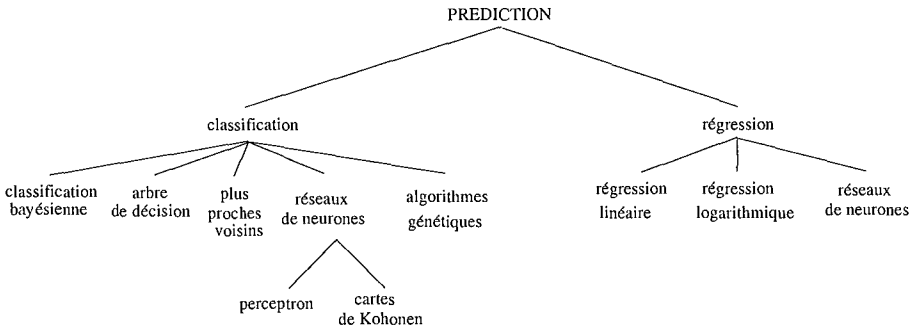


FIG. 5.5 – Quelques unes des méthodes dont le but est la prédiction de données.

En résumé, les techniques développées pour atteindre les buts de la fouille sont diverses et touchent des domaines tels que les statistiques, l'analyse de données, l'algorithmique, l'apprentissage automatique [Two Crows Corporation, 1999] (voir figure 5.1). Les figures 5.4 et 5.5 donnent un aperçu des méthodes employées respectivement pour des objectifs de description et de prévision.

5.3.1.2 Le choix d'une méthode

Parmi toutes les techniques existantes, il n'existe pas bien sûr de méthode supérieure à toutes les autres. Comme nous l'avons déjà dit la sélection d'une méthode se fera en fonction de la nature du problème mais aussi des données. Les critères de choix seront donc liés à :

- la tâche que l'on souhaite résoudre,
- l'utilisation finale du modèle qui doit être obtenu (complexité, performances, pérennité),
- le type des données accessibles,
- les connaissances et compétences disponibles,
- le contexte global.

Comme nous l'avons dit, le choix des méthodes à employer est la plupart du temps dirigé par la nature des données. Aussi, nous distinguerons les méthodes numériques qui travaillent sur des données de type quantitatif et les méthodes symboliques qui s'appliquent à des données de type qualitatif.

L'approche numérique
Elle correspond classiquement aux méthodes d'analyse de données. Les objets numériques sont caractérisés par un nombre fini de propriétés à valeurs numériques (par exemple, en chimie les conditions réactionnelles telles que la température, la pression peuvent être reportées sous forme numérique : température = 60°, pression = 2 bars) ou, de façon qualitative, transformables en données binaires (par exemple, en chimie, chauffage = oui). Les techniques numériques exploitent le fait qu'un objet caractérisé par p propriétés peut être représenté par un point dans un espace vectoriel de dimension p. Ces données sont traitées avec des méthodes numériques en utilisant l'algèbre linéaire et les outils de la statistique mais aussi les méthodes connexionistes. L'objectif des méthodes numériques est essentiellement de mettre en évidence les corrélations entre les attributs décrivant les objets. Les approches numériques sont efficaces mais pèchent par manque d'explication. Les résultats obtenus sont difficilement compréhensibles et exploitables par un non spécialiste de la méthode numérique qui a été employée.

L'approche symbolique
Un objet symbolique est un objet complexe dont les propriétés ne peuvent pas être exprimées uniquement de façon numérique ou binaire. Un objet symbolique se présente sous forme d'une conjonction ou d'une disjonction de propriétés [Polaillon, 1998] qui peuvent prendre des valeurs particulières (intervalles (pH compris entre 8 et 10), énumération marquant la multitude de choix (réactif = permanganate de potassium ou bichromate de sodium)). La difficulté est de rendre compte de la complexité des objets réels manipulés (incertitude, imprécision, variation). Bien qu'elles soient indispensables, les réductions qui sont faites doivent engendrer le moins de perte d'information possible. Les méthodes symboliques doivent être guidées par les connaissances du domaine qui rendent compte de la sémantique (avec par exemple, l'exploitation de taxonomies du domaine). Certaines techniques numériques ont été adaptées pour traiter les objets symboliques. Elles utilisent pour cela des métriques mais ne tiennent pas compte de la sémantique du domaine. L'approche symbolique a l'avantage d'être centrée sur la description des individus tenant compte de connaissances du domaine et d'être capable de fournir des explications [Kodratoff, 1991].

5.3.2 Les treillis, les motifs fréquents et l'extraction de règles

Nous décrivons de façon plus détaillée ce type de méthode car c'est celle que nous avons décidé d'employer et dont nous verrons l'application dans la suite de ce mémoire. Cette section est fortement inspirée de la thèse d'Yves Bastide [Bastide, 2000] à qui l'on doit, ainsi

qu'aux collaborateurs avec lesquels il a travaillé[4], les algorithmes et leurs implémentations que nous avons utilisés au cours de nos travaux.

L'utilisation pratique la plus citée de la recherche de règles d'association est certainement celle liée au problème de l'analyse du « panier de la ménagère »[5] qui intéresse beaucoup le secteur de la grande distribution. Le but est de mettre en évidence des profils de clients ou plutôt des associations d'achats pour ensuite organiser en conséquence la disposition des articles en magasin ou mettre en place des campagnes de promotion ciblées. Mais cette application n'est pas la seule. De nombreux domaines sont concernés par la découverte de corrélations dans leurs données.

Les règles obtenues ont l'avantage d'être facilement interprétables. Par contre, comme nous le verrons, le nombre de règles produites peut être très grand et l'intérêt des règles est très variable, des règles triviales ou inutiles étant également produites. Il faut également tenir compte des faiblesses des mesures de qualité telles que celle de la confiance dont nous parlerons dans ce qui suit. Cette technique est non supervisée.

Pour commencer, nous devons définir quelques uns des concepts mathématiques sur lesquels est basée cette technique de fouille, notamment ceux à propos des ensembles ordonnés.

5.3.2.1 La classification par treillis

Définitions

Ordre

Soit $\mathcal{X}$ un ensemble et $\leq$ une relation binaire sur $\mathcal{X}\times\mathcal{X}$. La relation $\leq$ est une relation d'ordre partiel si elle est transitive, antisymétrique et réflexive. Si, de plus, pour tout couple (x,y) de se $\mathcal{X}$, on a x$\leq$y ou y$\leq$x, alors on dit que l'on a une relation d'ordre total. On appelle ordre ou ensemble ordonné le couple $\mathcal{P} = (\mathcal{X},\leq)$.

Soit x et y deux éléments de $\mathcal{X}$. On note $x < y$ si et seulement si $x \leq y$ et $x \neq y$.

Éléments maximaux et minimaux

Soit $\mathcal{A}$ une partie arbitraire de l'ordre $\mathcal{P}$. Un élément x de $\mathcal{A}$ est *minimal* s'il n'existe pas d'élément y de $\mathcal{A}$ tel que $y < x$. De même, un élément x de $\mathcal{A}$ est *maximal* s'il n'existe pas d'élément y de $\mathcal{A}$ tel que $y > x$. La partie $\mathcal{A}$ de $\mathcal{P}$ peut éventuellement admettre plusieurs minimaux et maximaux. Les ensembles respectifs de ces éléments particuliers de $\mathcal{A}$ sont :

$$min\mathcal{A} = \{x \in \mathcal{A} \ / \ \nexists\, y \in \mathcal{A}, y < x\}$$

$$max\mathcal{A} = \{x \in \mathcal{A} \ / \ \nexists\, y \in \mathcal{A}, y > x\}$$

Ensembles des majorants et des minorants

Soit $\mathcal{A}$ une partie arbitraire de $\mathcal{P}$. L'ensemble des *minorants* de $\mathcal{A}$, noté $l(\mathcal{A})$, est l'ensemble

[4]Nicolas Pasquier, Rafik Taouil et Lotfi Lakhal.

[5]le fameux « market basket »

des éléments de $\mathcal{P}$ plus petits que tout élément de $\mathcal{A}$. De façon duale, l'ensemble des *majorants* de $\mathcal{A}$, noté u($\mathcal{A}$), est l'ensemble des éléments de $\mathcal{P}$ plus grands que tout élément de $\mathcal{A}$. Plus formellement :

$$l(\mathcal{A}) = \{x \in \mathcal{P} \ / \ \forall\, y \in \mathcal{A}, x \leq y\}$$

$$u(\mathcal{A}) = \{x \in \mathcal{P} \ / \ \forall\, y \in \mathcal{A}, x \geq y\}$$

Bornes d'un ensemble ordonné

Si l($\mathcal{A}$) a un plus grand élément unique, cet élément est appelé *borne inférieure* de $\mathcal{A}$ ou Meet($\mathcal{A}$) ou encore infinum ($\wedge$) ou Bottom ($\perp_A$). De même, si u($\mathcal{A}$) existe et a un plus petit élément unique, cet élément est appelé *borne supérieure* de $\mathcal{A}$ ou Join($\mathcal{A}$) ou encore supremum ($\vee$) ou Top ($\top_A$).

Treillis

Un *treillis* est un ensemble ordonné $\mathcal{P} = (\mathcal{X}, \leq)$ dans lequel tout couple (x,y) de $\mathcal{X}^2$ admet une borne inférieure et une borne supérieure. Si $\mathcal{P}$ est fini, alors c'est un treillis fini. Tout comme un ordre partiel, un treillis peut être représenté par un diagramme de Hasse[6].

L'ensemble des parties d'un ensemble $\mathcal{X}$ muni de la relation d'inclusion est un treillis complet, dit treillis des parties (voir figure 5.6).

Base de données ou contexte formel

Soit un ensemble fini d'objets $\mathcal{O}$, $\mathcal{P}$ un ensemble ordonné fini d'éléments ou items et $\mathcal{R}$ une relation binaire entre ces deux ensembles. Une *base de données* ou *contexte formel* correspond au triplet $\mathcal{D} = (\mathcal{O}, \mathcal{P}, \mathcal{R})$. Les éléments de $\mathcal{O}$ sont aussi appelés *individus* et ceux de $\mathcal{P}$ sont dits *propriétés* ou *attributs*. Quand un individu o est en relation avec une propriété p, on dit que l'individu o possède la propriété p.

Exemple :

Soit $\mathcal{D}$ une base de données sur des réactions décrites par des propriétés. Les objets de $\mathcal{D}$ sont au nombre de 6 (R1, R2, R3, R4, R5, R6). Les propriétés sont le milieu acide (A), le milieu basique (B), le chauffage (C), la pression (P), le reflux(R). L'exemple de base de données est décrit dans les tableaux de la figure 5.7.

Concept formel et treillis de Galois

Soit un contexte formel ($\mathcal{O}$, $\mathcal{P}$, $\mathcal{R}$). Un concept formel C de ($\mathcal{O}$, $\mathcal{P}$, $\mathcal{R}$) est défini comme un couple (Extension(C), Intension(C)) où Extension(C) est l'ensemble des individus de $\mathcal{O}$ possédant les propriétés de Intension(C) et Intension(C) l'ensemble des propriétés communes

[6] Un diagramme de Hasse se construit de la façon suivante : à chaque élément x de $\mathcal{P}$ on fait correspondre un point du plan euclidien ; pour chaque couple (x,y) de $\mathcal{P}^2$ tel que x < y et $\nexists$ z $\in \mathcal{P}$ / x < z < y, on relie les deux points du plan correspondant ; les conventions de dessin sont les suivantes : si x < y, le point correspond à x est au-dessous du point de y. Il ne doit pas y avoir d'intersection entre la représentation de z et l(x,y) quelque soit z$\neq$x et z$\neq$y. On élimine ainsi du dessin les arcs de transitivité.

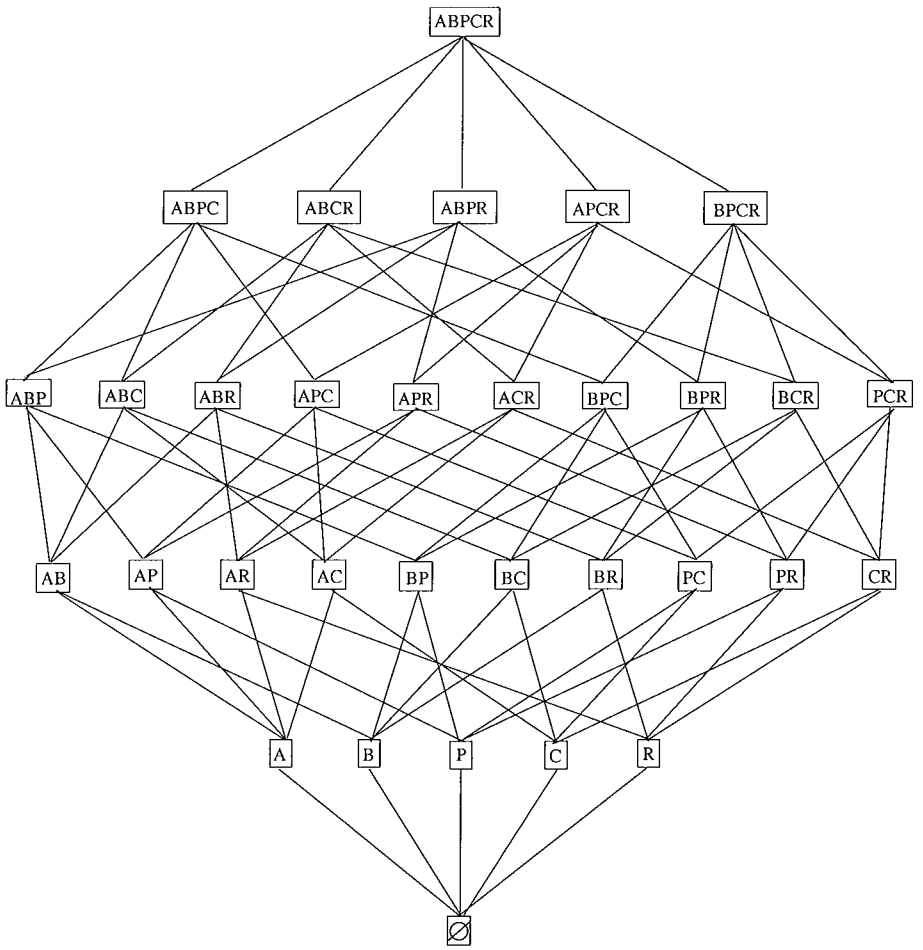


FIG. 5.6 – Diagramme de Hasse du treillis des parties de l'ensemble $\mathcal{P}$ = {A,B,P,C,R}. L'ensemble étant constitué de 5 éléments, le nombre de parties que l'on peut avoir à partir de ces éléments est 32 (2^5).

aux individus de Extension(C). Un concept formel rend donc compte d'un regroupement naturel d'individus possédant les mêmes propriétés. Il peut être décrit en extension et en intension. L'ensemble des concepts formels muni de la relation de subsomption[7] [Napoli, 1992] définit une structure de treillis, et plus précisément une structure de treillis de Galois (voir un exemple figure 5.8).

Opération de fermeture

Un opérateur de *fermeture* h sur un ensemble ordonné $\mathcal{P}$ est une application qui a les propriétés suivantes :

- monotone croissante, c'est-à-dire $\forall$ (x,y) $\in \mathcal{P}^2$, x $\geq$ y $\Longrightarrow$ h(x) $\geq$ h(y)
- extensive, c'est-à-dire $\forall$ x $\in \mathcal{P}$, x $\leq$ h(x)

[7]Une relation de subsomption, notée $\sqsubseteq$, est définie sur l'ensemble des concepts formels d'un treillis par : C1 $\sqsubseteq$ C2 $\Longleftrightarrow$ Extension(C1) $\subseteq$ Extension(C2) et Intension(C1) $\supseteq$ Intension(C2)

objets	liste de propriétés associées
R1	{A,P,C,R}
R2	{B,C}
R3	{B,P,C}
R4	{A,R}
R5	{B,P,C}
R6	{A,C}

$\Longrightarrow$

	A	B	P	C	R
R1	√		√	√	
R2		√		√	
R3		√	√	√	
R4	√				√
R5		√	√	√	
R6	√			√	

FIG. 5.7 – La base de données $\mathcal{D}$. À gauche, la base de données décrite en terme de liste d'attributs associée à chaque objet. À droite, le tableau représente la relation binaire existant entre les objets et les attributs de la base $\mathcal{D}$. Un √ dans la case (i,j) indique que l'individu de la ligne i possède la propriété de la colonne j.

– indempotente : $\forall\ x \in \mathcal{P}$, h(h(x)) = h(x)

Fermé

Un sous-ensemble X de $\mathcal{P}$ est dit fermé pour l'opérateur de fermeture h si et seulement si X = h(X).

La structure mathématique des treillis de concepts est intéressante à plusieurs titres et est donc utilisée par plusieurs communautés scientifiques [Leblanc, 2000; Polaillon, 1998] : acquisition et représentation des connaissances, formation de concepts [Hereth et al., 2000], raisonnement à partir de cas, recherche documentaire [Nauer, 2001], atelier de réutilisation logicielle, conception de hiérarchies de classes dans le développement orienté objet [Leblanc, 2000] et bien entendu le domaine de la fouille de données depuis les années 90. Dans ce dernier domaine, nous avons bénéficié, notamment, des travaux de l'équipe Stumme, Taouil, Bastide, Pasquier et Lakhal.

La construction d'un treillis de Galois permet de regrouper les individus et de mettre en évidence des concepts définis en intension et en extension. Cette technique de classification a l'avantage d'être *objective* (il n'y a pas d'alternative possible, seule la condition de fermeture doit être respectée) et de donner une classification *unique* (toutes les méthodes de construction du treillis de concepts mènent au même résultat) [Valtchev, 1999].

5.3.2.2 Motifs fréquents

La recherche de règles d'association se divise en deux étapes [Srikant and Agrawal, 1994] :

– la détermination des motifs fréquents (« large itemsets ») parmi l'ensemble des données dont on dispose est la phase préliminaire,

– le calcul des règles d'association se fait ensuite à partir des motifs fréquents générés.

Nous allons donc nous intéresser dans cette section à la première étape et donc aux motifs fréquents.

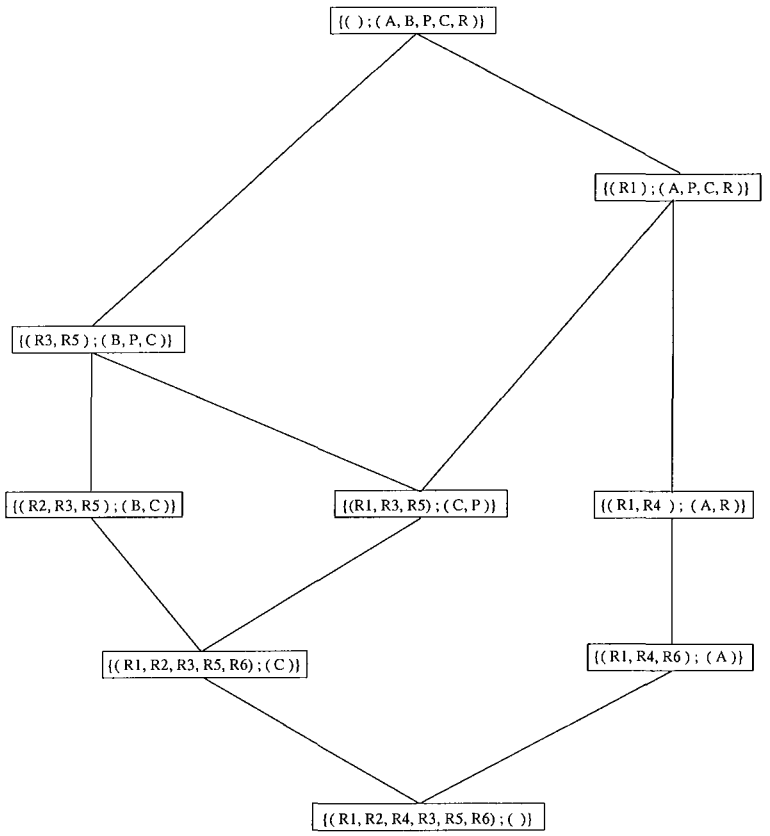


FIG. 5.8 – Diagramme de Hasse du treillis de concepts correspondant au contexte formel $\mathcal{D}$.

Définitions

Motif

Un *motif* est un sous-ensemble de $\mathcal{P}$.

Soit un objet o $\in \mathcal{O}$. On dit qu'un motif P est *inclus* dans l'objet o si P et o sont en relation, c'est-à-dire $\forall$ p $\in$ P, (o,p) $\in \mathcal{R}$. Un motif représente un ensemble (conjonction) de propriétés ou d'items (*itemsets*). Un motif de *taille* k est un motif comprenant k items, c'est un *k-motif*.

Exemple :

La base de données $\mathcal{D}$ contient, en autres, les motifs suivant :

- 1-motifs : A, B, P, C, R
- 2-motifs : (A, C), (A, R), (A, P), (C, R), (B, C)
- 3-motifs : (A, P, C), (B, C, P)
- 4-motifs : (A, P, C, R)

Image d'un motif - Image d'un ensemble d'objets

Soit f la fonction qui a tout motif P fait correspondre l'ensemble des objets qui le possèdent.

f est définie de la façon suivante :

$$\begin{aligned} f : \mathcal{P} &\longrightarrow \mathcal{O} \\ P &\longrightarrow f(P) = \{o \in \mathcal{O} / o \textit{ contient } P\} \end{aligned}$$

Soit g la fonction qui a tout ensemble d'objets O fait correspondre l'ensemble des propriétés communes aux objets. g est définie de la façon suivante :

$$\begin{aligned} g : \mathcal{P} &\longrightarrow \mathcal{O} \\ O &\longrightarrow g(O) = \{p \in \mathcal{P} / \forall o \in O, (o,p) \in \mathcal{R}\} \end{aligned}$$

Remarque : les compositions de ces deux fonctions (g o f et f o g) sont des opérateurs de fermeture. C'est sur la base de ces deux opérateurs qu'est construit le treillis de Galois de la relation $\mathcal{R}$.

Exemple :

Pour la base de données $\mathcal{D}$, on a :

f(R1) = {A, P, C, R}	g(A) = {R1, R4, R6}
f(R2) = {B, C}	g(B) = {R2, R3, R5}
f(R3) = {B, P, C}	g(P) = {R1, R3, R5}
f(R4) = {A, R}	g(C) = {R1, R2, R3, R5, R6}
f(R5) = {B, P, C}	g(R) = {R1, R4}
f(R6) = {A, C}	

Support d'un motif

Le *support* d'un motif P rend compte de la proportion des objets du contexte formel qui le possèdent :

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

Le support est parfois défini de façon absolue, on a alors sup(P) = card(f(P)).

Exemple :

Pour la base de données $\mathcal{D}$,

sup(A) = $\frac{1}{2}$	sup(A,C) = $\frac{1}{3}$	sup(A,P,C) = $\frac{1}{6}$	sup(A,P,C,R) = $\frac{1}{6}$
sup(B) = $\frac{1}{2}$	sup(A,P) = $\frac{1}{6}$	sup(B,C,P) = $\frac{1}{3}$	
sup(P) = $\frac{1}{2}$	sup(A,R) = $\frac{1}{3}$		
sup(C) = $\frac{5}{6}$	sup(C,R) = $\frac{1}{6}$		
sup(R) = $\frac{1}{3}$	sup(B,C) = $\frac{1}{2}$		

Motif fréquent

Soit un seuil minimum compris entre 0 et 1 inclus, appelé *support minimum* et noté *minsup*. Un motif est dit fréquent si son support est supérieur ou égal à *minsup*.

Algorithmes

La détermination des motifs fréquents d'une base de données peut être très coûteuse en temps car elle nécessite plusieurs passes sur la base de données. Si la base est de grande taille (ce qui est généralement le cas puisque l'extraction de motifs fréquents n'a de sens que sur de grands volumes de données), on comprend que le procédé est d'autant plus long. De plus, le nombre de motifs générés peut être énorme si le nombre de propriétés est important. Il existe des bases contenant la description de millions d'objets selon des centaines voir des milliers de propriétés. Il faut donc utiliser une méthode permettant de faire le moins d'accès possible à la base de données.

La principale approche repose sur une recherche par niveau [Bastide, 2000], c'est-à-dire que l'on va détecter d'abord les motifs de taille 1 puis ceux de taille 2 à partir de ceux de taille 1, etc. Cette démarche a pour but d'éviter l'explosion combinatoire de l'approche naïve qui consisterait à générer toutes les parties[8] du contexte formel et de compter combien de fois chacune est présente dans la base de données. Le processus de recherche de motif est de complexité exponentielle en la taille des données.

L'algorithme général de la recherche par niveau est simple : pour chaque niveau k, on génère l'ensemble des k-motifs potentiellement fréquents appelés motifs candidats. Les supports des motifs candidats sont ensuite calculés et seuls les motifs fréquents sont conservés. La génération des candidats se fait en tenant compte de la propriété suivante : « tout sous-motif d'un motif fréquent est fréquent »[9].

Un des premiers algorithmes à avoir connu le « succès » est APRIORI. Il a été développé par Agrawal et Srikant [Srikant and Agrawal, 1994] et génère tous les motifs fréquents. Il met en œuvre le principe de base que l'on vient d'énoncer et nécessite un nombre de passes sur la base de données égal à la taille des motifs les plus grands + 1. Beaucoup de candidats sont comptés sans que cela soit nécessaire.

Depuis, d'autres algorithmes plus efficaces ont été élaborés soit à partir de structures de données différentes, de parallélisation, de réductions de motifs à prendre en compte (motifs fréquents maximaux, algorithme MAXMINER), des propriétés structurelles des treillis de Galois ([Bastide, 2000], algorithmes CLOSE [Bastide et al., 2000], PASCAL [Bastide et al., 2002] et TITANIC [Stumme et al., 2002]). Généralement, les concepteurs de ces algorithmes essaient de minimiser le nombre de parcours de la base de données et/ou le nombre de candidats générés.

Ainsi, l'algorithme CLOSE extrait les motifs fréquents fermés. En effet, l'ensemble global des motifs fréquents peut être déduit des fermés sans besoin d'accéder à la base de données. CLOSE, tout comme APRIORI, énumère des candidats mais ici ce sont les candidats générateurs particuliers qui sont calculés d'après la fermeture.

Le principe de PASCAL repose sur le regroupement de motifs considérés comme équi-

[8] Un ensemble de n propriétés contient 2^n parties. Pour le petit exemple donné ci-dessus cela fait déjà un total de 32 parties.

[9] Démonstration : soit G un motif et P un sous-motif de G. L'inclusion des motifs nous donne sup(G) $\leq$ sup(P). Si G est fréquent alors sup(G) $\geq$ *minsup*, donc sup(P) $\geq$ *minsup*. On en déduit que P est également un motif fréquent. On peut démontrer, de façon analogue, la contraposée de cette propriété : « tout sur-motif d'un motif non fréquent est non fréquent ».

valents. Deux motifs font partie de la même classe d'équivalence si et seulement si leurs images sont égales. La détermination des classes d'équivalence permet de réduire le nombre de motifs dont la fréquence doit être évaluée, étant donné que tous les motifs appartenant à une même classe d'équivalence ont même support dans la base. De plus, dans l'algorithme par niveau, les premiers éléments d'une classe d'équivalence mis en évidence sont minimaux au sens de l'inclusion. Ce sont des motifs clés. PASCAL se contente de compter le support des *motifs clés* candidats, le support des motifs non clés sera ensuite déterminé grâce au principe de comptage par inférence.

Jean-François Boulicaut et col. [Boulicaut et al., 2000] proposent une formalisation plus générale, appelée δ-libre, qui permet de faire des extractions de motifs fréquents dans des temps raisonnables pour des cas difficiles. En effet, lorsque les données sont très corrélées, le nombre de motifs fréquents peut croître de telle façon que la détermination de leur fréquence devient impossible. Pour pallier ce problème d'explosion combinatoire, la technique avancée est celle de la détermination de la valeur approchée de la fréquence des motifs grâce à des représentations ϵ-adéquates des ensembles de motifs. Ces représentations sont condensées par rapport à celles d'origine et présentent une perte de précision proportionnelle au paramètre ϵ. L'extraction se fait à partir des motifs δ-libres. Si $\delta = 0$, les δ-libres sont équivalents aux clés extraites dans PASCAL. Plusieurs représentations condensées ont été étudiées et donnent des résultats très prometteurs dans des cas d'extraction très difficiles.

Motifs séquentiels

Le problème de la recherche de motifs séquentiels dans une base de données est une extension de celui de l'extraction des motifs fréquents que nous avons présenté dans cette section [Masseglia, 2002]. C'est une vue particulière de l'extraction de motifs dans laquelle l'ordre des propriétés est important. La découverte de motifs séquentiels implique la prise en compte de contraintes temporelles. Les relations entre les individus et les attributs de la base de données sont exprimées sous forme de transactions datées. Un même individu peut être présent plusieurs fois dans la base mais pour des transactions différentes. La recherche de motifs séquentiels a pour but de trouver des corrélations entre les transactions effectuées par les individus. On peut ainsi obtenir le fait qu'un individu qui achète une mallette du « parfait petit chimiste » achètera cinq jours plus tard une blouse blanche pour protéger ces vêtements. D'une utilisation principalement commerciale, les motifs séquentiels trouvent désormais des applications très diverses comme la détection de fraudes, l'identification de symptômes de maladies en médecine, la prévention des risques industriels et la surveillance d'ouvrages tels que les barrages, l'aide à la navigation sur Internet (*web usage mining* [Masseglia, 2002]).

5.3.2.3 Règles d'association

Définitions

Règle d'association

Une règle d'association est une règle d'implication probabiliste [Bastide, 2000] de la forme : si *condition* alors *conclusion* dans x% des cas. Elle est calculée à partir d'un motif `M` et d'un

des sous-motif A inclus dans M. Ainsi B = (M - A) représentant l'ensemble des propriétés contenues dans M mais pas dans A. Une règle impliquant ces deux motifs est notée A $\Rightarrow$ B dans un certain nombre de cas. Cette règle s'interprète de la façon suivante : si un objet possède les propriétés contenues dans A, alors, dans un certain nombre de cas, cet objet a également les propriétés contenues dans B. A est appelée prémisse ou partie gauche ou antécédent de la règle. B est appelée conclusion ou partie droite ou conséquent de la règle. A et B sont des conjonctions de propriétés.

Support d'une règle d'association

Le *support d'une règle* est égal au support de l'union de sa prémisse et de sa conclusion.

$$sup(\mathtt{A} \Rightarrow \mathtt{B}) = sup(\mathtt{A} \cup \mathtt{B})$$

Confiance d'une règle

La *confiance d'une règle* représente le rapport entre le support de la règle et le support du motif A :

$$conf(\mathtt{A} \Rightarrow \mathtt{B}) = \frac{sup(\mathtt{A} \cup \mathtt{B})}{sup(\mathtt{A})}$$

La confiance est aussi appelée précision de la règle.

Règle d'association valide

Une règle d'association est dite *valide* si son support et sa confiance sont supérieurs ou égaux respectivement à *minsup* et *minconf*[10]. Historiquement toutes les mesures de qualité reposent sur des extractions faites à partir des motifs fréquents. Une règle valide provient donc d'un motif fréquent.

Règle d'association exacte

Une règle d'association dont la confiance est égale à 100% est dite *exacte*. En effet, cette règle est toujours vérifiée car les supports de A et de A $\cup$ B sont égaux. Bien entendu, une règle exacte est forcément valide.

Règle d'association approximative

Une règle d'association dont la confiance est strictement inférieure à 100% est dite approximative ou partielle. Si le support de cette règle est supérieur à *minsup*, la règle est dite approximative valide.

Autres mesures statistiques sur les règles d'association

Face au grand nombre de règle susceptibles d'être générées, il peut être utile de disposer de moyens pour différencier les règles selon certains critères. L'association d'indices statistiques à chaque règle permet de les ordonner les unes par rapport aux autres [Cherfi and Toussaint, 2002]. Plusieurs mesures utilisant les probabilités sont ainsi envisageables

[10] *minconf* est appelé seuil de confiance

[Cherfi and Toussaint, 2002]. Soit A et B des motifs, on peut définir les mesures statistiques suivantes reposant sur diverses probabilités (notées P(X), par exemple, pour désigner la probabilité de l'évènement X) :

- intérêt :

$$int[\mathtt{A} \Rightarrow \mathtt{B}] = \frac{P(\mathtt{A},\mathtt{B})}{P(\mathtt{A}) \times P(\mathtt{B})}$$

- conviction :

$$conv[\mathtt{A} \Rightarrow \mathtt{B}] = \frac{P(\mathtt{A}) \times P(\neg\mathtt{B})}{P(\mathtt{A},\neg\mathtt{B})}$$

- dépendance :

$$dep[\mathtt{A} \Rightarrow \mathtt{B}] = |P(\mathtt{B}|\mathtt{A}) - P(\mathtt{B})|$$

- nouveauté :

$$nov[\mathtt{A} \Rightarrow \mathtt{B}] = P(\mathtt{A},\mathtt{B}) - P(\mathtt{A}) \times P(\mathtt{B})$$

- satisfaction :

$$sat[\mathtt{A} \Rightarrow \mathtt{B}] = \frac{(P(\neg\mathtt{B}) - P(\neg\mathtt{B}|\mathtt{A}))}{P(\neg\mathtt{B})}$$

- étonnement :

$$spr[\mathtt{A} \Rightarrow \mathtt{B}] = \frac{(P(\mathtt{A},\mathtt{B}) - P(\mathtt{A},\neg\mathtt{B})}{P(\mathtt{B}))}$$

Ces types de mesures peuvent être utiles à l'analyste dans son examen des règles. [Kodratoff, 2001] et [Cherfi and Toussaint, 2002] les utilisent dans des applications de fouille dans les textes.Dans sa thèse [Guillaume, 2000], Sylvie Guillaume propose une étude systématique des mesures d'intérêt. [Bastide, 2000] cite également un certain nombre de mesures probabilistes (fonction de Laplace, gain, gain d'entropie, gini, χ^2).

Règles d'association informatives

Parmi toutes les règles générées certaines sont redondantes. Pour réduire l'ensemble des règles à un échantillon pertinent, [Bastide, 2000] propose de se limiter à l'examen des règles informatives. Une règle *informative* est une règle dans laquelle la prémisse est minimale (en taille) tandis que la conclusion est maximale.

Il existe d'autres définitions des règles informatives.

Règles d'association informatives réduites

Il est encore possible de réduire l'ensemble des règles d'association supposées plus intéressantes que les autres en ne considérant que les règles d'association informatives réduites. Une règle informative réduite est une règle informative dont le support et la confiance sont supérieurs aux autres. Elle ne peut pas être déduite facilement d'une autre règle informative. [Bastide, 2000] démontre que l'ensemble des règles informatives approximatives réduites est suffisant pour générer toutes les règles approximatives valides avec leurs support et confiance.

Exemple :

Dans le tableau suivant, on trouvera quelques règles calculées sur la base de données $\mathcal{D}$.

règle exacte	règle approximative	règle informative
B $\Longrightarrow$ C (100%)	(B,C) $\Longrightarrow$ P (66,66%)	B $\Longrightarrow$ (P,C) (66,66%)
(B,P) $\Longrightarrow$ C (100%)	(A,C) $\Longrightarrow$ P (50%)	

Si l'on choisit un seuil de support de 30% et un seuil de confiance de 60%, les règles énoncées dans le tableau précédent sont valides à l'exception de la règle approximative dont la confiance est de 50%.

La règle (B,C) $\Longrightarrow$ P (66,66%) est considérée comme inutile car elle apporte moins d'informations que la règle informative présentée. Elle peut se déduire de cette règle informative et de la règle qui dit que B $\Longrightarrow$ C (100%).

Algorithmes

Les algorithmes permettant de rechercher des règles d'association sont très divers dans leur principe. En fait, ils varient essentiellement selon la nature des règles qu'ils extraient (règles exactes, règles partielles, règles informatives). Dans le cas de l'extraction des règles exactes, certains procédés conduisent à l'ensemble de toutes les règles (principe d'APRIORI) tandis que d'autres déterminent la base canonique des implications en éliminant les règles redondantes (règles pouvant être déduites par transitivité).

De nombreux algorithmes utilisent la structure du treillis de Galois pour extraire les règles. Dans [Bastide, 2000], Bastide indique trois stratégies différentes pour rechercher les règles d'association informatives réduites à partir de la structure des treillis :

- après la génération des motifs fermés fréquents,
- de façon simultanée à la génération des motifs fréquents,
- seulement génération des règles (possibilité introduite dans TITANIC).

5.3.3 Quelques autres méthodes de fouille de données

5.3.3.1 Les arbres de décision

L'histoire des arbres de décision est riche. Les différents domaines intéressés par de tels arbres les ont nommés et utilisés à des fins diverses. Ainsi, les premiers « arbres de décision » ont été conçus pour les besoins de la systématique. Les biologistes classent les espèces végétales et animales à l'aide de clés d'identification créées par les systématiciens [Lebbe, 1998]. Pour faciliter le travail des biologistes, les arbres dits d'identification se sont avérés pratiques à utiliser [Lebbe and Vignes, 1991]. Dans le domaine de l'analyse de données, ce type d'arbre a pris la dénomination d'arbre de segmentation. C'est dans le

cadre de l'apprentissage automatique que ces arbres ont été désignés par le terme d'arbre de décision.

Principe et définition

Un arbre de décision permet d'identifier un spécimen par rapport à une classification préexistante. L'identification est facilitée par le parcours de l'arbre de décision construit à partir de la classification.

Plus formellement [Simon, 2000], un arbre de décision est construit à partir d'un ensemble d'individus appelé ensemble d'apprentissage. Les individus partitionnés en classes C_i sont décrits par des attributs P_j de domaine D_j. Un arbre de décision est défini par un couple (S,A) où S est un ensemble de sommets et A un ensemble d'arcs entre les sommets de S. Étant donné que nous avons un arbre, les sommets sont de deux types : les nœuds et les sommets terminaux (ou feuilles). Un arbre de décision est un arbre étiqueté sur les nœuds par les attributs P_j, sur les feuilles par les classes C_i, et sur les arcs par un test portant sur la valeur de l'attribut P_j associé au noeud origine de l'arc. Cette définition d'un arbre de décision fait apparaître la possibilité de parcourir le graphe de façon orientée. L'identification d'un nouveau spécimen se fait en suivant le chemin correspondant aux propriétés de ce nouvel individu pour aboutir à une classe terminale.

La construction des arbres de décision est très sensible aux techniques employées ainsi qu'à l'échantillon d'individus choisi. La méthode de collecte des données peut être la cause du problème du « sur-apprentissage » ayant pour conséquence l'incapicité de classer de nouveaux individus.

Algorithme de construction d'un arbre de décision

Il existe plusieurs algorithmes de construction d'un arbre de décision. Le principe de base est un processus récursif qui consiste à sélectionner un attribut permettant de partitionner au mieux l'ensemble des individus puis à appliquer de nouveau ce procédé à chacun des sous-ensembles générés, une condition d'arrêt permet de détecter quand il faut interrompre le partitionnement. La difficulté réside dans le choix du test sur les attributs et de la condition d'arrêt, des arbres très différents pouvant être construits. De la qualité de ces choix dépendra celle de l'arbre obtenu. On a recours à des critères comme l'entropie pour choisir l'arbre qui « informe » le plus. Une heuristique générale dit que les arbres les plus simples (en quelque sorte les plus courts) seront préférables car ils sont censés conduire à l'identification des individus le plus rapidement possible. Pour cela, des méthodes d'élagage ont été mises au point [Lebbe and Vignes, 1991]. Construits sur le même principe général (Simon présente l'algorithme de base dans [Simon, 2000]), les algorithmes (ID3, CART, C4.5, MAKEY etc.) diffèrent du point de vue des heuristiques de construction et d'élagage choisies.

5.3.3.2 La classification hiérarchique incrémentale de concepts

Le terme de classification signifie ici le fait de regrouper les individus d'un ensemble en un certain nombre de classes différentes mais non forcément disjointes. Pour cela, il est nécessaire que les individus soient décrits en terme d'attributs et de valeurs associées à ces attributs (les valeurs pouvant être aussi bien numériques que symboliques). Ce type de

classification permet d'obtenir des classes décrites chacune par un concept et organisées en hiérarchie. C'est en ce sens que la classification conceptuelle ne propose pas le même point de vue que les classifications numériques de l'analyse de données. La classification peut être automatique ou semi-automatique. Dans l'approche automatique, la construction de la hiérarchie de classes est non supervisée et se fait de façon descendante. L'algorithme incrémental de référence est COBWEB [Simon, 2000]. Au départ, la hiérarchie est vide. Le premier individu constitue le premier concept initial. Les individus suivants sont présentés séquentiellement et comparés aux concepts existants. Un nouveau concept est créé si l'individu courant est différent des concepts déjà présents et ainsi de suite. Plusieurs opérations sur les concepts sont nécessaires pour mettre à jour la hiérarchie [Gennari et al., 1989] : création d'un concept, fusion de plusieurs concepts, division d'un concept, suppression d'un concept. Le processus est dirigé par une fonction d'évaluation de la proximité entre individu et concept.

5.3.3.3 Autres méthodes

Il existe de nombreuses autres méthodes s'appuyant sur des théories très variées. Parmi celles-ci, on peut citer à titre indicatif :

- la classification bayésienne qui utilise des travaux sur les probabilités et est fondée sur le théorème de Bayes [Rich and Knight, 1991],
- les réseaux de neurones qui se fondent sur sur une étude du fonctionnement du cerveau humain (neurone biologique) transposée à un comportement logique (neurone artificiel) [Medler, 1998],
- les algorithmes génétiques dont le principe s'appuie sur la théorie de l'évolution énoncée par Darwin [Eiben and Schoenauer, 2002].

5.4 Conclusion

Dans ce chapitre, nous avons décrit le processus d'extraction de connaissances dans les bases de données. Nous avons vu l'importance de chacune des étapes de cette activité éminemment itérative et interactive. C'est ce processus que nous avons appliqué aux bases de données de réactions chimiques dans le but d'enrichir les bases de connaissances du système d'aide à la planification de synthèse RESYN ASSISTANT.

Troisième partie

L'acquisition de connaissances à partir de bases de données de réactions

Chapitre 6

La modélisation conceptuelle de la planification en synthèse organique

La nature nous pousse à avoir un esprit ambitieux... qui voudrait bien comprendre la merveilleuse architecture du monde.

Christopher Marlowe

Sommaire

6.1 Introduction

Travaillant à l'acquisition de connaissances dans le cadre de la conception d'un système d'information chimique susceptible d'aider à élaborer des plans de synthèse, nous avons à modéliser les connaissances utiles dans ce domaine et la façon dont celles-ci sont utilisées. Ceci nous amène à présenter les principales caractéristiques du modèle d'un agent rationnel

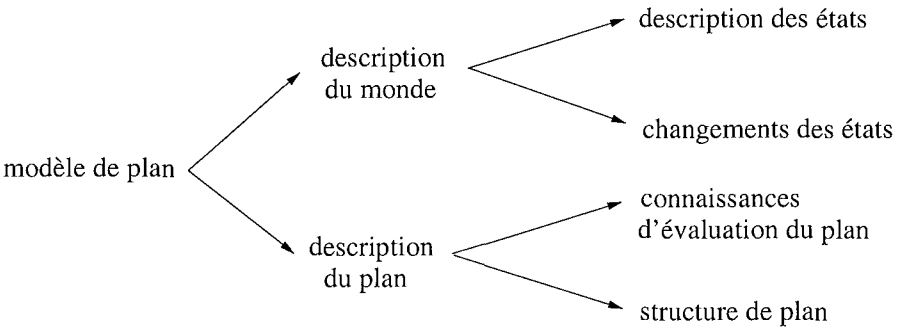


FIG. 6.1 – Modèle de plan selon Valente

de planification en synthèse organique dont les bases ont été établies par C. Laurenço et M. Py [Laurenço et al., 1990]. Ce modèle a été largement développé par le GDR TICCO, chaque nouvelle étude entreprise dans ce cadre ayant contribué à son raffinement. Son évolution se poursuit encore aujourd'hui.

L'agent qui est modélisé n'a pas pour objet de simuler le comportement complexe d'un chimiste pratiquant la synthèse organique mais de collaborer avec celui-ci en lui rendant « intelligemment »certains services. Aussi, cet agent est-il très interactif afin que le chimiste puisse garder le contrôle des opérations durant tout le processus de planification. Au niveau conceptuel, il se décrit en termes des connaissances et des modes de raisonnement dont il dispose pour répondre aux sollicitations du chimiste.

L'objectif d'un agent de planification est la construction d'un plan. Cet agent doit donc avoir des connaissances se rapportant à un tel plan dans son domaine d'activité. Celles-ci peuvent se définir à partir d'un modèle général de plan, proposé par Valente dans le contexte d'une analyse de la planification selon la méthodologie KADS [Valente, 1995], que la figure 6.1 présente sous une forme simplifiée.

Ce modèle est composé de la description du monde, c'est-à-dire le domaine dans lequel la planification s'effectue, et de la description du plan. La description du monde contient les connaissances nécessaires pour décrire les états de l'espace du problème (en synthèse organique, un stade du plan est décrit par des structures moléculaires), ainsi qu'à celle des opérateurs permettant le passage d'un état à un autre (les méthodes de synthèse ou les transformations correspondantes, par exemple). La description du plan recouvre à la fois la structure type du plan et les connaissances qui permettent l'évaluation d'un plan engendré et sa comparaison avec d'autres plans. À ce modèle de plan doivent être associées des connaissances stratégiques visant à contrôler le processus de planification, l'ensemble constituant les connaissances du domaine qui ont un rôle statique – car elles ne varient pas – dans le processus de planification. Par ailleurs, l'agent doit pouvoir disposer de connaissances ayant un rôle dynamique afin de décrire le stade d'avancement de la planification : l'état initial et l'état final de l'espace du problème posé, les parties du plan déjà construites, les objectifs restant à atteindre, l'état courant de l'espace à partir duquel on cherche à atteindre un objectif, etc.

Les modes de raisonnement exploitant les connaissances du domaine pour construire

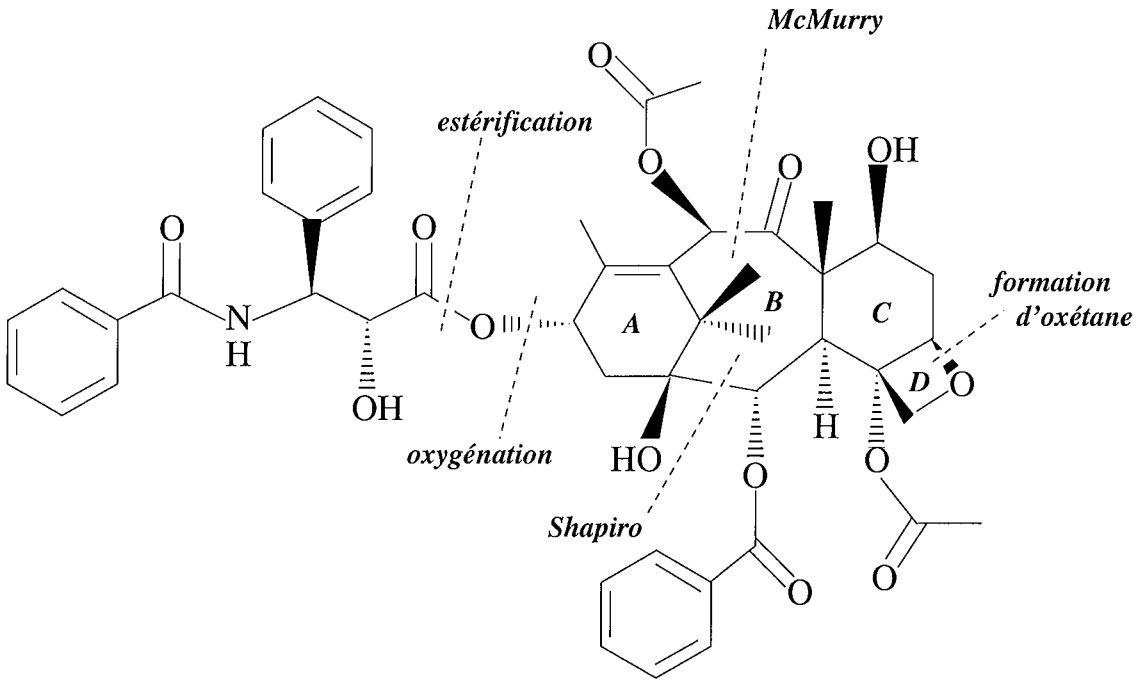


FIG. 6.2 – Déconnexions de liaisons stratégiques dans la molécule de Taxol selon Nicolaou et col. [Nicolaou et al., 1994]

un plan s'expriment en termes de méthodes de résolution de problème. Elles-mêmes sont décrites par des tâches et des méthodes susceptibles d'accomplir celles-ci, ainsi qu'en termes de structures de contrôle organisant les tâches. Dans ce modèle, au départ du processus de planification, on dispose d'une représentation de la structure cible et éventuellement d'une information sur ses propriétés et/ou sur des molécules analogues ainsi que des données sur le contexte dans lequel va se réaliser la synthèse, c'est-à-dire sur les motivations de la synthèse et les moyens pouvant être mobilisés. L'élaboration d'un plan de synthèse procède selon une démarche descendante, à travers une hiérarchie de niveaux. Elle commence avec des considérations des plus générales et se raffine progressivement jusqu'à examiner les détails les plus fins. Nous allons essayer de l'illustrer avec l'exemple du Taxol (figure 6.2), en parcourant les quatre niveaux de la hiérarchie sans distinction des actions accomplies respectivement par l'agent et par le chimiste.

1. **niveau politique** : c'est le niveau le plus général, on prend en compte le contexte et les buts de la synthèse ; par exemple, dans un contexte industriel, le Taxol est probablement trop complexe pour que soit recherchée une synthèse totale et on se tournera plutôt vers une synthèse partielle ; à l'inverse, dans un contexte académique, on recherchera une synthèse totale en profitant de la complexité la molécule pour étudier de nouvelles méthodes de synthèse. Les choix faits à ce niveau déterminent un certain nombre de contraintes qui vont se propager sur les niveaux inférieurs ; par exemple, dans le cas d'une synthèse partielle, la tâche prioritaire au niveau stratégique sera de rechercher un produit de départ aussi proche que possible de la cible, tout

en tenant compte de facteurs économiques, toxicologiques, etc. Il peut être spécifié que l'emploi de certaines classes de molécules ou de méthodes de synthèse est à privilégier ou à éviter voire à interdire. Ceci restreint l'espace dans lequel la recherche de solutions peut s'effectuer.

2. **niveau stratégique** : c'est le niveau où l'on détermine des objectifs stratégiques et on essaie de les ordonner. La démarche est guidée par les stratégies de planification et les heuristiques du domaine. Dans le cas d'une synthèse totale du Taxol, d'un point de vue rétrosynthétique, l'objectif le plus évident est de séparer les deux parties alcaloïdique et terpénique, ce qui implique de transformer le groupement ester qui les relie en déconnectant au moins une de ses liaisons. Ceci permet de décomposer le problème initial en deux sous-problèmes de nature très différente et répond au principe de convergence dont nous avons parlé en section 2.2. Si on considère ensuite la partie terpénique, on constate que les motifs acyle et benzoyle sont secondaires par rapport au système polycyclique complexe `ABCD` et qu'ils peuvent être ignorés à ce stade. On constate aussi que dans le système polycyclique c'est l'oxétane `D` qui est la partie la plus fragile et qu'il faudrait la transformer en une structure plus stable. Si le polycycle carboné `ABC` n'apparaît pas comme étant globalement accessible par des méthodes connues, ses différents sous-systèmes seront considérés, ceci jusqu'à l'examen individuel de chaque cycle. La transformation du cycle à 8 chaînons est un objectif incontournable tandis que les cycles à 6 chaînons peuvent être éventuellement recherchés dans des produits de départ potentiels. Et ainsi de suite. Les objectifs stratégiques sont déterminés sans préjuger des transformations particulières qui permettront de les atteindre ou des molécules disponibles particulières pouvant servir de produit de départ.
3. **niveau tactique** : c'est le niveau où l'on va rechercher les moyens permettant d'atteindre les objectifs fixés précédemment. Rétrosynthétiquement, ce sont les transformations correspondant aux méthodes de synthèse les mieux adaptées au contexte structural de la molécule cible. Si une transformation ne peut pas être appliquée directement, on engendrera des objectifs tactiques permettant d'adapter la structure ; par exemple, si le couplage de McMurry apparaît comme susceptible de construire le cycle à 8 chaînons du Taxol, la transformation correspondante ne peut pas lui être appliquée car elle nécessite la présence d'un motif diol 1,2 ; il faut donc d'abord transformer le motif acyloïne présent sur le cycle en diol-1,2. Si l'objectif est de ne pas transformer une structure, on recherchera des produits de départ pouvant la fournir.
4. **niveau opérationnel** : c'est le niveau où l'on examine les détails du franchissement des étapes. L'opérationnalisation du plan consiste à passer des transformations appliquées au niveau tactique aux réactions élémentaires constituant les méthodes de synthèse qui correspondent à ces transformations. On pourra spécifier des conditions réactionnelles et faire une simulation, dans le sens synthétique, des réactions mises en jeu afin de vérifier qu'il n'y a pas de réactions concurrentes ; on envisagera, si nécessaire, des protections de sites réactifs, etc.

À chaque niveau, il faut contrôler la validité et la pertinence de chaque action ainsi que

la cohérence globale du plan dans son état courant. La démarche d'élaboration d'un plan comporte nécessairement des retours en arrière. Si l'on rencontre un échec à un certain niveau, on pourra remonter à un niveau plus général pour repartir dans une autre direction. Le parcours dans cette structure hiérarchique n'est pas systématique, c'est-à-dire qu'on n'est pas obligé d'avoir exploré entièrement un niveau pour passer au suivant. Par exemple, si l'on reconnaît qu'une partie de la cible est une sous-structure correspondant à un objectif majeur de la synthèse, on pourra l'étudier en priorité et en profondeur, en faisant abstraction du reste. Par ailleurs, l'application d'une transformation peut modifier profondément la structure cible et faire apparaître de nouveaux objectifs plus pertinents que ceux qui restaient à atteindre, ce qui impose de ré-analyser le problème. De plus le processus n'est pas linéaire, les objectifs dépendant fréquemment les uns des autres. Chaque niveau nécessite des connaissances et des méthodes spécifiques ainsi qu'une représentation particulière du problème, la représentation globale pouvant être vue comme une hiérarchie des plans abstraits allant du plus général au plus particulier. Ce modèle a été décrit en détail dans la thèse de Gien [Gien, 1998]. Cependant, certains de ces aspects n'ont pas encore été approfondis, c'est particulièrement le cas du niveau opérationnel. La prédiction des réactions pose des problèmes complexes dont la résolution fait appel à un domaine de connaissances et des méthodes différents de ceux qui sont employés aux autres niveaux. Bien que des méthodes empiriques basées sur l'analogie soient utilisables, les méthodes théoriques, statistiques, etc. doivent jouer ici un rôle prépondérant. D'autres aspects, mieux définis, n'ont pas encore été implémentés et expérimentés sur les plates-formes ReSyn ou ReSyn Assistant. Dans ce chapitre, nous ne décrirons que les parties du modèle qui sont relatives à notre étude, c'est-à-dire qui correspondent aux niveaux stratégique et tactique.

6.2 Les objets du domaine : molécules, méthodes et plans de synthèse

Très généralement, le problème de la synthèse des molécules organiques peut se diviser en deux sous-problèmes, d'une part la construction de leur squelette carboné et d'autre part l'introduction, la modification et/ou l'élimination de divers groupements fonctionnels [Ireland, 1969]. Cette constatation amena Ireland à considérer qu'une catégorisation des réactions organiques en deux groupes, celles qui forment des liaisons carbone-carbone et celles qui changent seulement la fonctionnalité, devait être utile en synthèse. Il compléta cette vision un peu sommaire par l'observation que certaines caractéristiques de la plupart des molécules ont des analogies avec des systèmes moins complexes pour lesquels existent des méthodes de synthèse. Afin de rendre évidentes de telles méthodes au cours de la planification d'une synthèse, on devrait donc se concentrer moins sur les carbones individuels que sur des parties de la cible qui constituent des sous-unités distinctes. Enfin, beaucoup de molécules que l'on cherche à synthétiser ont plusieurs stéréo-isomères possibles et il est clair qu'un plan de synthèse convenable devrait conduire à une stéréochimie correcte tout autant qu'au squelette et à la fonctionnalité requis par la structure cible. Ces éléments

fournissent la base de la description des objets du domaine de la synthèse organique. Ils permettent de définir des représentations abstraites appropriées pour classifier ces objets et reconnaître des similitudes grâce auxquelles des problèmes pourront être résolus par analogie.

6.2.1 Les molécules

Fondamentalement, la chimie n'est pas la science des molécules mais celle de la transformation des substances [Brakel, 1997]. Une substance (pure) se définit par rapport à des propriétés macroscopiques physiques (température d'ébullition, densité, etc.), chimiques (produits de l'analyse ou de la synthèse) ou structurelles (réseau de relations entre des substances et les réactions chimiques dans lesquelles elles entrent). Cependant, elle ne peut pas se définir en terme de structure moléculaire suivant une voie réductionniste. De même, la molécule ne peut pas se définir en terme de mécanique quantique. Bien qu'on ne doive pas confondre les concepts de substance et de molécule, les chimistes pratiquant la synthèse admettent généralement qu'une substance pure se définit en termes de composition atomique et de structure moléculaire. Quand les circonstances l'exigent, ils peuvent adopter des représentations basées sur des modèles géométriques ou de chimie théorique mais celle qu'ils utilisent couramment est la formule structurale dérivée de la structure de Lewis. Dans ce modèle, les molécules sont des assemblages d'atomes reliés par des liaisons. La notion de liaison est difficile à cerner étant donné qu'elle ne représente pas une réalité matérielle [Chevreau et al., 2001]. C'est un concept qui a permis de rendre compte de la cohésion des édifices moléculaires et que Lewis a symbolisé par des traits tirés entre les atomes représentés par leurs symboles atomiques. Comme, par ailleurs, elle est aisément transposée en graphe, la structure de Lewis nous sert de base pour décrire l'objet principal de la synthèse. Néanmoins, ce n'est qu'un modèle très imparfait, plus proche de l'entité moléculaire que de la substance, aussi nous emploierons le terme « molécule » plutôt que « substance » pour nommer cet objet.

Le modèle « molécule » est représenté sur le diagramme de classes UML[1] de la figure 6.3. Une molécule est une structure. Dans notre cas, nous postulons qu'un graphe est composé d'au moins deux sommets ainsi que d'arêtes représentant des liens entre les sommets. Un sommet peut être relié à 1 ou n autres sommets. Les sommets et les arêtes d'une structure sont identifiés par leur numéro dans la structure. Une molécule est une structure résultant de l'agrégation de sommets et d'arêtes particuliers que l'on appelle respectivement atomes et liaisons. Un atome est un sommet, il possède un `nom`, un `symbole`, un `type`, une `charge` et une `valence`. Une liaison est une arête et a un `nom` et un `type`.

Dans notre propos, nous parlerons de :

- liaison simple au lieu de liaison de type σ,
- liaison double pour signifier deux liaisons, une de type σ et l'autre de type π, entre deux atomes,

[1]Pour une brève description des concepts présentés dans les diagrammes de classes UML et de leur représentation, voir l'annexe A.

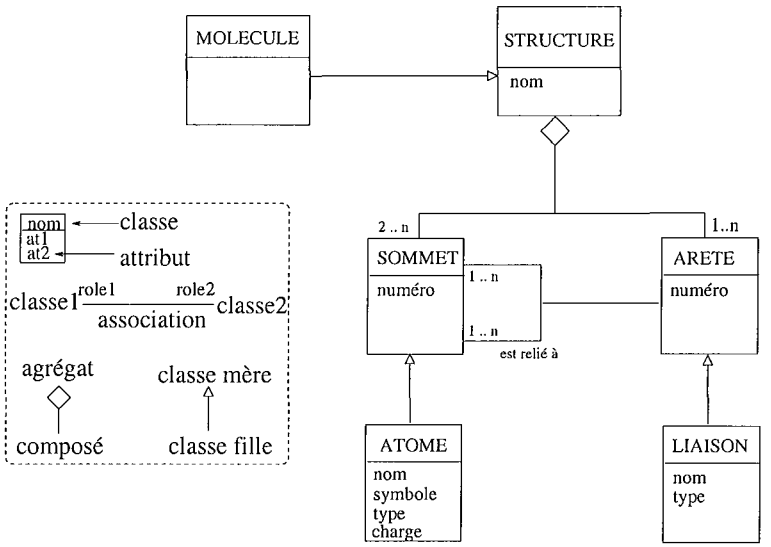


FIG. 6.3 – Le modèle des molécules. Une molécule est une structure qui est un agrégat d'atomes et de liaisons qui lient deux atomes entre eux.

- liaison triple pour décrire l'ensemble d'une liaison σ et de deux liaisons de type π entre deux atomes donnés.

Le type d'un atome permet de connaître l'élément chimique auquel est rattaché l'atome considéré (carbone, hydrogène, oxygène, azote, chlore, brome, etc.). Le symbole (`C`, `H`, `O`, `N`, `Cl`, `Br`, etc.) et le nom de l'atome dérivent du type de l'atome. Nous pouvons classer les atomes et les liaisons selon leur type. Ainsi nous travaillons sur une hiérarchie des atomes qui s'appuie en grande partie sur la classification des éléments chimiques de Mendeleev. Nous avons rajouté la différenciation entre les atomes des éléments carbone et hydrogène, d'une part, et les hétéroatomes, d'autre part. La figure 6.4 donne une petite partie de la hiérarchie des atomes sur laquelle nous nous basons

Une hiérarchie des liaisons a été également établie. La partie de la hiérarchie des liaisons présentée figure 6.4 donne un exemple d'héritage multiple. Au niveau des liaisons de covalence, nous distinguons les liaisons suivant leur type simple, double, triple ou aromatique. Pour rendre compte des propriétés particulières des liaisons constituant les cycles aromatiques de taille impaire tels que les pyrroles, les furanes, etc., on crée le type `Liaison simple aromatique` qui hérite à la fois des classes `Liaison aromatique` et `Liaison simple`. De même pour la classe `Liaison double aromatique` qui hérite à la fois des classes `Liaison double` et `Liaison aromatique`. Nous verrons dans la section 6.2.1.2 l'utilité et la signification des types de liaison aromatique, double aromatique et simple aromatique. Les liaisons de type simple avant et simple arrière correspondent aux liaisons qui sont indiquées hors du plan de la feuille sur laquelle est dessinée la structure moléculaire. Nous ne discuterons pas ici des liaisons ioniques et métalliques.

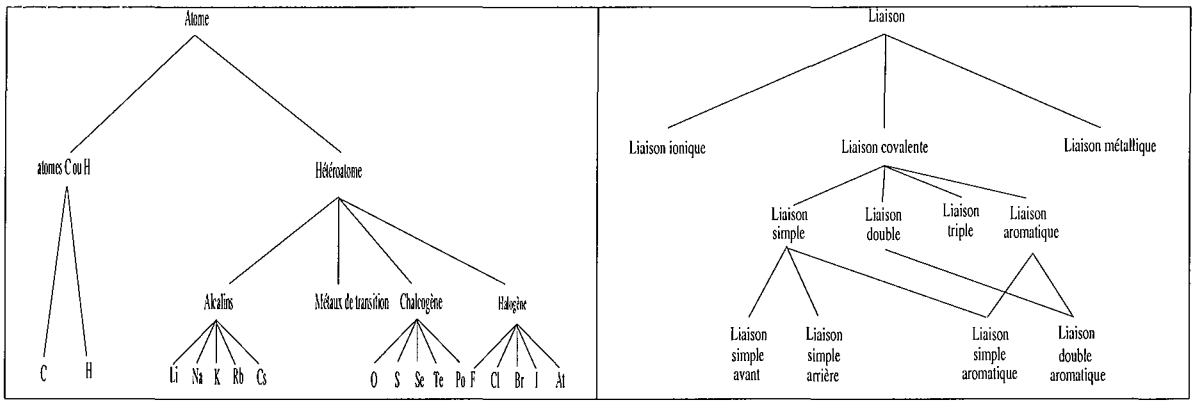


FIG. 6.4 – Une partie de la hiérarchie des atomes ainsi que de celles des liaisons.

6.2.1.1 Abstraction, perception, classification

La complexité des molécules est souvent trop grande pour qu'on ait une compréhension globale du problème de leur synthèse. Une partie capitale de la compréhension d'un tel problème résulte de la perception de la molécule cible. Partant de la formule structurale de la cible, on analyse celle-ci afin d'en reconnaître les composants et leurs relations. Au-delà des atomes et des liaisons, on reconnaît des motifs structuraux caractéristiques. Puis, employant ces informations, à l'aide des connaissances du domaine, on classifie la molécule. La classification se fait généralement dans plusieurs hiérarchies différentes ; par exemple, une molécule peut être à la fois un dérivé polycyclique, un alcool, un terpène, etc. La perception fournit une base pour l'action puisque l'appartenance d'une molécule à des familles chimiques détermine certaines de ses propriétés, en l'occurrence ses modes d'obtention, sa stabilité, etc. De la perception doit se dégager un ensemble structuré de représentations constituant une représentation globale de la molécule, elle-même partie intégrante de la représentation du problème. C'est en raisonnant sur cette représentation que les objectifs stratégiques à atteindre et les moyens tactiques à mettre en œuvre seront déterminés et que les solutions possibles seront développées. Le moteur de ce processus est l'abstraction. L'abstraction est une opération constamment employée en résolution de problèmes. Elle intervient dans la décomposition des problèmes, la reconnaissance d'une similitude entre des objets ou des situations, le raisonnement par analogie, l'acquisition de connaissances, etc. En planification, le parcours pour rechercher un chemin peut s'opérer dans une hiérarchie d'espaces abstraits, notre modèle en donne un exemple. Une représentation abstraite doit décrire l'essentiel d'un objet ou d'une situation. On peut obtenir de telles représentations à partir de représentations détaillées en employant différents types de procédés, notamment :

- la simplification par élimination de détails superflus,
- la généralisation par remplacement de termes particuliers par de plus génériques,
- la sélection par la prise en compte d'un seul aspect parmi d'autres,
- la schématisation par la décomposition puis la recomposition d'un système,
- la construction par ajout d'information expliquant les raisons d'une situation ou d'une action.

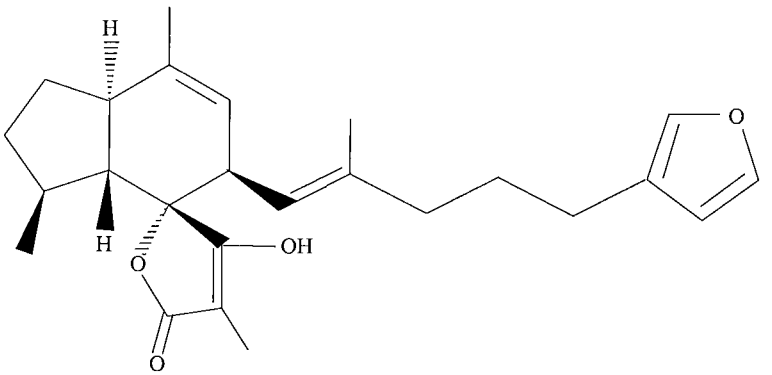


FIG. 6.5 – La structure moléculaire de la molécule d'Ircinianin.

Des hiérarchies de représentations abstraites peuvent être construites en réitérant plusieurs fois l'emploi de ces procédés. C'est une question qui a été beaucoup étudiée par le GDR 1093 TICCO, en relation avec la perception des molécules. L'objet « molécule » est décrit à la fois selon plusieurs points de vue et à plusieurs niveaux d'abstraction, le graphe moléculaire correspondant au modèle défini au paragraphe précédent faisant la jonction entre les différentes représentations abstraites.

6.2.1.2 Les points de vue

Les points de vue correspondent à une abstraction obtenue par la sélection de différents aspects de l'objet « molécule » considérés indépendamment. Ils exhibent chacun des caractéristiques pertinentes relatives à sa topologie, sa fonctionnalité ou sa stéréochimie.

Nous allons illustrer notre propos en analysant la molécule d'Ircinianin dont la structure moléculaire est représentée sur la figure 6.5.

La topologie

Selon la définition de D. H. Rouvray [Rouvray, 1986], la topologie est l'organisation des interconnexions entre les atomes qui détermine l'architecture fondamentale de la molécule. L'analyse topologique considère combien il y a d'atomes dans la molécule, à combien d'autres atomes chacun d'eux est connecté et, pour un ensemble d'atomes connectés, s'ils forment une simple chaîne, une chaîne ramifiée, un cycle (ou un polycycle), ou une combinaison de cycles et de chaînes. Rouvray précise que pour l'analyse topologique d'une molécule, la forme réelle tridimensionnelle, la longueur et la nature des liaisons, les angles de liaisons et parfois même le type des atomes sont sans importance.

La topologie est le point de vue qui s'attache à la description de la structure moléculaire en terme de voisinage des atomes. Les éléments de nature topologique sont perçus à partir de la structure correspondant au squelette. Au niveau topologique, on met en évidence, dans un premier temps, les chaînes et les cycles. En toute rigueur, à ce niveau, la nature des atomes et le type des liaisons sont des propriétés qui ne doivent pas être considérées. Cependant, les travaux du GDR TICCO ont montré qu'il peut être utile de

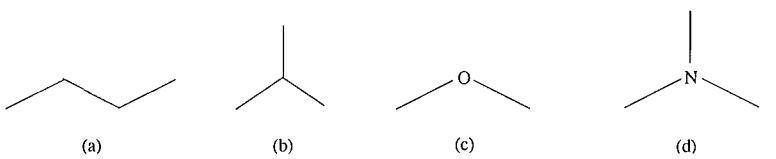


FIG. 6.6 – Les différents types de chaînes. (a) chaîne carbonée linéaire. (b) chaîne carbonée ramifiée. (c) chaîne hétéroatomique linéaire. (d) chaîne hétéroatomique ramifiée.

mettre en évidence les chaînes particulières (figure 6.6) telles que les chaînes hétéroatomiques[2] et celles qui lient les parties cycliques aux parties non cycliques ou bien à d'autres parties cycliques. Au niveau des chaînes, on distingue également les chaînes linéaires des chaînes ramifiées. Les chaînes hétéroatomiques linéaires et ramifiées sont englobées sous le concept de *lien hétéroatomique* car ces parties de la molécule constituent un lien entre deux parties carbonées. Ces liens ont un intérêt synthétique particulier. De même, pour les liaisons entre des parties cycliques et non cycliques que nous appelons *lien de cycle.* Les chaînes carbonées ramifiées contiennent un ou des atomes de carbone liés par trois ou au plus quatre liaisons à d'autres atomes de carbone. Ces atomes présentent, eux aussi, des caractéristiques intéressantes du point de vue synthétique : nous les appelons *atome de carbone de branchement* ou *carbone de branchement* par abus de langage. De par leur connectivité, ces parties du squelette sont des nœuds de complexité. Un atome de carbone de branchement est, lui aussi, un lien entre d'autres parties du squelette.

Si l'on considère les tensions présentes dans les structures cycliques, celles-ci peuvent être classées en fonction de leur taille :

- petits cycles : ce sont les cycles de taille trois et quatre. Les petits angles induisent de fortes contraintes et font de ces cycles des structures instables,
- cycles communs : ce sont les cycles de taille cinq, six et sept. Les tensions sont dans ce cas faibles. Les cycles de ce type sont très présents dans les molécules en particulier ceux de taille cinq et six,
- cycles moyens : ce sont les cycles dont la taille est comprise entre 8 et 11 inclus. Ces cycles se rencontrent assez rarement car les tensions y sont très grandes du fait des angles que font les liaisons entre elles,
- grands cycles : ce sont les cycles de taille supérieure ou égale à 12. Ces cycles ne sont pas ou peu contraints.

Pour chaque classe de cycle, il existe des méthodes de synthèse qui sont plus spécifiquement adaptées.

De même qu'au niveau des chaînes, une distinction peut être faite en considérant le type des atomes du cycle. Lorsqu'un cycle est constitué exclusivement d'atomes de carbone, on dit que l'on a un *cycle carboné* ou *carbocycle.* S'il est composé d'au moins un hétéroatome, on parle d'un *hétérocycle* (figure 6.7). Ainsi, par exemple, un cycle comportant un atome d'azote sera désigné plus précisément par le terme « cycle azoté ».

Si l'on considère la molécule d'Ircinianin (figure 6.5) du point de vue topologique, on

[2]c'est-à-dire des chaînes qui ne sont pas uniquement carbonées mais contiennent des hétéroatomes.

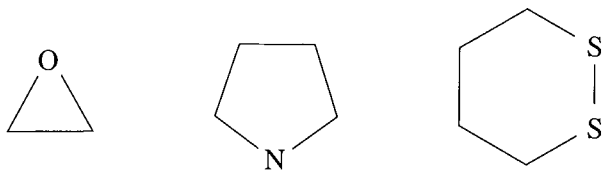


FIG. 6.7 – Trois exemples d'hétérocycles. (a) un cycle oxygéné de taille 3. (b) un cycle azoté de taille 5. (c) un cycle soufré de taille 6.

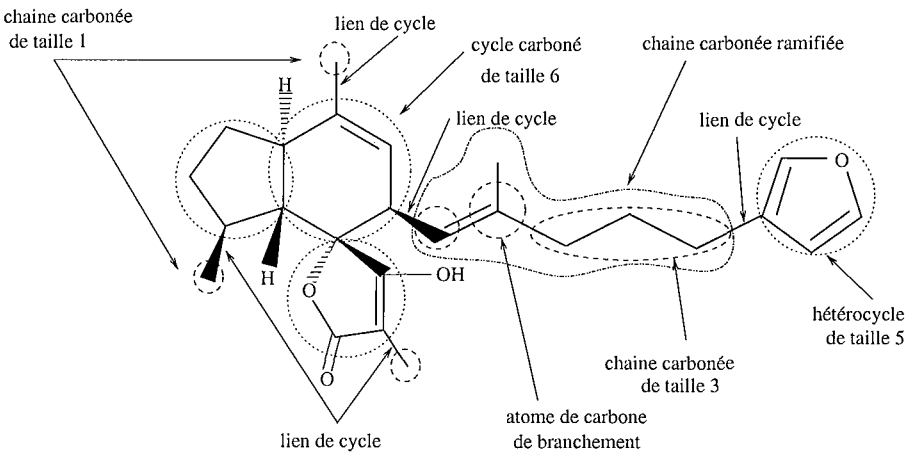


FIG. 6.8 – La molécule d'Ircinianin selon le point de vue de la topologie.

voit que cette molécule comporte deux parties cycliques séparées par une chaîne ramifiée (figure 6.8).

La fonctionnalité

Le terme de *groupement fonctionnel* est très souvent utilisé dans la littérature lorsqu'il s'agit de désigner les sous-structures moléculaires de nature fonctionnelle. Ainsi, IUPAC propose une définition des groupements fonctionnels (*functional group*). Dans la recommandation IUPAC, la notion de groupement fonctionnel désigne tout atome ou groupement d'atomes dont le comportement est similaire quel que soit les composés chimiques dans lesquels il apparaît. Si l'on suit ce raisonnement, tout groupe d'atomes peut finalement être considéré comme un groupement fonctionnel. Même les chaînes carbonées saturées[3] ont des propriétés données. Étant donné que nous voulons rendre compte, du point de vue de la fonctionnalité d'éléments bien précis, nous écarterons cette définition qui revient à dire que tout est fonctionnel dans une molécule.

Le point de vue de la fonctionnalité donne des indications sur les centres réactifs d'une molécule. En effet, comme dit précédemment, il existe une différence de stabilité entre les liaisons simples entre deux atomes de carbone et celles de type multiple et/ou entre

[3]ensembles d'atomes de carbone liés entre eux par des liaisons simples.

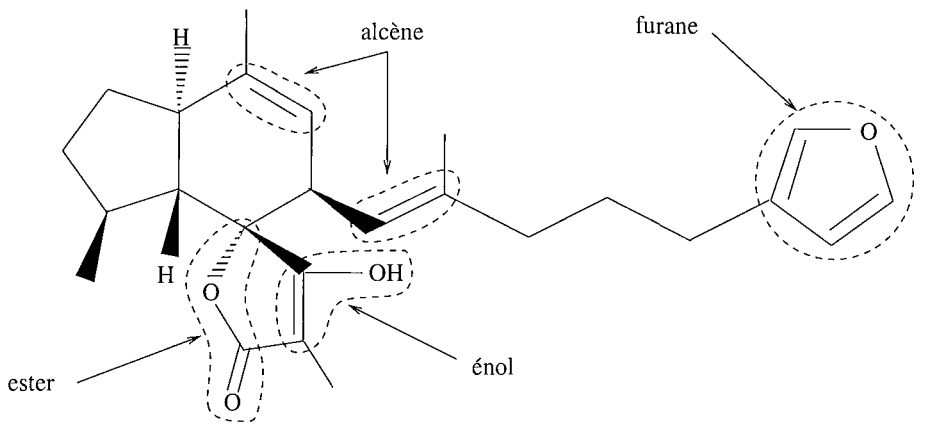


FIG. 6.9 – La molécule d'Ircinianin selon le point de vue de la fonctionnalité.

atomes de natures différentes ou hétéroatomiques. Ce deuxième type de liaison est dite *liaison fonctionnelle.* Nous appellerons, par la suite, *fonction* tout ensemble d'atomes liés par des liaisons fonctionnelles. De nature algorithmique, cette définition a le mérite d'être précise. La présence d'une fonction influence également la réactivité des liaisons placées à son voisinage.

La figure 6.9 montre les fonctions détectées au niveau de la molécule d'Ircinianin.

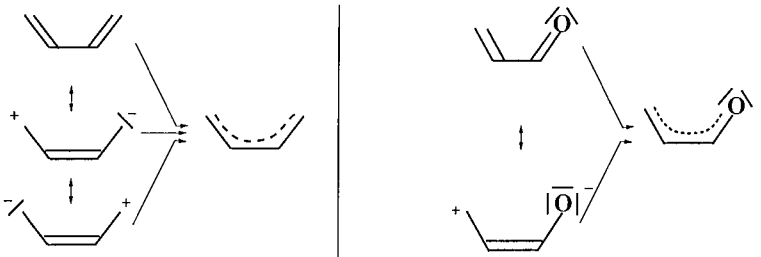
La mésomérie

Outre la définition, de type algorithmique, des fonctions qui a été posée, il est important de tenir compte de propriétés telles que la mésomérie. La mésomérie permet de décrire la délocalisation des électrons au travers des formules de Lewis classiques. Une même molécule est alors décrite par plusieurs formules, appelées formes limites ou formes mésomères, différant uniquement par la localisation de certains électrons. La structure réelle de la molécule ayant des propriétés de mésomérie est un hybride de résonance de l'ensemble des formes limites. En fait, ce qu'il est important de prendre en compte est le caractère particulier des liaisons du système conjugué[4]. Ces liaisons ne sont pas de type purement simple ou double car les électrons π participant à ces liaisons ne sont pas localisés. Les liaisons simples des systèmes conjugués ont donc un caractère partiel de liaison double. Sur la figure 6.10, nous avons représenté deux cas de mésomérie.

L'aromaticité

L'aromaticité est un cas particulier de la mésomérie. Cette propriété concerne un certain type de cycles à qui elle confère une stabilité particulière ainsi qu'une réactivité spécifique. On parle de « chimie des aromatiques ». Un cycle qui a la propriété d'aromaticité est appelé *cycle aromatique.* Les cycles carbonés aussi bien que les hétérocycles peuvent être aromatiques, cependant certains cycles n'ont pas un caractère pur de cycle aromatique et d'autres propriétés doivent être prises en compte parallèlement. C'est le cas pour les hétérocycles azotés de taille 5 tel que le pyrrole. On observera pour ce type de cycle à

[4]un système de liaisons conjuguées est un ensemble de liaisons qui sont alternativement de type simple et double. Dans un système conjugué, deux liaisons consécutives ont des types différents.



cas des diènes | cas des carbonyles α, β éthyléniques

FIG. 6.10 – Deux exemples de la mésomèrie : le cas des diènes et le cas des carbonyles α, β éthyléniques.

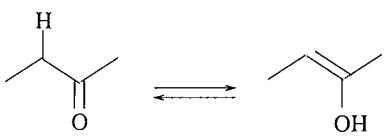


FIG. 6.11 – Un exemple particulier de la tautomérie : la prototopie dans l'équilibre céto-énolique.

la fois une réactivité spécifique de l'aromaticité ainsi que celle qui peut être associée aux liaisons non délocalisées. Pour rendre compte des caractéristiques particulières des liaisons des cycles aromatiques, un type `Liaison aromatique` a été créé. Le cas des liaisons des cycles de taille impaire est traité de telle façon que les caractères aromatique et de simple ou double liaison soient exprimés conjointement (type `Liaison simple aromatique` et type `Liaison double aromatique`).

La tautomérie

Un des exemples les plus cités du phénomène de la *tautomérie* est l'équilibre céto-énolique[5] (figure 6.11).

Contrairement à la mésomérie, et donc à l'aromaticité, la tautomérie se manifeste par un véritable équilibre chimique entre des isomères de constitution[6]. Les formes tautomères s'interconvertissent réversiblement les unes en les autres. Le problème soulevé par la tautomérie est celui de la représentation. À cause de la tautomérie, une même substance peut être associée à plusieurs structures moléculaires.

Si nous revenons à la molécule d'Ircinianin que nous avons pris pour exemple, nous constatons que cette molécule possède à la fois des propriétés de tautomérie (présence de l'énol) et de mésomérie (présence d'un carbonyle α, β éthylénique).

La stéréochimie

La stéréochimie est le point de vue tenant compte de l'aspect tridimensionnel des molé-

[5]l'équilibre céto-énolique est en fait un cas particulier de la tautomérie, c'est-à-dire la *prototropie*.

[6]deux structures moléculaires sont des isomères de constitution si elles ont la même composition en terme d'atomes. Elles diffèrent l'une de l'autre par l'arrangement des liaisons.

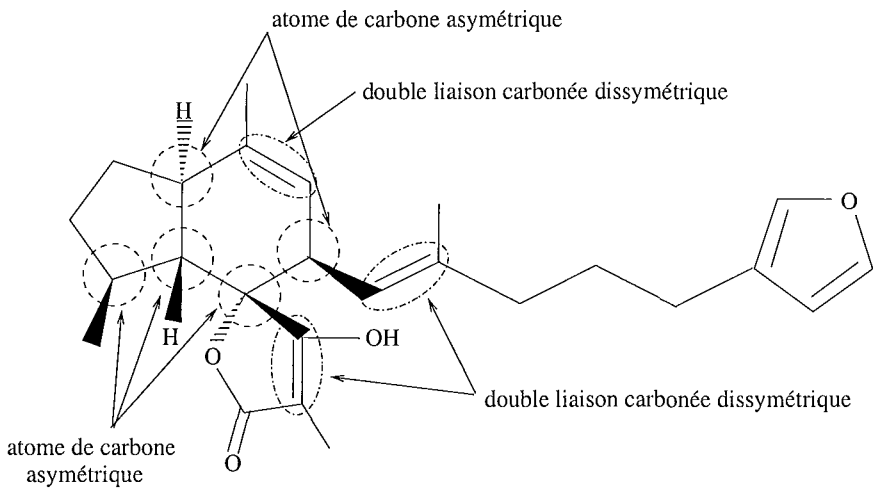


FIG. 6.12 – La molécule d'Ircinianin selon le point de vue de la stéréochimie.

cules. L'arrangement spatial des atomes des molécules a une grande importance. En effet, deux molécules topologiquement équivalentes peuvent différer uniquement du point de vue de la disposition relative dans l'espace des atomes qui les constituent. Cette seule différence est susceptible d'avoir énormément de conséquences sur les propriétés respectives des deux molécules, en particulier sur leur activité biologique[7]. Les motifs caractéristiques des propriétés stéréochimiques, les *stéréogènes*, sont soit des atomes asymétriques, soit des liaisons disymétriques, soit des conformations dues à des arrangements relatifs de groupements d'atomes les uns par rapport aux autres.

Dans la structure moléculaire de l'Ircinianin, on distingue plusieurs stéréogènes (sources de chiralité ou d'isomérie géométrique). Sur la figure 6.12, les différents stéréogènes (atomes asymétriques et liaisons disymétriques) sont mis en évidence.

L'existence de stéréogènes dans une molécule confère à celle-ci une complexité supplémentaire. Synthétiser de façon sélective une molécule donnée ayant des propriétés stéréochimiques est un exercice difficile mais essentiel. Les résultats de la perception selon ce point de vue doivent être pris en compte le plus tôt possible au niveau de la rétrosynthèse. Pour construire une molécule ayant des propriétés stéréochimiques, il faudra adopter des stratégies de résolution particulières et employer des tactiques adaptées. La détection et la gestion de ces motifs a donc été de première importance dans le cadre des travaux du GDR TICCO.

6.2.1.3 Les niveaux d'abstraction

Les différents points de vue, que nous venons de voir, peuvent être représentés à différents niveaux d'abstraction que nous appelons : *niveau micro*, *niveau méso*, *niveau macro*. Nous allons expliquer dans les paragraphes suivants quelle est la signification et la portée

[7]la nature des interactions entre molécules du milieu vivant est très dépendante de la géométrie de celles-ci.

de ces trois niveaux.

Le niveau micro

Le niveau micro est le niveau d'abstraction le plus bas. Il rend compte de la structure moléculaire au niveau des atomes et des liaisons. Ce niveau nous renseigne sur la nature des atomes et des liaisons qui les lient. À ce niveau, on peut savoir combien d'atomes de chaque type est présent dans la molécule, quels sont les centres potentiellement chiraux, etc.

Le niveau méso

Avec le niveau méso, nous nous plaçons à un niveau d'abstraction supérieur au niveau micro. Les atomes et les liaisons ne sont plus considérés individuellement mais par groupes correspondant à des motifs caractéristiques. Les motifs élémentaires, que nous appelons *blocs*, perçus sont les fonctions, les chaînes, les cycles, les liens, les stéréogènes, etc. Si l'on regarde la molécule d'Ircinianin selon chacun des points de vue dès le niveau méso, on peut construire des *représentations abstraites* de la molécule schématisées sur la figure 6.13. Ces représentations mettent en relief les objets perçus selon un point de vue ainsi que leurs relations structurales.

Le niveau macro

Au niveau macro, nous considérons des ensembles d'unités perçues au niveau méso et liées par des *chemins méso*. Nous appelons ces ensembles des *systèmes de blocs*.

Ainsi du point de vue de la topologie, on aura des polycycles. Selon l'arrangement relatif des cycles les uns par rapport aux autres, deux cycles peuvent être condensés (les cycles partagent une liaison et une seule), pontés (les cycles ont en commun au moins deux liaisons) ou en position spiranique (les cycles ont en commun seulement un atome) (figure 6.17). Bien entendu, un polycycle peut être à la fois condensé et/ou ponté et/ou spiranique comme dans la molécule d'Ircinianin qui comporte un polycycle à la fois condensé et spiranique.

Au niveau de la fonctionnalité, il faut considérer les enchaînements de fonctions, c'est-à-dire les fonctions et les chemins les liant les unes aux autres. Il faut cependant émettre une restriction car selon la distance séparant les fonctions, celles-ci auront plus ou moins, voire pas, d'influence les unes envers les autres. Dans le domaine de la chimie organique, on considère principalement les interactions pour des distances de 1, 2 et 3 liaisons (par exemple dans la molécule d'Ircinianin, la fonction pyrrole est à une distance de 4 liaisons de la fonction alcène la plus proche).

Au niveau de la stéréochimie, le fait que des unités stéréochimiques soient proches dans la structure moléculaire présente un grand intérêt du fait de la complexité. Dans la molécule d'Ircinianin, les atomes asymétriques et les liaisons disymétriques sont concentrées dans une partie restreinte de la structure.

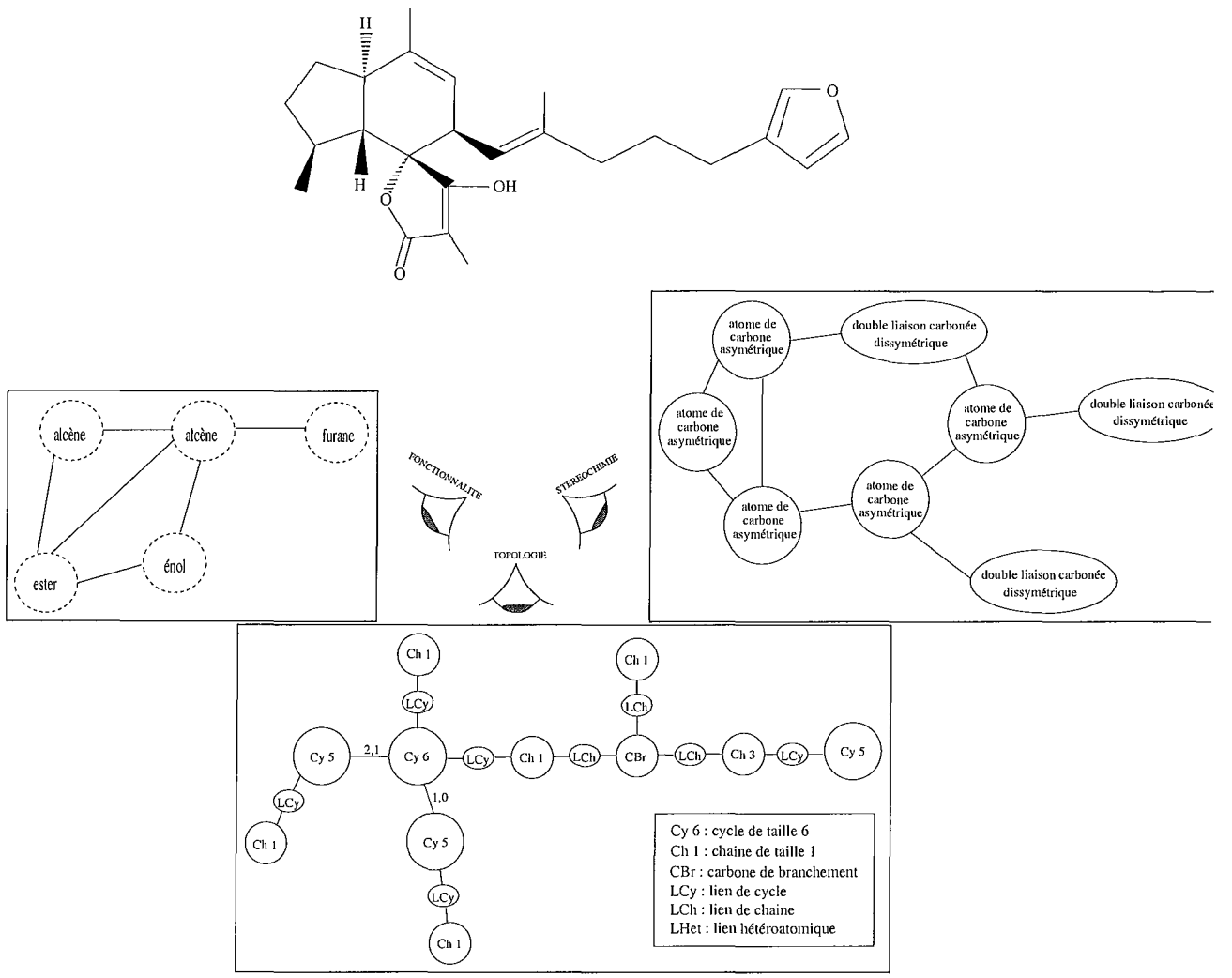


FIG. 6.13 – Représentations abstraites au niveau méso de la molécule d'Ircinianin vue selon chacun des points de vue.

6.2.1.4 Notion de squelette

Nous avons introduit précédemment le concept de squelette d'une molécule, notion très importante pour la recherche de stratégies de planification de synthèse. En effet, les chimistes s'accordent à dire que pour résoudre un problème de synthèse il faut trouver un bon « découpage » du squelette de la structure cible. Ce découpage sera bon si les fragments correspondent à des produits de départ convenables et si la fonctionnalité présente permet l'assemblage de ces fragments. Le squelette doit faire apparaître dans la structure cible la partie principale devant être construite. Néanmoins, cette notion est mal définie par les chimistes. Si ceux-ci parlent de « squelette carboné », on peut constater sur l'exemple de la molécule d'Ircinianin (figure 6.18) qu'il est très réducteur de considérer uniquement les atomes de carbone et les liaisons qui les relient. En particulier, cela fait disparaître les

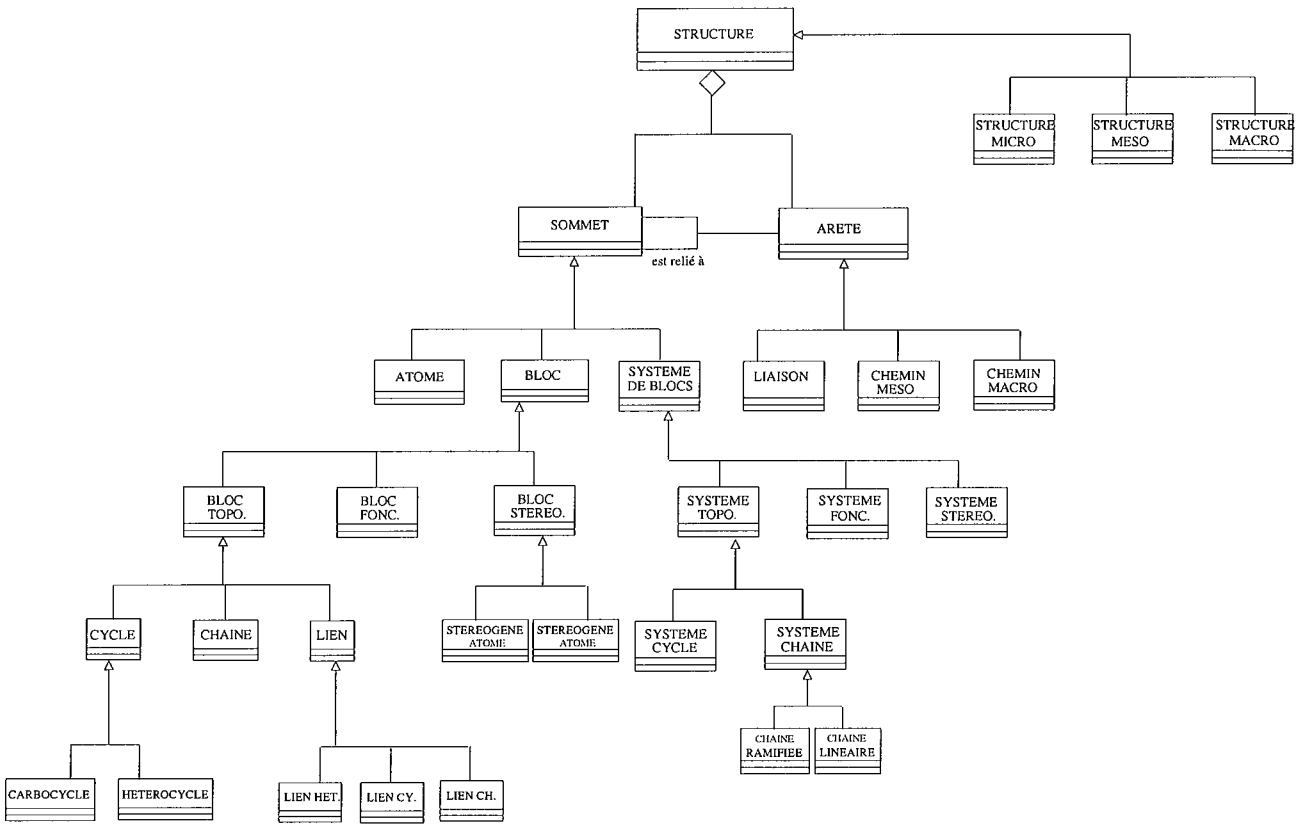


FIG. 6.14 – Le modèle conceptuel des structures.

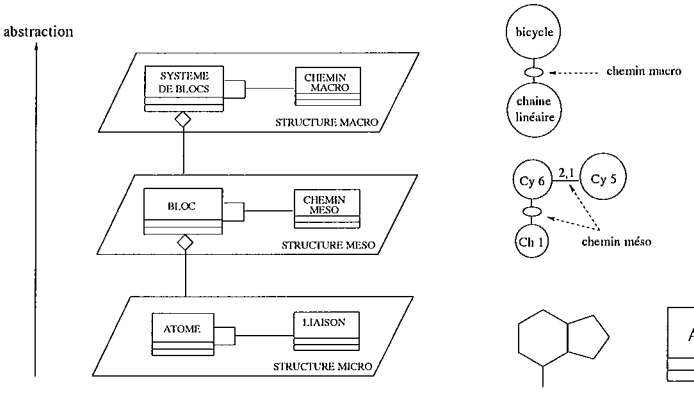


FIG. 6.15 – Les relations entre les niveaux d'abstraction.

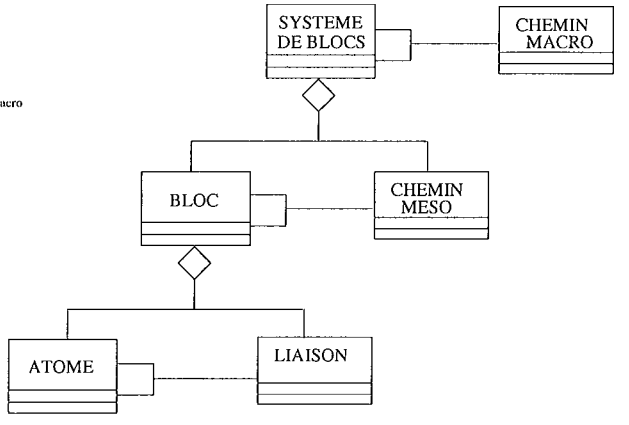
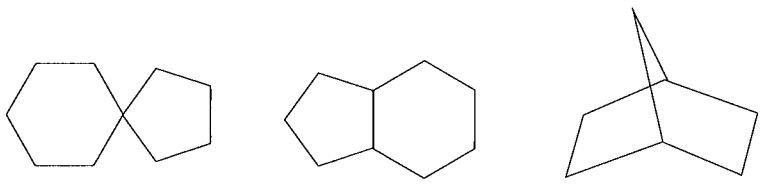
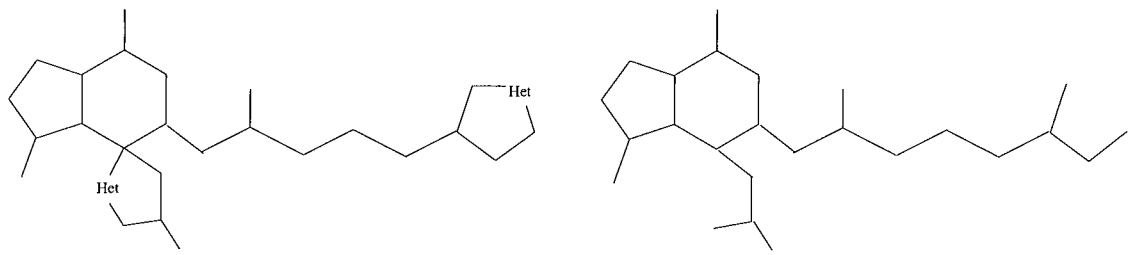


FIG. 6.16 – Relations d'agrégation entre les éléments caractéristiques des niveaux d'abstraction.



(a) (b) (c)

FIG. 6.17 – Les différents types de polycycles : (a) un polycycle spiranique, (b) un polycycle condensé, (c) un polycycle ponté.



squelette squelette carboné

FIG. 6.18 – Le squelette et le squelette carboné de la molécule d'Ircinianin.

cycles hétéroatomiques, lesquels constituent des éléments importants de la topologie.

Nous avons choisi de définir le squelette comme étant la structure moléculaire dépouillée de ses fonctions terminales[8], y compris celles qui sont masquées par un groupement protecteur. Du point de vue chimique, un groupement protecteur est un groupement d'atomes qui est insensible aux conditions réactionnelles que l'on désire employer et qui peut s'interchanger facilement avec la fonction qu'il protège, c'est-à-dire par des réactions que l'on appelle des réactions de protection ou de déprotection. Il n'existe pas de définition algorithmique d'un groupement protecteur. Le squelette peut être réduit à la topologie ou indiquer également la présence d'atomes fonctionnels et de stéréocentres à l'aide d'un marquage.

Le modèle conceptuel des structures est donné sur la figure 6.14. Les relations entre les différents niveaux d'abstraction sont illustrées par le schéma de la figure 6.15 tandis que les relations d'agrégation sont représentées sur le diagramme de classes UML de la figure 6.16. La figure 6.19 donne un exemple du passage entre les différentes représentations abstraites pour la molécule d'Ircinianin perçue selon le point de vue topologique. Dans ce cas, au niveau micro, c'est donc le squelette de la molécule qui est considéré.

6.2.1.5 Notion de famille

Les chimistes ont l'habitude de catégoriser les substances en familles chimiques. Les critères qu'ils emploient pour définir ces familles sont très variés, allant de propriétés phy-

[8]que nous appellerons également *hétéroatomes pendants*, c'est-à-dire les hétéoatomes liés par une seule liaison au reste de la structure moléculaire.

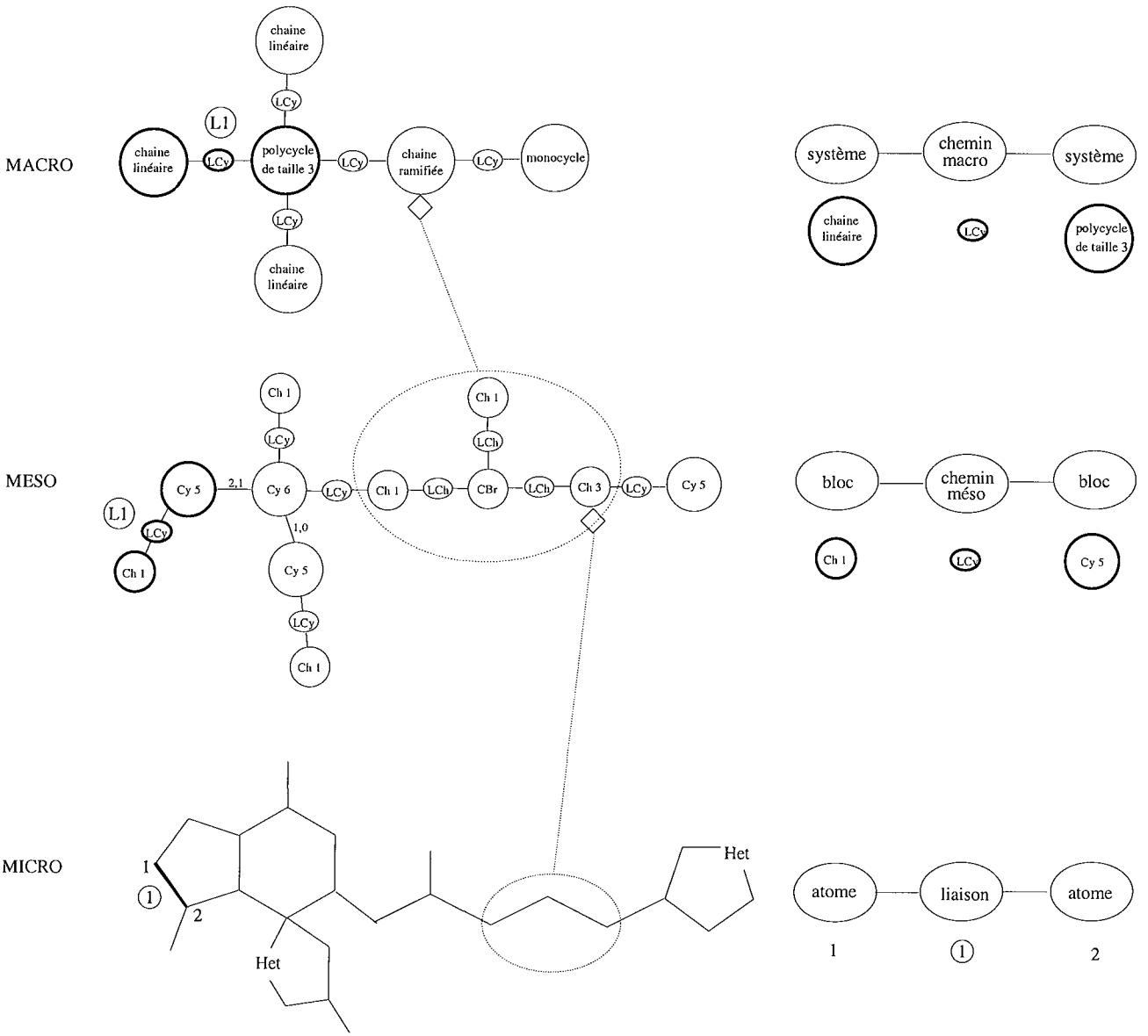


FIG. 6.19 – La molécule d'Ircinianin vue à divers niveaux d'abstraction selon le point de vue de la topologie.

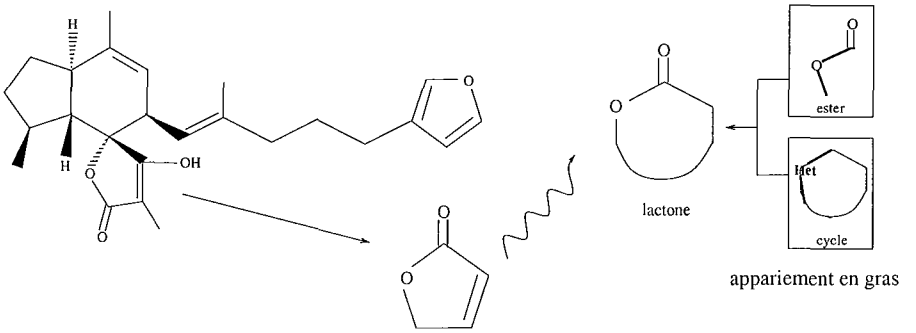


FIG. 6.20 – Une molécule de la famille des lactones : l'Ircinianin.

siques, chimiques ou biologiques aux applications ou à l'origine des substances en passant par des motifs structuraux. Les exemples sont très nombreux, parmi ceux-ci on peut citer les alcaloïdes, les stéroïdes, les solvants, les antibiotiques, les cétones, les halogénures, etc.

Les familles de molécules ayant des parties structurales communes ou semblables ont une importance en synthèse dans la mesure où, d'une part, des molécules apparentées à la cible sont des produits de départ potentiels et, d'autre part, des méthodes de synthèse plus ou moins générales permettent d'accéder à des membres de ces familles. Par exemple [Hassner and Stumer, 1994], la synthèse d'indoles peut se faire par les méthodes de Fisher, de Borsche-Drechsel ou de Madelung, la synthèse d'esters par la méthode de Baeyer-Villiger, la synthèse de macrolactones par les méthodes de Mukaiyama ou de Yamaguchi.

De telles familles chimiques ne sont pas définies formellement à partir de règles générales mais selon l'usage établi par les chimistes. Celui-ci est fondé sur l'importance de certaines parties structurales. Prenons par exemple, quelques définitions de familles particulières que donne l'IUPAC :

- Lactone : ester cyclique d'un acide hydroxycarboxylique ayant une structure de 1-oxacycloalcan-2-one, ou analogues comportant une insaturation ou des hétéroatomes remplaçant un ou plusieurs atomes de carbone du cycle.
- Stérol : produit naturel dérivé du squelette des stéroïdes et contenant un groupe hydroxyle en position 3, étroitement apparentés au cholestan-3-ol.
- Cétone : composé dans lequel un groupe carbonyle est lié à deux atomes de carbone.

On remarque que, dans le cas des lactones et des stérols, les définitions restent assez vagues et couvrent de très nombreuses possibilités.

Dans notre modèle, la notion de famille prend en compte cet usage et se définit par rapport aux trois points de vue sous lesquels une molécule est visible : topologie, fonctionnalité et stéréochimie. Ainsi, la molécule d'Ircinianin appartient, entres autres, à la famille des lactones qui se définit par un motif structural associant un cycle à une fonction ester (figure 6.20), le cycle étant hétéroatomique de tout type sauf aromatique. En d'autres termes, une molécule de la famille des lactones est caractérisée par la présence d'un ester cyclique. Nous avons marqué en gras l'appariement qui représente le recouvrement qu'il doit y avoir entre le cycle et l'ester. La structure de base partagée par les molécules d'une même famille peut être plus ou moins complexe. Ainsi la famille des alcools est une famille

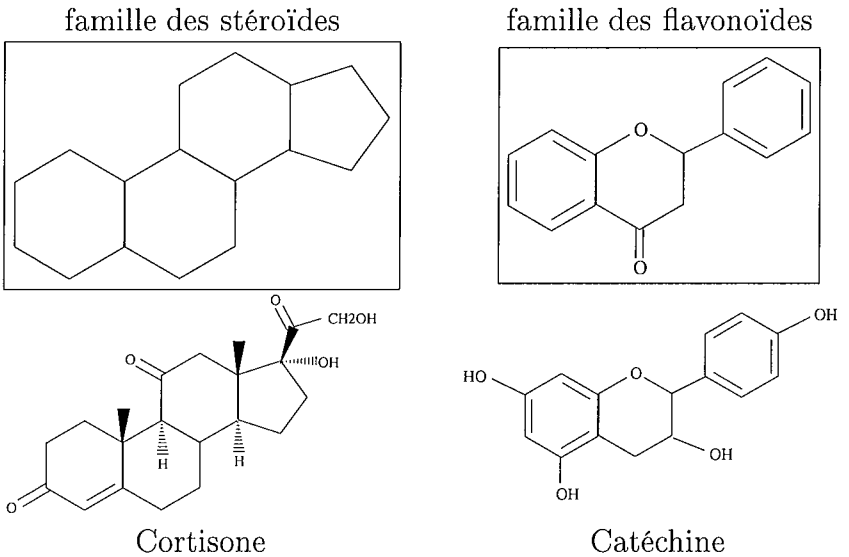


FIG. 6.21 – Deux des grandes familles de composés chimiques. En haut, la structure récurrente caractéristique de la famille. En bas, un exemple de molécule appartenant à la famille. On peut remarquer que la structure récurrente d'une famille se retrouve à la fonctionnalité près.

de composés à part entière mais on peut avoir des familles caractérisées par des motifs structuraux aussi complexes que ceux de la figure 6.21

Chaque point de vue détermine une hiérarchie de classes organisée selon une relation de spécialisation, c'est-à-dire que des sous-classes peuvent spécialiser une classe donnée. On dispose donc de trois taxonomies différentes dans lesquelles un objet molécule – une molécule particulière instance du concept molécule – peut se classer en se rattachant à une ou plusieurs classes. Cette organisation est librement inspirée des travaux réalisés sur la représentation de connaissances par objets dans le cadre du projet SHERPA [Euzenat, 1998]. La figure 6.22 montre que le concept Molécule est visible sous les trois points de vue, à chaque fois à la racine d'une taxonomie de classes. Un objet molécule tel que la δ-valérolactone est une instance de la classe des lactones, famille définie par son rattachement à la classe des molécules cycliques de la taxonomie « topologie » et à la classe des molécules ester de la taxonomie « fonctionnalité ». Il existe une relation d'appariement entre les motifs structuraux caractérisant ces deux classes (voir figure 6.20 pour l'appariement dans le cas des lactones). Le cholestérol quant à lui est une instance de la classe des stérols, famille définie par son rattachement à une classe de chacun des trois points de vue. Ici encore une relation d'appariement existe entre les motifs structuraux caractérisant ces trois classes. Le cas de la digitoxigénine est plus complexe puisque cette molécule est une instance à la fois de la classe lactone et de la classe stérol. La nécessité de créer une classe plus spécialisée peut se poser. Cette partie du modèle conceptuel n'est pas encore stabilisée, les différentes taxonomies ne sont pas encore complètement fixées. On peut remarquer qu'une famille peut résulter de classes génériques ou spécifiques, et d'une seule ou de plusieurs taxonomies

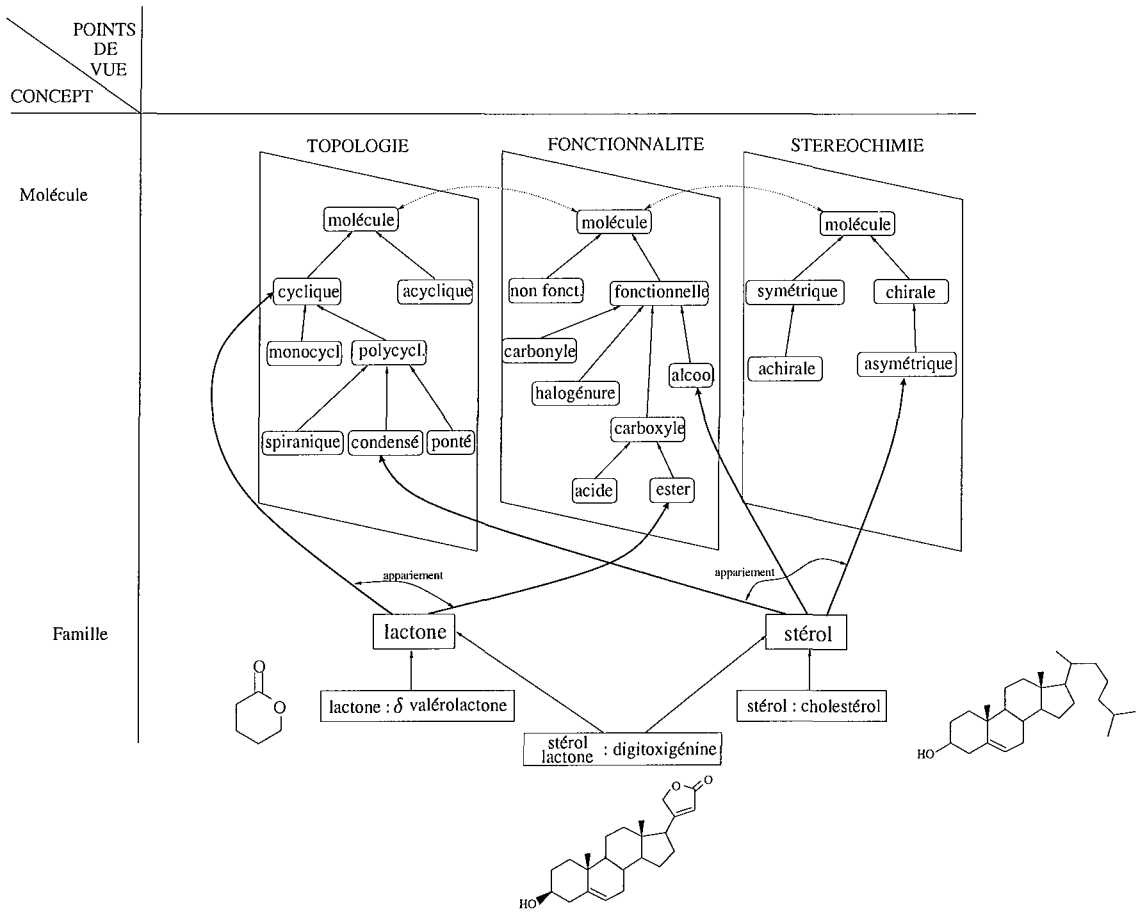


FIG. 6.22 – Les points de vue sur le concept Molécule.

La détection de l'appartenance d'une molécule à une famille chimique correspond à un processus de recomposition. Dans la reconnaissance d'un objet, à la phase d'analyse permettant de distinguer les différents éléments de l'objet, doit succéder une phase de synthèse où chacun des éléments est replacé dans l'ensemble pour obtenir une vue globale de l'objet. C'est ainsi que l'on peut procéder pour classifier de façon chimiquement pertinente un objet « molécule ». Après avoir décomposé la structure de celui-ci selon les trois points de vue et en avoir identifié les différents éléments et leurs relations, on recompose cette structure en faisant des associations significatives d'éléments ou motifs structuraux de différents points de vue.

6.2.2 Les méthodes de synthèse

Nous venons de voir que nous pouvons classer les molécules selon les familles de composés, plus ou moins spécifiques, auxquelles elles appartiennent. Cette classification a un double intérêt : celui de connaître certaines propriétés de la substance en question mais aussi celui de mettre en évidence les méthodes de synthèse susceptibles de conduire à une

telle substance.

6.2.2.1 Distinction entre méthodes de synthèse et réactions

Dans le langage habituel des chimistes, il y a une confusion permanente entre réaction et méthode de synthèse. Dans l'élaboration d'un modèle conceptuel de planification de synthèse, il y a lieu de distinguer ces deux notions. Une réaction appartient à la connaissance générale de la chimie tandis qu'une méthode de synthèse appartient à la connaissance spécifique de la synthèse.

Une réaction est une propriété chimique d'une substance au même titre que sa couleur est l'une de ses propriétés physiques. Elle correspond à la conversion de cette espèce chimique et de ses éventuels co-réactants en une ou plusieurs autres espèces chimiques. Elle s'effectue dans un milieu déterminé (stœchiométrie des réactants, nature de la phase, présence de solvants et d'additifs tels que des catalyseurs, etc.) soumis à certaines conditions de température et de pression et avec une certaine cinétique chimique. L'ensemble de ces indications devraient être présentes dans l'équation stœchiométrique représentant une réaction. Les réactions particulières, celles qui se rapportent à des substances individuelles, sont catégorisées et donc généralisées selon des caractéristiques très diverses : type des composés mis en jeu, type de conditions réactionnelles, mécanisme réactionnel, etc.).

De même qu'une substance est susceptible de trouver une application grâce à l'une de ses propriétés (par exemple, l'indigo en raison de sa couleur bleu foncé est appliquée depuis l'antiquité dans la teinture des textiles), une réaction isolée ou une suite de réactions peut trouver son application comme méthode de synthèse.

Rappelons qu'une méthode est un moyen de parvenir à un but. Le but d'une méthode de synthèse est de permettre l'obtention d'une substance. L'intérêt d'une méthode sera d'autant plus grand qu'elle donnera accès de manière la plus générale aux membres d'une (ou plusieurs) famille chimique ainsi que nous les avons définies précédemment. On ne peut cependant pas exclure les cas où une méthode est très spécifique d'un petit nombre de molécules voire même d'une seule. Une méthode de synthèse générale peut être représentée par une équation générique, pas nécessairement stoechiométrique, présentant le bilan de la méthode et l'étendue de ses possibilités. Par exemple, la méthode de Diels-Alder a pour but de construire un cyclohexényle, c'est-à-dire un cycle comportant une double liaison. Pour cela, cette méthode met en jeu une structure diénique[9] et une double liaison carbonée. La réaction sera favorisée par la présence de substituants électro-attracteurs sur la double liaison et celle de substituants électro-donneurs sur le diène. Sur la figure 6.23, les substituants électro-attracteurs sont notés par la lettre `Z` tandis que ceux aux propriétés d'électro-donneur sont symbolisés par la lettre `R`. Une liste des groupements électro-attracteurs est donnée à côté du schéma de la méthode. Dans l'exemple de réalisation de la méthode de Diels-Alder donné, on reconnaît dans les réactants le diène et la double liaison requis et on voit dans le produit le cycle de taille 6 construit.

[9] Un diène est une structure comportant deux doubles liaisons carbonées conjuguées, c'est-à-dire l'enchaînement carboné de deux doubles liaisons séparées par une simple liaison. Un diène est une fonction composée (position en α).

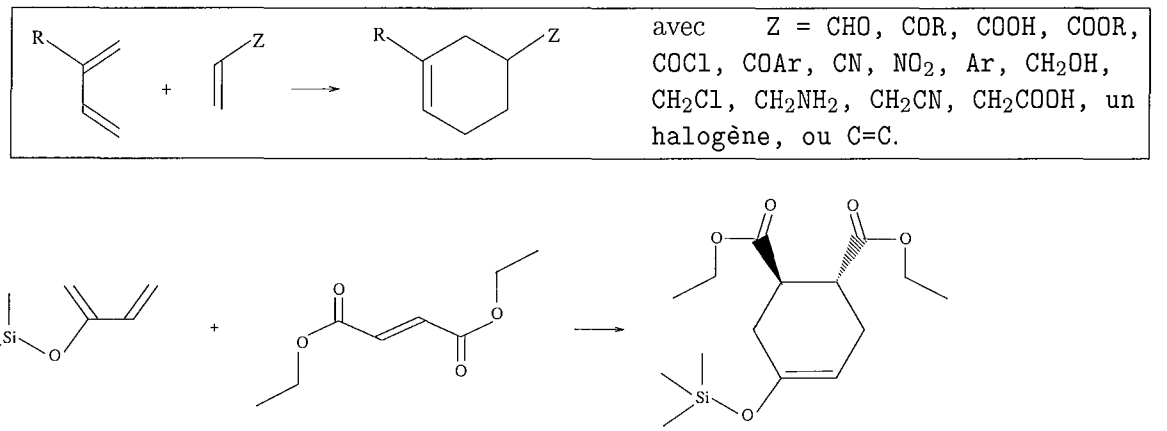


FIG. 6.23 – La méthode de Diels-Alder et un exemple de réaction associé.

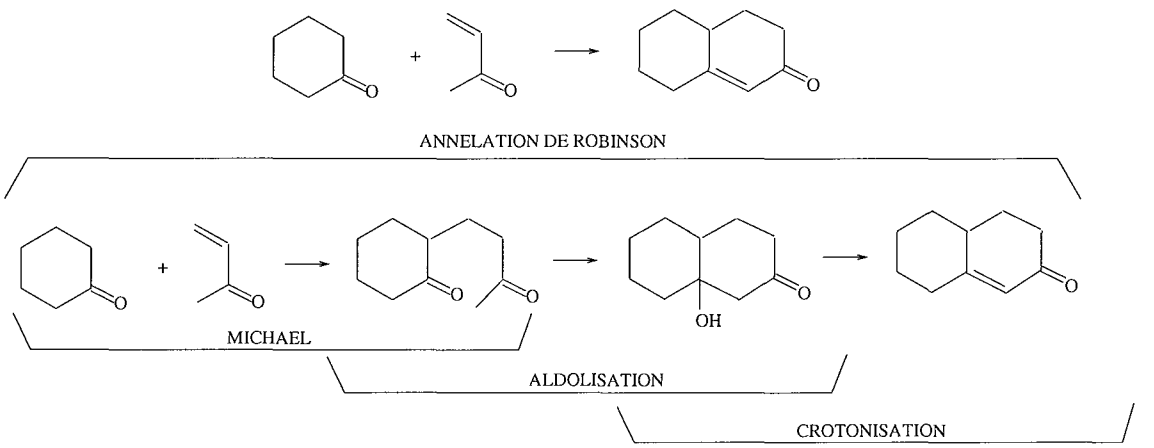


FIG. 6.24 – La méthode d'annélation de Robinson : une macro-méthode de synthèse

Si une méthode de synthèse correspond à une suite de réactions génériques, la séquence est linéaire mais on pourrait imaginer qu'elle se présente sous la forme d'un arbre de réactions. Ainsi, certaines méthodes peuvent être considérées comme des macro-méthodes de synthèse qu'il est possible de décomposer en sous-méthodes. Par exemple, dans le cas de l'annélation de Robinson (figure 6.24), on peut distinguer une première étape correspondant à la méthode de Michael, une deuxième étape illustrant la méthode d'aldolisation et enfin une troisième étape de crotonisation. Les méthodes de Michael, d'aldolisation et de crotonisation sont des méthodes à part entière tandis que l'annélation de Robinson est une macro-méthode. Cette décomposition possible nous permet de faire une analogie avec les opérateurs et macro-opérateurs [Rich and Knight, 1991]. Un macro-opérateur représente le bilan d'une séquence d'opérateurs plus ou moins élémentaires. Le coût du parcours de l'espace de recherche peut être diminué si l'on trouve des macro-opérateurs pour passer d'un état à un autre. La figure 6.25 présente la comparaison du parcours de l'espace de recherche selon le point de vue des réactions, des méthodes et des macro-méthodes et l'intérêt de la prise en compte des macro-méthodes.

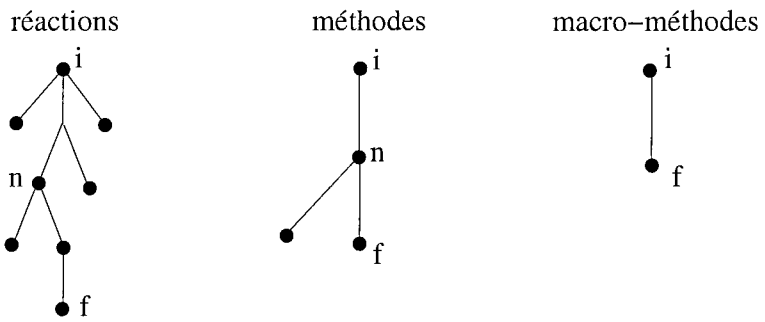


FIG. 6.25 – Macro-méthodes, méthodes et réactions : un parcours de l'espace de recherche différent.

La notion de méthode de synthèse est forcément associée à une exigence d'efficacité. Celle-ci se mesure en termes de :

- rendement : pour que l'on ait une méthode de synthèse, le rendement des réactions mise en œuvre par la méthode doit être, la plupart du temps, acceptable c'est-à-dire être supérieur à un seuil requis. Cependant une méthode qui a un faible rendement mais qui est la plus efficace connue pour atteindre un objectif sera considérée comme une méthode à part entière.
- sélectivité : une méthode de synthèse ne devrait pas mener à un mélange de produits, surtout si ceux-ci sont difficilement séparables. Une méthode de synthèse doit en revanche permettre de contrôler la stéréosélectivité, la régiosélectivité et la chimiosélectivité pour donner le résultat escompté.

6.2.2.2 La classification des méthodes de synthèse

Dans un système d'aide à la planification, ce qui nous préoccupe c'est de savoir ce que l'on doit faire et comment on va le faire. Ireland puis Hendrickson ont montré l'intérêt d'examiner les méthodes de synthèse selon leur bilan au niveau de l'élaboration de l'édifice moléculaire. À partir de ce principe, ils différencient les méthodes de construction du squelette de celles qui n'agissent que sur la fonctionnalité. D'autres experts de la synthèse tels que Corey, Trombini et Deslongchamps ont également considéré la dichotomie méthodes de construction/méthodes d'interchange fonctionnel. Dans [Trombini, 1987], Trombini rajoute les étapes de séparation des produits et celles d'aménagement fonctionnel telles que la protection/déprotection, l'activation/désactivation de fonctions, etc.). Corey propose une classification des transformations selon le changement structural observé :

- au niveau du squelette moléculaire,
- au niveau des fonctions,
- au niveau des stéréogènes.

Corey est resté au stade de la décomposition structurale. Nous essayons, pour notre part, de faire la recomposition car une « bonne » méthode de synthèse est celle qui est capable d'atteindre plusieurs objectifs en même temps. L'objectif d'une méthode de synthèse c'est de parvenir à produire des membres d'une famille. La méthode de Diels-Alder, par exemple,

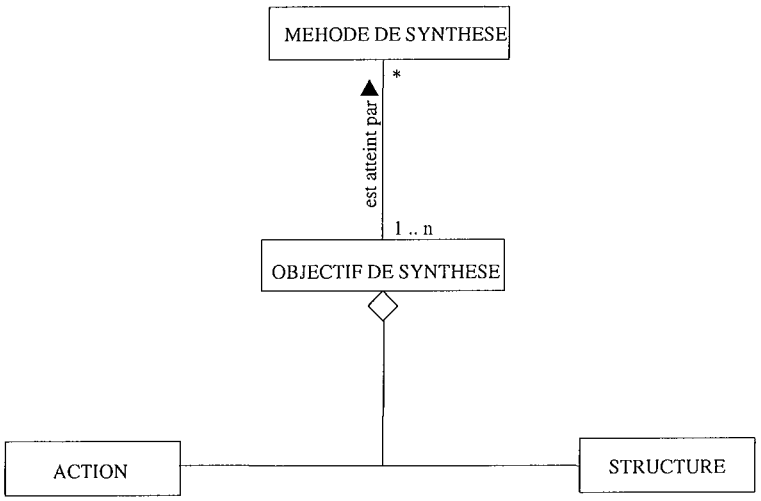


FIG. 6.26 – Modèle du concept d'objectif de synthèse.

a pour objectif de produire des molécules ayant non seulement un cycle de taille 6 mais également une fonction alcène et des stéréocentres.

6.2.2.3 Définitions

Objectif de synthèse

Les différentes étapes qui mènent progressivement à l'élaboration de la molécule cible correspondent à la réalisation d'un (ou de plusieurs) objectifs de synthèse. Dans le concept d'objectif de synthèse, sont comprises à la fois les notions de structure et d'action.

Nous définissons un objectif de synthèse comme une action sur une structure (figure 6.26) [Gien, 1998]. Les objectifs de synthèse sont déterminés par analyse de la structure moléculaire cible. Ainsi, nous souhaiterons « construire un cycle » ou bien « préserver un stéréocentre » ou encore « protéger une fonction », etc. Les actions « construction », « préservation », « aménagement fonctionnel » et « aménagement stéréochimique » sont les concepts les plus généraux d'action. Une action peut être beaucoup plus spécifique, par exemple une « extension d'un cycle de taille 6 » indiquera que l'on a construit un cycle à partir d'un cycle plus petit en taille. Les structures représentent l'ensemble des structures moléculaires possibles. Elles peuvent être, elles aussi, très générales (un cycle) ou très spécifiques (un cycle de taille 7 substitué 3 fois en 1,1,3).

Sur le diagramme de classes UML de la figure 6.26, nous avons modélisé un objectif de synthèse par l'agrégation d'une action et d'une structure.

Le schéma 6.27 donne une partie des actions et de leur organisation. Les actions de « construction » se rapportent à la construction de motifs topologiques. Nous distinguons les actions construisant des parties cycliques, et parmi celles-ci des monocyles, des actions construisant des chaînes. Au niveau des actions de modification de la fonctionnalité ou de la stéréochimie, nous différencions les actions de création, d'élimination et d'interchange entre deux motifs du point de vue considéré. Du point de vue de la stéréochimie, nous

considérons séparément les stéréogènes « atomes » et les stéréogènes « liaisons ».

Le choix des objectifs de synthèse est un choix de nature stratégique. Nous emploierons indifféremment les termes d'objectif de synthèse et d'objectif stratégique. La sélection des objectifs stratégiques se fait à partir de la perception de la molécule et grâce à des heuristiques.

Méthode de synthèse

Une méthode de synthèse sert à atteindre un ou plusieurs objectifs structuraux généralement bien définis. Sur la figure 6.26, nous avons représenté la relation entre les classes objectif et méthode de synthèse par une association binavigable « est atteint par ». Étant donné que l'intérêt d'une méthode de synthèse réside dans l'obtention d'un ou de plusieurs buts, la cardinalité de l'association indique qu'une méthode de synthèse atteint au moins un objectif de synthèse. Par contre, nous estimons qu'il est possible d'avoir un objectif de synthèse dont on ne connaît aucune méthode associée en l'état des connaissances du système.

Outre les objectifs qu'elle atteint, il est important de connaître les limites d'une méthode. Ainsi, pour une méthode de synthèse, il est possible de préciser des environnements favorables ou défavorables. Certains environnements seront favorables pour des méthodes alors qu'ils seront défavorables pour d'autres. La présence de groupements électro-attracteurs aux extrémités de la double liaison réagissant dans la méthode de Diels-Alder est un exemple d'environnement facilitant la réaction.

6.2.2.4 Un modèle orienté vers la rétrosynthèse

Si l'on parvient à classer les méthodes de synthèse selon les objectifs de synthèse qu'ils atteignent, on sera capable de proposer les solutions possibles pour un objectif donné. Les solutions pourront, de plus, être classées selon différents niveaux de généralité. Le problème réside dans la détermination des objectifs de synthèse du point de vue stratégique. Supposons que nous ayons pour objectif de synthèse de « construire un cycle carboné de taille 6 ». La méthode de Diels-Alder peut être proposée en première approximation mais il faudra vérifier que le cycle à construire et son environnement dans la structure moléculaire correspondent à un contexte favorable d'application de cette méthode.

Plus une molécule est complexe, plus la « chimie » qui sera employée pour la synthétiser sera spécifique. À partir de cette constatation, il devient évident qu'il est nécessaire de considérer les niveaux d'abstraction des plus hauts aux plus bas. En effet, supposons que l'on ait pour cible une molécule polycyclique. Si l'on connaît, au niveau tactique, des méthodes permettant de construire le type de polycycle détecté, ces méthodes seront celles qui seront sélectionnées en priorité. Par contre, si l'on ne connaît pas de méthode correspondant à cet objectif, il faut voir si l'on sait construire les sous-systèmes cycliques qui composent le polycycle. Si la base de connaissances des méthodes ne contient pas les méthodes recherchées, on se ramènera au niveau des liaisons et on recherchera des méthodes permettant de construire des liaisons. Le choix des méthodes sera alors certainement très

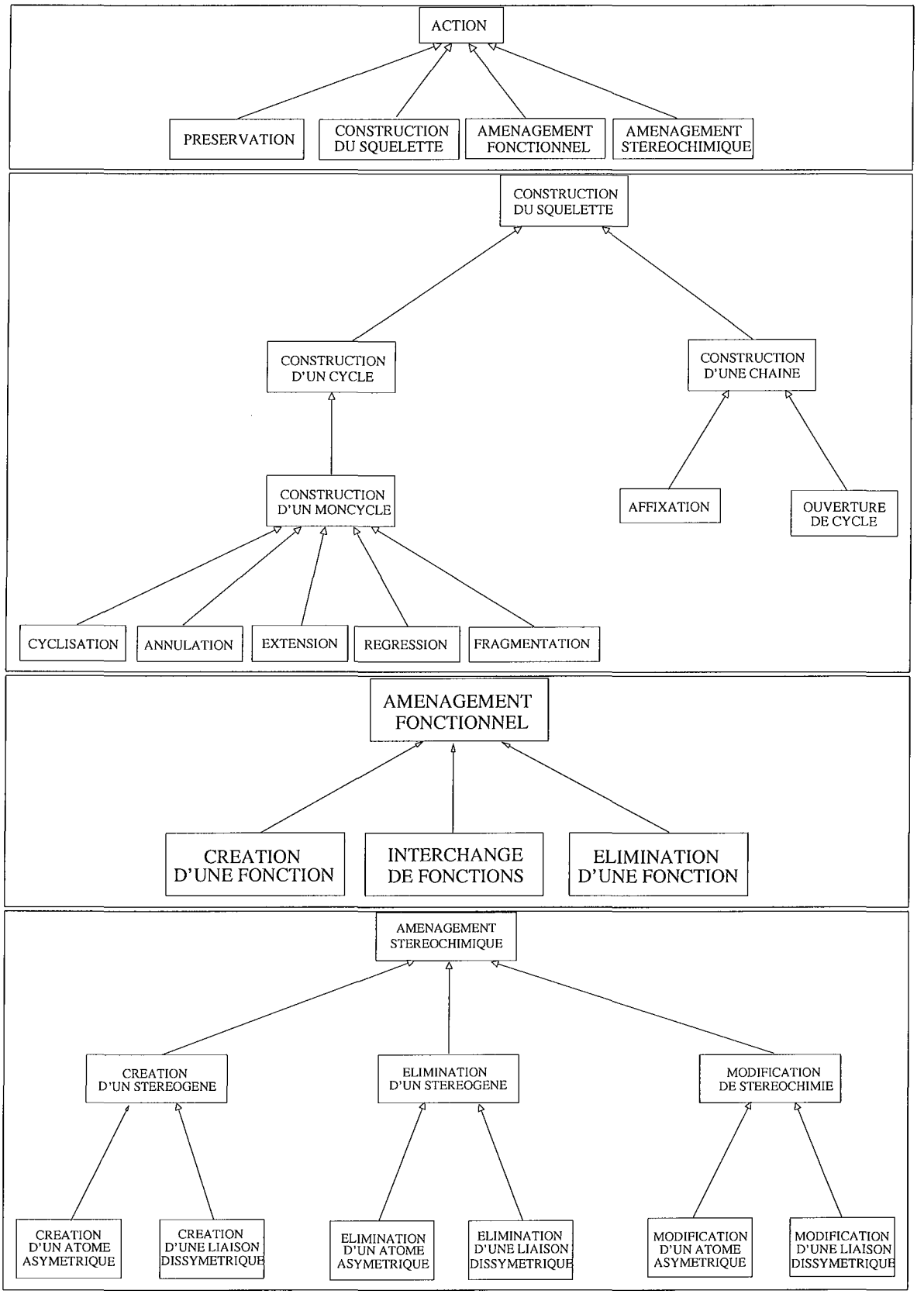


FIG. 6.27 – Modèle du concept d'action.

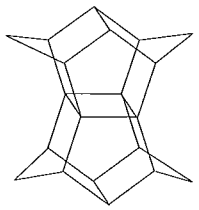


FIG. 6.28 – La molécule de Pagodane : une molécule polycyclique.

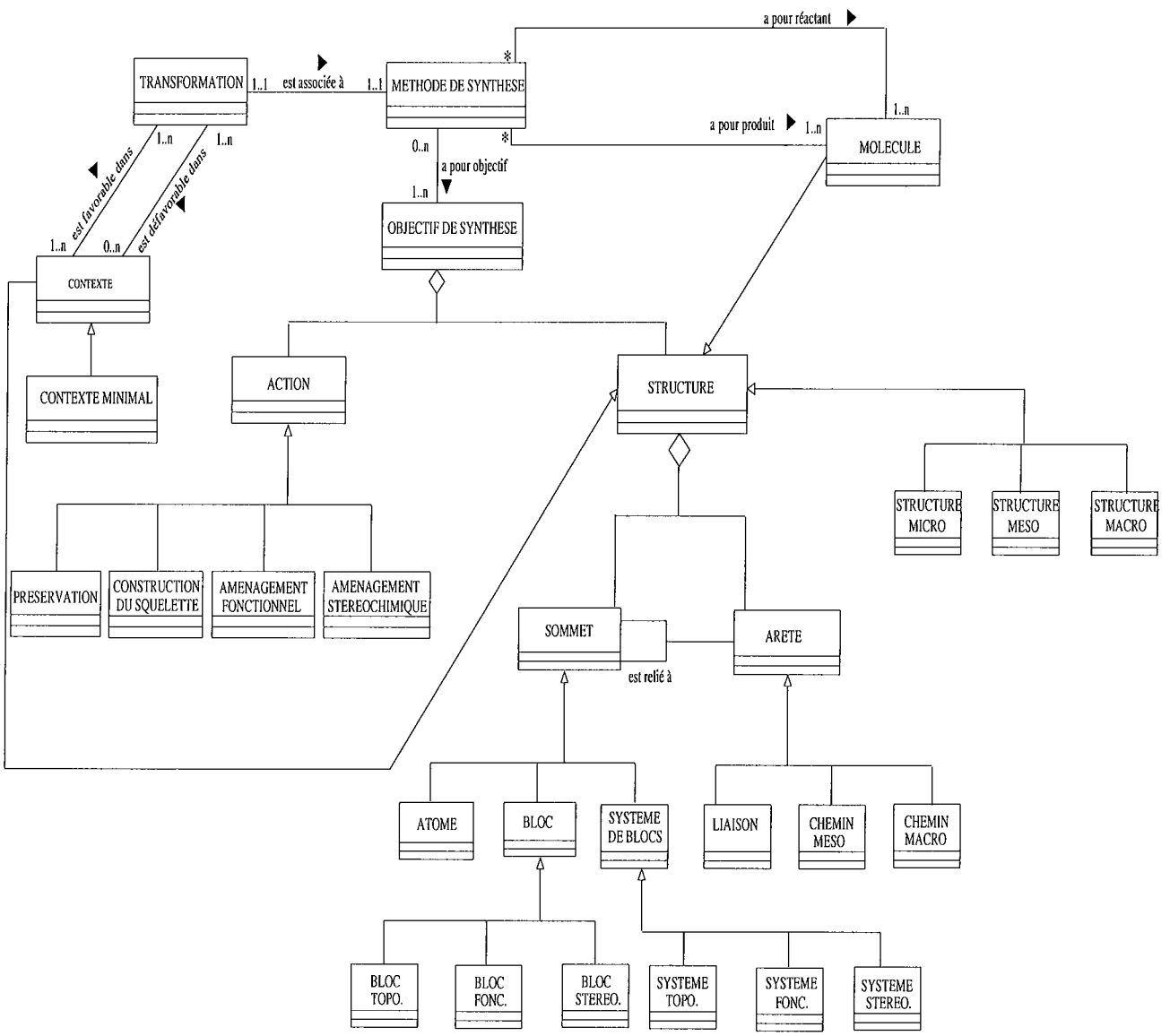


FIG. 6.29 – Une partie de notre modèle conceptuel : modèle du domaine

grand et rendra l'exploitation difficile voire impossible. En prenant l'exemple du Pagodane (figure 6.28), on se rend compte très vite que savoir synthétiser un cycle de taille 4 n'aidera pas à résoudre ce problème. Par contre, si l'on trouve une méthode construisant un des sous-systèmes cycliques du polycycle global, on aura grandement progressé[10].

Pour pouvoir raisonner dans le sens rétrosynthétique, nous devons définir la transformation correspondante à chaque méthode de synthèse. Pour rendre compte des informations que l'on est susceptible de connaître sur une méthode de synthèse, à une transformation sont associés :

- un contexte minimal d'application de la transformation c'est-à-dire la sous-structure construite par la méthode de synthèse,
- un ou des contextes favorables et/ou défavorables,
- un opérateur rendant compte des modifications à appliquer sur une structure moléculaire pour obtenir les précurseurs.

Deux transformations différentes peuvent avoir des contextes identiques (minimaux ou non). Un contexte peut être favorable pour une transformation donnée et défavorable pour une autre. C'est l'homologue des environnements dont il a été question au sujet des méthodes de synthèse (fin de section 6.2.2.3). Les contextes sont d'une importance primordiale car ils sont le support d'une information permettant un élagage de l'espace de recherche. Si l'on a déterminé un objectif stratégique dans la molécule cible correspondant au contexte défavorable d'une ou de plusieurs transformations, bien que le contexte minimal de celles-ci soit présent, ces transformations devront être écartées. La figure 6.31 donne quelques exemples de contextes favorables ou défavorables des méthodes de Wittig, de Diels-Alder et de Lindlar.

En conclusion, une vue globale du modèle du domaine que nous avons établi dans l'optique de la planification de synthèse est présenté sur la figure 6.29.

6.2.3 Les plans de synthèse

Selon le niveau où l'on se place, le plan de synthèse fera appel à l'une des différentes représentations du problème. In fine, le plan doit pouvoir être lu et suivi dans le sens synthétique. Il met alors en jeu des réactions et doit être le plus précis possible avec un maximum d'indications sur les conditions réactionnelles. Aux niveaux stratégique et tactique, le plan est à rapprocher d'un arbre de rétrosynthèse. Du point de vue stratégique, le plan expose les objectifs à atteindre et leur ordonnancement. Au niveau tactique, il indique les méthodes de synthèse envisageables pour atteindre les objectifs déterminés. Yolande Arhonovitz a travaillé sur les plans de synthèse tels qu'ils sont décrits dans la littérature. Elle propose dans [Arhonovitz, 1999] une première exploration de la modélisation de la structure des plans de synthèse sous forme d'arbre et de relations.

[10]Nous suivons de cette façon le précepte « Qui peut le plus, peut le moins ».

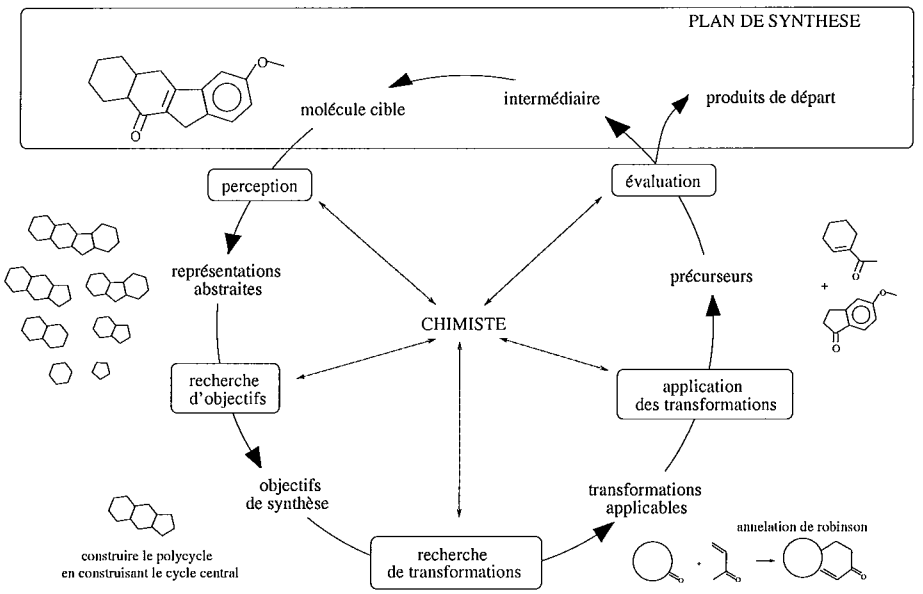


FIG. 6.30 – Le cycle de génération de plan de synthèse par rétrosynthèse.

6.3 Les raisonnements du domaine

Le cycle de génération d'un plan de synthèse selon l'approche de la rétrosynthèse est représenté sur la figure 6.30 [Laurenço, 1998]. Au centre du processus, se situe le chimiste qui intervient tout au long de l'élaboration du plan pour sélectionner, valider ou rejeter les propositions données par le système.

6.3.1 La classification hiérarchique

Les classifications qui sont faites sont la plupart du temps hiérarchiques. C'est à dire que les concepts sont organisés en hiérarchie des concepts les plus généraux à ceux plus spécifiques. Ces classifications suivent les principes énoncés par Platon et Aristote[11].

La classification est duale. D'une part, elle permet de construire des classes et, d'autre part, de classer de nouveaux individus par rapport aux classes existantes. Ce mode de raisonnement est très employé dans les travaux du GDR 1093 TICCO tant pour la mise à jour des connaissances que pour la résolution de problèmes. Nous pouvons prendre en exemple l'utilisation du réseau de contextes dans RESYN [Laurenço, 1998]. Le réseau des contextes associés aux transformations est organisé en une hiérarchie selon la relation de subsomption de graphe (figure 6.31). La mise à jour de la base des contextes se fait en ajoutant un nouveau contexte à la hiérarchie. De même, lorsque l'on a une molécule à synthétiser pour laquelle on a déterminé un objectif stratégique (par exemple, une liaison à construire), on classe sa structure dans le réseau en entrant par les contextes les plus généraux permettant d'atteindre cet objectif. Il sera possible d'appliquer toutes les transformations associées de

[11] Avec sa classification des êtres vivants, Aristote édifie une « échelle de la Nature ».

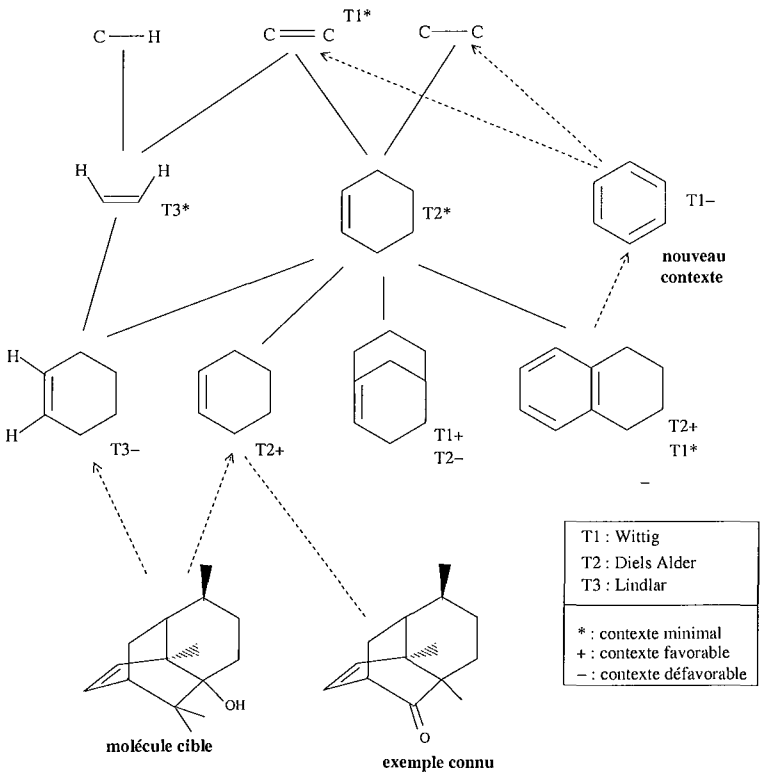


FIG. 6.31 – Mise à jour et résolution de problème grâce au raisonnement par classification dans le réseau de contextes.

manière favorable aux contextes subsumant la structure moléculaire cible. Ainsi pour la molécule cible de la figure 6.31, l'application de la méthode de Diels-Alder paraît favorable tandis que celle de la méthode de Lindlar ne l'est pas car la double liaison est cyclique. Pour ce qui est de l'application de la méthode de Wittig, seul le contexte minimal de la transformation est présent.

6.3.2 Le raisonnement par analogie

Dans le cadre du GDR 1093 TICCO, plusieurs travaux ont été menés sur l'exploitation du raisonnement par analogie dans le cadre de la planification de synthèse. Les premiers travaux ont été effectués au cours de la thèse de Michel Py [Py, 1992]. Ceux-ci ont été suivis les travaux de Jean Lieber qui a développé un système de raisonnement à partir de cas, cas particulier du raisonnement par analogie. Le raisonnement à partir de cas part de l'hypothèse que « deux problèmes similaires ont des solutions similaires ». Comme son nom l'indique, le raisonnement à partir de cas est fondé sur l'utilisation de cas, appelés *cas sources*. Un cas source est un couple formé par l'énoncé d'un problème et la solution connue de ce problème. Le problème courant à résoudre, appelé *problème cible*, est comparé à l'ensemble des problèmes sources. Dans [Lieber, 1997], la mise en évidence d'un chemin de similarité entre le problème cible et un ou des problèmes sources permet la remémoration

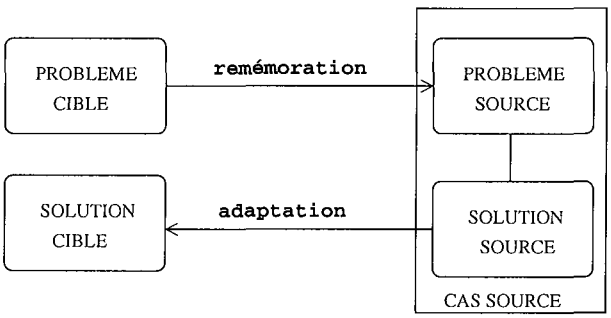


FIG. 6.32 – Un raisonnement de type analogique : raisonnement à partir de cas.

des cas sources pouvant aider à la résolution du problème cible. Ensuite une adaptation qui se fait selon les chemins de similarité est nécessaire pour construire la solution du problème cible. La difficulté réside dans la définition de la similitude entre deux problèmes, et donc de l'étape de remémoration de cas similaires. Pour diminuer la complexité de la remémoration, les cas sources doivent être classés hiérarchiquement, du cas le plus général au cas le plus spécifique.

Dans le cadre de la synthèse, un cas est un couple formé d'une structure moléculaire à synthétiser et d'un plan de synthèse solution. La thèse de Jean Lieber [Lieber, 1997] a permis d'établir les principes du raisonnement à partir de cas dans le domaine de la planification de synthèse. L'une des difficultés dans les applications de raisonnements à partir de cas sur des données réelles est la définition d'une mesure de similitude. Dans le cas de la rétrosynthèse, étant donné que les problèmes se matérialisent sous la forme de structures moléculaires, il est nécessaire de posséder une mesure de similitude entre molécules, plus précisément entre structures moléculaires. De nombreux travaux ont abouti à la détermination de métriques[12] qui ne sont pas utilisables. En effet, ces techniques ne permettent en aucun cas de fournir une explication et donc d'obtenir des indications sur le chemin de similitude qu'il existe entre deux molécules. Dans le cadre du GDR TICCO, l'approche est toute autre et a conduit à la thèse de Yannic Tognetti qui a, en partie, porté sur la *classification approchée* des structures moléculaires. Supposons que l'on ait à synthétiser un carbocycle de taille 6 complètement saturé. A priori, la classification exacte élimine la possibilité d'application de la méthode de Diels-Alder alors qu'une classification floue devrait pouvoir sélectionner cette méthode en indiquant quelle opération supplémentaire est nécessaire pour que la méthode soit effectivement applicable. Cela sous-entend que l'on possède une base de transformations suffisamment riche [Tognetti, 2002].

6.3.3 La négociation

La décomposition entre différents points de vue et niveaux qui apparaît dans ce modèle est une vision qui ne correspond pas à la façon exacte dont les chimistes raisonnent. Ceux-ci adoptent généralement une démarche prenant compte en même temps tous les points de vue

[12] nous avons déjà abordé ce sujet au chapitre 3 section 3.2.2.2.

et les niveaux d'abstraction, trop complexe à modéliser. Les travaux de Pierre Jambaud ont permis d'approcher quelque peu cette façon de raisonner par la mise au point d'un système distribué au niveau duquel des agents correspondant à chacun des trois points de vue résolvent le problème par un dialogue. Les conflits sur le choix des objectifs de synthèse sont réglés par un processus de négociation entre les divers agents.

6.4 Conclusion

Dans ce chapitre, nous avons exposé le modèle d'ordre chimique sur lequel s'appuie ces travaux de thèse. Ce modèle ne s'est pas construit en un jour mais est le fruit de multiples raffinements effectués en grande partie dans le cadre du GDR TICCO. Ces améliorations successives ce sont en partie faites grâce aux expérimentations possibles à partir de prototypes implémentant le modèle informatique que nous allons voir dans le chapitre suivant. L'application à des données réelles est particulièrement révélatrice des faiblesses d'un modèle.

Chapitre 7

La modélisation et le traitement informatique des objets de la synthèse organique

All models are imperfect approximations.

Mark. A. Musen, 1993

Sommaire

Nous allons aborder dans ce qui suit les problèmes de représentation des structures chimiques d'un point de vue informatique. Dans ce cadre, le modèle qui a retenu notre attention est celui offert par la théorie des graphes. De façon générale, les graphes constituent un remarquable outil de modélisation de la connaissance. L'application aux objets chimiques que sont les molécules et les réactions a considérablement aidé au traitement informatique de l'information chimique. Un certain nombre d'algorithmes de manipulation des graphes moléculaires a été conçu et implémenté dans REsyn Assistant.

7.1 Représentation des molécules

À partir du moment où les chimistes ont modélisé les molécules comme des agrégats d'atomes plus ou moins proches les uns des autres, ils ont voulu représenter cette « réalité » pour mieux appréhender ces objets microscopiques. Le mathématicien Sylvester est le premier à proposer le terme de graphe pour décrire la représentation des structures moléculaires aussi appelées formules structurales ou formules constitutionnelles [Rouvray, 1995]. Les travaux de Cayley portant sur l'énumération des isomères d'une série d'homologues furent les précurseurs de nombreux développements de la théorie des graphes et de recherches en combinatoire. Nous ne présenterons pas ici les bases de la théorie des graphes. Les rudiments de cette théorie ont très bien été traités dans la thèse de Philippe Vismara [Vismara, 1995]. Le livre de Claude Berge [Berge, 1973] est une référence dans le domaine.

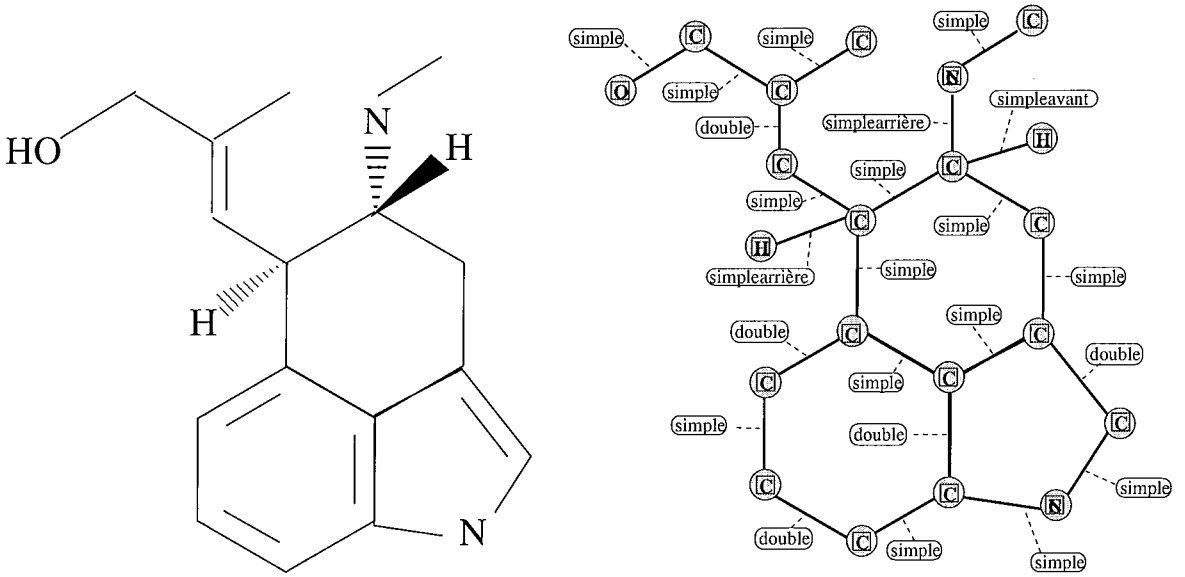
7.1.1 Le niveau micro : les graphes moléculaires

Les graphes moléculaires permettent de rendre compte de la constitution des molécules. Ce sont des graphes particuliers, non orientés et étiquetés sur les sommets et les arêtes. Chacun des sommets du graphe est étiqueté par la nature de l'atome auquel il est associé (par exemple `C` pour un atome de carbone, `O` pour un atome d'oxygène, `N` pour un atome d'azote, etc.). L'étiquette d'une arête sert à indiquer le type de la liaison à laquelle elle correspond (par exemple, `simple` pour une liaison simple, `double` pour une liaison double etc.). De façon plus formelle, si $\mathtt{G}_M$ est le graphe moléculaire de la molécule *mol*, nous pouvons écrire que :

$\mathtt{G}_M$ = (X,E) où X est l'ensemble des sommets de $\mathtt{G}_M$ et E est l'ensemble des arêtes de $\mathtt{G}_M$. Les éléments de E sont des couples de sommets $(\mathrm{x}_i,\mathrm{x}_j) \in \mathtt{X}^2$.

Les fonctions d'étiquetage sont définies de la façon suivante :
Soit $\mathcal{T}_A$ l'ensemble des symboles atomiques : $\mathcal{T}_A$ = {C, H, N, O, Cl, etc.}. La fonction d'étiquetage des sommets est :

$$\begin{aligned} f_A : \ \mathtt{X} &\rightarrow \ \mathcal{T}_A \\ \mathrm{x} &\mapsto \ f_A(x) = \text{symbole de l'atome associé à x} \end{aligned}$$



structure moléculaire graphe moléculaire

FIG. 7.1 – La molécule de Chanoclavine I : une représentation de la structure moléculaire et le graphe moléculaire associé.

$\mathcal{T}_{\mathcal{L}}$ ensemble des types de liaisons : $\mathcal{T}_{\mathcal{L}} = \{\text{simple, double, triple, aromatique, etc.}\}$

$$\begin{aligned} f_{\mathcal{L}} : \quad E &\rightarrow \quad \mathcal{T}_{\mathcal{L}} \\ e &\mapsto \quad f_{\mathcal{L}}(e) = \text{type de la liaison associée à e} \end{aligned}$$

Le graphe moléculaire correspondant à la molécule *mol* est complètement défini à partir de $\mathtt{G}_M$ et des fonctions d'étiquetage $\mathcal{T}_{\mathcal{A}}$ et $\mathcal{T}_{\mathcal{L}}$.

Sur la figure 7.1, nous avons représenté la structure moléculaire de la molécule de Chanoclavine et le graphe moléculaire associé.

Le traitement des atomes d'hydrogène

Les molécules organiques sont principalement constituées d'atomes de carbone et d'atomes d'hydrogène. Pour des raisons de lisibilité, les atomes d'hydrogène sont généralement omis de la représentation structurale. Ceci est possible car, connaissant le reste de la structure, on est capable de calculer le nombre d'atomes d'hydrogène qui devraient être ajoutés pour que la structure soit complète du point de vue chimique. Les seuls atomes d'hydrogène qui sont indiqués explicitement sont ceux qui permettent de préciser la stéréochimie d'un centre asymétrique. Dans la molécule de Chanoclavine, deux atomes d'hydrogène sont indiqués car ils précisent chacun la configuration d'un atome de carbone. Sur la structure moléculaire peuvent aussi figurer des atomes d'hydrogène dont l'indication n'est pas fondamentalement importante mais correspond plus à un usage des chimistes (par exemple, le `HO` en haut à gauche dans le dessin de la structure moléculaire de la Chanoclavine). Dans un graphe moléculaire, nous avons donc éventuellement des sommets de type hydrogène qui sont

la plupart du temps liés par une arête étiquetée par un type de liaison simple hors du plan (`simple avant` ou `simple arrière`). Pour calculer le nombre de sommets de type hydrogène adjacents à un sommet quelconque, on dispose d'un attribut sur les atomes qui précise, selon leur type et leur charge, le nombre de liaisons incidentes possibles. Cet attribut s'appelle la valence. Par exemple, un atome de type azote à l'état neutre a une valence de 3. Le nombre de sommets de type hydrogène liés à un sommet est donné par la différence entre la valence et le nombre effectif d'arêtes incidentes au sommet, sachant qu'une liaison simple compte pour 1, une liaison double pour 2, etc. Dans la molécule de Chanoclavine (voir figure 7.1), l'atome de type azote faisant partie du cycle de taille 5 est lié par une liaison de type simple à deux atomes de type carbone. Il est donc également lié à un atome de type hydrogène qui n'est pas représenté sur la figure. Si le sommet n'a que des arêtes incidentes de type simple (hors du plan oun non), ce nombre correspond au degré du sommet dans le graphe, c'est-à-dire au nombre d'arêtes incidentes au sommet (si les arêtes ne sont pas de type simple, il est nécessaire de tenir compte du type de l'arête).

Quelques précisions sur le vocabulaire associé aux graphes moléculaires

Nous emploierons par la suite certaines expressions pour désigner des éléments du graphe moléculaire.

- *sommet de type carbone* pour indiquer un sommet dont l'étiquette est `C`,
- *arête carbonée* pour signaler une arête dont les deux extrémités sont des sommets de type carbone,
- *sommet de type hétéroatomique* pour indiquer un sommet étiqueté par un type atomique autre que `C` ou `H`,
- *arête hétéroatomique* pour désigner une arête dont au moins une des extrémités est un sommet de type hétéroatomique.
- *arête de type multiple* ce qui correspond aux arêtes dont le type n'est pas le type `simple`, `simple avant` ou `simple arrière`

Un exemple de manipulation des graphes : la recherche d'isomorphismes de graphes moléculaires

À partir de la représentation sous forme de graphes des structures moléculaires, de nombreux traitements informatiques sont possibles. Parmi ceux-ci la recherche d'isomorphismes de graphes répond à des besoins apparaissant dans la gestion de bases de graphes moléculaires. Ainsi, on peut vouloir savoir si un graphe donné est contenu dans une base, le problème à résoudre est un problème d'isomorphisme de graphes. Si l'on veut savoir si un graphe moléculaire contient un sous-graphe donné : le problème est celui de l'isomorphisme de sous-graphes (partiels). Enfin, au niveau de la recherche de similitudes chimiques, le problème est celui de la recherche de plus grands sous-graphes partiels communs. Du point de vue de la théorie des graphes, ces différents problèmes s'énoncent de la façon suivante, en posant deux graphes $\mathtt{G} = (\mathtt{X},\mathtt{E})$ et $\mathtt{H} = (\mathtt{Y},\mathtt{F})$:

- `G` et `H` sont isomorphes si et seulement si $\exists$ une bijection $\mathtt{f} : \mathtt{X} \rightarrow \mathtt{Y}$ / $\forall\ \mathtt{x}_i$ et $\mathtt{x}_j \in \mathtt{X}$, $(\mathtt{x}_i,\mathtt{x}_j) \in \mathtt{E} \Longleftrightarrow (\mathtt{f}(\mathtt{x}_i), \mathtt{f}(\mathtt{x}_j)) \in \mathtt{F}$,

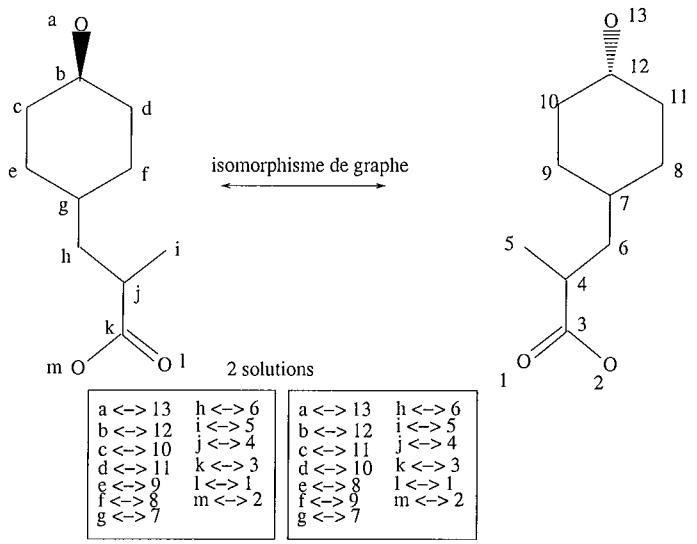


FIG. 7.2 – Un exemple de recherche d'isomorphisme de graphes moléculaires.

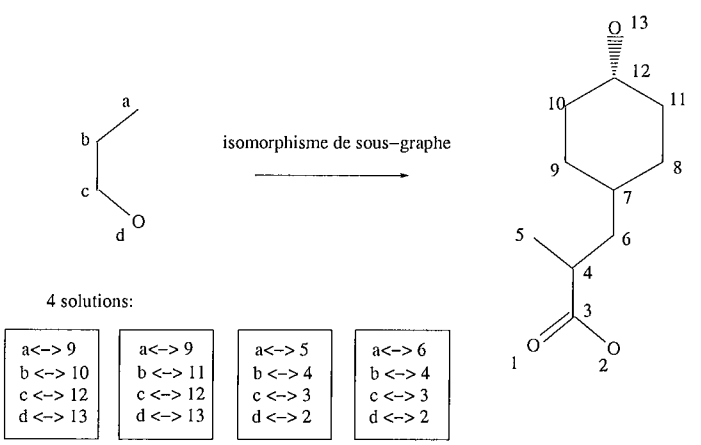


FIG. 7.3 – Un exemple de recherche d'isomorphisme de sous-graphes moléculaires.

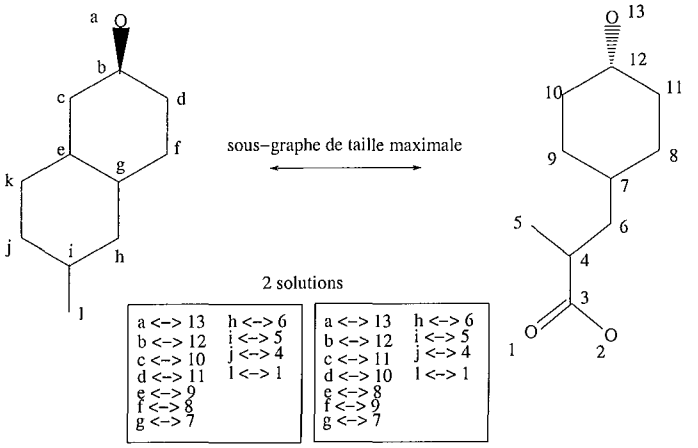


FIG. 7.4 – Un exemple de recherche de recherche de plus grands sous-graphes commun.

- H est sous-graphe de G si et seulement si $Y \subseteq X$ et $F \subseteq E$ et $\exists$ une bijection $f : X \rightarrow Y$ / $\forall\ x_i$ et $x_j \in X$, $(x_i,x_j) \in E \Longrightarrow (f(x_i), f(x_j)) \in F$,
- le plus grand sous-graphe commun est calculé par rapport à un entier k. La question est de savoir s'il existe $R \subseteq X \times Y$ avec $|R| \geq k$ / $\forall \{x_i,x_j\}$ et $\{y_k,y_l\} \in R$, $\{x_i,x_j\} \in E \Longleftrightarrow \{y_k,y_l\} \in F$.

Les figures 7.2, 7.3 et7.4 donnent chacune un exemple de recherche d'isomorphismes de graphes moléculaires correspondant respectivement à l'un des trois problèmes évoqués ci-dessus. Étant donné qu'à ce niveau nous ne prenons pas en compte les considérations spatiales, nous avons une classe d'équivalence entre les étiquettes `simple`, `simple avant` et `simple arrière`. Celle-ci nous autorise à apparier ensemble deux arêtes de types simple, simple avant ou simple arrière.

La recherche sous-structurale entre graphes moléculaires implique la recherche d'isomorphisme de sous-graphes partiels avec prise en compte de propriétés chimiques. Les connaissances chimiques à respecter sont exprimées par l'intermédiaire des étiquettes des sommets et des arêtes. On dispose de hiérarchies de types sur ces ensembles d'étiquettes. La relation établie entre deux graphes dont l'un est inclus dans l'autre est appelée relation de co-subsomption dans [Napoli, 1992] car elle doit à la fois vérifier l'inclusion de graphe ainsi que la compatibilité des étiquettes (relation de généralisation/spécialisation sur les types d'atomes et de liaisons). La recherche d'isomorphisme d'un sous-graphe (partiel) g dans un graphe G est très coûteuse en temps. Durant sa thèse effectuée au sein du GDR 1093 TICCO, Jean-Charles Régin a réussi à diminuer le coût de cette opération par utilisation de techniques de résolution de problèmes par satisfaction de contraintes ou CSP[1]. La construction d'un *réseau de contraintes* permet d'élaguer l'espace de recherche en affectant des domaines (des sous-ensembles de sommets de G) à des variables (les sommets de g) et en instanciant les variables en respectant un certain nombre de contraintes. Dans un réseau de contraintes associé à des graphes moléculaires, il est possible de fixer des contraintes de

[1]CSP : Constraint Satisfaction Problem

différents ordres :

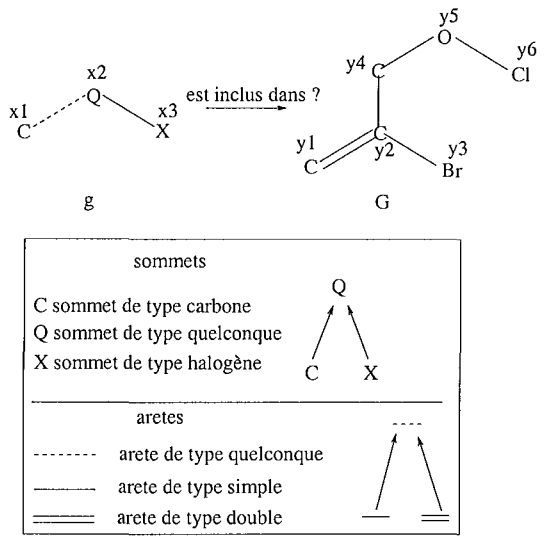
- les contraintes unaires qui imposent des conditions sur les étiquettes et les degrés des sommets[2] que l'on apparie. Ainsi, l'on peut apparier un sommet x_g de g avec un sommet x_G de G si l'étiquette de x_g est la même ou est plus générale que l'étiquette de x_G. On n'apparie pas un sommet x_g de g avec un sommet x_G de G si le degré de x_g est supérieur au degré de x_G,
- les contraintes binaires qui imposent d'apparier deux sommets adjacents de g avec deux sommets de G eux aussi adjacents,
- une contraintes n-aire de différence qui impose à tout moment que tout couple de sommets de g soit apparié avec des sommets de G différents.

La figure 7.5 illustre sur un exemple le principe des CSP. La résolution d'un CSP se fait en utilisant des algorithmes de type *backtrack*. Le principe est d'instancier les variables dans un ordre prédéfini en vérifiant les contraintes au fur et à mesure. Lorsqu'une incompatibilité est détectée, on effectue un retour arrière pour essayer une autre possibilité et ainsi de suite jusqu'à la fin de l'exploration de l'espace de recherche. Diverses solutions ont été proposées pour réduire l'espace de recherche. Parmi celles-ci le maintien de la consistance d'arc (*arc consistance* en anglais) s'est révélée très efficace [Régin, 1995]. Dans ReSyn Assistant, de nombreuses fonctions impliquant la recherche sous-structurale ont été implémentées à partir d'algorithmes de type CSP développés au cours des travaux de thèse de Jean-Charles Régin [Régin, 1995] effectués en collaboration avec Christian Bessière. Cette implantation a été réalisée par Pierre Jambaud et a été ensuite étendue. Yannic Tognetti et Philippe Vismara [Tognetti and Vismara, 2000] ont mis au point un algorithme original à base de CSP permettant de découvrir les symétries explicites ou implicites (que nous appelons respectivement *symétries exactes* et *symétries cachées*) dans les graphes moléculaires. En effet, le problème de la décomposition symétrique d'un graphe consiste à isoler dans un graphe deux sous-graphes partiels disjoints, connexes et isomorphes. Yannic Tognetti propose dans sa thèse [Tognetti, 2002] une étude des problèmes posés par la recherche sous-structurale dans les graphes moléculaires ainsi qu'un aperçu des fonctions basées sur ce principe et implémentées dans ReSyn Assistant.

7.1.2 Le niveau méso : les blocs

Nous pouvons représenter une molécule à différents niveaux. Pour cela, il est nécessaire de s'abstraire de la représentation atomique du graphe moléculaire pour mettre en évidence des motifs pertinents, c'est-à-dire des sous-graphes connexes particuliers du graphe moléculaire. Comme nous l'avons signalé dans le chapitre précédent, nous employons le terme de *bloc* pour désigner chacun de ces sous-graphes. Selon le point de vue à partir duquel un bloc aura été perçu, le bloc sera qualifié de topologique, fonctionnel ou stéréochimique. Dans les représentations abstraites, les blocs deviennent eux-mêmes des sommets qui font référence aux sous-graphes correspondants, l'ensemble des représentations a une structure

[2]degré d'un sommet : nombre de voisins du sommet considéré (moyennant la prise en compte des types des arêtes).



CONTRAINTES UNAIRES	
variables	domaines
x1	y1, y2, y4
x2	y2, y4, y5
x3	y3, y6

CONTRAINTES BINAIRES	
contraintes	couples autorisés
(x1,x2)	(y1,y2), (y2,y4),(y4,y5)
(x2,x3)	(y2,y3), (y5,y6)

graphe biparti associé
à la contrainte n-aire de différence

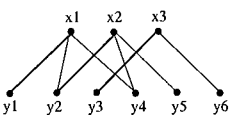


FIG. 7.5 – Principe des CSP : contraintes unaires, contraintes binaires et contraintes n-aires de différence.

de graphes imbriqués [Poulovassilis and Levene, 1994]. Nous appelons *graphe des blocs* un graphe dont les sommets sont des blocs.

7.1.2.1 Les blocs topologiques

Que l'on se place du point de vue de la chimie (structure moléculaire) ou de l'informatique (graphe), les deux principaux types d'éléments de nature topologique qui sont distingués sont les motifs cycliques et les motifs non cycliques, c'est-à-dire respectivement les cycles et les chaînes[3] dans le cadre des graphes non orientés.

Ceci est une première approximation qui ne satisfait pas forcément tous les besoins du domaine. Elle engendre une dichotomie qui ne permet pas de mettre en évidence des éléments très importants en synthèse. C'est pourquoi un découpage plus fin a été fait dans le cadre du GDR TICCO. Cette partition, qui ne correspond pas à une partition du graphe, ni en termes de sommets ou en termes d'arêtes au sens strict[4], a été décidée pour rendre compte de la démarche stratégique des chimistes.

Les différents types de blocs topologiques

- chaîne

[3]Il s'agit en fait ici des chaînes « non cycliques » car en toute rigueur un cycle est une chaîne pour laquelle les sommets placés aux extrémités coïncident. Dans notre cas, une chaîne est donc un arbre.

[4]Les éléments topologiques tels que les liens hétéroatomiques et les liens de cycles peuvent se recouvrir.

- cycle
- lien de cycle
- lien hétéroatomique
- carbone de branchement

Pourquoi ce découpage ?
- pour mettre en évidence les chaînes carbonées linéaires et ramifiées (chaîne et carbone de branchement)
- pour mettre en évidence les liens entre parties cycliques et non cycliques mais aussi entre plusieurs parties cycliques (lien de cycle)
- pour mettre en évidence les liens entre parties carbonées (lien hétéroatomique)

Notion de squelette

Le calcul du squelette a pour but de mettre en évidence les liaisons qui contribuent de façon essentielle à l'architecture de la molécule. Nous avons vu dans le chapitre précédent l'importance de la notion de squelette ainsi que les différents types de squelette qui peuvent être pris en compte (squelette carboné, squelette connexe). Dans les cas où le squelette carboné est non connexe, le squelette représente incomplètement l'édifice moléculaire. Dans [Vismara, 1995], Vismara propose la définition d'un squelette connexe rendant compte de façon aussi complète que possible de l'architecture de la structure moléculaire correspondante. En terme de graphe, il procède tout d'abord à une partition des sommets du graphe moléculaire G=(X,E) en fonction de leur étiquette c'est-à-dire du type atomique auquel ils sont associés. Il définit ainsi trois ensembles de sommets :
- l'ensemble des sommets de type carbone : $X_{carb.} = \{x \in X \ / \ f_{\mathcal{A}}(x) = C\}$
- l'ensemble des sommets de type hydrogène : $X_{hydr.} = \{x \in X \ / \ f_{\mathcal{A}}(x) = H\}$
- l'ensemble des sommets de type hétéroatomique : $X_{het.} = \{x \in X \ / \ f_{\mathcal{A}}(x) = T_{\mathcal{A}} - \{C,H\}\}$

Le graphe du squelette carboné est engendré par les sommets de $X_{carb.}$. Dans le squelette, il n'y aura pas de sommets de type hydrogène car ceux-ci ne contribuent pas à la connexité du graphe étant donné leur degré constant de 1. Les atomes d'hydrogène sont dits *terminaux* en chimie, les sommets associés sont qualifiés de *pendants* en théorie des graphes. Pour obtenir un graphe connexe préservant toutes les propriétés intéressantes du graphe moléculaire de départ, aux sommets du squelette carboné sont ajoutés les hétéroatomes cycliques ainsi que ceux qui lient deux parties carbonées du graphe moléculaire. En résumé, l'ensemble des sommets X_S permettant d'engendrer le sous-graphe de G_M correspondant au squelette G_S est :

$$X_S = X_{carb.} \cup X_{cycl.} \cup X_{het.conn.}$$

avec $X_{cycl.}$: ensemble des sommets appartenant au moins à un cycle de G_M et $X_{het.conn.}$: ensemble des sommets hétéroatomiques x tels qu'il existe une chaîne élémentaire contenant x dont les extrémités appartiennent à $X_{carb.}$.

La détermination du squelette est donc précédée par celle des cycles.

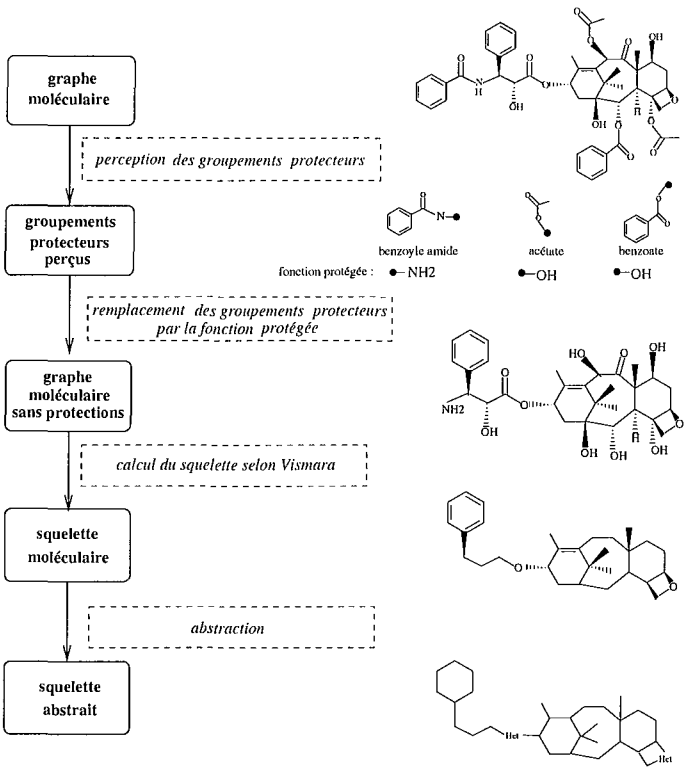


FIG. 7.6 – La détermination du graphe du squelette d'une molécule : exemple du Taxol

Outre les fonctions *pendantes* qui sont éliminées dans la définition du squelette présentée par Vismara, les groupements protecteurs pourraient être enlevés du squelette. Pour cela, possédant une base de groupements protecteurs construite par des experts, il faut d'abord percevoir les groupements protecteurs présents dans le graphe moléculaire puis remplacer chacun des groupements protecteurs par le bloc fonctionnel qu'il protège. Le squelette est calculé à partir du graphe obtenu après l'opération précédente. Finalement, le squelette est une abstraction du graphe moléculaire qui ne rend compte que de notions intéressantes du point de vue topologique. Dans cette représentation abstraite, les étiquettes des arêtes ne sont pas nécessaires mais il peut être important de souligner la présence des hétéroatomes en typant de façon générale les sommets concernés par l'étiquette `Het`. La figure 7.6 schématise cette démarche et prend en exemple le cas du Taxol.

Avant d'en donner la définition, il faut préciser que les blocs topologiques sont déterminés à partir du graphe du squelette moléculaire dont on vient de détailler la construction. Les blocs topologiques sont définis comme suit :

⇒ cycle : la définition est celle donnée dans la théorie des graphes. L'inconvénient est que le nombre de cycles qui peuvent être détectés de cette façon peut être très grand. Il est nécessaire de définir une base minimale de cycles qui permette de rendre compte de la réalité chimique tout en réduisant au maximum le nombre d'informa-

tions. Le modèle adopté est celui des cycles pertinents que nous allons décrire dans le paragraphe suivant,

⇒ carbone de branchement : un sommet de type carbone de branchement est un sommet de type carbone non cyclique ayant au moins trois sommets de type carbone voisins non cycliques,

⇒ chaîne (carbonée) : une chaîne carbonée est un sous-graphe connexe composé
- de sommets de type carbone non cycliques et ayant au plus deux sommets de type carbone voisins non cycliques et des arêtes qui les relient entre eux ainsi qu'à des sommets classés comme carbone de branchement,
- ou bien d'un sommet de type carbone lié à aucun autre sommet de type carbone non cyclique,
- ou bien une arête entre deux carbones de branchement,

⇒ lien de cycle : un lien de cycle est une arête non cyclique dont au moins une extrémité est dans un cycle et la (ou les) extrémité(s) cyclique(s),

⇒ lien hétéroatomique est un sous-graphe connexe composé de sommets de type hétéroatomique et non cycliques du squelette et des arêtes (hétéroatomiques) qui les lient entre eux ou à des sommets de type carbone.

Détermination des cycles pertinents

Une partie de la thèse de Philippe Vismara [Vismara, 1995] a été consacrée à l'analyse structurale des graphes moléculaires et plus particulièrement à la détermination d'un ensemble de cycles conduisant à une représentation canonique des systèmes polycycliques. Ce travail a abouti à la définition du concept de *cycle pertinent*. Selon cette définition, tout cycle qui ne constitue pas la somme de cycles plus petits que lui est un cycle pertinent d'un point de vue chimique. L'ensemble des cycles pertinents d'un graphe moléculaire est constitué par l'union des bases de cycles de taille minimale. Cet ensemble peut être beaucoup plus petit que celui des cycles élémentaires d'une molécule. Vismara propose un algorithme polynômial permettant de calculer l'ensemble des cycles pertinents. Sur la figure 7.7, nous avons représenté un système polycyclique et sa décomposition en cycles élémentaires, d'une part, et en cycles pertinents, d'autre part. On remarque effectivement que le nombre de cycles élémentaires (ici, 7) est pratiquement le double du nombre de cycles pertinents (ici, 4), qui, rappelons-le, suffisent à générer tous les autres cycles du système. Par exemple, le cycle de taille 9, `c6`, peut être déduit de l'addition des deux cycles pertinents `C1` et `C2`, l'addition de cycles correspondant à l'union moins l'intersection de leurs arêtes.

Représentation abstraite d'une molécule selon le point de vue topologique au niveau méso

Le résultat de la perception des différents motifs topologiques est donné dans le tableau 7.1. Dans un premier temps, on détecte les blocs topologiques selon les définitions que l'on a posées. Les relations entre les blocs perçus que l'on observe dans le graphe moléculaire nous permettent de construire un graphe des blocs topologiques. On remarque que les relations entre les cycles sont particulières car deux cycles peuvent partager plus d'un sommet. Ainsi

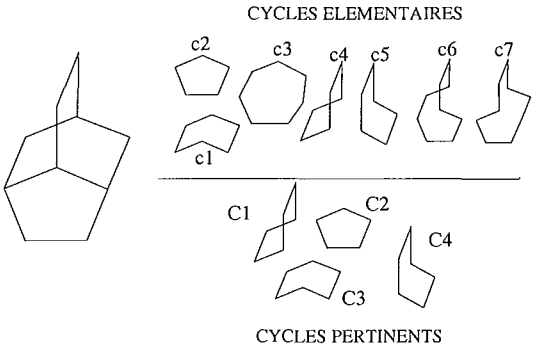
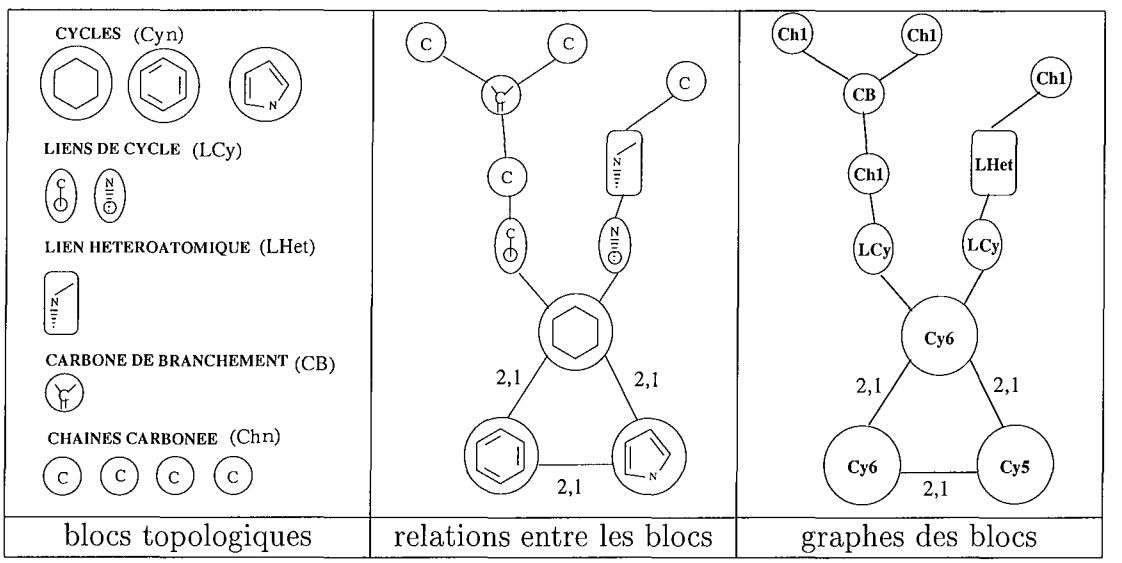


FIG. 7.7 – La perception des cycles dans une molécule polycyclique.



TAB. 7.1 – La perception de la topologie au niveau méso du graphe moléculaire de la Chanoclavine.

pour tenir compte de cette particularité propre aux systèmes cycliques, nous indiquons le nombre de sommets et le nombre d'arêtes partagés au niveau des relations entre blocs cycliques.

7.1.2.2 Les blocs fonctionnels

Définition algorithmique d'un bloc fonctionnel

Nous reprenons ici la définition de type algorithmique des fonctions présentée au chapitre 6. Un bloc fonctionnel correspond donc à un sous-graphe connexe dont les sommets sont de type carbone et/ou hétéroatomique et liés par ce que nous appelons des arêtes fonctionnelles. Une *arête fonctionnelle* est une arête de type multiple entre sommets de type carbone ou une arête hétéroatomique, c'est-à-dire une arête dont au moins une des extrémités est un sommet de type hétéroatomique.

Cette définition algorithmique est fondée sur la représentation des structures molécu-

laires selon le modèle de Lewis. Or ce modèle ne permet pas de prendre en compte les phénomènes de délocalisation des électrons existant dans les systèmes conjugués et qui confèrent à ceux-ci des propriétés particulières. Le phénomène de résonance présent dans ce type de structure donne aux liaisons un caractère différent de celui des liaisons localisées (`simple`, `double`, etc.) : les liaisons carbonées de type simple des systèmes conjugués sont en effet fonctionnelles. La prise en compte de la propriété de mésomérie, et plus particulièrement d'aromaticité, est très importante vu la quantité de molécules pour lesquelles on observe de tels phénomènes. À titre indicatif, en 1980 [Mockus and Stobaugh, 1980] dans le registre des CAS, il y avait 70% de molécules aromatiques.

La mésomérie

La mésomérie concerne potentiellement tout le graphe moléculaire. Détecter la mésomérie consiste à observer s'il existe dans le graphe moléculaire des cas d'alternance entre arêtes de type multiple et arêtes de type simple. Dans le chapitre précédent nous avons montré deux exemples de systèmes conjugués, celui des diènes et celui des carbonyle α,β éthylénique. Dans la plupart des systèmes, la détection des liaisons alternées sert ultimement à la détection de l'aromaticité.

L'aromaticité

L'aromaticité est un cas particulier de la mésomérie. La prise en compte de la propriété d'aromaticité peut se faire par une détection automatique au niveau du graphe moléculaire. Au cours de sa thèse [Métivier, 1987], Pascal Métivier a mis au point un algorithme[5] permettant de détecter la propriété d'aromaticité dans les graphes moléculaires. Contrairement aux précédents travaux sur le sujet, cet algorithme ne dépend pas de la façon dont les graphes ont été saisis. Le principe de l'algorithme se décompose en trois étapes :

1. élimination des cycles ne pouvant pas être aromatiques c'est-à-dire des cycles de taille supérieure ou égale à 8, des cycles contenant une fonction sulfoxyde, une fonction sulfone ou un atome de carbone « sp^3 »[6],
2. génération récursive des sous-systèmes cycliques qui doivent être examinés,
3. évaluation de l'aromaticité pour chacun des sous-systèmes cycliques.

En ce qui concerne l'évaluation de l'aromaticité des sous-systèmes cycliques, il est possible d'utiliser la règle de Hückel qui dit qu'un cycle est aromatique s'il comporte $4n + 2$ électrons π avec n entier naturel différent de zéro. Cette règle ne s'appliquant rigoureusement qu'aux monocycles, Métivier a donné tout un ensemble de règles plus ou moins complexes pour déterminer si un système cyclique est aromatique ou non.

Le résultat de la perception de l'aromaticité sur la molécule de Chanoclavine I permet de préciser le type des arêtes du graphe moléculaire (figure 7.8). Le cycle à 6 chaînons

[5]Cet algorithme a été implanté dans le prototype RESYN ASSISTANT par Anne Laurenço au cours d'un stage de 2^{eme} année d'ENSEEIHT.

[6]Un atome de carbone hybridé sp^3 est un atome de carbone ne faisant avec ces voisins que des liaisons simples. Au niveau du graphe moléculaire, cela se traduit par un sommet de type Carbone, n'ayant que des arêtes incidentes de type simple.

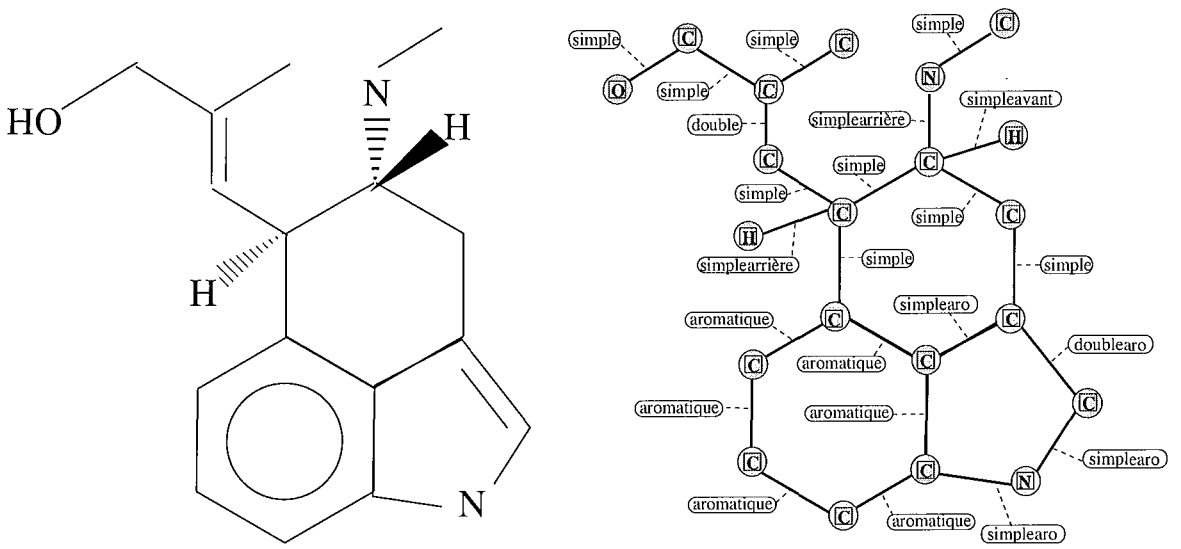


FIG. 7.8 – La perception de la molécule de Chanoclavine I : une représentation de la structure moléculaire et le graphe moléculaire associé.

est un cycle aromatique tandis que les arêtes du cycle azoté de taille 5 possèdent à la fois leur caractère initial tout en étant aromatiques, ce que nous spécifions en leur affectant une étiquette `simple aro` ou `double aro`. Dans notre modèle, le type aromatique pur (`aromatique`) prime sur les autres types de liaison, et notamment sur les types `simple aro` et `double aro`. L'arête commune au cycle aromatique de taille 6 et au cycle de taille 5 est donc étiquetée `aromatique`. Sur la représentation graphique de la structure moléculaire, l'aromaticité du cycle de taille 6 est symbolisée par un cercle illustrant la délocalisation des électrons sur l'ensemble du cycle.

La tautomérie

La prise en compte de la tautomérie est un problème qui a été étudié dans les systèmes d'information chimique [Barnard, 1991]. Deux structures tautomères peuvent s'interchanger l'une en l'autre par un équilibre chimique. Ceci implique de connaître les différentes transformations qui permettent de passer d'une forme tautomère à l'autre. Les motifs structuraux généraux participant à l'équilibre tautomère sont définis sur la figure 7.9 [Mockus and Stobaugh, 1980]. L'équilibre décrit est celui d'un cas particulier de la tautomérie, c'est-à-dire la prototropie. Mais il est signalé que le `H` mobile peut être un atome d'hydrogène mais aussi une charge négative. Sur ce schéma sont représentés les motifs structuraux minimaux nécessaires mais il existe des formes tautomères plus étendues, linéaires, ramifiées ou bien cycliques, correspondant à l'enchaînement de motifs successifs.

Deux problèmes principaux se posent :

- est-ce que deux formes tautomères doivent être regardées comme la même substance ou comme deux composés différents ?
- comment les formes tautomères d'une même substance peuvent être liées automati-

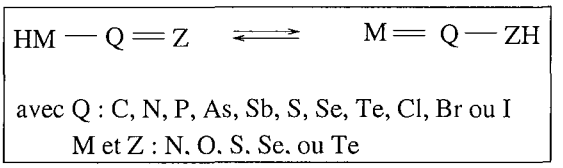


FIG. 7.9 – Les structures tautomères élémentaires.

quement les unes aux autres ?
La plus simple des approches du point de vue de la programmation est d'établir des conventions sur la façon de saisir un tautomère. Des règles de ce type ont été utilisées pour le codage linéaire. Dans le système CAS [Mockus and Stobaugh, 1980], à chaque fois qu'une structure est ajoutée à la base, le graphe moléculaire correspondant est normalisé. Un algorithme de dénormalisation est également employé pour nommer les structures ou retrouver une représentation sous forme de liaisons localisées simples ou doubles. Cependant, les formes céto-énoliques ne sont pas détectées comme tautomères. Dans BEILSTEIN [Welford, 1988], la stratégie adoptée est similaire. Toutes les formes tautomères d'un composé sont stockées individuellement. Un code identique leur est associé. La recherche d'une structure débute par une normalisation qui conduit au calcul du code qui permettra de retrouver toutes les formes tautomères correspondantes présentes dans la base. Dans ORAC [Hopkinson et al., 1990], par contre, le principe utilisé est vraiment différent. L'approche adoptée est basée sur l'utilisation des transformations permettant de passer d'une forme tautomère à l'autre. Quand l'utilisateur le désire, la recherche exacte se fait sur les différentes formes tautomères obtenues après application de chacune des transformations possibles stockées dans une base. Pour notre part, nous n'avons pas encore adopté une position nette face à ce problème difficile. Il est cependant certain qu'il faudrait doter le système d'un module de détection des molécules ayant la propriété de s'interchanger en diverses formules. La connaissance des transformations permettant d'effectuer les interchanges entre les différentes formes nous paraît relativement incontournable dans le traitement de ce problème.

La base de connaissances des fonctions
Une fois tous les blocs fonctionnels détectés dans le graphe moléculaire, il est nécessaire de nommer les motifs mis en évidence. Cela se fait par une classification dans une hiérarchie de graphes de fonctions. Cette hiérarchie fait partie de la base de connaissances du système et nous sert de base d'index pour catégoriser les structures moléculaires. Les graphes moléculaires correspondant aux fonctions qui ont été décrits par des experts sont organisés par une relation d'inclusion de graphes. Dans la hiérarchie, cette relation liant les graphes n'a pas d'autre sens chimique que celui de la spécialisation des structures. En effet, du point de vue de l'inclusion de graphes, une fonction ester (`O=C-O-C`) est plus spécifique qu'une fonction éther (`C-O-C`), cependant les propriétés chimiques de ces deux fonctions sont extrêmement différentes. L'organisation de la hiérarchie permet uniquement d'accélérer la classification et non de rendre compte d'une proximité des propriétés chimiques. Celle-ci doit être établie dans une autre hiérarchie, au niveau des familles chimiques telles que nous les avons définies précédemment. Dans la hiérarchie des fonctions, l'héritage est

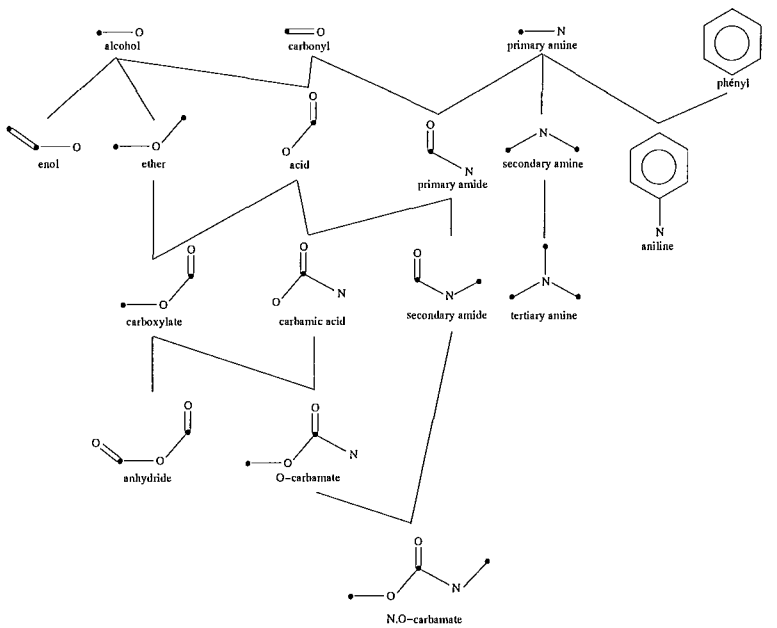


FIG. 7.10 – Un petite partie de la hiérarchie des fonctions de RESYN ASSISTANT.

multiple. Par exemple, un ester est à la fois plus spécifique qu'un éther mais aussi qu'un acide du point de vue de l'inclusion de graphe.

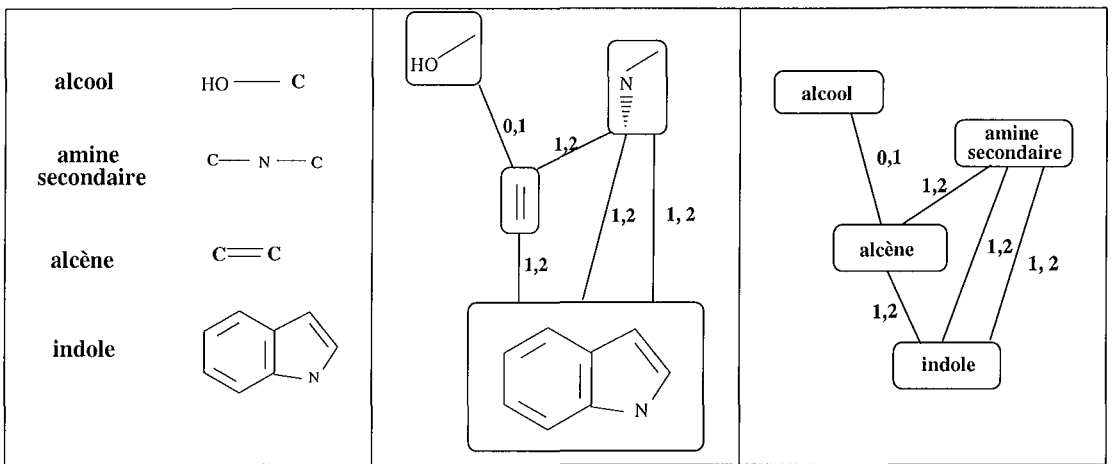
La classification d'un bloc fonctionnel perçu consiste à rechercher dans la hiérarchie de fonctions un graphe isomorphe. Si un tel graphe existe, le bloc fonctionnel portera le nom de la fonction à laquelle il correspond.

Le tableau 7.2 donne les résultats de la perception de la fonctionnalité dans la molécule de Chanoclavine. Tous les blocs fonctionnels détectés de façon automatique ont été nommés car ils correspondent à une fonction de la base de connaissances du système. Sur le graphe des blocs fonctionnels nous avons indiqué la longueur des chemins carbonés séparant les blocs fonctionnels les uns des autres en termes de nombre de sommets puis de nombre d'arêtes. Bien entendu, ces deux informations se déduisent l'une de l'autre. Si deux atomes séparent un bloc fonctionnel d'un autre bloc fonctionnel, cela veut dire que trois liaisons les séparent et réciproquement.

7.1.2.3 Les blocs stéréochimiques

Nous avons vu qu'au niveau du graphe moléculaire il est possible de trouver des indications de nature stéréochimique sur les liaisons. Ainsi pour indiquer les centres de chiralité, les arêtes associées au sommet sont susceptibles d'être étiquetées par un type de liaison hors du plan (`simple avant` ou `simple arrière`). Pour les liaisons dissymétriques, seules les coordonnées dans le plan sont nécessaires.

Du point de vue topologique, rien ne distingue deux graphes ayant mêmes relations de connexité. La condition d'isomorphie basée sur les critères topologiques des graphes ne suffit donc pas pour vérifier que deux graphes sont réellement identiques l'un de l'autre.



blocs fonctionnels	relations entre les blocs	graphes des blocs

TAB. 7.2 – La perception au niveau méso de la Chanoclavine I selon le point de vue de la fonctionnalité.

Sur chacun des schémas de la figure 7.11, les trois structures moléculaires représentées correspondent à des graphes isomorphes entre eux du point de vue topologique, étant donné que l'on a créé une classe d'équivalence entre les types d'arête simple, simple avant et simple arrière. En fait un graphe peut être considéré dans un espace topologique à deux ou trois dimensions. Ceci ne permet pas de donner des coordonnées précises (c'est-à-dire proportionnelles à la réalité) mais conduit tout de même à connaître les positions relatives entre les sommets du graphe (par exemple, on peut savoir qu'un sommet est « à côté », « à gauche », « devant » par rapport à un autre sommet). La détermination des représentations du point de vue de la stéréochimie que nous présentons dans ce qui suit nécessite la prise en compte des coordonnées des sommets du graphe moléculaire.

Représentation implicite

Pour reconnaître de façon non ambiguë si deux structures stéréogéniques sont identiques, les chimistes utilisent les règles établies par Cahn, Ingold et Prelog, connues sous le nom de règles CIP [Cahn et al., 1966]. La première règle CIP ordonne les voisins d'un atome selon leur numéro atomique. Les règles suivantes permettent de résoudre le problème de l'existence de voisins identiques (c'est-à-dire ayant le même numéro atomique) par établissement d'un ordre par voisinage. Nous ne considérons ici que les cas des atomes de carbone asymétriques et des liaisons doubles carbonées dissymétriques mais le modèle peut être étendu à d'autres types d'atomes comme par exemple les atomes d'azote trivalents.

Le cas des atomes asymétriques est traité en considérant l'atome de carbone au centre d'un tétraèdre, les atomes liés à cet atome sont placés aux quatre coins du tétraèdre. Après avoir établi l'ordre de priorité des substituants (sommets voisins du centre d'asymétrie) et placé celui de priorité la plus faible en arrière du plan, il est possible de déterminer le sens de rotation des trois autres substituants suivant leur ordre de priorité. Selon le sens de rotation observé, la configuration de l'atome sera dite R (sens de rotation des aiguilles

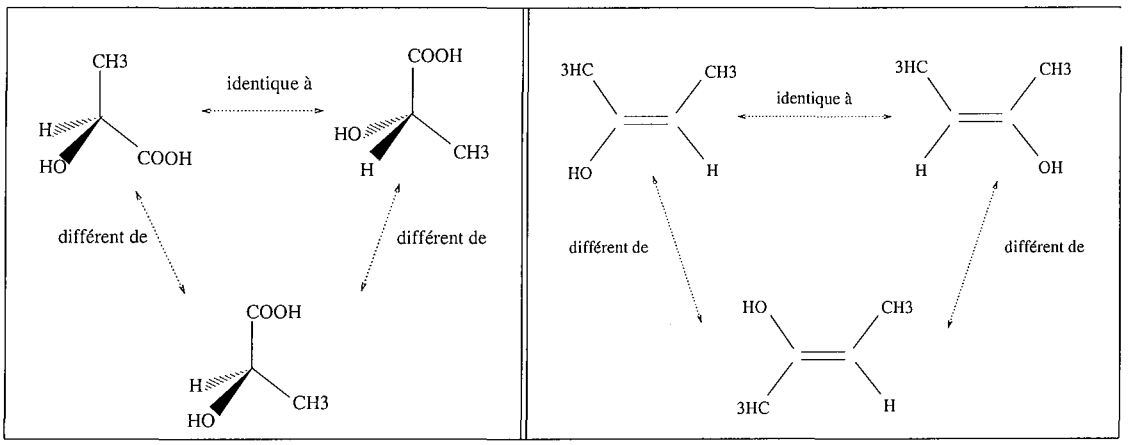


FIG. 7.11 – Exemple de structures topologiquement équivalentes mais pas forcément identiques du point de vue stéréochimique.

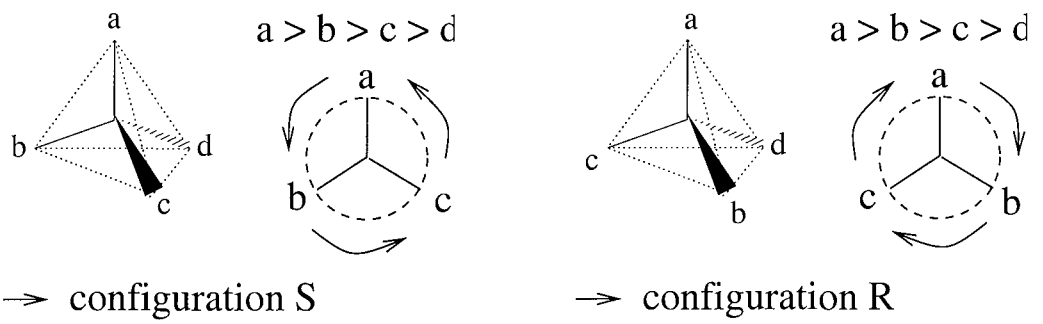


FIG. 7.12 – Le calcul des descripteurs CIP correspondant aux atomes.

d'une montre) ou S (sens de rotation trigonométrique) (figure 7.12).

Pour ce qui est de l'isomérie géométrique, la détermination de celle-ci repose également sur l'établissement de l'ordre de priorité sur les sommets voisins des extrémités de l'arête de type double considérée. La configuration de la double liaison est obtenue en regardant la position relative des substituants de plus haute priorité de chacune des deux extrémités par rapport à la droite contenant la double liaison. Si ces deux substituants sont du même « côté » de la double liaison, la configuration est Z[7], sinon la configuration est notée E[8] (figure 7.13).

Au niveau de la molécule de Chanoclavine, la perception de la stéréochimie (figure 7.14) met en évidence à la fois une double liaison dissymétrique et deux atomes asymétriques. On remarque que ces stéréogènes sont intimement liés les uns aux autres car ils partagent soit un atome soit une liaison.

Le passage par la détermination d'un ordre de priorité sur les substituants du stéréogène, que ce soit un atome ou une liaison, implique que la structure moléculaire sur laquelle on travaille soit une structure complète et non une sous-structure.

[7]Z pour *zusammen* c'est-à-dire ensemble en allemand.

[8]E pour *entgegen* c'est-à-dire opposé en allemand.

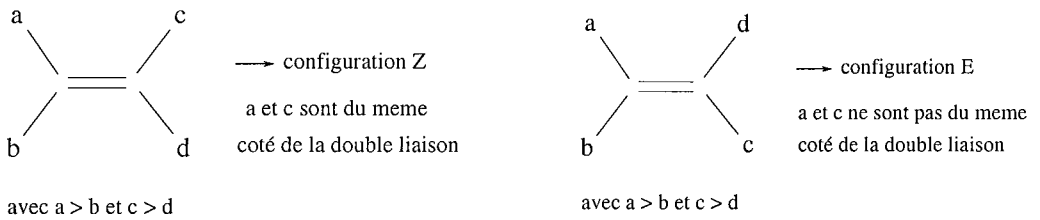


FIG. 7.13 – Le calcul des descripteurs CIP correspondant aux liaisons.

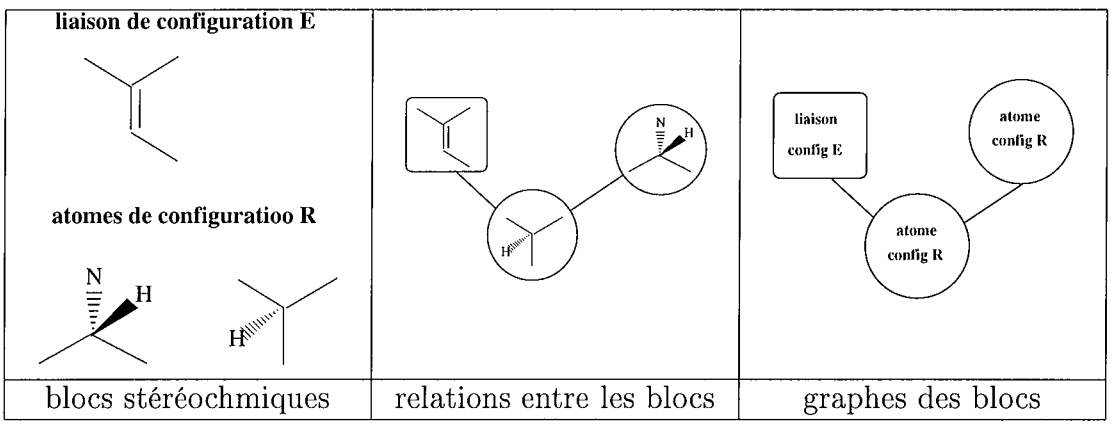


FIG. 7.14 – La perception au niveau méso de la Chanoclavine I selon le point de vue de la stéréochimie.

La représentation implicite donnée par les règles CIP permet de faire de la recherche structurale exacte entre deux graphes G_a et G_b . Il suffit pour cela de rechercher tous les isomorphismes de graphes de G_a dans G_b. Ensuite, on teste pour chaque isomorphisme donné si les descripteurs des atomes appariés sont identiques. Les isomorphismes retenus seront ceux pour lesquels chaque atome de G_a aura un descripteur équivalent à celui de son atome homologue dans G_b.

Représentation explicite

La représentation précédente ne peut être faite que sur des structures moléculaires complètes. Elle ne peut donc pas servir à faire de la recherche sous-structurale. Pour cela, il est nécessaire de développer des modèles généraux de représentation comme l'a fait en particulier Andreas Dietz. Les travaux de Dietz l'ont amené à modéliser de deux façons différentes tout type de centre chiral. Ainsi, il a utilisé un modèle dit des « roues à aubes » [Dietz, 1995] puis s'est appuyé sur des descriptions à base de cartes combinatoires [Dietz et al., 2000]. Sans expliquer ici le principe de ces deux modèles, nous pouvons dire que leur point commun est de s'affranchir de la représentation géométrique en termes de coordonnées et d'angles. Le but est, in fine, d'obtenir une représentation topologique indiquant les relations dans l'espace des atomes ou groupements d'atomes les uns avec les autres. Nous avons repris ce principe pour développer des descriptions plus simples mais moins puissantes que celles proposées par Dietz. Pour cela, nous ne tenons pas compte d'un quel-

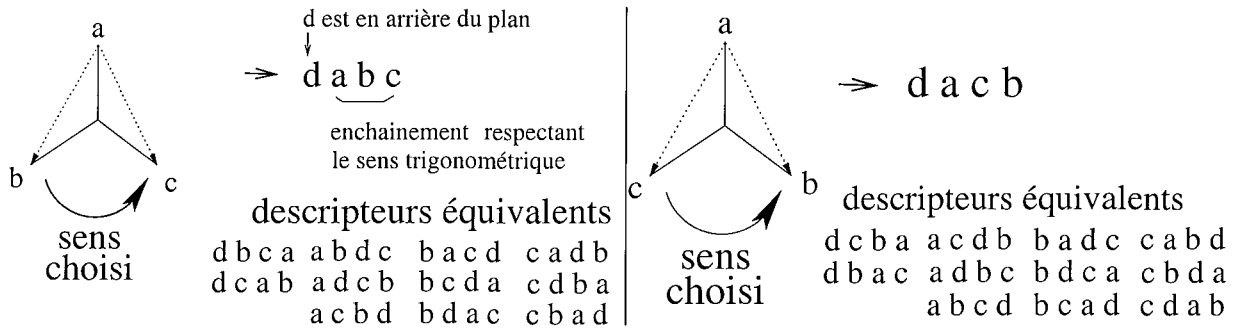


FIG. 7.15 – Le calcul des descripteurs correspondant aux atomes.

conque ordre de priorité sur les sommets adjacents au stéréogène potentiel. Le but est de décrire l'enchaînement des substituants par une liste ordonnée composée de leurs numéros dans le graphe moléculaire. Pour un même stéréogène plusieurs descripteurs peuvent être déterminés car leur calcul dépend de la façon dont le graphe moléculaire a été saisi. L'équivalence de deux descripteurs sera obtenue après des permutations circulaires effectuées sur un des descripteurs.

Cas des atomes asymétriques

Comme nous l'avons vu, un atome de type carbone ayant quatre atomes voisins peut être placé au centre d'un tétraèdre. Si l'on regarde une des faces du tétraèdre (schéma de gauche de la figure 7.15), nous sommes capables de déterminer une liste ordonnée rendant compte de l'enchaînement des atomes selon un sens de rotation fixé. Le sens de rotation positif choisi est le sens trigonométrique, l'enchaînement est déduit du signe du produit vectoriel de deux des vecteurs formés par les arêtes du tétraèdre.

La formule du produit vectoriel est : $\overrightarrow{ab} \wedge \overrightarrow{ac} = ((\|\overrightarrow{ab}\| \times \|\overrightarrow{ac}\|) \times \sin \mathrm{a}).\overrightarrow{N}$, avec $\overrightarrow{N}$ vecteur normal au plan défini par les vecteurs $\overrightarrow{ab}$ et $\overrightarrow{ac}$.
Le calcul du produit vectoriel revient à évaluer le sinus de l'angle entre les deux vecteurs. Ce que nous appellerons désormais produit vectoriel est la valeur algébrique de $\overrightarrow{ab} \wedge \overrightarrow{ac}$. On considère le sens trigonométrique comme sens de rotation positif. Si le produit vectoriel est positif alors on aura l'enchaînement `abc`, alors que si le produit vectoriel est de signe négatif on aura `acb`. Si de plus, par défaut, on convient de compléter la liste par l'indication de l'atome que l'on a mis en arrière en début de liste on obtient : `dabc` pour le produit vectoriel positif et `dacb` pour un produit vectoriel négatif. La liste ordonnée des atomes voisins du stéréocentre obtenue nous donne un descripteur de la configuration de l'environnement spatial du carbone asymétrique. Pour une même configuration, on peut obtenir plusieurs listes (12 au total) équivalentes entre elles. Cela dépend de l'atome choisi mis en arrière et des vecteurs à partir desquels on calcule le produit vectoriel. Il suffit d'avoir une méthode pour déterminer toutes les configurations équivalentes à une liste donnée de façon à pouvoir faire des comparaisons. Sur la figure 7.15, nous avons indiqué les descripteurs équivalents que l'on peut générer à partir d'une configuration.

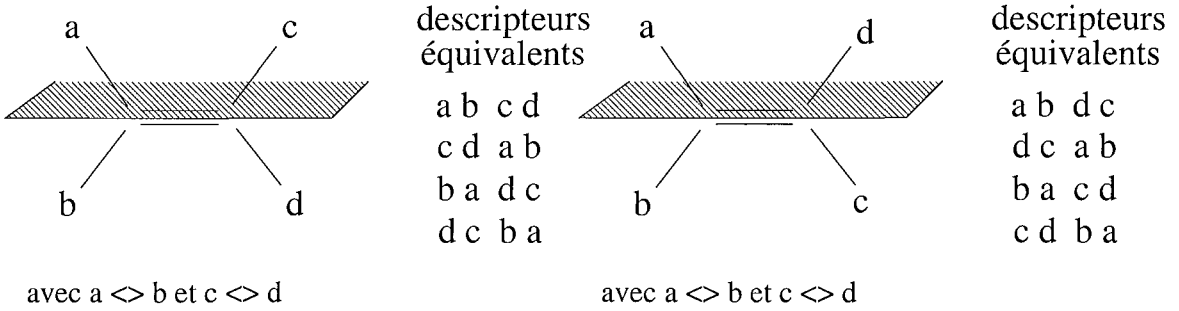


FIG. 7.16 – Le calcul des descripteurs correspondant aux liaisons.

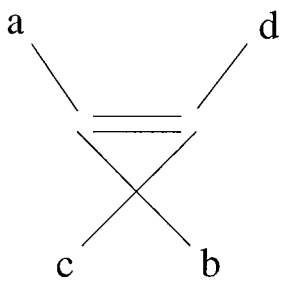


FIG. 7.17 – Cas de dessin ambigu d'une double liaison.

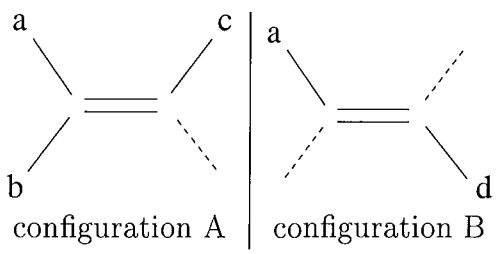


FIG. 7.18 – Des doubles liaisons au voisinage incomplet.

Cas des doubles liaisons dissymétriques

Au niveau des liaisons doubles, si l'on considère la droite dont un segment est la liaison double, on peut constater que les voisins des extrémités de la double liaison se situent de part et d'autre de cette droite. Pour savoir si deux atomes se trouvent du même côté de la droite, il suffit d'évaluer dans quel demi-plan délimité par cette droite se trouve chacun des atomes donnés.

Soit `E : y - ax - b = 0` l'équation de la droite,

l'un des demi-plans ouverts admet pour inéquation `I : y -ax -b > 0`

et l'autre `II : y -ax -b < 0`.

Pour déterminer à quel demi-plan un atome de coordonnées (x,y) appartient, il suffit de calculer la valeur de l'expression `(y - ax - b)` et d'en regarder le signe. Pour construire le descripteur, on décrit chacun des couples de voisins des extrémités de façon à ce que les premiers éléments de chacun des deux couples soient du même côté de la double liaison. Ainsi, si l'on regarde la double liaison de gauche de la figure 7.16, on a les couples de voisins (a,b) et (c,d). Or a et c sont dans le même demi-plan délimité par la double liaison. Le descripteur sera donc `ab cd` ou un de ses trois équivalents obtenus par permutation représentés sur la figure 7.16.

Il faut considérer les cas où le dessin de la double liaison et de son voisinage n'est pas absolument conventionnel. Par exemple, on doit traiter le cas particulier où les liaisons simples partant des atomes extrémités de la double liaison se croisent comme sur le schéma de gauche de la figure 7.17. Pour prendre en compte ces situations, il faut détecter les

liaisons qui se croisent et permuter les coordonnées des atomes. Dans l'exemple de la figure 7.18, il faut échanger les coordonnées des atomes d et c. Avec ces nouvelles coordonnées, on peut faire les calculs sur les distances et rendre compte de la configuration effective de la double liaison.

Il peut y avoir des cas où les extrémités de la double liaison n'ont qu'un voisin d'indiqué explicitement (si c'est dans une molécule, un atome d'hydrogène est implicite ; si c'est dans une sous-structure, il y a un atome fantôme implicite, c'est-à-dire qu'un atome peut se placer dans l'emplacement libre). Dans ces cas là, il faut imaginer que les atomes fantômes viennent se placer dans les positions restantes et sont liés à l'extrémité correspondante par une liaison imaginaire (schémas du milieu et de droite de la figure 7.18). Il est possible de déterminer l'arrangement spatial des stéréocentres dans ces cas particuliers, les atomes fantômes étant indiqués par convention par un -1 dans le descripteur. Par exemple, la configuration A pourra être notée par `abc(-1)` tandis que la configuration B sera notée par `a(-1)(-1)d`. La condition minimale pour décrire une double liaison est que chacune des deux extrémités ait un atome voisin explicite autre que l'autre extrémité de la double liaison ou autre qu'un atome d'hydrogène.

Cette approche est très similaire à celle adoptée par Petrarca et collaborateurs [Petrarca et al., 1967]. La différence est que ceux-ci se basent sur une numérotation des atomes de type Morgan. Nous retrouvons également dans le traitement de la stéréochimie au niveau du système SECS [Wipke and Dyott, 1974] une démarche relativement proche.

La représentation implicite des stéréogènes à l'aide des règles CIP permet de faire de la recherche structurale exacte mais ne peut être utilisée dans la recherche sous-structurale. La solution à ce problème est d'utiliser la représentation explicite que nous venons de présenter.

Recherche des isomorphismes de sous-graphes respectant la stéréochimie

Le principe est le même que celui énoncé dans le cas de recherche de structure exacte à l'aide de la représentation implicite de la stéréoisomérie. Nous partons de l'ensemble de tous les isomorphismes de sous-graphes existant entre les deux graphes à apparier et nous testons chacun des appariements (algorithme 1). Pour chacun des isomorphismes de sous-graphes, on vérifie pour chaque atome du sous-graphe que le descripteur obtenu est équivalent à celui de l'atome avec lequel il est apparié (algorithme 2) et pour chaque liaison double carbonée du sous-graphe que le descripteur est équivalent à celui de la liaison avec laquelle elle est appariée (algorithme 3).

Gestion des atomes

D'une part, on parcourt tous les atomes de la sous-structure et grâce à la correspondance établie, on accède à chacun des atomes homologues dans la structure. Soit `atom_x` l'atome de la sous-structure et `atom_y` l'atome de la structure correspondant. Plusieurs cas peuvent se présenter :

- soit `atom_x` et `atom_y` ne sont pas des stéréocentres : c'est un cas trivial de compatibilité,

Entrée : Graphes moléculaires struct et sousstruc
Sortie : Vecteur contenant tous les appariements compatibles tenant compte de la stéréochimie

```
begin
    vecteur_app_sym ⟵ calcul_app_sym(struct,sousstruct);
    perception_stereo(struct);
    perception_stereo(sousstruct);
    vecteur_result ⟵ vecteur_matching_sym ;
    si struct.stereo ≠ null et sousstruct.stereo ≠ null alors
        pour tout app du vecteur_app_sym faire
            si app_atome_stereo(struct,soustruct,app) rend faux alors
                vecteur_result ⟵ vecteur_result − app;
            si app_liaison_stereo(struct,soustruct,app) rend faux alors
                vecteur_result ⟵ vecteur_result − app;
    retourner vecteur_result;
end
```

Algorithme 1: AppariementStereo(struct,sousstruct)

- soit `atom_x` est stéréocentre et `atom_y` n'est pas un stéréocentre : c'est un cas trivial d'incompatibilité,
- soit `atom_x` n'est pas un stéréocentre et `atom_y` est un stéréocentre : c'est un cas trivial de compatibilité,
- soit `atom_x` est stéréocentre et `atom_y` est un stéréocentre : c'est un cas non trivial, il faut vérifier la compatibilité entre les configurations.

Les deuxième et troisième cas doivent être différenciés car le problème de recherche sous-structurale n'est pas symétrique. En effet, on peut admettre qu'un atome stéréocentre de la structure soit apparié avec un atome non stéréogène de la sous-structure tandis que l'inverse est impossible. La méthode `compare_atom_stereo(atom_struct,atom_sousstruct,app)` donne le résultat de la comparaison des descripteurs associés aux atomes passés en paramètre.

Gestion des liaisons

Comme les liaisons potentiellement stéréochimiques ne sont pas détectées, il faut faire des tests différents de ceux faits pour les atomes et ne pas se baser sur les résultats donnés par la perception de la stéréochimie.

La méthode `compare_liaison_stereo(liaison_struct,liaison_sousstruct,app)` donne le résultat de la comparaison des descripteurs des liaisons passées en paramètre.

7.1.3 Le niveau macro : les systèmes de blocs

Le niveau macro correspond à un niveau d'abstraction supérieur au niveau méso. À ce niveau, on ne s'intéresse plus aux blocs mais aux enchaînements de blocs que nous avons

Entrée : Molécules struct et de soustruct , Appariement app
Sortie : Booléen indiquant si la stéréochimie est respectée par l'appariement donné
begin
  pour *tout atom_sstruct de sousstruct* **faire**
    atom_struct ⟵ appariement(atom_soustruct);
    si *atom_sousstruct.stereo = null et atom_struct.stereo = null* **alors**
      // cas trivial de compatibilité, on peut continuer
    si *atom_sousstruct.stereo ≠ null et atom_struct.stereo = null* **alors**
      // cas trivial d'incompatibilité
      retourner faux;
    si *atom_sousstruct.stereo = null et atom_struct.stereo ≠ null* **alors**
      // cas trivial de compatibilité, on peut continuer
    si *atom_sousstruct.stereo ≠ null et atom_struct.stereo ≠ null* **alors**
      si *compare_atom_stereo(atom_struct,atom_sousstruct,app) rend faux* **alors**
        retourner *faux;*
  retourner *vrai;*
end

Algorithme 2: Sont_Apparies_Stereo_Atomes(struct,soustruct,app)

Entrée : Graphes moléculaires struct et soustruct, Appariement app
Sortie : Booléen indiquant si l'appariement respecte ou non la stéréochimie au niveau des liaisons doubles
begin
  pour *toute liaison_sousstruct de soustruct* **faire**
    liaison_struct ⟵ appariement(liaison_sousstruct) ;
    si *compare_liaison_stereo(liaison_struct,liaison_sousstruct,app) rend faux* **alors**
      retourner *faux*
  retourner *vrai*
end

Algorithme 3: Sont_Appariees_Stereo_Liaisons(struct,sousstructure,app)

appelés *systèmes de blocs* ou de façon plus concise *systèmes*. Les systèmes de blocs peuvent être déterminés à partir du graphe des blocs selon un point de vue donné. On obtiendra ainsi des :

- systèmes topologiques,
- systèmes fonctionnels,
- systèmes stéréochimiques.

Nous allons voir quelles sont les définitions de ces différentes classes de systèmes de blocs.

7.1.3.1 Les systèmes topologiques

Au niveau topologique, nous considérons trois types de systèmes caractéristiques :

- les systèmes cycliques,
- les systèmes carbonés linéaires,
- les systèmes carbonés ramifiés.

Deux cycles partageant au moins un sommet forment un système de cycles. Un cycle isolé, c'est-à-dire ne partageant aucun sommet avec un autre cycle, est un système de cycles à part entière. Comme nous l'avons dit dans le chapitre précédent, il existe plusieurs sortes de combinaisons de deux cycles :

- deux cycles forment un système spiranique s'ils partagent seulement un sommet,
- deux cycles sont condensés s'ils partagent une et une seule arête, donc ont deux sommets en commun,
- deux cycles forment un système ponté s'ils partagent au moins trois sommets et ont donc en commun au moins deux arêtes.

Toute chaîne perçue est un système carboné linéaire.

Les systèmes carbonés ramifiés consistent en des assemblages de chaînes et d'atomes de carbone de branchement consécutifs.

Si l'on perçoit les systèmes topologiques dans le graphe moléculaire de la molécule de Chanoclavine, on remarque un système cyclique, une chaîne carbonée ramifiée ainsi qu'une chaîne carbonée linéaire réduite à un seul sommet (figure 7.19). De même que l'on construit un graphe des blocs, il est possible de déterminer un graphe des systèmes de blocs décrivant les relations entre les différents systèmes topologiques.

7.1.3.2 Les systèmes fonctionnels

De par la définition des fonctions que nous avons adoptée, deux blocs fonctionnels ne peuvent être adjacents et sont donc séparés par au moins une arête carbonée étiquetée `simple`, `simple avant` et/ou `simple arrière`. En conséquence, les systèmes de blocs fonctionnels sont des enchaînements de blocs fonctionnels séparés les uns des autres par un chemin carboné. Dans le but de générer des systèmes de blocs fonctionnels pertinents du point de vue chimique, les travaux du GDR TICCO ont conduit à imposer une taille maximale au chemin carboné au delà de laquelle l'enchaînement de fonctions n'est pas considéré en tant que tel. Cette taille a été fixée à trois en terme des sommets de type

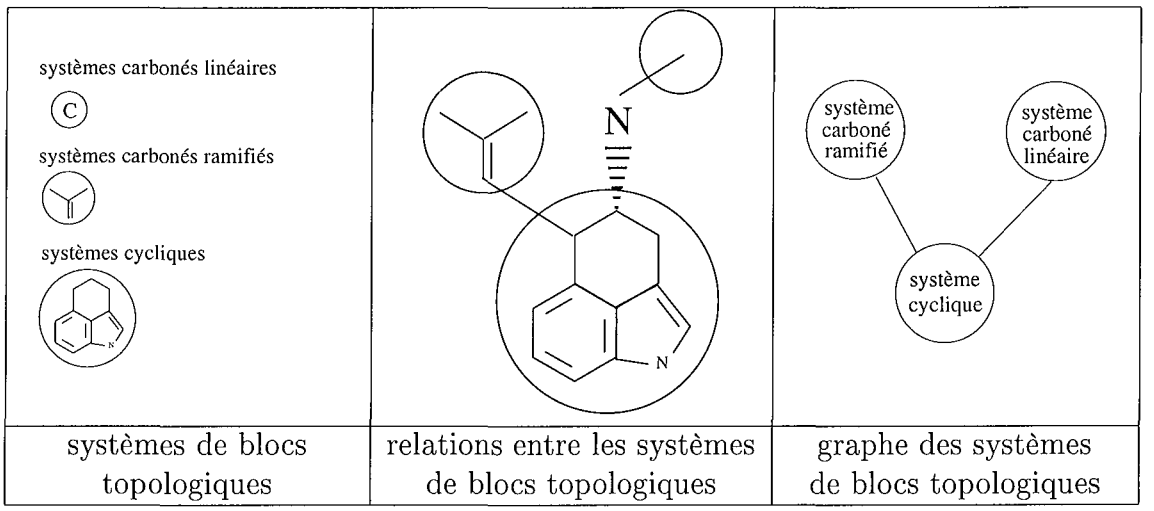


FIG. 7.19 – Les systèmes de blocs topologiques du graphe moléculaire de la Chanoclavine I.

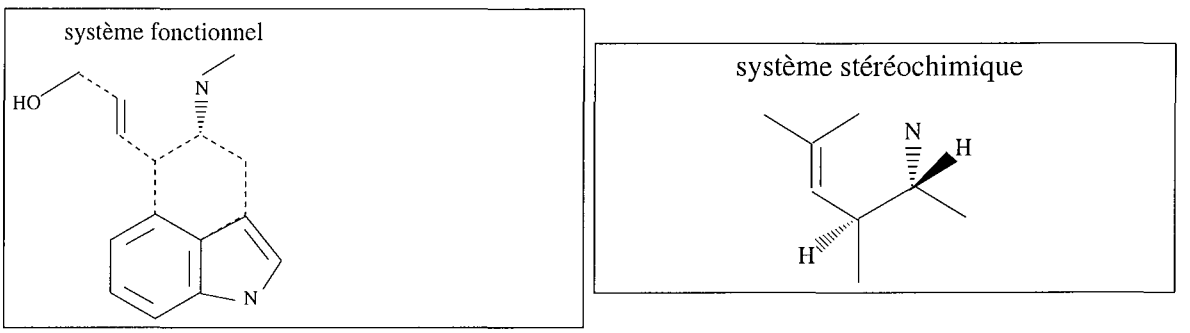


FIG. 7.20 – Le système de blocs fonctionnels et le système de blocs stéréochimiques perçus dans le graphe moléculaire de la molécule de Chanoclavine. Les chemins entre les blocs fonctionnels sont représentés par des lignes en pointillés.

carbone, c'est-à-dire que le chemin carboné entre deux blocs est constitué d'au plus quatre arêtes carbonées consécutives. Deux fonctions éloignées d'une distance supérieure ont peu, voire pas, d'interaction l'une envers l'autre.

7.1.3.3 Les systèmes stéréochimiques

Les systèmes stéréochimiques sont constitués de blocs stéréochimiques partageant un ou plusieurs sommets, voire des arêtes. Dans la molécule de Chanoclavine, on obtient un seul système de blocs stéréochimiques regroupant tous les blocs stéréochimiques perçus (figure 7.20).

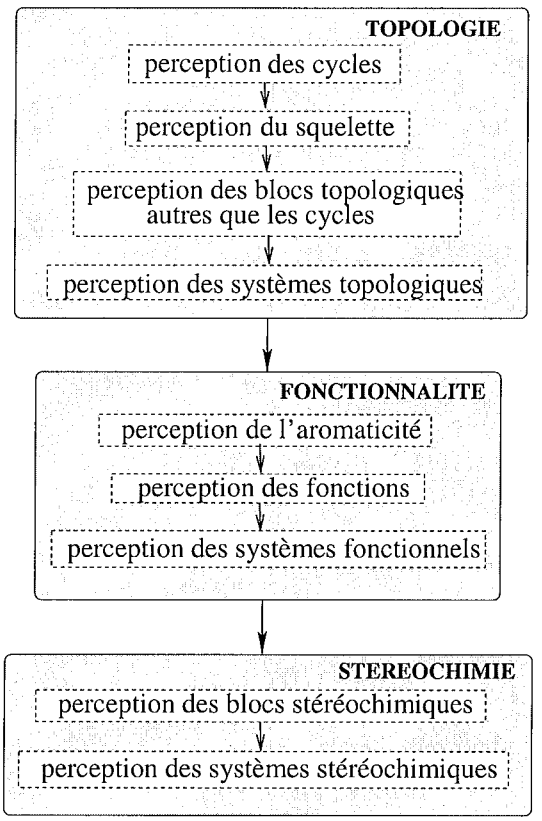


FIG. 7.21 – Le processus de perception des molécules.

7.1.4 Le processus de perception

La perception se fait selon l'enchaînement indiqué dans la figure 7.21. La détermination du squelette moléculaire nécessite la perception préalable des cycles du graphe moléculaire. On détermine ensuite les autres blocs topologiques à partir du graphe du squelette ainsi que les systèmes topologiques. À partir de la perception des systèmes cycliques, les propriétés d'aromaticité des différents cycles peuvent être calculées. Ensuite, les blocs fonctionnels sont déterminés de façon algorithmique et les systèmes fonctionnels sont construits par le calcul des chemins entre les blocs fonctionnels. Le point de vue stéréochimique peut alors être considéré avec la détermination des blocs stéréochimiques puis des systèmes stéréochimiques.

7.2 Représentation des méthodes de synthèse

Nous allons voir dans cette section comment nous analysons les méthodes de synthèse. Nous illustrerons nos propos par l'étude de la réaction représentée sur la figure 7.22. Cette réaction est un exemple d'utilisation de la méthode de Diels-Alder issu de la littérature.

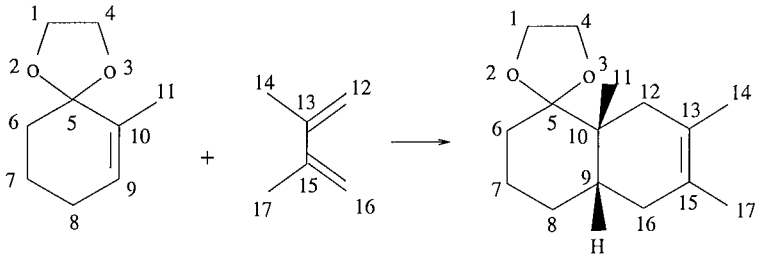


FIG. 7.22 – Un exemple d'utilisation de la méthode de Diels-Alder. Les résultats sur cette réaction ont été publiés en 1995 dans le journal Synlett par P. A. Grieco, J. L. Collins et S. T. Handy. La réaction se fait en milieu anhydride en présence d'éther et de THF. Dans ces conditions, les auteurs ont obtenu le produit avec un rendement de 96%. Nous avons indiqué l'appariement qui est nécessaire à l'analyse de la méthode.

7.2.1 Le niveau micro

Selon ce que l'on a présenté dans le chapitre précédent, le niveau micro est celui des atomes et des liaisons. Considérer une méthode de synthèse au niveau micro consiste à observer les modifications qui ont eu lieu au niveau des liaisons. C'est ce principe que Vleduts propose dans [Vleduts, 1963] pour classer et retrouver les réactions dans un système informatique d'aide à la résolution de problèmes synthèse. Cette façon de représenter les réactions a été par la suite reprise et a donné lieu à la construction de cœur, centre ou site de réactions, selon la terminologie adoptée ici et dont nous avons déjà brièvement parlé dans le chapitre 3. Pour déterminer ce qui se passe au cours de la méthode de synthèse, il faut disposer de la correspondance entre atomes (figure 7.22). Sur l'exemple choisi, l'appariement donné respecte le bilan d'une méthode de type Diels-Alder.

Grâce à cet appariement, nous pouvons déduire, entre autres, que :

- la liaison simple entre l'atome de carbone numéroté 17 et l'atome de carbone numéroté 15 reste inchangée,
- la liaison simple entre les deux atomes de carbone numérotés respectivement 15 et 13 a changé de type. Elle est devenue insaturée en passant du type simple au type double,
- la liaison entre les deux atomes de carbone numérotés respectivement 9 et 16 est créée au niveau du produit, elle n'existe pas dans les réactants.

Nous pouvons présenter de manière condensée et non redondante le bilan de la méthode de synthèse au moyen d'un graphe dont les types d'étiquettes des arêtes rendent compte de la dynamique de la réaction. Pour cela, Vladutz a proposé ce qu'il a appelé le *surimposed reaction graph* ou SRG puis le *surimposed reaction skeleton graph* ou SRSG [Vladutz, 1986]. Le SRSG est une réduction du SRG aux liaisons qui sont modifiées (créées, détruites ou dont le type a changé) au cours de la réaction. La construction du SRG consiste à superposer les sommets des graphes des réactants et leur homologue dans les graphes des produits de façon à ce que chaque atome ne soit représenté qu'une seule fois. De nombreuses équipes de recherche ont repris cette idée [Wilcox and Levinson, 1986] [Fujita, 1988] [Laurenço et al.,

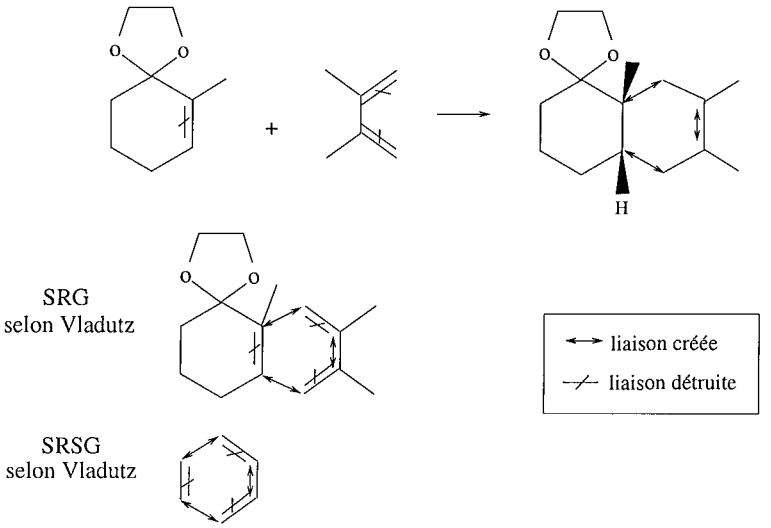


FIG. 7.23 – Une représentation condensée des réactions : SRG et SRSG selon Vladutz.

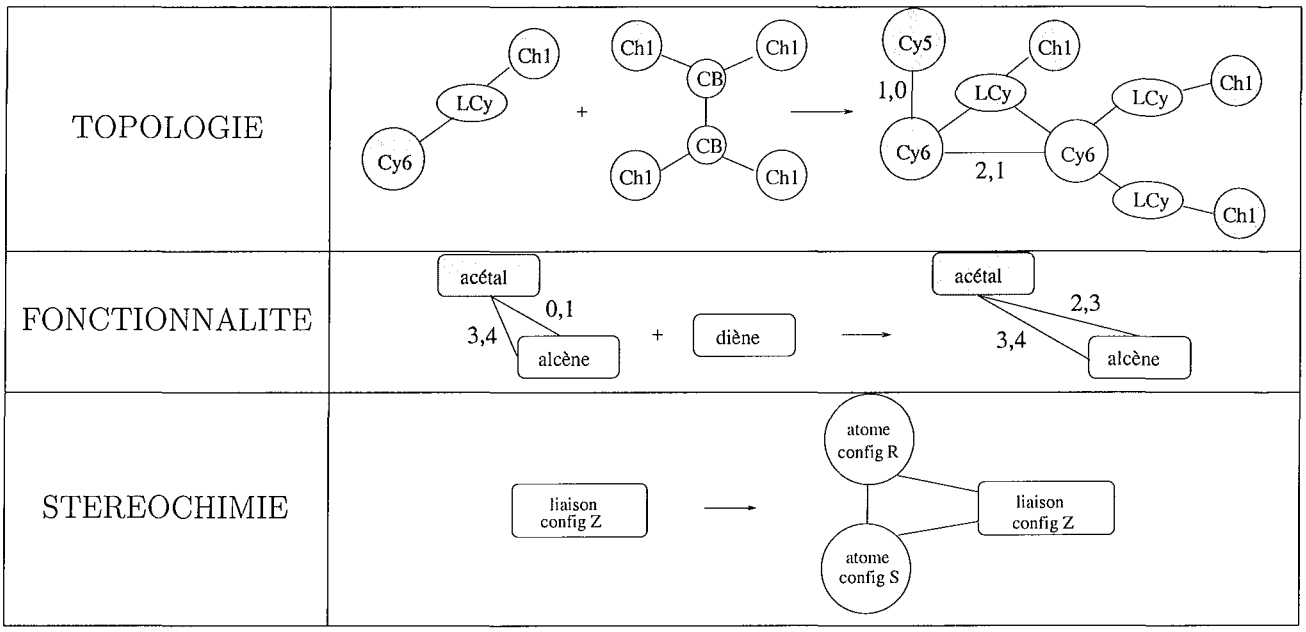
1990]. Ainsi Philippe Jauffret et collaborateurs construisent un *graphe condensé de réaction* ou GCR pour Graphe Condensé de Réaction [Hanser et al., 1995] qui est la superposition de *graphes moléculaires enrichis* (GME) des réactants et des produits. Dans leur modèle, les graphes moléculaires sont enrichis par l'ajout d'informations spécifiques telles que la configuration CIP des stéréogènes ou la charge des atomes (représentée par un pseudo-atome relié à l'atome qu'il qualifie). Les représentations condensées des réactions présentent donc une partie statique et une partie dynamique. L'ensemble des liaisons modifiées au cours de la réaction constitue le *cœur de la réaction*, c'est-à-dire le SRSG selon Vladutz.

7.2.2 Le niveau méso

Le niveau micro décrit les méthodes de synthèse au moyen des modifications intervenues au niveau des liaisons. Ce niveau d'abstraction peut être utile mais ne permet pas une différenciation suffisante des méthodes de synthèse. Tout comme nous percevons les molécules, nous sommes capables de percevoir les méthodes de synthèse à divers niveaux d'abstraction. Au niveau méso, nous nous servons de la perception des molécules ainsi que de la correspondance des atomes pour analyser les méthodes de synthèse, ceci selon les trois points de vue dont nous avons précédemment discuté.

7.2.2.1 La correspondance des blocs

Notre analyse au niveau méso repose sur une correspondance entre les différents blocs perçus. Nous travaillons donc à partir des représentations abstraites des graphes correspondant aux réactants et aux produits. Connaissant les atomes associés à chacun des blocs perçus, nous établissons la correspondance grâce à l'appariement des atomes dont nous disposons. Pour cela, nous avons posé la définition suivante : *deux blocs se correspondent*



TAB. 7.3 – Les représentations abstraites des molécules participant à une méthode de synthèse.

s'ils ont en commun au moins un atome. Cette correspondance nous permet de mettre en évidence différentes situations. En effet, au cours d'une méthode de synthèse, un bloc peut avoir été modifié ou bien peut être resté inchangé. Un bloc restant inchangé peut tout de même avoir participé à la construction d'un autre bloc. Nous allons donner des exemples de ces diverses possibilités dans ce qui suit. Nous déterminons une correspondance des blocs pour chacun des trois points de vue, c'est-à-dire que nous n'établissons de correspondance qu'entre blocs de même nature (topologique, fonctionnelle ou stéréochimique). Il n'est pas difficile de faire une correspondance prenant en compte tous les points de vue à la fois mais une analyse de ce type pose le problème de l'interprétation et de l'utilité d'une telle représentation. Dans les sections suivantes, nous allons voir la description abstraite des méthodes de synthèse que la correspondance des blocs nous permet d'établir selon chacun des points de vue.

7.2.2.2 La perception des méthodes de synthèse selon le point de vue de la topologie

La première ligne du tableau 7.3 correspond à la perception de la méthode de synthèse selon le point de vue de la topologie. Les blocs grisés sont ceux qui sont inchangés au cours de la réaction. Bien qu'inchangé, le cycle de taille 6 participe à la création du cycle de taille 6 qui est créé au cours de la réaction pour former un bicycle. Les atomes de branchement du deuxième réactant sont à l'origine des liens de cycle du cycle formé. La figure 7.24 récapitule le détail de la correspondance de blocs que l'on établit. Le cycle hétéroatomique de taille 5 n'est pas pris en compte dans les blocs topologiques car ce bloc correspond au

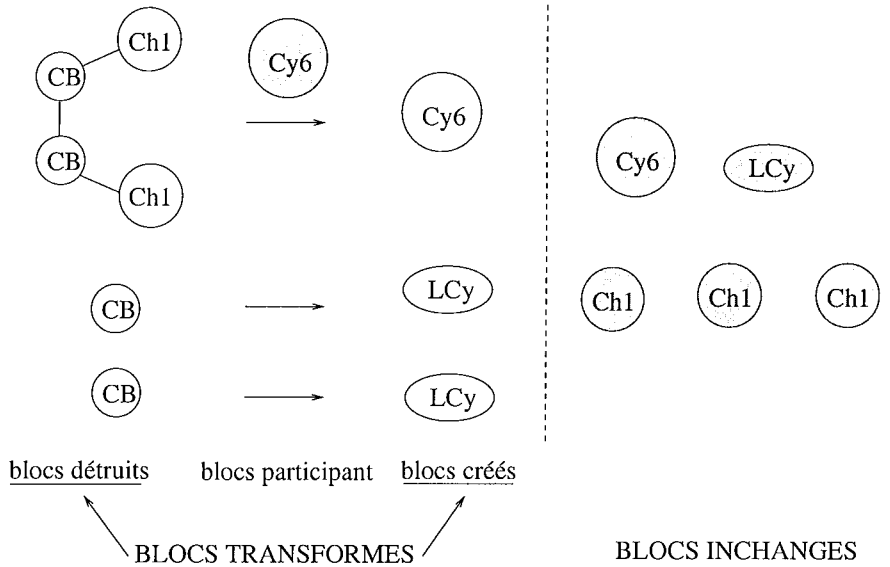


FIG. 7.24 – La correspondance des blocs topologiques de la méthode de synthèse.

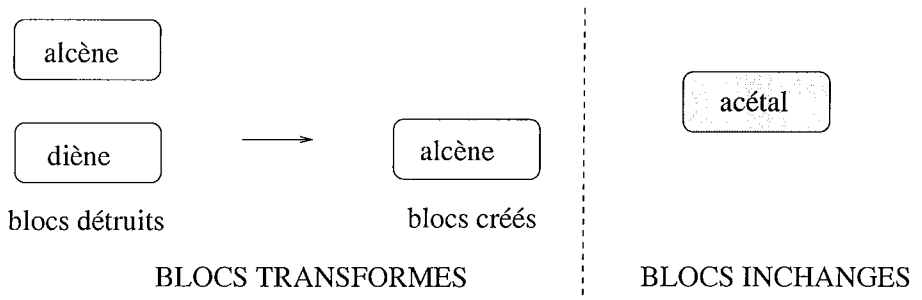


FIG. 7.25 – La correspondance des blocs fonctionnels de la méthode de synthèse.

1,3 dioxolane qui est un groupement protecteur de la fonction carbonyle. Ce bloc ne fait donc pas partie du squelette moléculaire à partir duquel est déterminée la représentation abstraite d'ordre topologique.

7.2.2.3 La perception des méthodes de synthèse selon le point de vue de la fonctionnalité

La correspondance des blocs nous indique qu'une fonction alcène a été créée à partir d'un diène tandis que la fonction alcène du premier réactant est détruite. Ces modifications de fonctionnalité s'opèrent alors qu'une fonction acétal (grisée sur la figure correspondant à la perception de la fonctionnalité dans le tableau 7.3) demeure inchangée. Dans le cas de la fonctionnalité, on n'a pas de bloc inchangé de type « participant » au sens où nous l'avons employé pour les blocs topologiques. Un bloc fonctionnel qui participe à la création d'un autre bloc fonctionnel est forcément détruit au cours de la réaction.

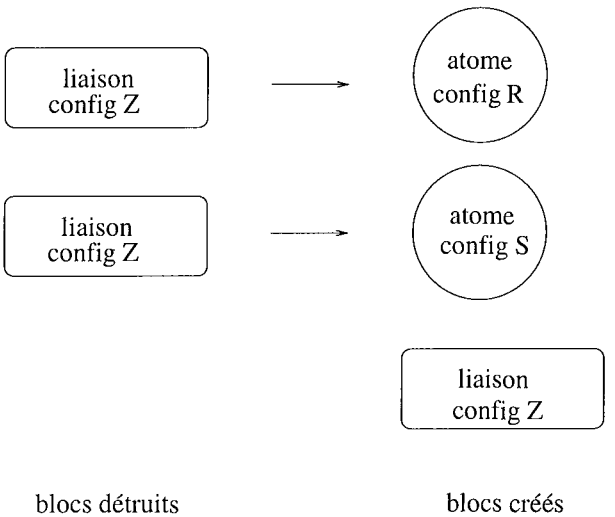


FIG. 7.26 – La correspondance des blocs stéréochimiques de la méthode de synthèse.

7.2.2.4 La perception des méthodes de synthèse selon le point de vue de la stéréochimie

Au niveau de la stéréochimie, tous les blocs stéréochimiques ont été modifiés au cours de la méthode de synthèse. La liaison double présente dans les réactants est réduite et n'est donc plus un stéréogène dans le produit. Deux stéréocentres sont créés à la jonction des deux cycles condensés et la double liaison créée est dissymétrique.

7.2.3 Le niveau macro

Nous venons de voir l'analyse des méthodes de synthèse que nous faisons au niveau méso. De façon analogue, nous pouvons faire une analyse au niveau macro. Pour cela, nous établissons une correspondance entre systèmes de blocs selon le principe suivant : *deux systèmes de blocs se correspondent si et seulement s'ils partagent des atomes.* Un système de blocs sera considéré comme inchangé si chacun des blocs le composant reste inchangé au cours de la réaction et chacun des chemins entre les blocs reste également inchangé.

L'analyse de la méthode prise en exemple selon chacun des points de vue est représentée ci-dessous. Du point de vue topologique, il est intéressant de voir qu'il y a eu création d'un nouveau système cyclique à partir d'un autre système cyclique et d'une chaîne ramifiée (figure 7.24). Au niveau fonctionnel (figure 7.28), deux systèmes s'interchangent pour en donner un nouveau tandis qu'au niveau stéréochimique (figure 7.29), un système réduit à une double liaison participe à la création d'un système stéréochimique complexe.

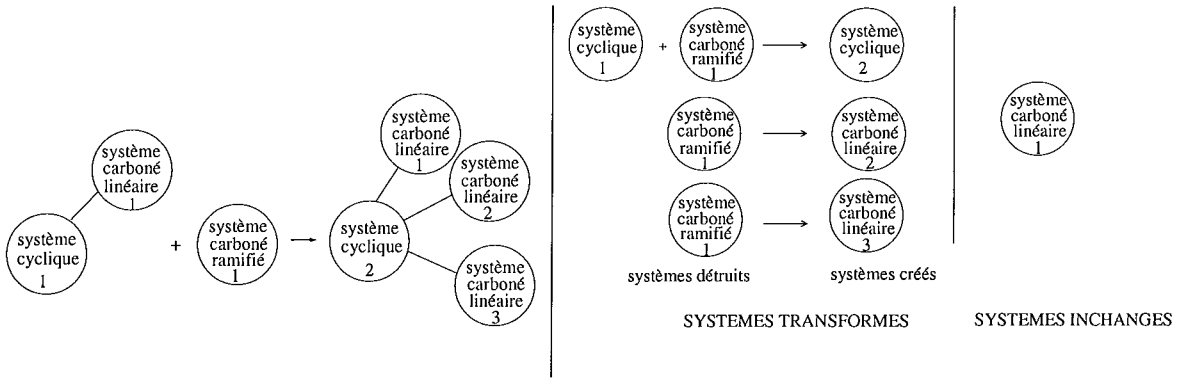


FIG. 7.27 – La correspondance des systèmes de blocs topologiques de la méthode de synthèse.

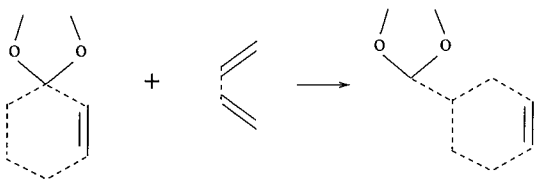


FIG. 7.28 – La correspondance des systèmes de blocs fonctionnels de la méthode de synthèse.

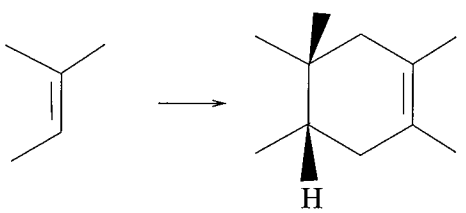


FIG. 7.29 – La correspondance des systèmes de blocs stéréochimiques de la méthode de synthèse.

7.2.4 Conclusion

La perception des méthodes de synthèse que nous effectuons nous permet de les décrire en terme de blocs ou de systèmes modifiés ou inchangés. Cette analyse des méthodes de synthèse correspond bien à l'expression des objectifs atteints par la méthode de synthèse. Dans le chapitre 8, nous verrons que dans certains cas nous utiliserons directement les résultats donnés par cette perception tandis que dans d'autres cas nous procèderons à une interprétation de la correspondance mise en évidence.

7.3 Transformations et objets associés

Du point de vue de la rétrosynthèse, nous devons considérer les méthodes de synthèse sous l'angle des transformations. Conformément à notre modèle conceptuel, à une méthode de synthèse correspond une transformation.

7.3.1 Le cœur d'une transformation

Le cœur d'une transformation est le contexte minimal d'application d'une transformation auquel on rajoute l'information sur les modifications de liaisons observées ainsi que les transformations stéréochimiques. Pour exprimer les modifications des liaisons au niveau

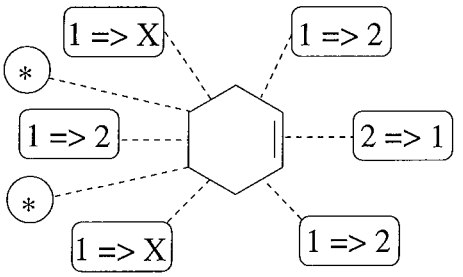


FIG. 7.30 – La cœur de la transformation associée à un exemple d'utilisation de la méthode de Diels-Alder.

du graphe du cœur, nous avons créé un type de liaison dynamique qui contient l'information du type de liaison avant (type de liaison dans le produit) et après (type de liaison dans les réactants) l'application de la transformation. Cette notion de *liaison dynamique* est présente depuis longtemps dans les travaux du GDR TICCO [Laurenço et al., 1990]. La construction du cœur d'une transformation est basée sur la correspondance des atomes des précurseurs et de la cible. Cette correspondance entre atomes permet de déterminer la correspondance entre les liaisons qui elle-même conduit à la comparaison des types de liaisons avant et après la transformation. Nous rappelons, si cela est nécessaire, que la prise en compte des changements stéréochimiques effectifs observés nous a amené à créer une classe d'équivalence entre les types `simple`, `simple avant` et `simple arrière`. En effet, le fait de remarquer qu'au cours d'une méthode de synthèse une liaison est passée en avant du plan n'indique pas de façon correcte la modification observée et ne permet pas d'envisager une application de la transformation. Celle-ci doit se faire grâce à une description explicite qui ne dépend pas de la façon dont le graphe moléculaire a été saisi. Pour signaler qu'un sommet ou une arête est le centre d'une modification de nature stéréochimique, nous les marquons à l'aide d'un drapeau (astérisque sur le schéma de la figure 7.30).

7.3.2 Les contextes d'une transformation

Dans le chapitre 6, nous avons associé à la notion de transformation celle de contexte, c'est-à-dire d'environnement favorable ou défavorable à la réalisation de la méthode de synthèse. La définition des contextes d'une transformation fait intervenir de nombreuses connaissances chimiques. Dans le prototype RESYN, tout un réseau de contextes a été construit par des experts de la synthèse. Bien que la mise à jour de ce réseau soit facilitée par l'interface graphique proposée, cette tâche n'en reste pas moins longue et fastidieuse car elle nécessite une étude précise de la littérature. Au cours de cette thèse nous avons essayé d'automatiser la mise à jour du réseau de contextes en extrayant de façon algorithmique les contextes associés aux transformations. La question s'est donc posée de savoir comment définir un contexte de façon à pouvoir l'extraire de manière automatique des données dont on dispose à propos des méthodes de synthèse. Les chimistes utilisent des critères reposant sur des connaissances qui sont difficiles à formaliser. En revanche, étant donné que les environnements influençant la faisabilité de la méthode de synthèse sont censés être

Sortie : le cœur de la transformation
begin
typeLiaisonDynam : type de liaison dynamique ;
atomeProd, atomeReact : sommets respectivement du produit ou d'un des réactants ; cœur ← graphe_vide ;
pour *chaque arête liaisonProd du produit* **faire**
liaisonReact ← correspondance_arête(liaisonProd) ;
etiqProd = étiquette(liaisonProd) ;
etiqReact = étiquette(liaisonReact) ;
si *etiqProd* ≠ *etiqReact* **alors**
calcul des types de liaison dynamique et ajout des liaisons
si *(etiqlProd = Liaison simple avant* ∨ *etiqlProd = Liaison simple arrière)* **alors**
etiqProd ← Liaison simple ;
si *(etiqlReact = Liaison simple avant* ∨ *etiqlReact = Liaison simple arrière)* **alors**
etiqReact ← Liaison simple ;
si *(etiqProd* ≠ *etiqReact)* **alors**
typeLiaisonDynam ← calculTypeDynamique(etiqProd,etiqReact);
ajouter(liaisonProd,typeLiaisonDynam,cœur);
marquage des sommets pour lesquels on observe des changements d'ordre stéréochimique
pour *chaque sommet atomeCoeur du cœur* **faire**
atomeProd ← sommet correspondant à atomeCoeur ;
atomeReact ← sommet correspondant à atomeCoeur ;
si *(descripteur_stereo(atomeProd) non équivalent à descripteur_stereo(atomeProd))* **alors**
marquer(atomeCoeur) ;
retourner *cœur*;
end

Algorithme 4: construction−coeur

proches du centre de réaction, il nous a paru judicieux d'examiner les motifs structuraux adjacents au centre de réaction et ceci selon les trois points de vue. Un motif, c'est-à-dire un bloc, est adjacent au centre de réaction s'il a au moins un sommet en commun avec le sous-graphe correspondant au centre de réaction. Selon ce principe, nous sommes capables de construire :

- un contexte topologique qui comprend tous les blocs topologiques adjacents au centre de réaction,
- un contexte fonctionnel qui comprend tous les blocs fonctionnels adjacents au centre de réaction,
- un contexte stéréochimique qui comprend tous les blocs de nature stéréochmique adjacents au centre de réaction.

Au premier abord, le contexte fonctionnel peut apparaître le plus intéressant des contextes déterminés. Il indique les fonctions proches des liaisons réactives. Du fait de la proximité, ces fonctions ont certainement participé au déroulement de la méthode de synthèse. Les contextes topologiques et stéréochimiques sont également importants. Pour une méthode du type Diels-Alder, par exemple, on constate que le cycle créé est très souvent substitué (dans le contexte topologique, on retrouve les liens de cycle) et que les substitutions sont le siège de stéréocentres.

Idéalement, les connaissances sur les contextes devraient couvrir à la fois ceux des exemples favorables et des exemples défavorables connus. Cependant, étant donné l'origine des données que nous exploitons, nous ne pouvons récupérer de cette façon que des contextes favorables. Vu qu'il n'existe pas de base répertoriant les exemples négatifs, l'apport de contextes défavorables à la base de connaissances du système dépend de la saisie manuelle d'exemples négatifs par des experts de la synthèse.

Sortie : le contexte minimal de la transformation
begin
  contexteMinimal ← graphe_vide ;
  pour *chaque arête bondProd du produit* **faire**
    bondReact ← correspondance_arête(bondProd) ;
    si *étiquette(bondProd)* ≠ *étiquette(bondReact)* **alors**
      ajouter(bondProd,ontexteMinimalr) ;
  retourner *contexteMinimal ;*
end

Algorithme 5: construction−contexte−minimal

7.3.3 Le transformateur d'une transformation

Une fois le travail de sélection des transformations applicables effectué, il faut être capable d'appliquer ces transformations à la molécule cible pour déterminer les précurseurs potentiels. La transformation s'effectue par l'intermédiaire d'un opérateur que nous

Entrée : un point de vue POV
Sortie : le contexte correspondant au point de vue POV
begin
  contextePOV ← contexteMinimal ;
  pour *chaque arête bondContexteMin de contexteMinimal* **faire**
    pour *chaque blocPOV de POV auquel bondContexteMin appartient* **faire**
      contextePOV ← fusionner(contextePOV,blocPOV) ;
  retourner *contextePOV ;*
end

Algorithme 6: construction1−contexte−POV

Entrée : un point de vue POV
Sortie : le contexte de la transformation correspondant au point de vue POV
begin
  contextePOV ← contexteMinimal ;
  pour *chaque blocPOV perçu dans le produit selon POV* **faire**
    pour *chaque blocPOV auquel bondContexteMin appartient* **faire**
      si *BlocPOV est créé* **alors**
        contextePOV ← fusionner(contextePOV,blocPOV) ;
      sinon
        si *BlocPOV est inchangé mais participe* **alors**
          contextePOV ← fusionner(contextePOV,blocPOV) ;
  retourner *contextePOV ;*
end

Algorithme 7: construction2−contexte−POV

Sortie : le contexte global de la transformation
begin
  contexteGlobal ← contexteMinimal ;
  pour *chacun des points de vue POV* **faire**
    contextePOV ← contexte du point de vue POV ;
    contexteGlobal ← fusionner(contexteGlobal,contextePOV) ;
  retourner *contexteGlobal ;*
end

Algorithme 8: construction−contexte−global

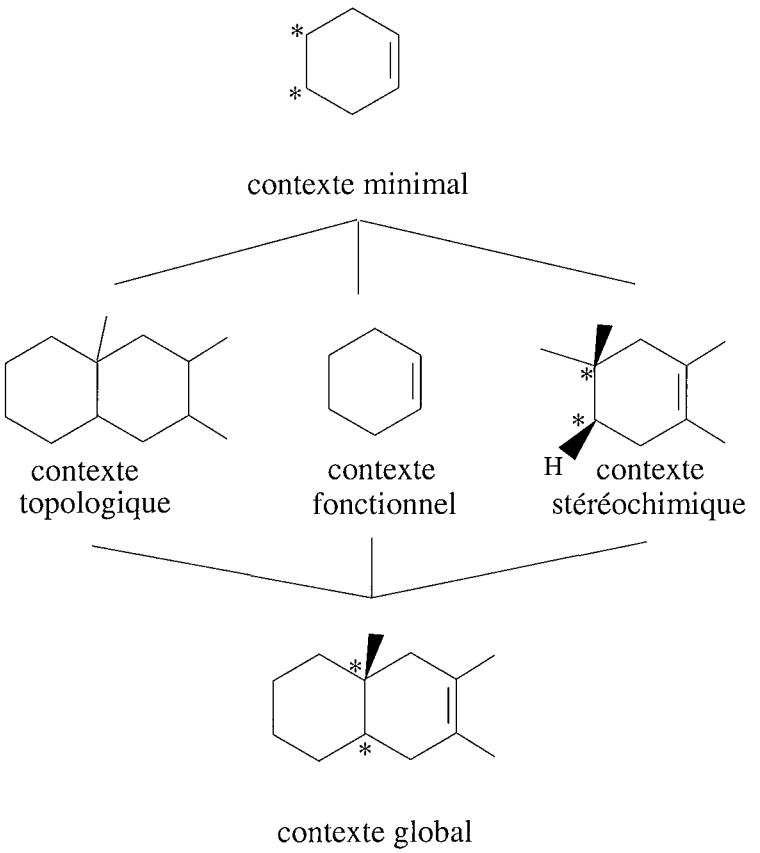


FIG. 7.31 – Les contextes de la transformation associée à un exemple d'utilisation de la méthode de Diels-Alder.

appelons le *transformateur*. Le transformateur associé à une transformation décrit les modifications à effectuer au niveau du graphe moléculaire pour obtenir les graphes moléculaires correspondant aux précurseurs. Nous stockons toutes les informations sur les opérations à effectuer dans un graphe. Le transformateur est construit à partir du cœur de la transformation. Nous pouvons remarquer à ce niveau que si la méthode de synthèse est saisie de façon stœchiométrique, le transformateur est identique au cœur. Si ce n'est pas le cas, on ajoute en plus les sommets ainsi que les arêtes uniquement présents dans les graphes moléculaires des réactants. Au niveau du transformateur, la difficulté est de savoir correctement exprimer les modifications respectant l'aspect stéréochimique de la méthode de synthèse. En effet, on ne peut se servir directement des modifications de liaisons dans ce cas là. À ce niveau, ce seront donc des descripteurs tels que ceux présentés 7.1.2.3 qui devront être utilisés. Dans le système SECS, Wipke et Dyott [Wipke and Dyott, 1974] proposent une génération des précurseurs basés sur l'utilisation de tels descripteurs. Dietz et collaborateurs [Dietz et al., 1997] proposent une méthode originale de représentation informatique des méthodes de synthèse stéréosélectives.

L'application d'une transformation pose également le problème du dessin des graphes moléculaires. Ce problème est compliqué et est loin d'être résolu.

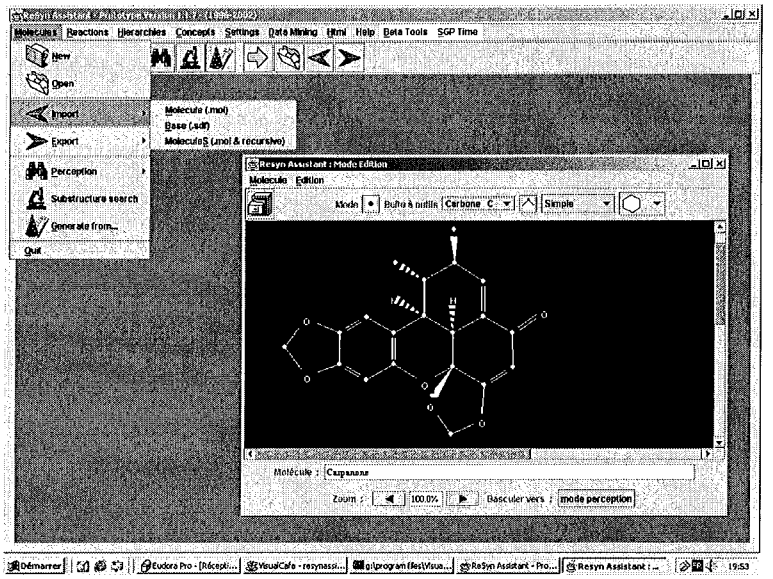


FIG. 7.32 – L'éditeur de molécules de ReSyn Assistant

7.4 Quelques mots sur l'implémentation de ReSyn Assistant

7.4.1 Les fonctionnalités sur les molécules

L'éditeur de molécule nous permet de saisir des molécules ou d'ouvrir des fichiers de divers formats. En effet, nous avons développé un format qui nous est propre pour la sauvegarde des informations sur une molécule mais nous sommes également capables d'importer des données provenant des bases de données. Nous avons, à l'inverse, un module d'exportation de nos données au format Mdl (voir annexe B).

Avant la perception d'une molécule, le dessin saisi est l'objet de plusieurs vérifications. La connexité du graphe, le nombre de voisins des atomes en tenant compte de leur charge ainsi que le respect de certaines conventions de dessin sont testés. Ainsi, nous ne percevons pas une molécule :

- non connexe,
- dont le nombre de voisins par atomes ne correspond pas à l'une des valences possibles pour chacun des atomes,
- dont deux liaisons consécutives (au sens du voisinage dans l'espace comme sur la figure 7.33) hors du plan sont exactement du même type (deux liaisons simples avant ou deux liaisons simple arrière).

L'application d'une transformation à une molécule cible est possible. Après vérification de l'applicabilité de la transformation sélectionnée selon le contexte minimal, les différentes possibilités d'application sont proposées à l'utilisateur. Pour l'instant, la génération des précuseurs ne prend pas en compte l'aspect stéréochimique mais cet aspect du transformateur est à l'étude.

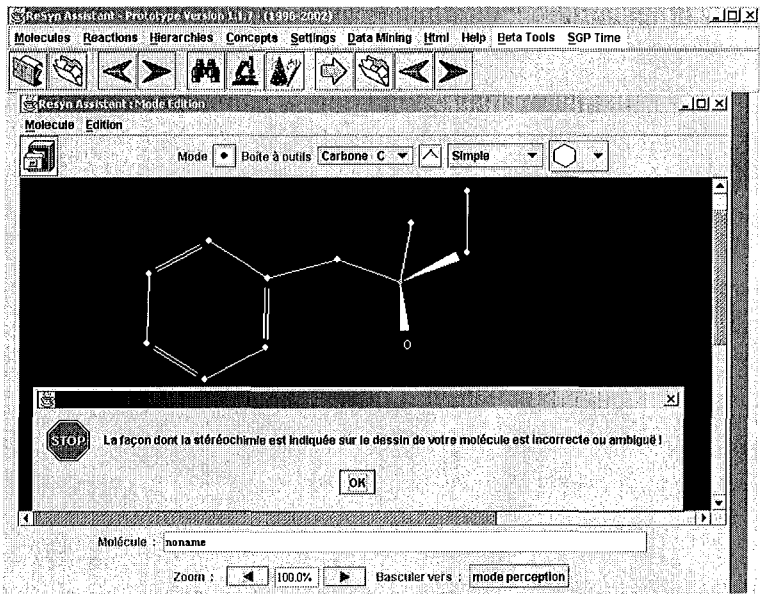


FIG. 7.33 – Un test sur le dessin dans l'éditeur de molécules de RESYN ASSISTANT.

Pour ce qui est de la recherche sous-structurale, un mode permet de vérifier quels sont les appariements topologiques qui sont corrects du point de vue de la stéréochimie.

7.4.2 Les fonctionnalités sur les méthodes de synthèse

Lorsque ce travail de thèse a débuté, le prototype RESYN ASSISTANT possédait de très nombreuses fonctionnalités de manipulation des graphes moléculaires dont il a été question dans la première partie de ce chapitre. En effet, le projet à l'origine de la conception de RESYN ASSISTANT était d'élaborer un outil d'aide à la compréhension des problèmes de synthèse organique grâce à la perception des molécules. Le traitement des méthodes de synthèse n'a été envisagé et réalisé qu'à partir de la fin de l'année 1999. Étant donné que notre but était d'exploiter les données présentes dans les bases de données de réactions, une des premières préoccupations fût de munir RESYN ASSISTANT de fonctions d'importation des données disponibles dans le format propre à RESYN ASSISTANT. Le traitement des données issues des bases de données de réactions nous a confronté à la difficulté du traitement de données réelles. Pour permettre la manipulation et la visualisation des méthodes de synthèse, un éditeur de réactions a été développé initialement par Pierre Jambaud. Cette interface permettait de saisir graphiquement les structures moléculaires impliquées dans la réaction, d'indiquer l'appariement des atomes et un certain nombre de données textuelles. Cet éditeur a ensuite été complété et amélioré pour tenir compte des avancées des travaux effectués au cours de cette thèse. Divers modules d'analyse des réactions ont enrichi l'éditeur : détermination et visualisation des cœur et contextes, détermination et visualisation de la correspondance des blocs et des systèmes de blocs, analyse des méthodes en termes d'objectifs fonctionnels et topologiques atteints.

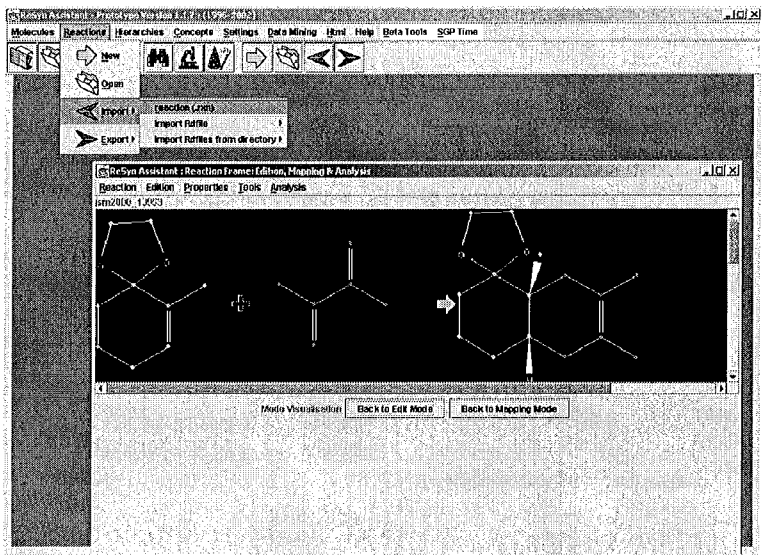


FIG. 7.34 – L'éditeur de réactions de ReSyn Assistant.

Dans l'éditeur de réactions (figure 7.34), on a la possibilité de :
- saisir une nouvelle réaction avec l'item `New` du menu `Reaction`,
- ouvrir une réaction au format ReSyn Assistant (fichier .rct) avec l'item `Open` du menu `Reaction`,
- importer des données au format Mdl dans le format ReSyn Assistant avec l'item `Import` du menu `Reaction`; la figure 7.34 montre les diverses possibilités, c'est-à-dire importer une seule réaction (fichier au format `.rxn` ou importer plusieurs réactions contenues dans un fichier au format `.rdf` ou encore importer plusieurs fichiers au format `.rdf` par parcours récursif d'un répertoire les contenant).
- exporter des données au format ReSyn Assistant dans le format `.rxn`; il est possible d'exporter un seul fichier ou un répertoire de fichiers au format `.rxn`.

Lorsqu'une réaction est saisie ou ouverte dans l'éditeur de réactions et que l'on dispose de l'appariement des atomes, la réaction peut être analysée. Les résultats de l'analyse des réactions peuvent être visualisés en sélectionnant le type d'analyse désiré (figure 7.35). Pour ce qui concerne la visualisation de la correspondance des blocs, nous avons répété le schéma de la méthode auquel nous avons adjoint de chaque côté les listes des blocs perçus dans les réactants (listes à gauche) et dans les produits (listes de droite). De chaque côté, on a une liste des blocs transformés pour chacun des points de vue, dans la quatrième liste se trouvent les blocs inchangés tous points de vue confondus. En cliquant sur les items des listes, le bloc sélectionné ainsi que les blocs lui correspondant sont mis en évidence dans le schéma de la réaction (par une couleur différente de celle prise par défaut).

Le prototype ReSyn Assistant est téléchargeable à partir du site `http://www.lirmm.fr/~resyn`. L'inconvénient est qu'il n'existe pas de manuel d'utilisation du logiciel pour l'instant, ce à quoi nous comptons remédier dans les mois qui suivent. Une description plus précise de certaines fonctionnalités de ReSyn Assistant pourra être trouvée dans

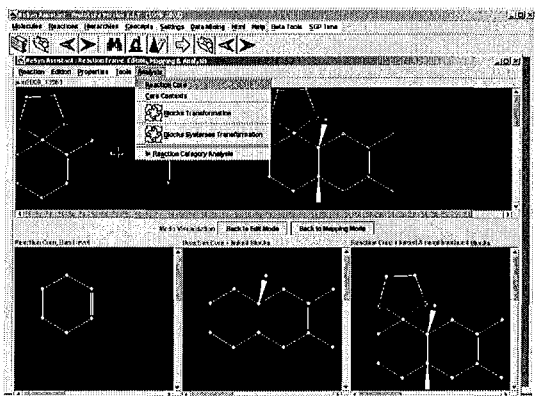


FIG. 7.35 – L'éditeur de réactions de RESYN ASSISTANT : analyse selon contextes minimal et global.

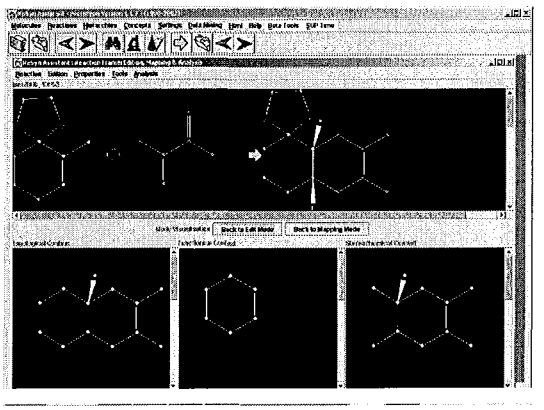


FIG. 7.36 – L'éditeur de réactions de RESYN ASSISTANT : analyse selon les contextes topologique, fonctionnel et stéréochique.

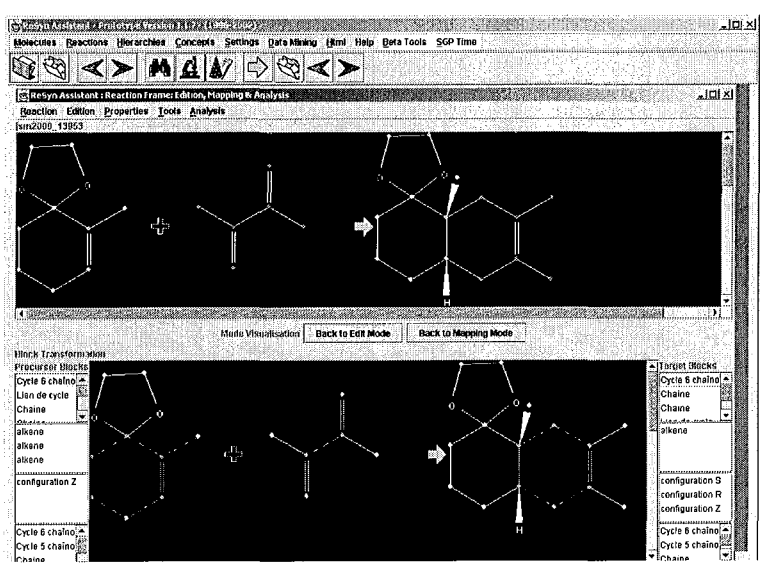


FIG. 7.37 – L'éditeur de réactions de RESYN ASSISTANT : analyse selon la correspondance des blocs.

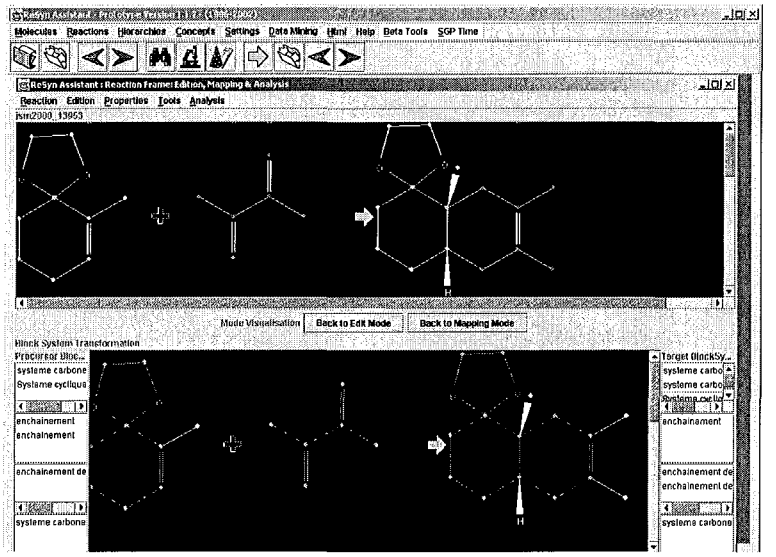


FIG. 7.38 – L'éditeur de réactions de ReSyn Assistant : analyse selon la correspondance des systèmes de blocs.

le thèse de Yannic Tognetti [Tognetti, 2002].

Chapitre 8

Extraction de connaissances à partir de bases de données de réactions

Sommaire

8.1 Introduction

Dans ce chapitre, nous allons montrer quels types de connaissances nous avons voulu extraire des bases de données de réactions ainsi que comment nous avons procédé pour remplir nos objectifs. Cette expérience n'est pas la première en son genre. D'autres équipes de recherche se sont intéressées à l'application de techniques de fouille de données dans les bases de données de réactions. Ces travaux ont pour but de faire émerger des généralisations potentiellement pertinentes et utiles d'un point de vue chimique. Ces objectifs, une fois atteints, permettent de dépasser la simple consultation de réactions individuelles qui est l'utilisation habituelle des bases de données de réactions.

Dans le cadre du GDR 1093 TICCO, l'utilisation des techniques d'apprentissage automatique a été effectuée à propos de la détermination des liaisons stratégiques dans les molécules [Régin, 1995]. Bien que concluante, cette expérimentation a mis en évidence le

fait que le caractère stratégique d'une liaison est fortement lié à son environnement et que l'important est de mettre en évidence non pas les liaisons stratégiques mais plutôt les objectifs stratégiques décelables dans une molécule.

Dans ce qui suit nous allons présenter l'utilisation des méthodes de fouille de données appliquées à l'analyse des méthodes de synthèse selon leurs objectifs stratégiques. Nous nous sommes intéressés à des méthodes de construction du squelette, d'une part, et d'aménagement fonctionnel, d'autre part. Ainsi, nous présentons ici les résultats concernant les méthodes de construction de cycle ainsi que celles conduisant à des interchanges fonctionnels. Les connaissances extraites le sont sous la forme de règles heuristiques donnant les tendances des méthodes employées en synthèse chimique. Parmi, toutes les techniques de fouille de données disponibles, nous avons utilisé celles d'extraction de motifs fréquents et de règles d'association grâce à la mise à disposition de programmes implémentés par Y. Bastide et dont nous avons parlé en section 5.3.2.

Dans les deux cas, l'exploitation directe des données structurales issues de l'analyse des réactions par la correspondance des blocs est impossible car les algorithmes de fouille, que nous avons utilisés, ne traitent pas les données de type structural. Nous avons donc procédé à une transformation des données. Bien entendu, cette opération implique une perte d'information.

Pour nos expérimentations, nous utilisons les bases de données accessibles via le système de gestion ISIS dont nous avons parlé au chapitre 3. Ces bases de données de réactions sont des bases orientées vers la méthodologie de synthèse (la base REACCS-JSM est, par exemple, la version électronique du Journal of Synthetic Methods). Aussi, nous considérons les données présentes dans ces bases comme celles sur des méthodes de synthèse potentielles et nous les étudions en tant que telles.

8.2 Les méthodes de synthèse vues selon la fonctionnalité

La fonctionnalité est à la base de la réactivité. Beaucoup de réactions ont pour résultat une modification de la fonctionnalité. Le contrôle des parties fonctionnelles des molécules est un souci permanent dans l'activité de synthèse. De plus, l'intérêt de certaines méthodes de synthèse réside dans le changement de fonctionnalité qu'elles permettent. Comme nous l'avons vu dans le chapitre 6, au niveau des actions, nous distinguons trois grandes classes de méthodes d'aménagement fonctionnel : les méthodes conduisant à la création d'une fonction, les méthodes d'élimination d'une fonction et les interchanges fonctionnels permettant le passage d'une fonctionnalité à une autre fonctionnalité. C'est ce troisième type de méthodes que nous avons étudié. Ainsi, nous avons voulu, avec cette expérience, essayer de répondre à des requêtes du type :

`Q1` : Quelles sont les fonctions à partir desquelles peut être obtenue telle autre fonction ?

`Q2` : Quelles sont les méthodes 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 ?

Parallèlement à l'analyse des interchanges fonctionnels, nous avons mené une étude sur le contenu des bases de données de réactions du point de vue de la fonctionnalité, le but étant de répondre à des questions du type :

Q4 : quelles sont les fonctions les plus présentes dans les bases de données de réactions ?

Q5 : quand elles sont présentes, sont-elles souvent détruites, créées et/ou inchangées ?

8.2.1 Traitements des données

8.2.1.1 Sélection

Nous avons entrepris plusieurs études. Nous présentons ici nos résultats les plus récents sur les bases de données ORGSYN version 2000 et REACCS-JSM version 2002 diffusées par MDL. Les expérimentations antérieures ont été présentées dans [Berasaluce et al., 2002], les résultats sont très similaires.

Nous avons sélectionné ces deux bases, parmi celles qui ont été mises à notre disposition, pour des raisons de qualité des données et de sélectivité des méthodes. En effet, ORGSYN [ORGSYN, url] est la version électronique de la série complète des Organic Syntheses et contient des informations sur des « schémas de synthèse » bien connus ayant généralement de bons rendements. Chacune des réactions a été vérifiée expérimentalement dans les laboratoires experts avant d'être incluse dans la base. La version de l'année 2000 de cette base contient 5486 fiches c'est-à-dire la totalité des méthodes de synthèse publiées par Organic Synthesis depuis 1921. L'étude de la base REACCS-JSM nous a permis d'augmenter par un facteur de plus de 10 la taille de l'échantillon d'analyse. En effet, REACCS-JSM couvre la littérature publiée de 1980 à aujourd'hui par Derwent, soit 75304 fiches de méthodes de synthèse. Dans REACCS-JSM [REACCS-JSM, url], les méthodes sont sélectionnées sur des critères de nouveauté ou parce qu'elles présentent un avantage particulier par rapport aux méthodes existantes.

Pour écarter un maximum de réactions dont l'appariement est sujet à caution, nous nous sommes limités à l'étude des réactions saisies en tant que « monoétape »[1]. Comme nous l'avons précisé dans le chapitre précédent, nous créons une réaction individuelle pour chaque produit formé pour isoler chacune des réactions concurrentes.

8.2.1.2 Pré-traitement

Le pré-traitement débute par une importation de données présentes dans les bases de données diffusées par la société MDL. En annexe B, nous avons reporté un extrait d'un

[1]Le sens de réaction monoétape est particulier dans les bases de données de réactions. Une réaction est monoétape si elle n'est pas décomposable en d'autres réactions présentes dans la base de données. Dans les bases de données de réactions gérées par le système ISIS, une réaction monoétape ne présente pas de valeur pour le champ NSTEPS au niveau de l'enregistrement correspondant à la réaction (cf. ANNEXE B).

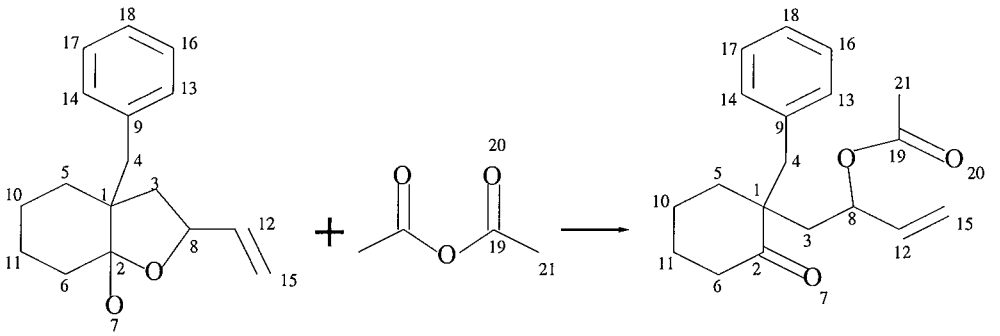


FIG. 8.1 – La réaction 13426 de la base REACCS-JSM version 2000 avec la correspondance entre atomes indiquée dans la base.

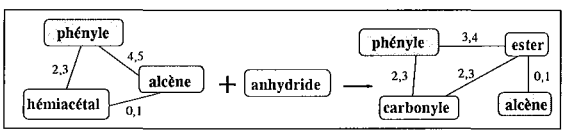


FIG. 8.2 – Analyse fonctionnelle de la réaction 13426 de la base REACCS-JSM version 2000.

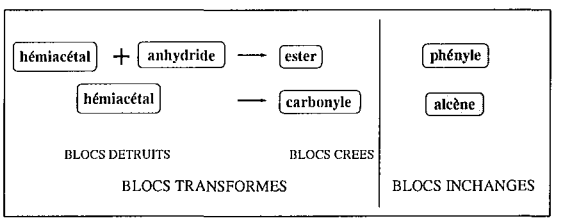


FIG. 8.3 – La correspondance des blocs résultant de la perception de la réaction 13426 de la base REACCS-JSM 2000 selon le point de vue fonctionnel.

fichier issu des bases de données et la structure de ce type de fichiers. Ensuite le pré-traitement consiste à faire l'analyse de chaque réaction en appliquant le principe de la correspondance des blocs fonctionnels que nous avons détaillée au chapitre 7. Considérons la réaction 13426 de la base REACCS-JSM version 2000 décrite figure 8.1. On peut remarquer que l'appariement des atomes donné dans la base de données est incomplet. La perception de la réaction au niveau méso et selon le point de vue de la fonctionnalité nous donne la représentation abstraite de la figure 8.2. Les blocs fonctionnels qui sont inchangés au cours de la réaction sont grisés. La figure 8.3 donne les précisions sur la correspondance des blocs déterminée. Au cours de cette réaction, on observe des changements de fonctionnalité tandis que certaines fonctions restent inchangées.

Au niveau du pré-traitement, nous avons dû prendre en compte de nombreux problèmes issus de la façon dont les données sont décrites dans les bases de données de réactions. Parmi ceux-ci, notons que :

- Une fiche sur une méthode de synthèse peut rassembler l'information sur des réactions concurrentes. Sur l'exemple de la figure 8.4, trois produits peuvent être obtenus par la combinaison des deux réactants dans les conditions données. L'obtention de ces trois produits correspond à trois transformations différentes, deux d'entre elles se distinguent l'une de l'autre uniquement du point de vue de la stéréochimie. La figure 8.5 montre les trois transformations créées à partir des données de la fiche sélectionnée.

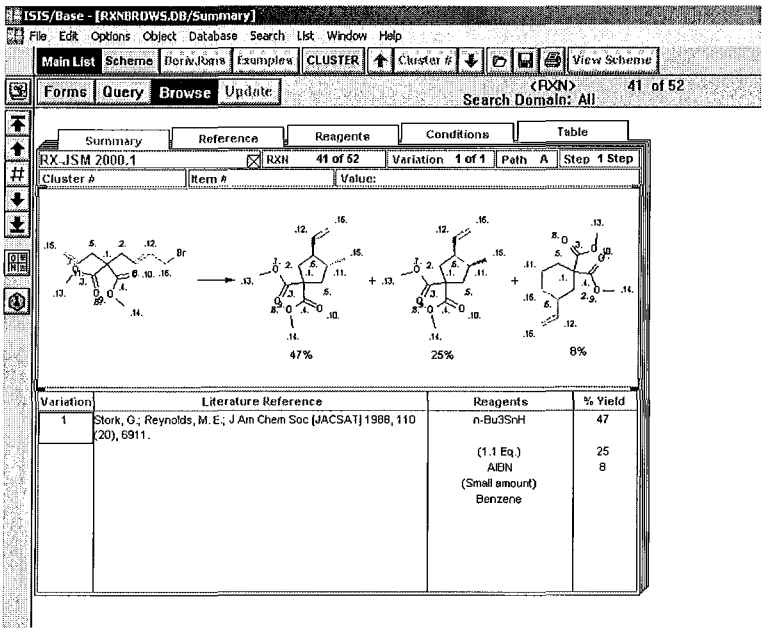


FIG. 8.4 – Réactions concurrentes dans une fiche de réaction issue de la base de données REACCS-JSM version 2000.

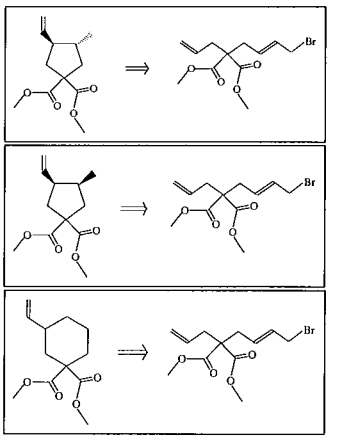


FIG. 8.5 – Les transformations créées pour chaque réaction concurrente.

- La figure 8.6 montre un exemple dans lequel les atomes d'un même réactant ont été appariés plusieurs fois dans le produit de la réaction. Contrairement à ce qui est fait dans les bases de données, nous avons décidé de prendre en compte explicitement la stœchiométrie par création d'autant de réactants que nécessaire Pour cela, nous créons autant d'équivalents d'un réactant que de nombre de fois où les atomes de ce réactant se retrouvent dans les atomes du produit (figure 8.7).
- Comme nous l'avons déjà souligné, il existe de nombreux problèmes d'appariement des atomes. La figure 8.8 est un exemple d'un appariement erroné. L'appariement des atomes correct est donné sur la figure 8.9. La méthode d'appariement implémentée dans RESYN ASSISTANT par Yannic Tognetti ne permet pas forcément d'apparier tous les atomes car cet appariement automatique est basé sur la détermination des sous-graphes maximums. Cependant, dans la meilleure solution proposée, tous les atomes qui sont appariés le sont correctement (figure 8.10). L'heuristique d'évaluation d'un appariement, proposée par Claude Laurenço, postule qu'au cours des réactions les atomes ne changent pas ou peu de degré de liaison sachant que, pour ce calcul, nous ne comptons pas les liaisons de l'atome avec les atomes d'hydrogène. Par exemple, sur la figure 8.8, l'atome numéroté 6 a un degré de 3 dans le réactant et un degré de 2 dans le produit. L'heuristique utilisée dans les bases est celle du changement minimal du type des liaisons. On remarque qu'au niveau de l'appariement exact et de celui proposé dans RESYN ASSISTANT, l'atome numéroté 6 a un degré de liaison de 3 dans le réactant ainsi que dans le produit. L'extension de l'algorithme à la recherche des sous-graphes maximaux est à envisager si l'on veut apparier un maximum d'atomes

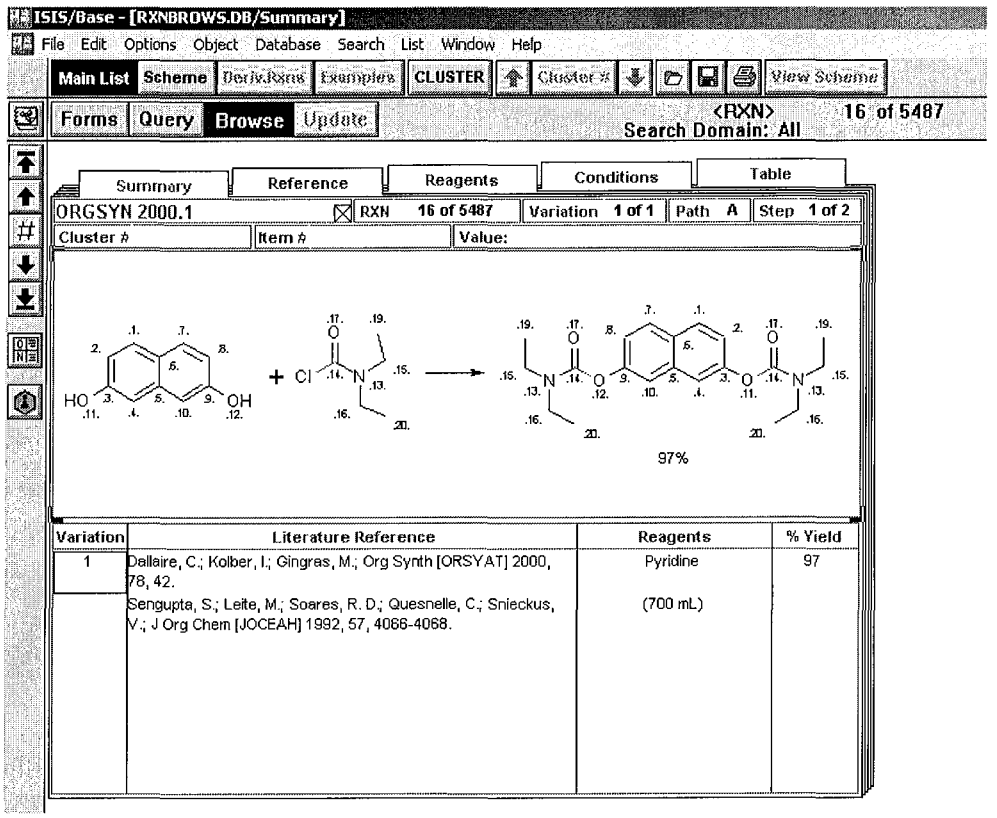


FIG. 8.6 – Non respect de la stœchiométrie dans une fiche de réaction issue de la base de données ORGSYN version 2000.

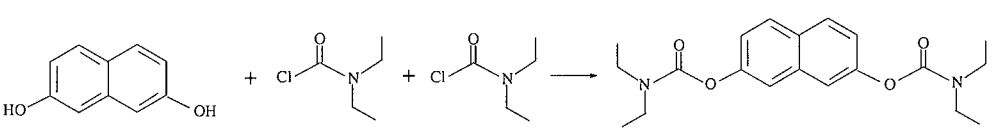


FIG. 8.7 – Rétablissement de la stœchiométrie.

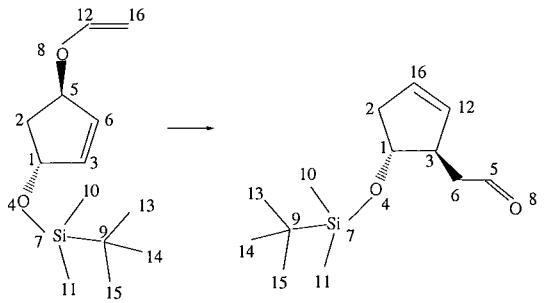


FIG. 8.8 – Un exemple d'appariement faux des atomes donné dans la base CIRX 97 diffusée par la société MDL sur un exemple d'utilisation de la méthode de Claisen.

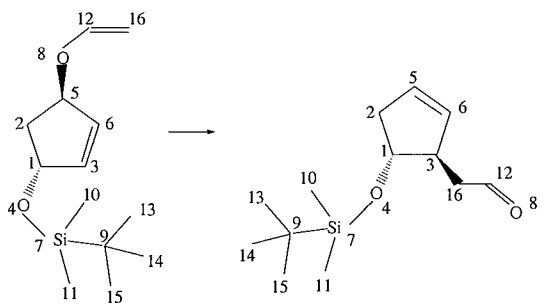


FIG. 8.9 – L'appariement effectif d'un exemple d'utilisation de la méthode de Claisen.

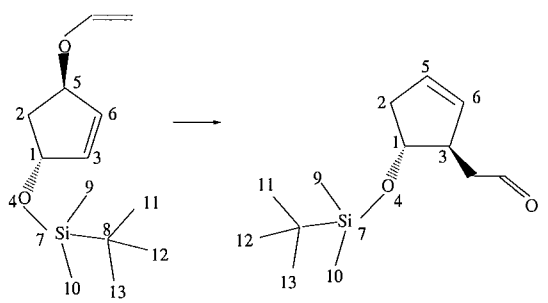


FIG. 8.10 – L'appariement automatique proposé dans RESYN ASSISTANT.

[Tognetti, 2002].
Les appariements faux ne sont pas décelables de façon automatique. Cependant, même si la détermination de l'appariement est un problème difficile, il serait très profitable d'étudier cet aspect des données et de développer d'autres heuristiques permettant de prendre en compte d'autres cas.

8.2.1.3 Transformation des données pour la fouille

Nous ne pouvons pas envisager de traiter directement les données structurales. Il est nécessaire de convertir ces données dans un format adapté à l'application des algorithmes de fouille. La solution que nous avons adoptée consiste à décrire la réaction à l'aide des noms des fonctions perçues et de leur statut au cours de la réaction (détruite, créée ou inchangée). Par ce passage de données structurales à des données symboliques, nous perdons (malheureusement) les informations de proximité entre les blocs.

Dans le cadre de la fouille de données, nous devons construire un contexte formel mettant en relation des objets et des propriétés. Les objets du contexte formel que nous considérons sont les réactions et les propriétés sont les fonctions détruites, créées ou inchangées. Deux transformations des données ont été envisagées :

- une première transformation par laquelle nous décrivons la réaction dans sa totalité

	blocs détruits	*blocs créés*	*blocs inchangés*
prise en compte implicite de la correspondance --→ 1 objet	hémiacétal ; anhydride	ester ; carbonyle	phényle ; alcène
prise en compte explicite de la correspondance --→ 2 objets	hémiacétal ; anhydride	ester	phényle ; alcène
	hémiacétal	carbonyle	phényle ; alcène

FIG. 8.11 – Les données sur la réaction 13426 de la base REACCS-JSM version 2000 transformées pour la fouille de données.

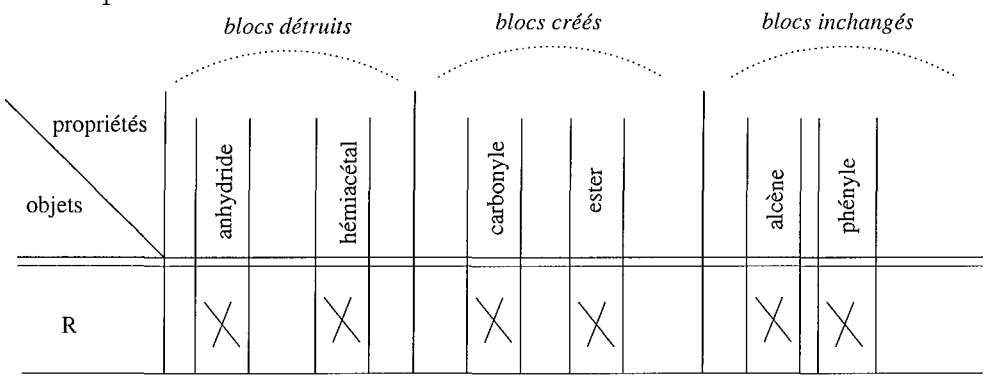


FIG. 8.12 – Tableau booléen construit après transformation sans prise en compte explicite de la correspondance.

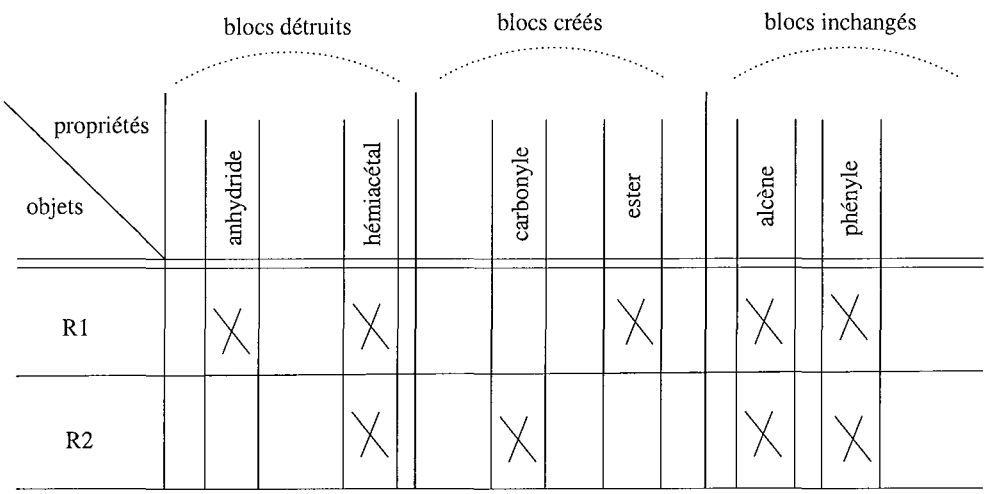


FIG. 8.13 – Tableau booléen construit après transformation avec prise en compte explicite de la correspondance.

en énumérant toutes les fonctions détruites, créées et inchangées,
- une deuxième transformation par laquelle nous créons pour une même réaction autant d'objets qu'il y a de fonctions créées au cours de la réaction. Chaque objet décrit alors la formation d'une fonction en détaillant les fonctions qui ont permis de la créer et en indiquant quelles fonctions restent inchangées.

La première transformation consiste à prendre en compte la correspondance des blocs de façon implicite car nous nous en servons tout de même pour reconnaître le statut des fonctions (détruite, créée ou inchangée). Dans la deuxième transformation, la correspondance est explicitement exprimée. Chacune de ces deux transformations rend compte de façon partielle de la réalité.

Si l'on reprend l'exemple de l'analyse de la réaction 13426 de la base REACCS-JSM version 2000, la première transformation crée un seul objet (figure 8.11). Nous ne pouvons pas déduire de cette description quelle fonction détruite a participé à la création de l'ester ou du carbonyle. Nous ne savons pas que le carbonyle provient uniquement de la destruction de la fonction hémiacétal. Par contre, nous avons l'information que cette réaction a permis de créer conjointement une fonction ester et une fonction carbonyle. Pour ce qui est de la deuxième transformation (figure 8.11), deux objets sont créés, un par fonction créée. Cette fois nous avons le détail de la formation de chacune des fonctions créées. Par contre nous perdons l'information indiquant que les deux fonctions se forment simultanément. Chacune de ces deux transformations devra donc être utilisée dans un but précis. La deuxième permet de connaître les conditions de formation d'une fonction donnée alors que la première pourra nous donner des indications sur la chimiosélectivité des méthodes, c'est-à-dire le fait que les conditions de formation d'une fonction induisent ou non la formation d'autres fonctions.

Ces deux transformations ont donné lieu à la création de deux tableaux booléens (figures 8.12 et 8.13) indiquant pour chacun des objets considérés les propriétés possédées.

8.2.2 Résultats

8.2.2.1 Étude de la réactivité des fonctions dans les bases de données de réactions

Nous nous sommes intéressés à l'étude de la réactivité des fonctions dans les bases de données de réactions. Pour répondre aux questions du type `Q4` et `Q5` que nous nous posons, nous avons utilisé les résultats donnés par la première transformation effectuée sur toutes les réactions quel que soit leur type. En effet, la fréquence des motifs de taille 1 issus de cette transformation correspond au nombre de fois où une fonction, selon un statut (créé, détruit ou inchangé), est trouvée dans l'ensemble des réactions. Pour notre étude, plusieurs chiffres nous intéressent. Pour chaque fonction, nous voulons connaître le nombre de fois où la fonction est :

- présente dans les réactants,
- présente dans les produits,
- détruite,

- créée,
- inchangée.

À partir de ces chiffres, nous avons fait quelques calculs élémentaires permettant d'examiner 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,

alors, à partir des motifs de taille 1, on récupère[2] :

- $\sup(F_c)$: nombre de fois où F est créée,
- $\sup(F_d)$: nombre de fois où F est détruite,
- $\sup(F_i)$: nombre de fois où F est inchangée,
- n_{Freact} : nombre de fois où F est présente dans les réactants,
- n_{Fprod} : nombre de fois où F est présente dans les produits,
- n_{Ftot} : nombre de fois total où la fonction est présente dans les molécules participant à la réaction.

n_{Freact}, n_{Fprod}, n_{Ftot} sont respectivement égaux à :

- $n_{Freact} = \sup(F_d) + \sup(F_i)$
- $n_{Fprod} = \sup(F_c) + \sup(F_i)$
- $n_{Ftot} = n_{Freact} + n_{Fprod} = \sup(F_d) + \sup(F_c) + 2\times\sup(F_i)$

À partir de ces différentes données sur la fréquence des fonctions, nous avons défini des pourcentages utiles pour répondre à nos interrogations sur :

1. la présence relative de la fonction dans les réactants et dans les produits : $p_r = n_{Freact}/n_{Ftot}$ et $p_p = n_{Fprod}/n_{Ftot}$,
2. la réactivité de la fonction : $p_{dr} = \sup(F_d)/n_{Freact}$ et $p_{ir} = \sup(F_i)/n_{Freact}$,
3. la facilité de création de la fonction : $p_{cp} = \sup(F_c)/n_{Fprod}$ et $p_{ip} = \sup(F_i)/n_{Fprod}$.

Quelques résultats

Nous travaillons à partir d'une base de connaissances de 441 fonctions. Sur ces 441 fonctions nommées, 170 sont détectées dans la base ORGSYN version 2000 tandis que 297 sont rencontrées dans REACCS-JSM version 2002. La différence s'explique par le fait que REACCS-JSM comporte plus de 10 fois plus de réactions qu'ORGSYN. Il y a donc dans notre base de connaissances des fonctions un ensemble de fonctions qui ne sont jamais rencontrées dans les bases de données de réactions étudiées. Nous reviendrons sur ce sujet dans la section suivante.

L'examen des pourcentages que l'on a définis ci-dessus nous permet de dégager deux règles sur les :

- *fonctions très stables* : les fonctions très stables apparaissent relativement fréquemment dans les bases et sont la plupart du temps inchangées quand elles sont présentes

[2]les supports sont exprimés en terme de nombre de fois où les motifs sont rencontrés dans le contexte formel.

dans une réaction. Cela se traduit par des valeurs élevées (aux environs de 90%) des pourcentages p_{ir} et p_{ip}, et donc, en contrepartie des valeurs faibles des pourcentages p_{dr} et p_{cp}. On remarque, par exemple, que la fonction phényle se retrouve dans énormément de molécules. Dans une étude sur CASREACT [Blower et al., 1997], les auteurs ont également mis en évidence que cette fonction est présente dans les trois quarts des molécules produites dans la base de données. Notre analyse nous permet d'ajouter que cette fonction reste très souvent inchangée que ce soit dans ORGSYN (p_{ir} = 86,5% et p_{ip} = 90,4%) ou dans REACCS-JSM (p_{ir} = 90,1% et p_{ip} = 93,5%). Cette observation correspond bien à la propriété de stabilité des fonctions aromatiques et est donc en accord avec les connaissances du domaine.

- *fonctions très réactives* : à l'inverse des fonctions très stables, les fonctions très réactives sont transformées pratiquement à chaque fois qu'elles sont présentes dans les réactants ou dans le produit. Contrairement à la règle précédente, les pourcentages p_{dr} et p_{cp} sont élevés (plus de 90%) tandis que les p_{ir} et p_{ip} sont faibles pour de telles fonctions. Il peut paraître paradoxal que ces fonctions soient majoritairement créées lorsqu'elles sont présentes dans les produits. L'explication que nous avançons est que, comme ces fonctions sont très réactives, elles ne se trouvent pas dans les produits de départ et il est donc nécessaire de les créer lorsque l'on en a besoin pour une étape ultérieure. D'après nos critères nous retrouvons des fonctions connues pour leur forte réactivité. Par exemple, on peut citer le chlorure d'acyle (p_{dr} = 99% et p_{cp} = 80,4% dans REACCS-JSM) et l'anhydride qui sont souvent employés pour améliorer le rendement des réactions dans lesquelles ils peuvent être utilisés. De même, entre autres, les fonctions iodure d'alkyle et thiol se distinguent.

8.2.2.2 Proposition pour un enrichissement de la base de connaissances sur les fonctions

Jusqu'à présent la base de connaissances des fonctions avait été constituée par des experts de la synthèse. À ce propos, nous pouvons souligner le travail de Sylvie Mouné et de Gilles Niel qui, en tant qu'experts de la synthèse organique, ont constitué cette base de connaissances. La base de fonctions nommées, que nous appelons aussi *fonctions connues*, comprenait, avant nos dernières expérimentations, la définition structurale de 441 fonctions. Étant donné que la base de connaissances n'est pas exhaustive, certains blocs fonctionnels détectés de façon algorithmique ne se voient pas attribuer de nom vu qu'ils ne correspondent structuralement à aucune des fonctions présentes dans la base de connaissances. Pour aider à recenser ces nouvelles fonctions, nous nous sommes proposés de conserver la trace des fonctions « inconnues » perçues dans les bases de données de réactions. Il n'est pas facile de nommer automatiquement une nouvelle fonction par un terme de la nomenclature chimique, il est par contre facile de lui attribuer un numéro permettant de la distinguer. Le principe de notre approche consiste à compter le nombre de fois où chaque fonction inconnue est perçue dans les molécules saisies dans les bases de données de réactions étudiées. Ce nombre donnera l'importance relative des fonctions encore inconnues. Il est certainement

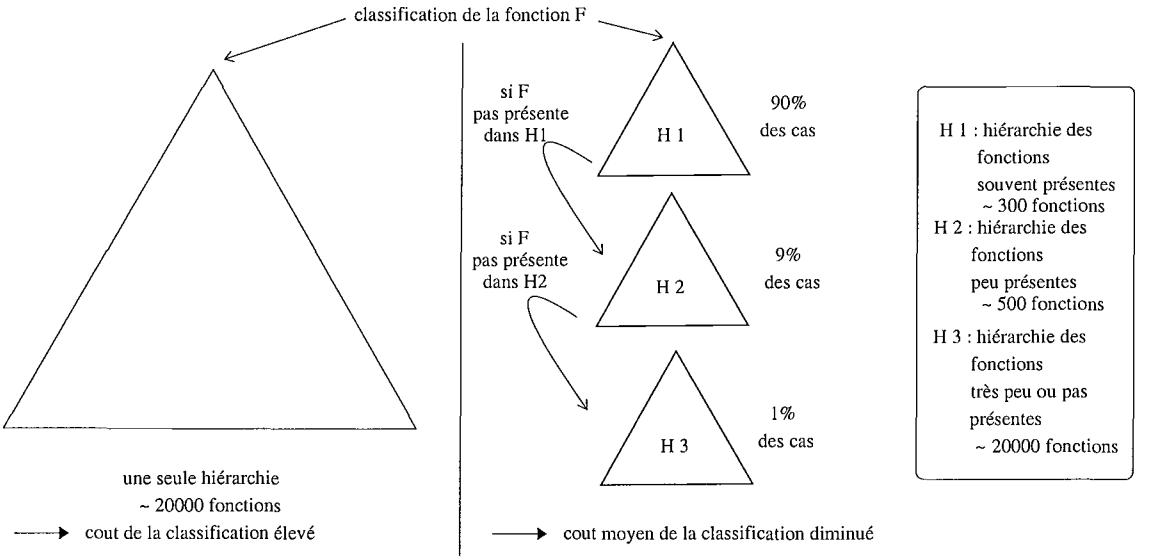


FIG. 8.14 – Une solution au problème du coût de la classification : la décomposition en plusieurs hiérarchies successives.

judicieux de rajouter à la base de connaissances celles qui sont les plus présentes. De façon analogue, si l'on compte les fréquences d'apparition des fonctions contenues dans la base de connaissances d'origine, cette étude montre que certaines fonctions sont soit présentes dans les bases de données de réactions de façon très anecdotique, soit en sont complètement absentes comme nous l'avons évoqué dans le paragraphe précédent.

Le recensement des fonctions inconnues dans la base ORGSYN nous a conduit à la construction d'une hiérarchie de 1271 éléments. Il y a moins d'une cinquantaine de fonctions encore inconnues qui sont présentes au moins 20 fois. Parmi celles-ci, nous retrouvons le ligand carbonyle, des ions tels que l'ion carbonate, le phenyl diazonium, de nombreux dérivés du phényl tels le diakylphtalate, le 2-naphtol, aniline tertiaire. Dans le tableau en annexe C, nous avons reporté les 10 fonctions inconnues trouvées le plus fréquemment dans la base ORGSYN. La base REACCS-JSM comporte un très grand nombre de fonctions inconnues (plus de 18000). Nous n'avons pas encore eu le temps à ce jour d'examiner les structures fonctionnelles encore inconnues de notre base de connaissances mais ceci est une de nos perspectives. Une remarque que l'on peut déjà faire est que parmi les fonctions encore inconnues se trouvent énormément de fonctions dérivées du phényle ou d'autres fonctions aromatiques.

À terme, il est nécessaire de prendre en compte toutes les fonctions que nous détectons dans les molécules participant aux réactions. Cependant la hiérarchie deviendra alors très grande : la prise en compte des fonctions inconnues d'ORGSYN et de REACCS-JSM nous conduit déjà à une hiérarchie de fonctions d'environ 20000 graphes. Dans le but de diminuer le coût de l'opération de classification, il serait judicieux de créer au moins deux bases de connaissances sur les fonctions, l'une contenant les fonctions les plus fréquemment rencontrées dans les molécules, l'autre contenant les fonctions le moins fréquemment ren-

contrées. Le processus de classification commencerait par la première hiérarchie, celle dans laquelle on a a priori le plus de chances de classer le bloc fonctionnel. Si celui-ci n'est pas présent dans cette première hiérarchie, on passe alors à la classification dans la deuxième hiérarchie. Ce découpage peut paraître grossier et il est envisageable, bien entendu, d'en définir de plus fin, c'est-à-dire en n hiérarchies successives, des fonctions trouvées le plus fréquemment à celles qui ont été peu souvent (voire pas encore) rencontrées. Sur la figure 8.14, nous proposons une décomposition en trois hiérarchies. Dans 90% des cas une fonction F sera classée dans la hiérarchie H1 qui ne comporte que 300 graphes. Si ce n'est pas le cas, on essaiera de classer F dans H2 puis dans H3 si l'on est dans les 10% de cas où F n'est pas présente dans H2.

Dans ce qui suit nous avons travaillé sur une base de connaissances de fonctions enrichie des 43 fonctions encore inconnues qui apparaissent au moins 20 fois dans la base ORGSYN. Ces fonctions ont été nommées par Gilles Niel. La base de connaissances de fonctions sur laquelle nous travaillons comporte désormais 484 éléments (441 + 43).

8.2.2.3 Les connaissances que l'on peut extraire des motifs

L'exploitation des motifs fréquents qui nous intéresse le plus, dans un système d'aide à la synthèse, 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).

Le premier travail consiste à déterminer la forme générale des motifs correspondant à la requête que l'on désire poser au système. Une fois les motifs récupérés, nous interprétons les résultats obtenus.

Vu le type de question que l'on peut se poser dans ce cadre, il faut regarder les motifs obtenus par exploitation des bases de données dans lesquelles on tient compte explicitement de la correspondance des blocs.

Contenu des bases : les interchanges fonctionnels les plus observés

Lorsque l'on tient compte de la correspondance entre les blocs, la recherche de motifs fréquents sur chacune des deux 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 sont celles qui ont pratiquement la plus grande fréquence d'apparition (environ 2%). Ce résultat montrant que les réactions d'estérification[3] et d'hydrolyse des esters[4] sont les interchanges fonctionnels les plus représentés dans les bases de données a déjà été remarqué dans une étude faite en 1996 sur CASREACT [Blower et al., 1997]. L'estérification ainsi que la réaction inverse d'hydrolyse des esters sont des réactions très employées.

Les interchanges fonctionnels qui apparaissent plus souvent que les deux méthodes précédemment citées correspondent à la conversion d'un alcène en un autre alcène (pour REACCS-JSM) et à la transformation d'un carbonyle en alcool ou en alcène, d'éther en

[3]L'estérification est la formation d'un ester à partir d'un alcool et d'un acide ou d'un dérivé d'acide.

[4]L'hydrolyse d'un ester correspond au passage d'un ester à l'alcool et l'acide correspondants.

fonction détruite	ORGSYN 2000	REACCS-JSM 2002
alcool	139	1030
acide carboxylique	95	660
anhydride	87	419
alcène	36	334
ester	32	651
chlorure d'acyle	28	175
ion carbonate	17	79
chloro carbonate	16	48
carbonyle	15	567
orthoester	13	63
vinyloxycarbonyle	10	140

TAB. 8.1 – Quelques unes des principales fonctions à partir desquelles on peut obtenir une fonction ester dans les bases ORGSYN version 2000 et REACCS-JSM version 2002. Le support est indiqué en nombre d'objets contenant le motif.

alcool et d'alcyne en alcool (pour REACCS-JSM). Ces différents interchanges correspondent à des changements de fonctionnalités très fréquemment observés au cours des réactions chimiques. Ils font intervenir des fonctions que l'on peut qualifier d'élémentaires et qui sont très souvent présentes dans les molécules. On n'est donc pas étonné d'observer de très nombreux interchanges entre ces « fonctions de base ».

Premier exemple de requête

Par exemple, 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 : $F_d \wedge ester_c$.

Dans notre cas, si l'on ne considère que les fonctions les plus fréquemment rencontrées, on a la liste de fonctions du tableau de la figure 8.1.

On remarque que les principales fonctions permettant d'obtenir un ester sont les mêmes dans les deux bases de données de réactions considérées. Par contre, elles ne sont pas ordonnées de la même façon selon leurs fréquences respectives. Il peut paraître étrange qu'une fonction ester réagisse pour donner une fonction de même nom mais cela est courant en chimie. Dans le cas que nous étudions (formation d'un ester), nous aurons une transestérification [March, 1992]. Une transestérification est une transformation d'un ester en un autre ester du même acide, c'est-à-dire que la majeure partie de la fonction ester se conserve et que seule une partie, que nous qualifions d'environnementale, change. La figure 8.15 donne le schéma général d'une transestérification. La méthode de transestérification est particulièrement présente dans la base REACCS-JSM.

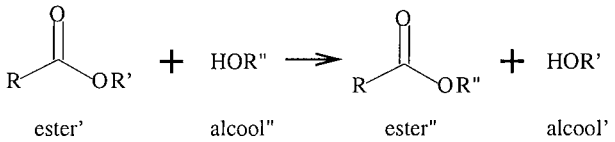


FIG. 8.15 – Schéma général d'une transestérification. La seule différence entre ester' et ester" est le groupe lié à l'oxygène. Un « groupe » désigne ici un groupement d'atomes liés entre eux par des liaisons. En général, le groupe est constitué d'atomes de carbone et d'hydrogène. En tous les cas, l'atome du groupe qui est lié à l'oxygène doit être un carbone ainsi que les atomes du groupe directement liés à celui-ci.

combinaison de fonctions détruites	ORGSYN 2000	REACCS-JSM 2002
alcool + anhydride	66	143
alcool + chlorure d'acyle	29	50
alcool + acide carboxylique	20	122
acide carboxylique + alcène	12	130
alcool + vinyloxycarbonyl	6	37
alcool + ester	4	140
carbonyle + ester	3	122
alcène + ion carbonate	3	10

TAB. 8.2 – Quelques unes des combinaisons de fonctions parmi les plus fréquentes permettant d'obtenir une fonction ester. Le support est indiqué en nombre d'objets contenant le motif.

Deuxième type de requête

Nous savons que les fonctions mises en évidence par la requête précédente sont dans la majeure partie des cas combinées deux à deux pour produire un ester. Nous voudrions connaître quelles sont les combinaisons de deux fonctions qui permettent d'obtenir une fonction ester. Dans ce cas, les motifs que nous recherchons sont de la forme : $F_{d1} \wedge F_{d2} \wedge ester_c$

Si l'on regarde les motifs montrant les combinaisons les plus fréquentes des fonctions permettant de construire un ester, on obtient le tableau 8.2.

Tout d'abord, on peut voir que l'on retrouve les mêmes motifs pour la question donnée d'une base de données à l'autre. Par contre, si l'on ordonne les réponses pour chacune des deux bases en fonction des fréquences d'apparition respectives, on n'obtient pas la même disposition des motifs les uns par rapport aux autres.

On peut souligner que ces fonctions et ces combinaisons de fonctions sont effectivement connues comme modes de formation des esters.

Le motif qui implique la combinaison d'un acide carboxylique avec un alcène conduit la plupart du temps à la création d'esters tertiaires. Un ester tertiaire est un ester dont l'atome de carbone lié par une liaison simple à l'oxygène est obligatoirement également lié à trois autres atomes de carbone. Ceci ne correspond pas à notre définition d'un bloc

fonction inchangée	ORGSYN 2000	REACCS-JSM 2002
phényle	59	1633
alcène	40	576
ester	30	495
alcool	29	213
acétal	19	173
carbonyle	17	321
acide carboxylique	13	22
bromure d'alkyle	7	10
chlorure d'alkyle	6	44
éther	6	217

TAB. 8.3 – Quelques unes des fonctions inchangées alors qu'une fonction ester est créée dans les bases de données ORGSYN 2000 et REACCS-JSM 2002.

fonctionnel mais à la définition d'une famille chimique. Il serait intéressant de vérifier, si l'on appliquait le même processus de fouille aux familles, que dans la majeure partie des cas un composé de la famille des esters tertiaires est fait à partir d'un acide carboxylique et d'un alcène. Pour cela, il nous reste à parfaire la modélisation des familles au préalable.

Dans les deux bases, la fonction ester se combine avec un alcool pour former un ester. On retrouve bien là le schéma général de la méthode de transestérification supposé précédemment. Nous remarquons que cette méthode est très représentée dans la base REACCS-JSM.

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 connaît des exemples de méthodes remplissant les conditions désirées. En examinant les motifs fréquents, on peut déjà savoir s'il existe des réactions dans la base de données construisant la fonction voulue en laissant inchangées d'autres fonctions et, si oui, lesquelles. Dans le tableau 8.3, on présente une partie de la liste des fonctions inchangées alors qu'une fonction ester est créée et le nombre d'exemples connus dans les deux bases. Les motifs sont de la forme générale : $\text{ester}_c \wedge \text{éther}_i \wedge \text{F}_d$.

On remarque que la fonction phényle inchangée se retrouve dans un nombre relativement très important de réactions formant un ester. Nous avons déjà remarqué précédemment que cette fonction est très souvent présente dans les molécules et de plus est très stable. Ce type de motif a donc peu d'intérêt pour nous par rapport aux autres. D'une manière générale, les motifs et les règles d'association contenant l'item phényle inchangé ne nous intéresseront pas.

8.2.2.4 Les connaissances que l'on peut extraire des règles

Les règles selon le type de fonctions (créées, détruites et/ou inchangées) qu'elles mettent en jeu permettent d'avoir des informations quantitatives sur les réactions les unes par

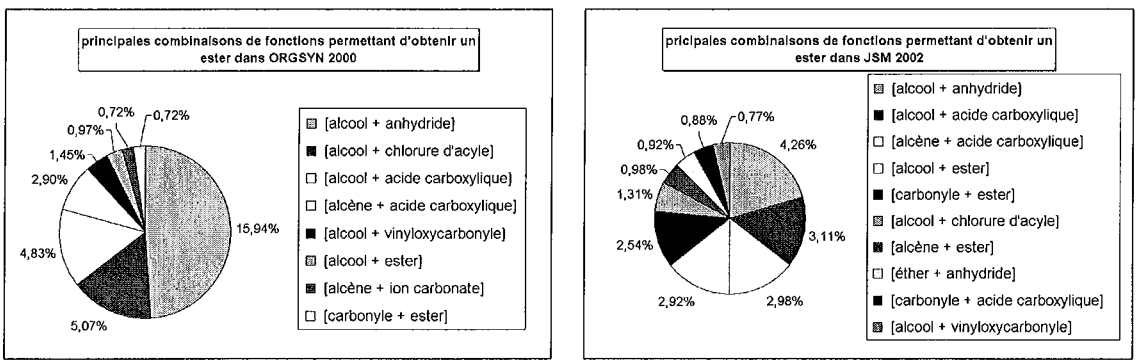


FIG. 8.16 – Les combinaisons de fonctions permettant d'obtenir un ester dans ORGSYN 2000 etREACCS-JSM 2002.

rapport aux autres.

Premier exemple de requête

De la même manière que précédemment, supposons que nous voulions obtenir une molécule appartenant à la classe des esters et que nous désirons savoir quelle est la meilleure façon d'obtenir des molécules de cette catégorie. Il faut regarder les règles obtenues en tenant compte de la correspondance. On doit alors d'abord s'intéresser aux règles dont la prémisse contient ester_c et la conclusion est composée d'items associés à des fonctions détruites. Ce sont des règles de la forme : $\text{ester}_c \Rightarrow \{\text{F}_d\}$ avec $\{\text{F}_d\}$ ensemble de fonctions détruites. Nous rappelons que la confiance d'une telle règle sera de la forme pour un ensemble $\{\text{F}_d\}$ de deux fonctions :

$$conf = \frac{sup(ester_c \wedge F_{d1} \wedge F_{d2})}{sup(ester_c)}$$

En comparant les valeurs des confiances de chacune des règles, on peut envisager de classer les possibilités de combinaison de fonctions qui permettent d'obtenir une fonction ester. Les principales combinaisons de fonctions sont données dans les graphiques de la figure 8.16.

Deuxième exemple de requête

Bien souvent, on forme une fonction à partir de deux autres fonctions. Ayant sélectionné la première fonction réagissant F_{d1}, on veut connaître et classer la liste des fonctions F_{d2} avec lesquelles F_{d1} réagit pour donner F_c. On doit comparer les confiances des règles d'association de la forme : $\text{F}_{d1} \wedge \text{F}_c => \text{F}_{d2}$
La figure 8.17 donne les principales fonctions qui réagissent avec un alcool pour donner un ester. La somme des pourcentages n'est pas égale à 100% car il existe d'autres fonctions qui se combinent à un alcool pour former un ester mais nous nous sommes restreints ici aux principales combinaisons.

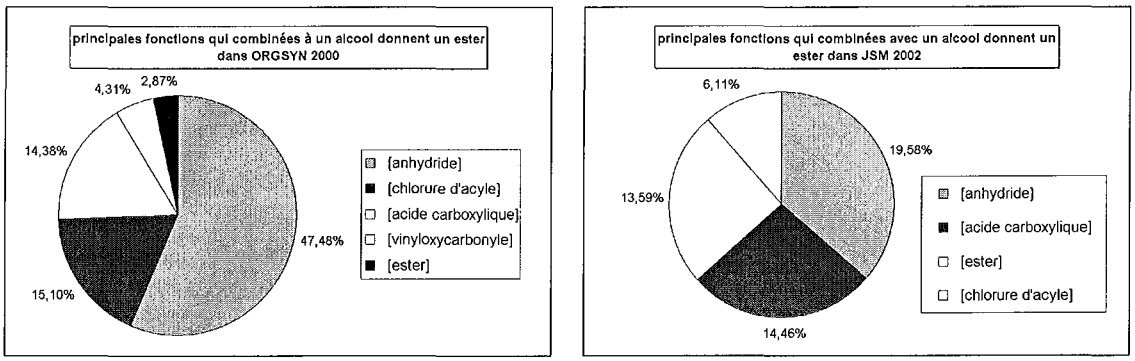


FIG. 8.17 – Les principales fonctions qui, avec un alcool, donnent un ester sur les bases ORGSYN 2000 et REACCS-JSM 2002.

On remarque dans cet exemple que, dans les deux bases, pour obtenir un ester, la méthode de synthèse la plus fréquemment employée est celle qui utilise un alcool et un anhydride. Par contre, on observe une différence pour les autres fonctions possibles. Dans ORGSYN, c'est un chlorure d'acyle qui est plus employé qu'un acide carboxylique. Or d'après ce que nous savons de la réactivité comparée de ces trois fonctions, un chlorure d'acyle est plus réactif qu'un anhydride, lui-même plus réactif qu'un acide carboxylique [March, 1992]. Cependant comme on peut le remarquer en examinant les bases de données de réactions, ce sont les anhydrides qui sont majoritairement employés ce que l'on peut retrouver dans [March, 1992]. Ceci s'explique par le fait que les réactivités relatives des fonctions ne sont pas les seuls paramètres pris en compte lors du choix des réactants à utiliser. En effet, des facteurs tels que la facilité d'utilisation, le coût ou les considérations environnementales et de sécurité influencent aussi la sélection. Dans REACCS-JSM, nous remarquons toujours l'importance de l'emploi de la méthode de transestérification tandis que dans ORGSYN cette méthode est peu représentée.

Troisième exemple de requête

Les fonctions étant des parties très réactives des molécules, il est aussi important de savoir quelles sont les fonctions qui ne réagissent pas alors que telle autre fonction est détruite. Si l'on veut savoir s'il existe des possibilités de faire F_c en laissant inchangée F_i et comparer les solutions connues (F_d), on pourra étudier les règles de la forme :

$F_c \wedge F_i \Rightarrow F_d$

Par exemple, si l'on veut connaître les exemples de méthodes qui laissent inchangée une fonction éther alors qu'un ester est formé, on regarde les règles ayant pour modèle général : $\text{ester}_c \wedge \text{éther}_i \Rightarrow F_d$. L'examen des résultats sur les bases de données ORGSYN 2000 et REACCS-JSM 2002 donne la liste de fonctions de la figure 8.18.

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 devons trouver des règles de la forme : $\text{ester}_c \wedge \text{éther}_i \Rightarrow \{F_d\}$ avec $\{F_d\}$ ensemble de fonctions

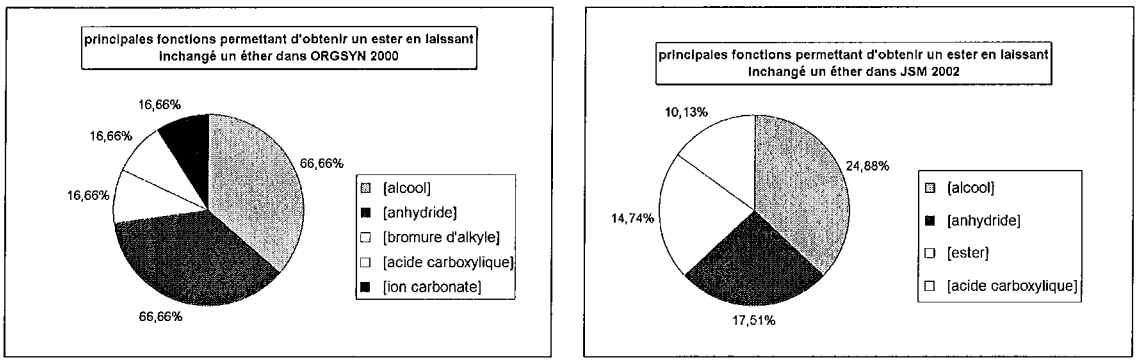


FIG. 8.18 – Quelques fonctions permettant d'obtenir un ester tout en laissant inchangé un éther dans ORGSYN 2000 et REACCS-JSM 2002.

détruites. Parmi celles-ci, se dégage la règle $\text{ester}_c \wedge \text{éther}_i \Rightarrow \text{alcool}_d \wedge \text{anhydride}_d$. Dans la base ORGSYN, cette règle a une confiance importante (67%) alors que dans REACCS-JSM, cette règle est beaucoup moins vérifiée (8,75%). Cette différence s'explique par le fait que dans ORGSYN nous avons un nombre limité de combinaisons de fonctions permettant d'obtenir un ester tandis que dans REACCS-JSM les possibilités sont plus nombreuses. En effet, REACCS-JSM est une base sélective c'est-à-dire que ses concepteurs ont décidé d'y représenter un maximum de méthodes de synthèse mais en décrivant chacune d'elles avec peu d'exemples.

De plus, si après examen de la figure 8.18, on décide d'employer un alcool et que l'on veut savoir quelle fonction l'on va faire réagir avec cet alcool, la règle $\text{ester}_c \wedge \text{éther}_i \wedge \text{alcool}_d \Rightarrow \text{anhydride}_d$ nous indique dans combien de cas un anhydride a été employé. Dans la base ORGSYN cette règle est vérifiée dans la totalité des cas (la confiance est égale à 100%) tandis que dans REACCS-JSM la règle est partielle car elle est vraie dans 35% des cas ce qui constitue une valeur tout de même élevée de confiance pour le type de données que nous explorons. En effet, les bases de données sélectionnées sont sélectives et ont pour priorité de représenter la diversité des méthodes possibles.

8.2.3 Discussion

Les résultats que nous avons obtenus nous permettent de dégager des tendances qui sont en accord avec nos connaissances chimiques. Tout au long du processus d'extraction de connaissances nous avons mis en évidence l'importance de l'analyste, qui étant donné ses connaissances et les buts poursuivis, est la seule personne à même de définir les transformations des données potentiellement utiles et d'interpréter les résultats issus de la fouille. Nous avons rencontré plusieurs problèmes au cours de cette application des techniques de fouille de données.

- Nous avons vu que les transformations envisagées induisent une perte d'information car nous ne pouvons exploiter directement les données de type structural. De plus, la prise en compte explicite de la correspondance duplique l'information sur les fonctions

restant inchangées tandis que la relation entre les objets créés n'est pas maintenue.
- En examinant les motifs, nous avons pu remarquer que les fréquences de ceux-ci étaient faibles par rapport à celles qui sont généralement observées dans les domaines d'application habituels de la fouille de données. C'est pour cela que nous n'avons pratiquement pas imposé de seuil minimum de support. De même, pour la confiance des règles. Il faut remarquer que, dans le domaine qui nous intéresse, savoir qu'il existe des motifs de faible fréquence correspondant à notre requête peut être potentiellement intéressant. En effet, un exemple unique de méthode permettant de passer d'un ensemble de molécules à un autre ensemble indique que la transformation est possible même si elle n'est pas fréquente. La faible fréquence des motifs observés est en grande partie dûe au fait que les bases de données sélectionnées sont de nature sélective : la collecte de données a donc été supervisée.
- Certaines fonctions sont très fortement représentées dans les bases de données de réactions. En général, les motifs et les règles impliquant ces fonctions ne nous intéressent pas ou peu.
- Considérer les règles dites informatives est intéressant dans certains cas particuliers. Par exemple, la recherche des combinaisons de fonctions permettant d'obtenir une fonction donnée utilise ce principe. Nous n'avons cependant pas recherché systématiquement ce type de règle. En effet, la traduction sous forme de modèle de règles des requêtes précises que nous nous posons conduit la plupart du temps à des règles ayant une prémisse plus importante en taille que la conclusion. Dans ces cas là, la prémisse exprime toutes les conditions posées.

8.2.4 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 [Ganter and Rudolph, 2001] où l'analyse de concepts formels est utilisée pour étudier la représentation et la classification 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 [Bastide, 2000]) et que sont extraites des règles d'association.

Notons également qu'il existe des travaux portant sur la recherche de sous-graphes fréquents appliqués, en particulier, à la découverte de relations entre structure chimique et activité biologique des molécules [Dehaspe et al., 1998] [Chittimoori et al., 1999] [Inokuchi et al., 2000] [Kuramochi and Karypis, 2002]. Les deux méthodes les plus récentes utilisent un codage canonique des graphes et procèdent à une recherche des sous-graphes communs par niveau, c'est-à-dire à dire en commençant par les sous-graphes de taille 1 (en terme de nombre d'atomes) puis en passant à ceux de taille 2 et ainsi de suite.

8.3 Analyse des méthodes de synthèse de construction de cycles

Au niveau méso et du point de vue topologique, on peut distinguer trois types de méthodes de construction du squelette moléculaire : les méthodes de construction de motifs cycliques, les méthodes de construction de chaînes carbonées et les méthodes de construction de liens. Un grand nombre de molécules comporte des cycles notamment lorsqu'elles sont d'origine naturelle [Kocovský et al., 1990]. La connaissance des méthodes de construction présente un grand intérêt du point de vue de la rétrosynthèse. Nous avons évoqué dans le chapitre 6, notre modélisation des réactions de construction de cycles. C'est à partir de cette description que nous avons extrait des connaissances sur les méthodes de synthèse permettant de construire un monocycle carboné. Ce travail de modélisation a été fait en collaboration avec Gilles Niel, qui s'est intéressé aux méthodes de construction des cycles de taille 7 au cours de ses recherches précédentes. Les résultats que nous exposons ici ont fait l'objet de communications lors de plusieurs conférences internationales [Laurenço et al., 2002], [Niel et al., 2002a], [Niel et al., 2002b].

8.3.1 Traitements des données

Avant de préciser comment nous avons sélectionné, pré-traité et transformé les données à partir desquelles nous avons travaillé, nous allons expliciter le modèle des données adopté. L'étude des méthodes du point de vue topologique et notamment des méthodes de construction de cycles nous a conduit à définir une interprétation de la correspondance des blocs particulière. Celle-ci nous permet de décrire l'objectif structural d'une méthode de construction de monocycles.

8.3.1.1 Définition de l'objectif de synthèse d'une méthode de construction de monocycle

Ainsi que nous l'avons défini, l'objectif stratégique atteint par une méthode correspond à une action sur une structure. Lorsque l'on considère les méthodes de construction de monocycles, on doit donc être capable de décrire

- les structures construites, c'est-à-dire les monocycles,
- les actions, c'est-à-dire les types de construction de monocycles.

Description des monocycles

Plusieurs paramètres permettent de décrire un monocycle : sa taille, sa nature exclusivement carbonée ou non, le nombre et la position relative des substituants sur le monocycle. On appelle *substituant* d'un cycle un atome autre qu'un atome d'hydrogène lié directement à l'un des atomes cycliques. Un monocycle peut être 0-substitué, 1-substitué, 2-substitué, etc., selon le nombre de substituants portés par ses atomes. Il existe une limite au nombre de substituants d'un cycle : celle-ci est donnée par le nombre restreint de liaisons que les

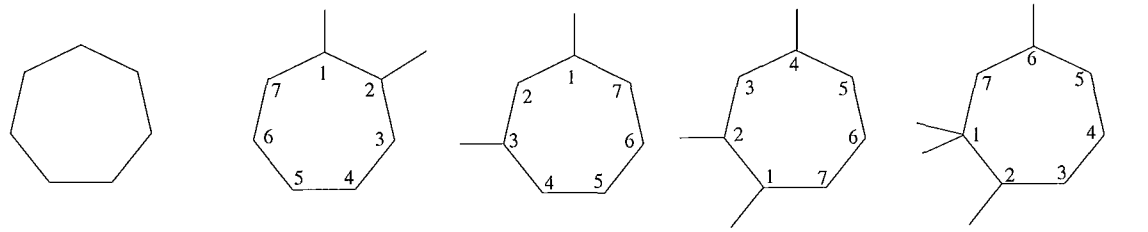


FIG. 8.19 – La description des monocycles. Les structures représentées ne tiennent pas compte du type des atomes, ni du type des liaisons.

atomes peuvent faire. Par exemple, un atome de carbone cyclique pourra au plus faire deux liaisons supplémentaires. Un monocycle carboné de taille 7 pourra donc au plus être 14-substitué. Outre le nombre de substituants, les positions relatives des substituants sur le monocycle représentent une information intéressante. Ainsi deux substituants peuvent être liés au même atome du cycle, ou bien à deux atomes voisins du cycle, ou encore à des atomes du cycle séparés d'une distance `n` en terme de liaisons, `n` étant compris entre 1 et la partie entière de (`taille du cycle / 2`). Dans le premier cas, on dira que les deux substituants sont en position `1,1`, dans le deuxième cas en position `1,2` et dans le troisième cas en position `1,n+1`. La description se complique lorsque l'on veut décrire la disposition de `k` substituants. De plus, il est nécessaire de mettre au point une méthode de description canonique pour qu'une même disposition soit toujours décrite de la même façon. Soient `i,j ..., p` les indices composant la description, le problème consiste à définir une numérotation des atomes du monocycle. Un substituant lié à l'atome numéroté `i` sera indicé par `i` dans la description. La méthode de détermination de ces indices (figure 8.19) conduit à la notation correspondant à un minimum de la somme des indices obtenus. Sur la figure 8.19, sont représentés quelques exemples de monocycles décrits selon nos conventions. Dans cette description, nous ne tenons pas compte du type atomique des substituants, ni de considérations stéréochimiques.

Description du type de construction de monocycles

Nous allons expliquer dans cette partie quels sont les grands types de construction de monocycles que nous avons choisis de mettre en évidence et la façon dont nous les détectons. Les classes topologiques générales de construction de cycles, comme nous les avons nommées, représentent les types de méthodes couramment envisagées par les chimistes quand ils parlent de méthodes de construction de cycles. Si l'on énumère les termes couramment utilisés, on a la liste suivante :

- extension ou expansion de cycle,
- régression ou contraction de cycle,
- cyclisation ou fermeture,

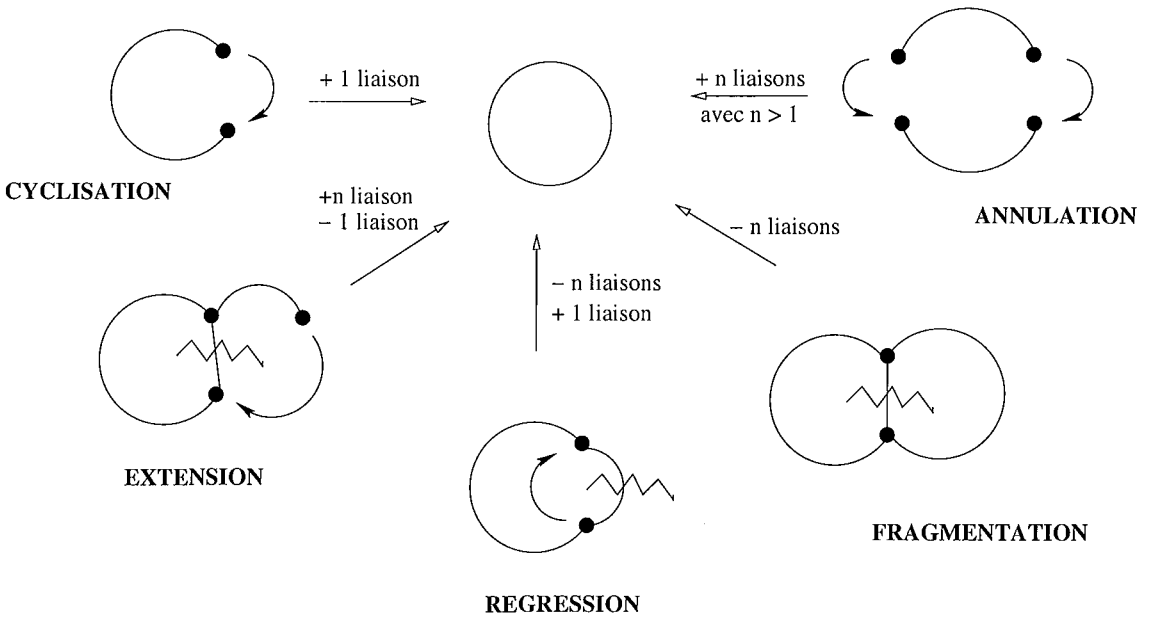


FIG. 8.20 – Les différents types de réactions de construction de cycles.

- annulation (ou dans certains cas particuliers cycloaddition[5]),
- fragmentation.

Cette liste ne prétend pas être exhaustive mais permet de couvrir la majeure partie des possibilités de méthodes formant des cycles. La figure 8.20, inspirée d'une présentation de J. Cossy d'après un cours (initialement sur les cycles de taille 7) écrit de P.A. Wender, schématise les différentes classes de réactions de construction de cycle selon un point de vue dynamique.

Ce schéma permet de visualiser les éléments (cycles et/ou chaînes) qui doivent être nécessairement présents dans les réactants ainsi que la dynamique des liaisons (cassure et/ou création) qui doit être observée. Nous sommes conscients que ce schéma est très imparfait. Nous allons donc faire quelques commentaires à son sujet et donner quelques précisions.

Pour commencer, le schéma ci-dessus n'a pas été conçu dans un but de formalisation stricte. Pourtant tout chimiste le comprendra sans problème car il en connaît les sous-entendus. Seul le dessin sur la cyclisation ne comporte pas d'ambiguïté.

- Pour la figure illustrant les méthodes de type annulation, on peut s'interroger sur la raison pour laquelle il est indiqué que l'on crée n liaisons (`+n liaisons`) alors que le dessin indique qu'une annulation consiste à connecter deux chaînes (dans ce cas, rigoureusement, seulement deux liaisons sont créées). En fait, ce qui est sous-entendu est qu'une annulation peut engager plus que deux chaînes et donc peut impliquer la création de plus de deux liaisons (par exemple, si ce sont trois chaînes qui sont à la

[5]une cycloaddition correspond à une annulation au cours de laquelle des liaisons insaturées sont réduites [IUPAC, 1994].

base de la formation du cycle, trois liaisons seront créées).
- De même, la figure décrivant les extensions ne représente que partiellement la réalité. La chaîne qui permet d'agrandir le cycle n'est pas forcément liée au cycle, elle peut faire partie d'un autre réactant. De plus, elle peut être de taille supérieure à un. Tous ces éléments expliquent pourquoi plusieurs liaisons sont susceptibles d'être créées au cours d'une extension.
- Des remarques semblables aux précédentes peuvent être faites au niveau des méthodes de type régression. La chaîne qui est extraite du cycle d'origine n'est pas forcément de taille un.
- Pour les méthodes de fragmentation, il est également possible de casser plusieurs liaisons quand les cycles sont pontés (possibilité à envisager contrairement à ce que le schéma représente).

Après ces quelques remarques, les règles que nous avons déterminées paraîtrons plus évidentes. Naturellement, nous nous servons de la correspondance des blocs topologiques que nous avons établie.

- extension ou régression de cycle : au niveau de la correspondance un cycle C_r a été transformé en un autre cycle C_p (différent). Soit n le nombre d'atomes communs aux deux cycles C_r et C_p.
  - on a une extension de cycle si n = taille de C_r et taille de C_r < taille de C_p,
  - on a une régression de cycle si n = taille de C_p et taille de C_r > taille de C_p.
- fragmentation : la correspondance de blocs établie indique que deux cycles ont été transformés en un autre cycle. Il faut vérifier que ces cycles partagent au moins une liaison et que la (ou les) liaison(s) commune(s) aux deux cycles a(ont) été cassée(s) (détruite(s)) au cours de la réaction.
- cyclisation : ce type de méthode résulte de la fermeture d'une chaîne sur elle-même. Pour construire le cycle, une (et une seule) liaison a été créée. Il faut donc vérifier au niveau du cycle créé qu'une seule de ses liaisons n'existait pas dans les réactants. Les autres peuvent avoir été modifiées.
- annulation : dans ce cas, deux ou plusieurs chaînes des réactants sont à l'origine du cycle. Pour construire le cycle, plusieurs liaisons ont été créées. Il faut donc vérifier qu'au moins deux liaisons du cycle construit ont été créées.

La méthode employée est d'abord de déterminer si la réaction considérée est une extension, une fragmentation ou une régression de cycle. Puis, si la réaction n'appartient pas à l'une de ces trois classes, on regarde si l'on a une cyclisation ou une annulation. L'algorithme 9 prend en entrée un bloc de type monocycle que nous appellerons cycleCible et dont tous les atomes doivent être appariés. En sortie, on obtient la catégorie de construction associée au monocycle formé. Au niveau du pseudo code de l'algorithme 9 que nous avons développé, les variables utilisées sont les suivantes :

- liste_cyclePrecurseurs : liste des cycles des précurseurs correspondant à cycleCible,
- nb_atoms_communs : nombre d'atomes communs entre cycleCible et cyclePrec,
- taille_cyCible : taille en terme d'atomes du cycle cycleCible,
- taille_cyPrec : taille en terme d'atomes du cycle cyclePrec courant,
- liste_CyclesExtFrag : liste des blocs cycles des précurseurs qui sont susceptibles

d'être à l'origine d'une extension ou d'une fragmentation, liste_CyclesExtFrag est initialement vide,
- nb_liaisons_creees : nombre des liaisons créées au niveau de cycleCible au cours de la réaction,
- nb_liaisons_détruites : nombre des liaisons détruites en relation avec la construction du cycle,
- categorie : chaîne de caractère, categorie est initialement la chaine vide,
- extension, fragmentation, regression, cyclisation, annulation : booléens indiquant à quelle catégorie appartient la formation d'un cycle. extension, fragmentation, regression, cyclisation, annulation sont initialisés à faux.

Outre la description du type général de construction de monocycle, il est possible de spécifier, en termes de nombre et de taille, les éléments structuraux qui interviennent dans la formation du cycle. Nous dirons, par exemple, qu'un cycle de taille 6 a été construit par double connexion de deux chaînes, l'une de taille 4, l'autre de taille 2. Dans notre notation, cela correspond à l'expression `Ch_4_PLUS_2` si les deux chaînes viennent de deux réactants différents (*réaction intermoléculaire*), `Ch_4_AND_2` si les deux chaînes proviennent du même réactant (*réaction intramoléculaire*). Il nous a paru en effet important de différencier ces deux types de cas. Nous avons mis au point une méthode qui permet d'obtenir automatiquement une expression normalisée du type spécifique de construction de monocycle. Le type spécifique de construction précise dans le cas des :
- annulations : le nombre et la taille des chaînes qui se connectent,
- extensions : le nombre et la taille des cycles d'origine (par exemple, une extension d'un cycle de taille 6 s'écrira `ECy_6`),
- régressions : la taille du cycle d'origine (par exemple, une régression d'un cycle de taille 6 s'écrira `RCy_6`),
- fragmentations : le nombre et la taille des cycles intervenant (par exemple, une fragmentation impliquant deux cycles de taille respective 6 et 3 s'écrira `FCy_6_AND_3`).

Pour obtenir la description des annulations, nous ne pouvons pas nous baser uniquement sur les blocs perçus du point de vue topologique. En effet le modèle des blocs topologiques adopté ne permet pas de rendre compte des chemins qui existent dans les structures. Nous sommes donc obligés de redescendre au niveau micro pour décrire les chaînes qui sont impliquées dans une annulation.

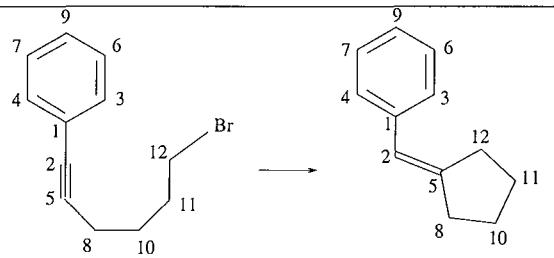
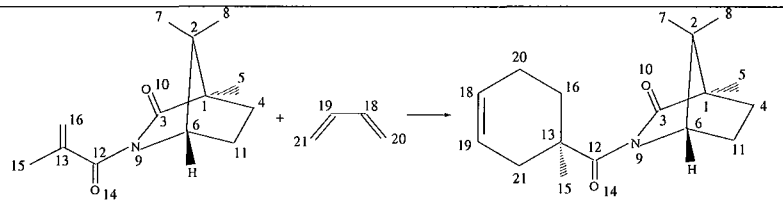
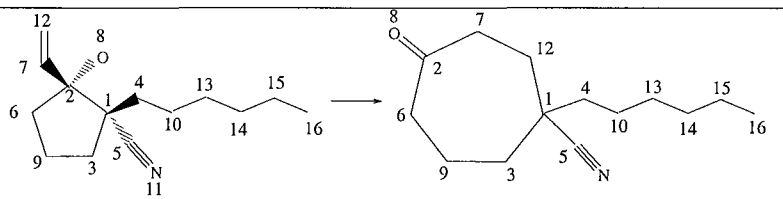
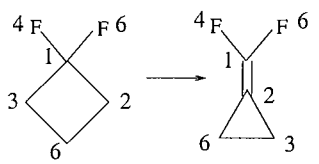
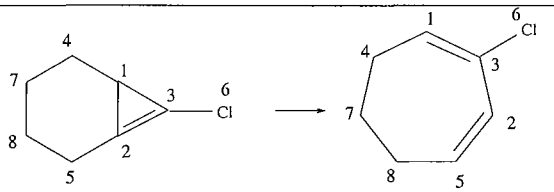
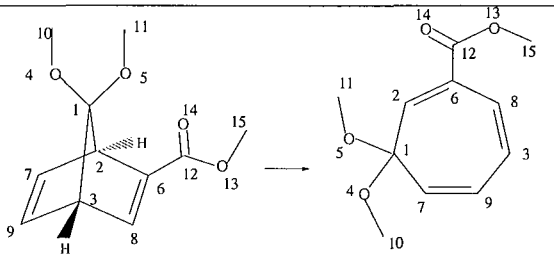
Le tableau 8.4 présente quelques exemples de méthodes de construction de monocycles et la description que nous leur associons. L'appariement des atomes indiqué dans la base de données dont est issu chacun des exemples a été reporté sur les schémas.

8.3.1.2 Sélection

En ce qui concerne les sources de données que nous avons considérées, nous nous sommes restreints aux bases CIRX (versions disponibles de 1992 à 2001), REFLIB et SPORE. Nous avons choisi cet ensemble de bases en raison de la large couverture qu'il permet, tout en évitant la récupération de doublons que l'on aurait pu avoir si l'on avait également intégré des bases telles que REACCS-JSM ou ORGSYN. Cette sélection de bases de données nous

```
begin
    pour tout cyclePrec de liste_cyclePrecurseurs faire
        calcul de nb_atoms_communs ; si nb_atoms_communs≤taille_cyPrec et
        nb_atoms_communs<taille_cyCible alors
            ajouter cyclePrec à liste_CyclesExtFrag ;
        sinon
            si nb_atoms_communs=taille_cyCible et nb_atoms_communs<taille_cyPrec
            alors
                regression ← vrai ; categorie ← ≪ regression ≫ ;
    si ¬ regression et taille de liste_CyclesExtFrag ≥ 1 alors
        si taille de liste_CyclesExtFrag = 1 alors
            si tous les atomes du bloc de liste_CyclesExtFrag sont aussi dans cycleCible
            alors
                extension ← vrai ; categorie ← ≪ extension ≫ ;
        sinon
            si tous les blocs de liste_CyclesExtFrag n'appartiennent pas au même pré-
            curseur alors
                si tous les atomes d'un des blocs de liste_CyclesExtFrag sont aussi dans
                cycleCible alors
                    extension ← vrai ; categorie ← ≪ extension ≫ ;
            sinon
                si les cycles de liste_CyclesExtFrag n'ont pas de liaisons en commun
                alors
                    si tous les atomes d'un des blocs de liste_CyclesExtFrag sont aussi
                    dans cycleCible alors
                        extension ← vrai ; categorie ← ≪ extension ≫ ;
                sinon
                    fragmentation ← vrai ; categorie ← ≪ fragmentation ≫ ;
    si ¬ extension et ¬ fragmentation et ¬ regression alors
        calcul de nb_liaisons_creees ; calcul de nb_liaisons_détruites ;
        si nb_liaisons_creees = 1 et nb_liaisons_detruites = 0 alors
            cyclisation ← vrai ; categorie ← ≪ cyclisation ≫ ;
        si nb_liaisons_creees > 1 et nb_liaisons_detruites = 0 alors
            annulation ← vrai ; categorie ← ≪ annulation ≫ ;
    retourner categorie ;
end
```

Algorithme 9: monocycle−construction−type

	monocycle de taille 5 1-substitué 1 cyclisation
	monocycle de taille 6 2-substitué 1,1 annulation Ch_4_PLUS_2
	monocycle de taille 7 3-substitué 1,1,4 extension ECy_5
	monocycle de taille 3 1-substitué 1 régression RCy_3
	monocycle de taille 7 1-substitué 1 fragmentation FCy_6_AND_3
	monocycle de taille 7 3-substitué 1,1,3 fragmentation FCy_5_AND_5

TAB. 8.4 – Quelques exemples de réactions de construction de monocycle.

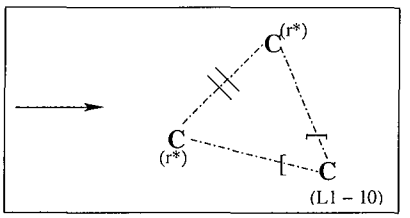


FIG. 8.21 – Requête posée pour récupérer des réactions correspondant à des extensions, régressions, cyclisations ou annulations

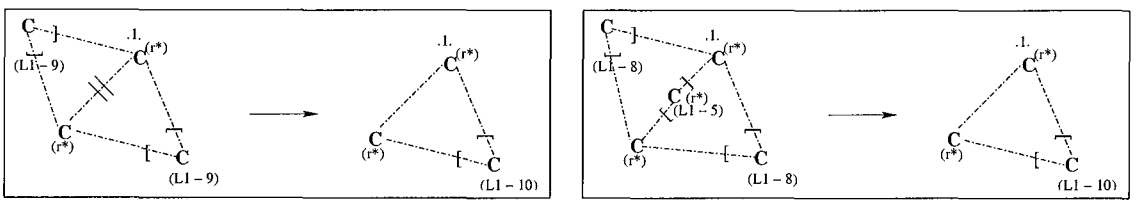


FIG. 8.22 – Les deux requêtes posées dans le but de récupérer des réactions de type fragmentation

donne accès à 855250 fiches de réactions. Ne pouvant raisonnablement pas envisager d'analyser directement ce nombre de fiches et forts du fait que la plupart ne correspond pas au type de méthodes que nous voulons étudier, nous avons posé des requêtes sur cet ensemble via ISIS. Pour récupérer un maximum de méthodes construisant des monocycles carbonés de taille comprise entre 3 et 12, nous avons procédé à trois interrogations différentes et avons fait ensuite l'union des listes données en réponse. Une requête assez générale nous permet de retrouver un ensemble de taille significative de méthodes correspondant à des cyclisations, annulations, extensions ou régressions de cycle (figure 8.21). Cette requête nous donne 16629 réponses. Par contre, nous avons dû poser deux autres requêtes afin de récupérer des réactions de type fragmentation (figure 8.22). La requête de gauche permet de retrouver des réactions de fragmentation entre cycles condensés (372 réponses) alors que la deuxième donne des exemples faisant intervenir des cycles pontés (1547 réponses). Dans tous les cas, nous avons spécifié au niveau du produit que l'on doit avoir un monocycle carboné en ajoutant à chacun des atomes de carbone du cycle tracé l'option (`r*`) qui impose que l'atome doit appartenir au nombre de cycles dessinés dans la requête. Pour récupérer en une seule requête les monocycles carbonés de taille comprise entre 3 et 12, nous avons indiqué que l'un des trois atomes du cycle formé peut correspondre à une liste de 1 à 10 atomes de carbone (notation (`L1-10`)). De plus, nous avons signalé que le type des liaisons nous est indifférent en cochant l'option `Any` pour chacune des liaisons. Dans la première requête, on impose de plus qu'au moins une des liaisons du cycle soit créée.

Nous avons effectué un croisement logique de ces listes afin d'éliminer les doublons. La liste finale comprend 18457 fiches.

Nous récupérons indifféremment par ces interrogations des réactions monoétapes et multiétapes

	taille du cycle					nombre de substituants					position des substituants						type général de construcion					type spécifique de construction					
propriétés / objets	cycle à 3	cycle à 4	cycle à 5	cycle à 6	cycle à 7	0–substitué	1–substitué	2–substitué	3–substitué	4–substitué	1	1,1	1,2	1,3	1,1,3	1,1,4	cyclisation	annulation	extension	régression	fragmentation	Ch_4_PLUS_2	ECy_5	FCy_6_AND_3	FCy_5_AND_5	RCy_3	
M1			X				X				X						X										
M2				X				X				X						X				X					
M3					X				X							X			X				X				
M4	X						X				X									X						X	
M5					X		X					X									X			X			
M6					X				X						X						X				X		

FIG. 8.23 – Le tableau booléen résultant de la transformation des données sur les méthodes de construction de monocycles.

8.3.1.3 Pré-traitement des données

Bien que nous ayons posé des requêtes permettant de récupérer des réactions de construction de monocycles carbonés, toutes les constructions de cycles observées ne concernent pas forcément ce type de cycle particulier. Ceci est dû, en premier lieu, au fait que l'on peut avoir une réaction au cours de laquelle plusieurs cycles sont formés, dont notamment un monocycle carboné. La seconde cause est que, vu la façon dont sont saisies les réactions dans ces bases de données, on peut récupérer des réactions concurrentes dont l'une au moins conduit à la formation d'un monocycle. Pour pallier ce problème, nous avons mis des filtres en place qui permettent d'isoler les analyses sur les méthodes de construction de monocycles carbonés.

8.3.1.4 Transformation des données pour la fouille de données

La transformation des données consiste à considérer chacune des méthodes de formation d'un monocycle donné comme un objet du contexte formel et d'associer à cet objet les propriétés possédées correspondant à chacun des détails de la description. D'après le modèle établi, nous décrivons une réaction de la façon suivante :

- description du cycle formé par trois propriétés,
  - taille du cycle,
  - nombre de substituants sur le cycle,
  - positions relatives de ces substituants sur le cycle,
- description de la façon dont le cycle a été construit par deux propriétés,

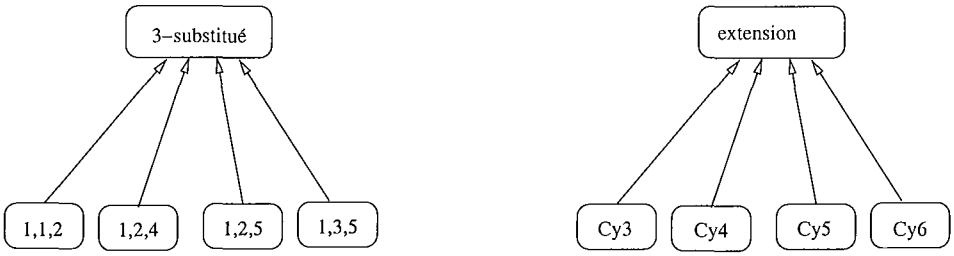


FIG. 8.24 – Les hiérarchies de concepts des propriétés décrivant les réactions de construction de cycle.

- catégorie générale,
- catégorie plus spécifique.

Ainsi, pour l'ensemble de méthodes du tableau 8.4, nous créons 6 objets dont la représentation en terme de tableau booléen est schématisée sur la figure 8.23.

Après sélection, pré-traitement et transformation, nous obtenons un contexte formel contenant 20565 descriptions de méthodes de formation d'un monocycle carboné.

Certaines des propriétés sont dépendantes les unes des autres. En effet, la propriété décrivant les positions relatives des substituants est directement liée à celle indiquant le nombre de substituants sur le cycle. De façon similaire, la catégorie plus spécifique résulte de la catégorie générale. Il y a donc des corrélations fortes dans notre base de données. En fait, dans les couples de propriétés :

- nombre de substituants et position des substituants,
- catégorie générale et catégorie spécifique,

il existe une relation de généralisation/spécialisation entre les membres des couples (figure 8.24). En effet, le nombre de substituants est une notion plus générale que celle des positions. De même pour la catégorie spécifique qui, comme son nom l'indique, spécialise la catégorie générale. Ces relations hiérarchiques entre propriétés sont intéressantes et nous conduisent à la génération de motifs de différents niveau selon les concepts impliqués tout comme cela est proposé dans [Yen and Chen, 2001]. Ainsi, le motif `cycle_a_7` $\wedge$ `3` est plus général que le motif `cycle_a_7` $\wedge$ `1_2_5`. Le support du premier motif est supérieur ou égal au deuxième motif étant donné que le premier comprend l'ensemble de tous les cycles de taille 7 substitués trois fois alors que le deuxième couvre les cycle de taille 7 substitués trois fois en position 1,2,5.

8.3.2 Résultats

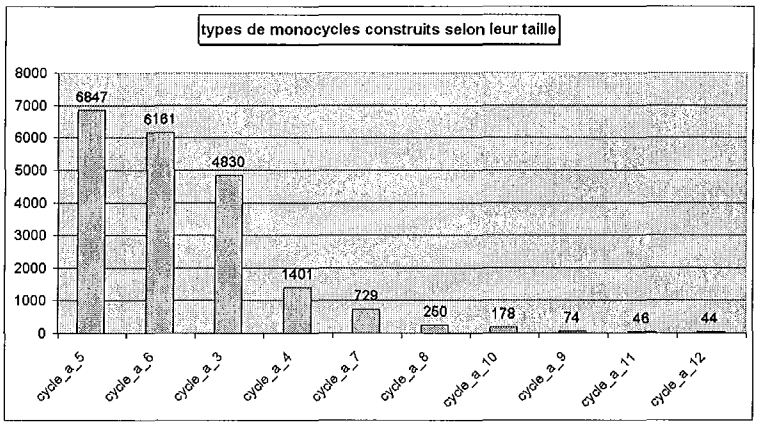
Suivant le même principe que pour les interchanges de fonctions, nous nous sommes intéressés au contenu des bases (section 8.3.2.1) ainsi qu'à la formulation de requêtes pertinentes du point de vue de la rétrosynthèse (section 8.3.2.2).

8.3.2.1 Contenu des bases

Types des cycles construits

On considère ici les monocycles selon leur taille (nombre d'atomes) puis le nombre de substituants.

Taille des monocycles construits

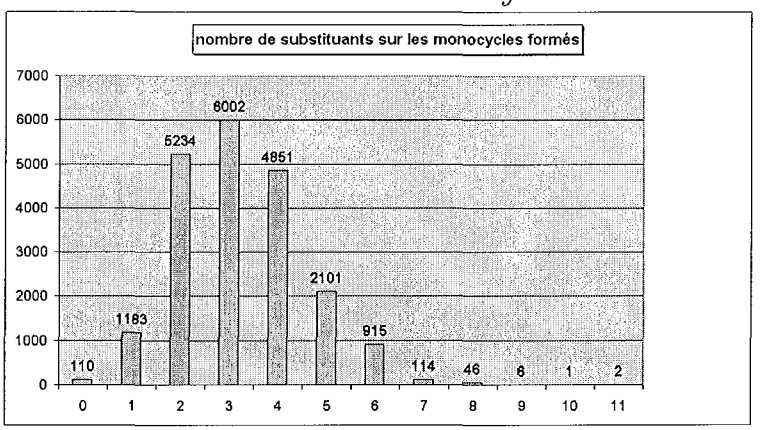


Dans le graphique ci-contre, l'axe des abscisses représente les monocycles selon leur taille, l'axe des ordonnées donne le nombre de monocycles formés pour une taille de monocycle donnée.

Si l'on regarde les fréquences respectives de formation des monocycles, on peut dire que :

- Le grand nombre de cycles à 3 atomes formés peut être expliqué par le fait que ces petits cycles sont très utilisés en tant qu'intermédiaires dans les synthèses. En effet, la formation des cycles à 3 atomes est cinétiquement favorisée mais thermodynamiquement défavorisée. Ils apparaissent néanmoins aussi beaucoup dans les produits naturels, comme par exemple les acides chrysanthémiques.
- Les cycles à 4 atomes, bien que notablement moins souvent formés, le sont avec une fréquence non négligeable. Ils sont difficiles à obtenir du fait de la tension de cycle de leur structure mais, tout comme les cycles à 3 atomes, leur formation est une étape dans une séquence de réactions.
- Les cycles à 5 atomes sont très souvent formés. Ce sont des cycles que l'on retrouve très souvent dans les molécules, en particulier naturelles telles que les prostanglandines. Ce sont des cycles énergétiquement faciles à construire.
- Les cycles à 6 atomes sont relativement peu formés par rapport à leur fréquence d'apparition dans les molécules. En effet, les cycles de taille 6 sont les monocycles les plus couramment présents dans les structures moléculaires mais sont relativement peu formés et restent souvent inchangés. De très nombreux terpènes comportent des cycles de taille 6.
- Les cycles à 7 atomes apparaissent relativement moins souvent que les cycles précédents mais leur fréquence n'est tout de même pas négligeable. Ceci s'explique par le fait que ce sont des cycles importants (présence dans les molécules naturelles) et dont des méthodes de synthèse sont connues.
- Les cycles à 8 et 10 atomes sont ensuite observés respectivement plus de 200 et 100 fois.
- Les monocycles à 9, 11, 12 sont construits quelques dizaines de fois sur l'ensemble de réactions choisi. Les cycles de taille 11 ne sont jamais inchangés.

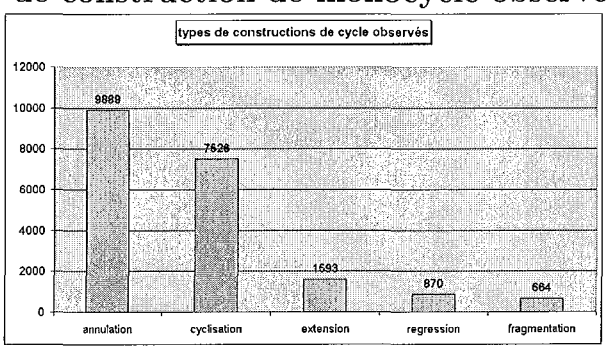
Nombre de substituants sur les monocycles construits



Dans le graphique ci-contre, l'axe des abscisses représente le nombre de substituants sur les monocycles créés, l'axe des ordonnées donne le nombre de monocycles formés pour un nombre de substituants donné.

Les monocycles créés sont donc très souvent substitués 2, 3 ou 4 fois. Selon la taille des monocycles, on observe des différences de nombres de substituants. Les grands monocycles construits sont relativement peu substitués. Les cycles à 3 et 4 atomes ont le plus souvent respectivement deux et quatre substituants. Sur les cycles de taille 5, 6 et 7, on trouve le plus fréquemment trois substituants.

Types de construction de monocycle observées



Dans le graphique ci-contre, l'axe des abscisses représente les types généraux de construction de monocycles, l'axe des ordonnées donne le nombre de monocycles formés pour un type général de construction donné.

En examinant les fréquences relatives des motifs de taille 1 concernant chacun des types généraux de construction de cycles, nous avons pu comparer les méthodes les plus fréquemment employées. Les résultats suivant ne tiennent donc pas compte de la taille des monocycles formés. On remarque que les annulations sont très représentées.

Étude croisée des types de monocycles formés et de la façon dont ils sont construits

Dans la section précédente, nous avons regardé l'importance relative de chacune des catégories de réactions de construction de monocycles les unes par rapport aux autres sans nous préoccuper de la taille des monocycles construits. Or on peut observer des différences notables si l'on considère ce paramètre (figure 8.25). Dans ce qui suit, nous allons comparer les cas selon la taille des monocycles.

On remarque que, dans la plupart des cas, les réactions de construction de monocycles

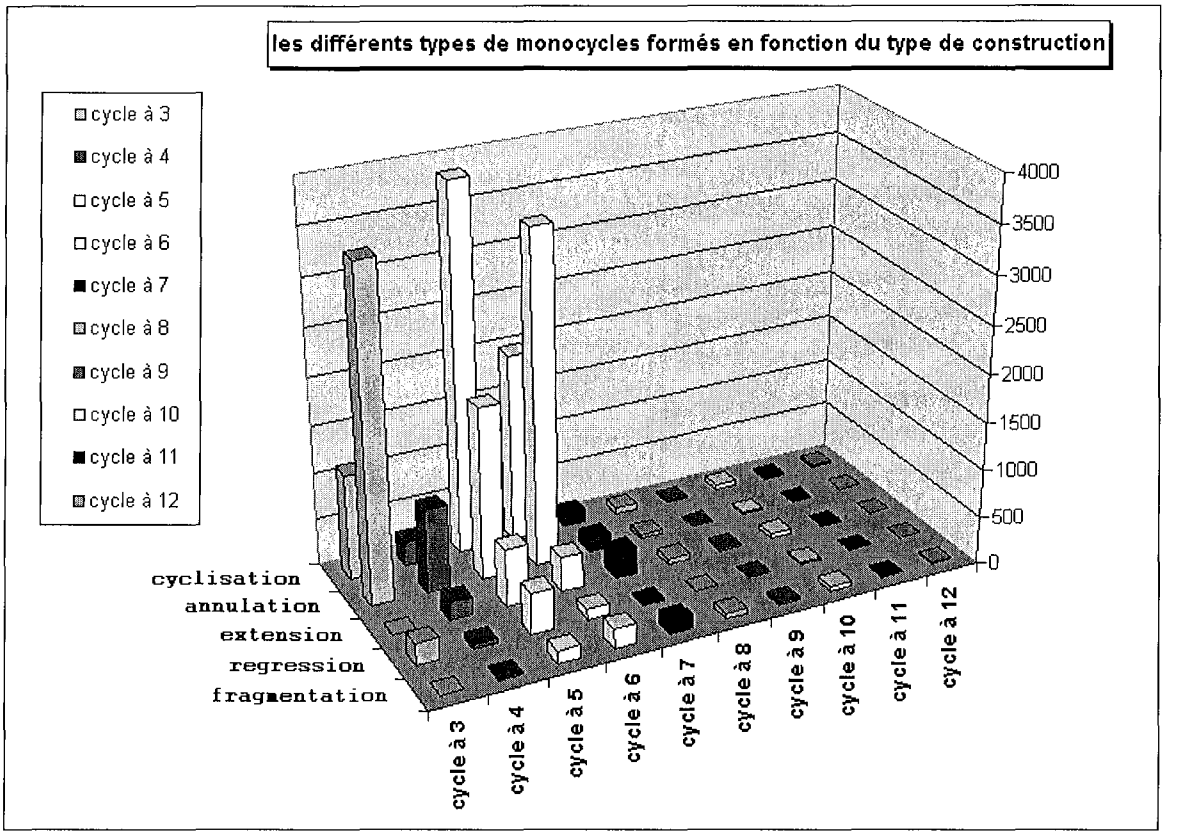
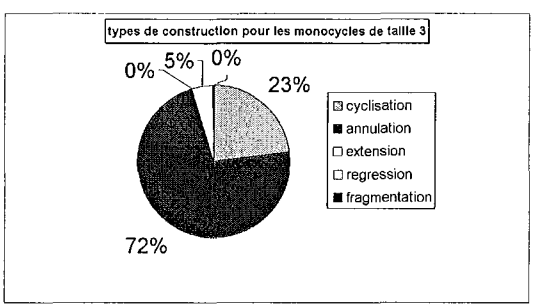


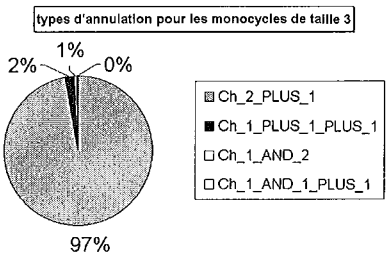
FIG. 8.25 – Étude des types de monocycles selon le type de construction.

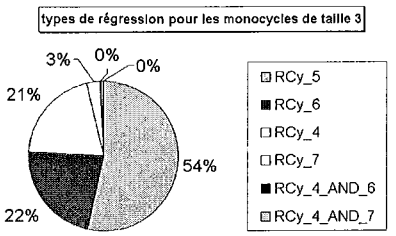
font partie d'une des catégories : cyclisation, annulation, extension. Les fragmentations sont très peu observées tandis que les régressions sont essentiellement utilisées pour faire des monocycles à 3, à 5, à 6. Les monocycles à 5 atomes sont majoritairement construits par cyclisation contrairement aux monocycles à 3, 4 et 6 pour lesquels la méthode principalement employée est une annulation. Les monocycles de taille 7 se distinguent par le fait qu'ils sont surtout construits par extension de cycle. Examinons plus en détail la formation des différents types de monocycle selon leur taille.

- **monocycles de taille 3 :**

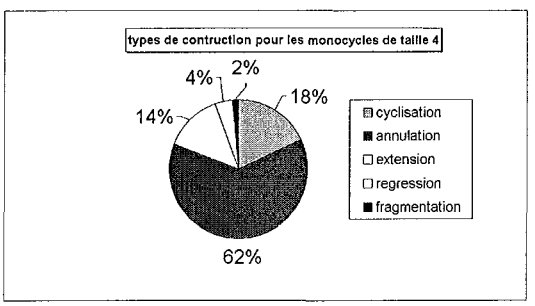


Les annulations impliquant deux chaînes de taille 2 et 1 sont le mode de construction majoritaire des monocycles de taille 3. Un monocycle de taille 3 n'est évidemment jamais construit par une extension de cycle.

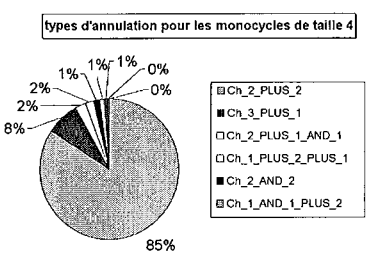


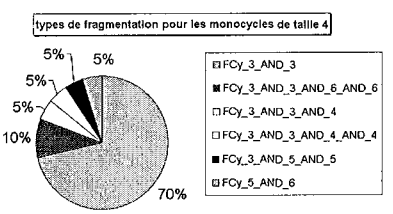


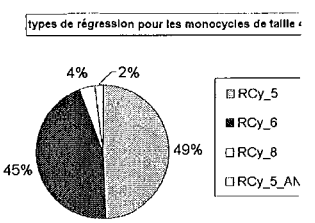
- **monocycles de taille 4 :**



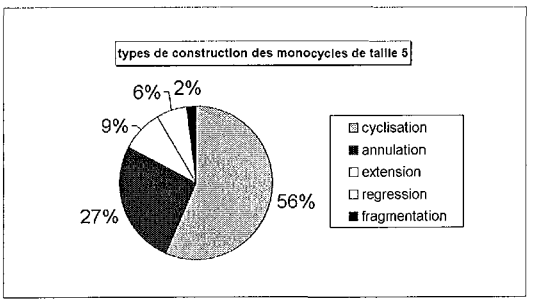
Les annulations impliquant la connection de deux chaînes de taille 2 sont le mode de construction majoritaire des monocycles de taille 4. Comme on peut s'y attendre, la seule extension de cycle formant un monocycle de taille 4 est une extension de cycle de taille 3.



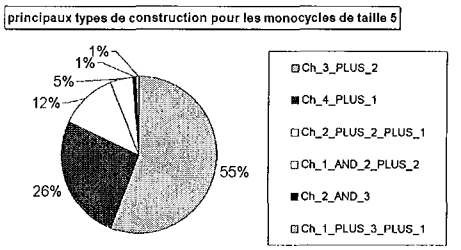


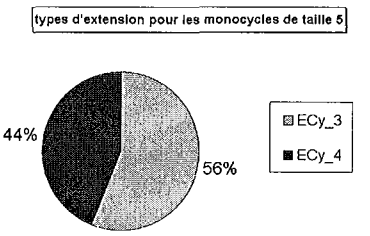


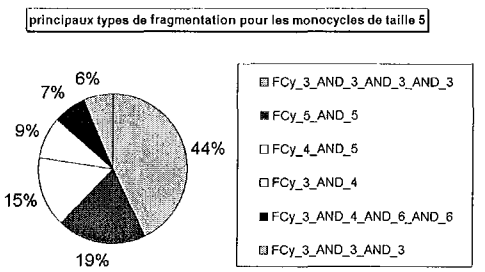
- **monocycles de taille 5 :**

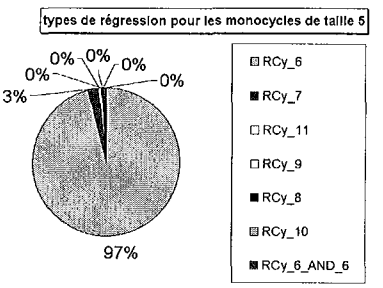


Comme nous l'avons remarqué précédemment, les monocycles de taille 5 sont majoritairement construits par cyclisation.

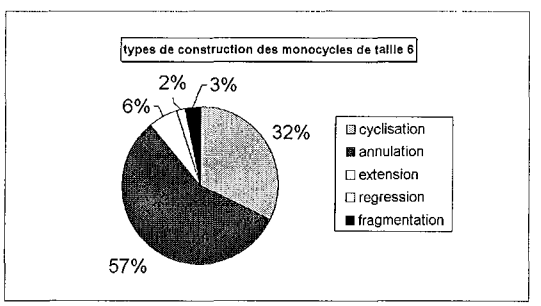




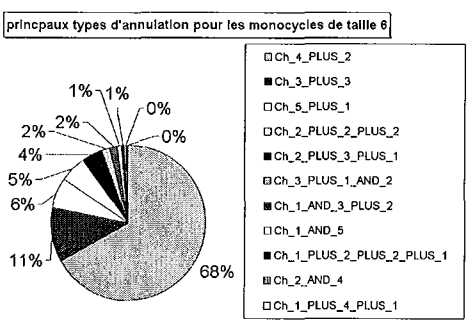


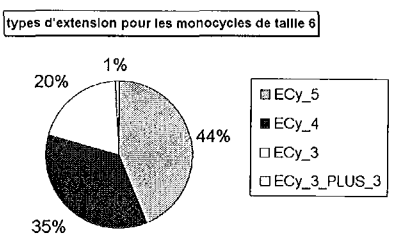


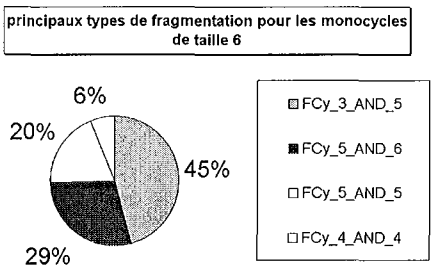
- **monocycles de taille 6 :**

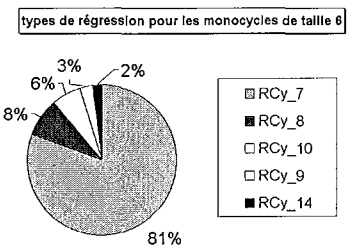


Les monocycles de taille 6 sont majoritairement formés par double connexion d'une chaîne de taille 4 avec une chaîne de taille 2.

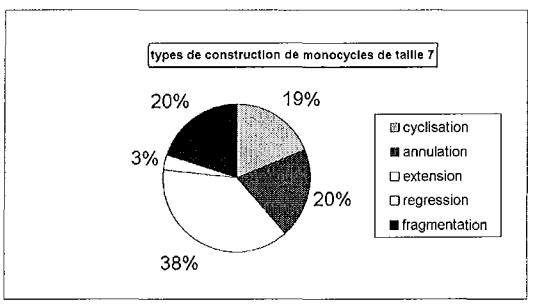




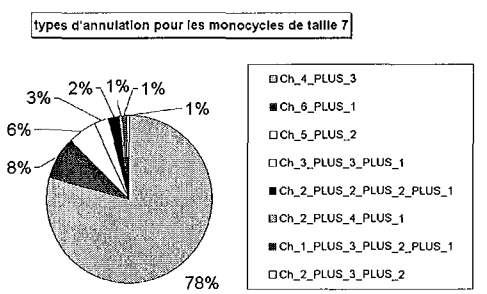


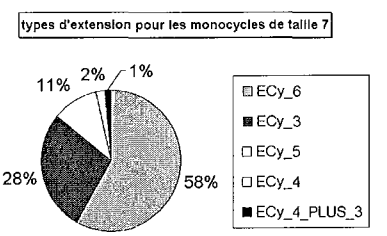


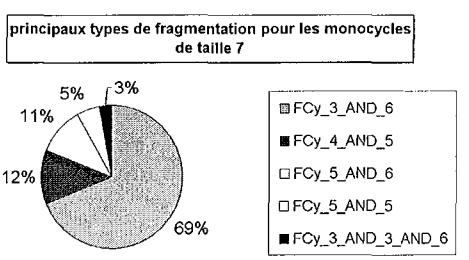
- **monocycles de taille 7 :**

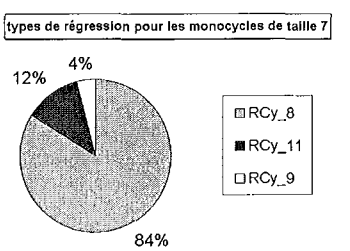


Les monocycles de taille 7 sont les seuls types de monocycles à être construits de préférence par extension de cycle. Dans ce cas, les extensions les plus fréquentes sont celles de cycles de taille 6.

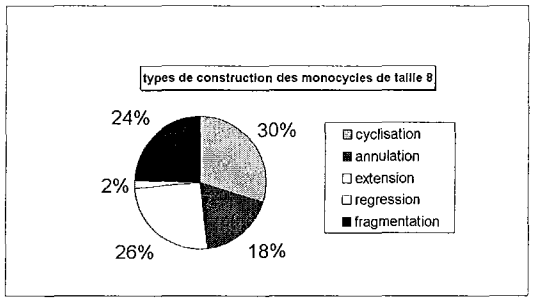




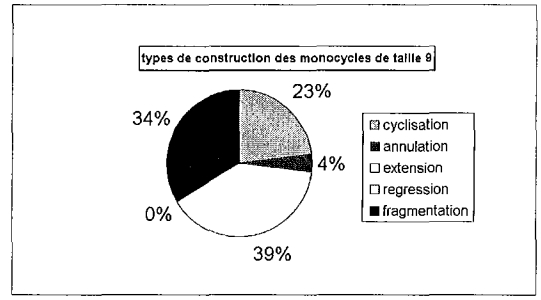




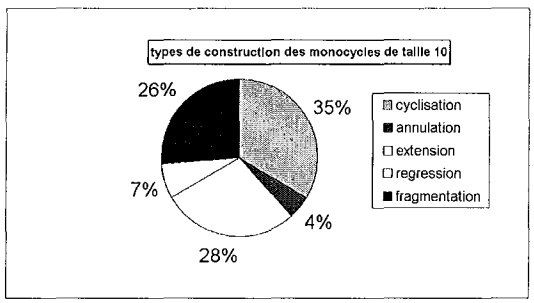
- **monocycles de taille 8 :**



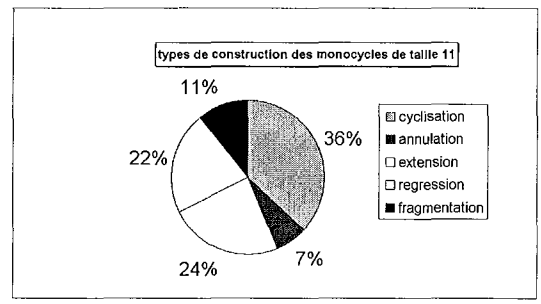
- **monocycles de taille 9 :**



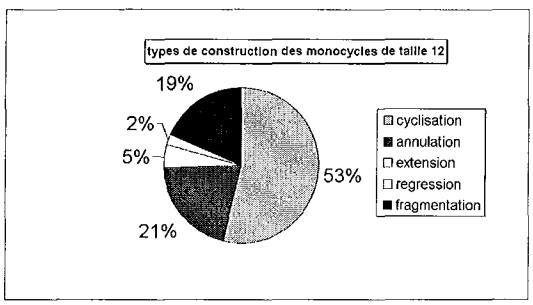
- **monocycles à 10 :**



- **monocycles à 11 :**



- **monocycles à 12 :**



Les tendances générales observées sur les bases de données de réactions sont en bon accord avec les connaissances du domaine. On peut notamment retrouver des similitudes avec ce qui est indiqué dans les chapitres traitant de la formation des divers types de cycle du livre de S. Warren [Warren, 1986].

8.3.2.2 Exploitation des motifs fréquents et des règles d'association face à un problème de synthèse donné

Les motifs fréquents et les règles d'association obtenus par le processus de fouille de données peuvent nous aider à répondre à des questions que l'on se pose devant un problème de synthèse. Si parmi les objectifs à atteindre dans la molécule cible se trouvent un ou des monocycles à construire, il est possible de trouver dans les modèles extraits des réponses donnant des tendances (quelles méthodes sont employées ?, quelles sont les méthodes les plus employées ?). En effet, une fois un tel objectif choisi, on connaît la description du cycle à construire. Il suffit alors de parcourir la liste des motifs et des règles incluant cette description (pour les règles, la description devra être présente dans la prémisse).

annulation	46,51%
extension	25,58%
fragmentation	20.93%
cyclisation	6.98%

TAB. 8.5 – Les catégories générales de méthodes de construction de cycle permettant d'obtenir un cycle de taille 7 substitué trois fois en 1,2,4, classées selon leur fréquence d'emploi relatif.

Cy_6	45.45%
Cy_5	27.27%
Cy_3	27,27%

TAB. 8.6 – Les types d'extension observés lors de la formation de cycle de taille 7 à trois substituants en position 1,2,4.

Premier exemple

Par exemple, considérons que notre objectif stratégique soit celui de construire un monocycle carboné de taille 7 substitué trois fois en position 1,2,4.

Examen des motifs fréquents

L'examen des motifs fréquents nous indique, dans un premier temps, s'il existe des exemples de réactions correspondant à la construction d'un tel objectif. Nous devons regarder le motif : `cycle_a_7` $\wedge$ `1_2_4`. Nous n'avons pas besoin de regarder le motif : `cycle_a_7` $\wedge$ `3` $\wedge$ `1_2_4` car celui-ci a le même support étant donné qu'un cycle substitué en 1,2,4 a forcément trois substituants (ces deux éléments de la description du cycle sont corrélés). Il y a 43 réactions qui répondent à cette exigence.

Si l'on veut savoir à quelles catégories sont associés ces exemples, on peut également regarder les motifs de la forme : `cycle_a_7` $\wedge$ `1_2_4` $\wedge$ `cat_gen`, `cat_gen` étant une catégorie générale de réactions de construction de cycle. On constate que l'on a détecté des exemples pour chacune des catégories sauf pour la régression.

Examen des règles d'association

Les règles d'association permettent de comparer les possibilités les unes par rapport aux autres.

Pour savoir quelles sont les catégories de réactions préférentiellement employées pour former un cycle de taille 7 trois fois substitué en position 1,2,4, on regarde les règles de la forme : `cycle_a_7` $\wedge$ `1_2_4` $\Rightarrow$ `cat_gen`, `cat_gen` étant une catégorie générale. On obtient les résultats consignés dans le tableau 8.5.

Ce sont donc les annulations qui sont les plus employées. Ceci est contraire à l'observation générale qui nous indique que les cycles de taille 7 sont préférentiellement construits par extension. On peut donc s'interroger sur le type de généralisations possibles.

Parmi les annulations, les possibilités sont :

- connexion de deux chaînes, l'une de taille 4 et l'autre de taille 3
- connexion de deux chaînes, l'une de taille 5 et l'autre de taille 2

Dans le tableau 8.7, on peut voir les règles associées.

L'interprétation de la règle 1 est la suivante : quand on forme un monocycle de taille 7 substitué en 1,2,4 par annulation, dans 90% des cas cela se fait par connection de deux chaines respectivement de taille 4 et 3.

règle 1	cycle_a_7 ∧ 1_2_4 ∧ annulation ⇒ Ch_4_PLUS_3	90%
règle 2	cycle_a_7 ∧ 1_2_4 ∧ annulation ⇒ Ch_5_PLUS_2	10%

TAB. 8.7 – Les types d'annulation permettant d'obtenir un cycle de taille 7 trois fois substitué en 1,2,4.

extension	58,97%
fragmentation	30,77%
annulation	7,70%
cyclisation	2,56%

TAB. 8.8 – Les catégories générales de réactions qui permettent d'obtenir un cycle de taille 7 substitué trois fois en 1,2,5

Cy_6	45.45%
Cy_5	27.27%
Cy_3	27,27%

TAB. 8.9 – Les types de extension permettant de construire un cycle de taille 7 substitué trois fois en 1,2,5.

Pour les types d'extension observés dans le cas de la construction des cycles de taille 7 substitués trois fois en 1,2,4, on obtient le tableau 8.6.

Deuxième exemple

Par exemple, supposons que nous voulons construire un monocycle carboné de taille 7 substitué trois 3 fois en position 1,2,5.

Examen des motifs fréquents

L'examen des motifs fréquents nous indique, dans un premier temps, s'il existe des exemples de réactions correspondant à la construction d'un tel objectif. Nous devons regarder le motif : cycle_a_7 ∧ 1_2_5. Nous n'avons pas besoin de regarder le motif : cycle_a_7 ∧ 3 ∧ 1_2_5 car celui-ci a le même support étant donné qu'un cycle substitué en 1,2,5 a forcément trois substituants (ces deux éléments de la description du cycle sont corrélés). Il y a 39 réactions qui répondent à cette exigence. Si l'on veut savoir à quelles catégories sont associés ces exemples, on peut aussi regarder les motifs de la forme : cycle_a_7 ∧ 1_2_5 ∧ cat_gen, cat_gen étant une catégorie générale. Comme dans l'exemple précédent, on constate que l'on a détecté des exemples pour chacune des catégories sauf pour la régression.

Examen des règles d'association

Les règles d'association permettent de comparer les possibilités les unes par rapport aux autres. Les règles de la forme : cycle_a_7 ∧ 1_2_5 ⇒ cat_gen, cat_gen étant une catégorie générale, indiquent à partir de quels types de méthode les cycles de taille 7 substitués trois fois en 1,2,5 sont formés et dans quelle proportion. Le tableau 8.8 illustre les résultats obtenus dans le cas de la formation d'un cycle de taille 7 substitué trois fois en 1,2,5.

On remarque que, cette fois, ce sont les extensions qui sont les plus employées. Ceci vérifie la tendance générale qui dit que les cycles à 7 sont préférentiellement construits par extension, tendance que l'on a pu déjà observer sur l'ensemble des cycles de taille 7 construits dans l'ensemble de réactions étudiés.

Pour les réactions de type annulation, il y a un seul type détecté et qui implique la connexion de deux chaînes, une de taille 4 et l'autre de taille 3 (règle d'association :

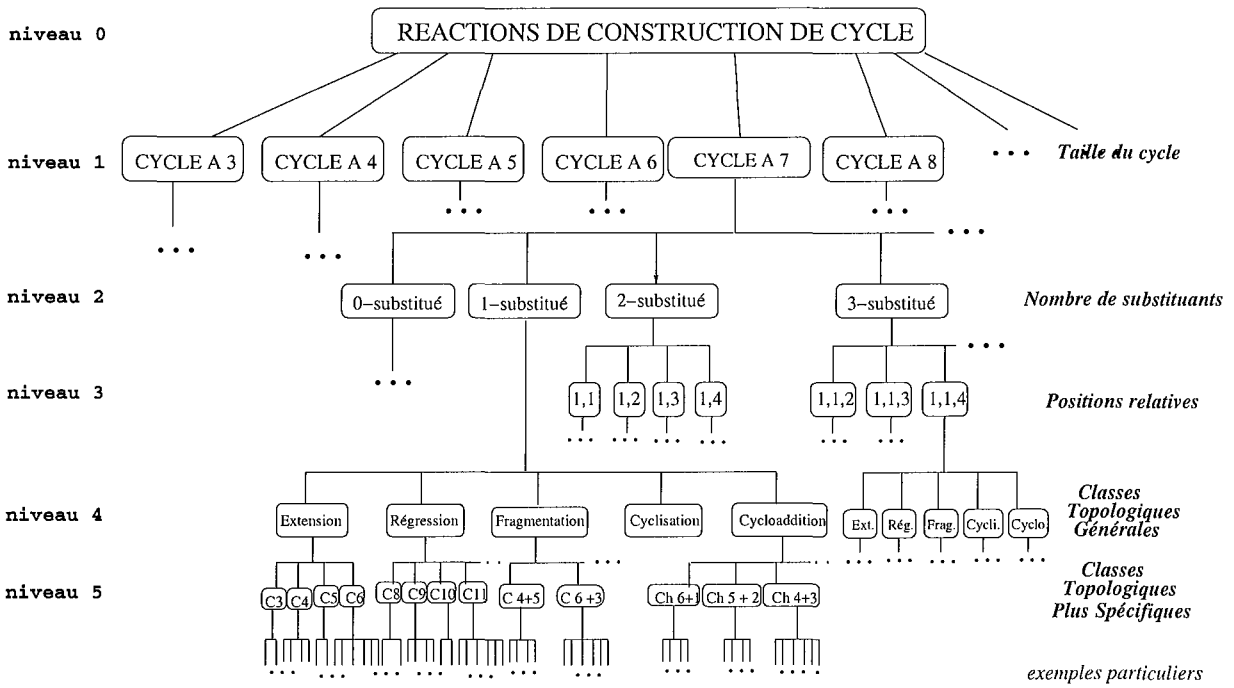


FIG. 8.26 – L'arborescence des méthodes de construction de cycle développée partiellement.

`cycle_a_7` $\wedge$ `1_2_5` $\wedge$ `annulation` $\Rightarrow$ `Ch_4_PLUS_3` dans 100% des cas).

Parmi les extensions qui permettent d'obtenir un cycle de taille 7 substitué trois fois en 1,2,5, on obtient les résultats du tableau 8.9.

8.3.2.3 Visualisation et navigation dans les données

Les résultats bruts que l'on obtient par le processus d'extraction de connaissances ne sont pas forcément facilement interprétables par un spécialiste de la synthèse. Pour faciliter l'analyse il est nécessaire de mettre au point des modules de visualisation.

À partir de la description des objectifs stratégiques atteints par les méthodes de synthèse construisant des monocycles, nous pouvons générer automatiquement une arborescence de fichiers dont les feuilles sont les exemples de méthodes et les nœuds sont les divers niveaux de description. Il y a plusieurs possibilités d'ordonnancement de ces niveaux. Nous avons choisi de commencer par la description de la structure puis celle de l'action (figure 8.26).

Construction et mise à jour

L'arborescence de fichiers n'est pas définie a priori mais elle est construite pas à pas, au fur et à mesure que l'on explore les exemples particuliers d'utilisation des méthodes que l'on récupère dans les bases de données. En effet, au départ l'arborescence est vide. Puis dès que l'on rencontre une réaction qui, après l'analyse de la correspondance de blocs, est une

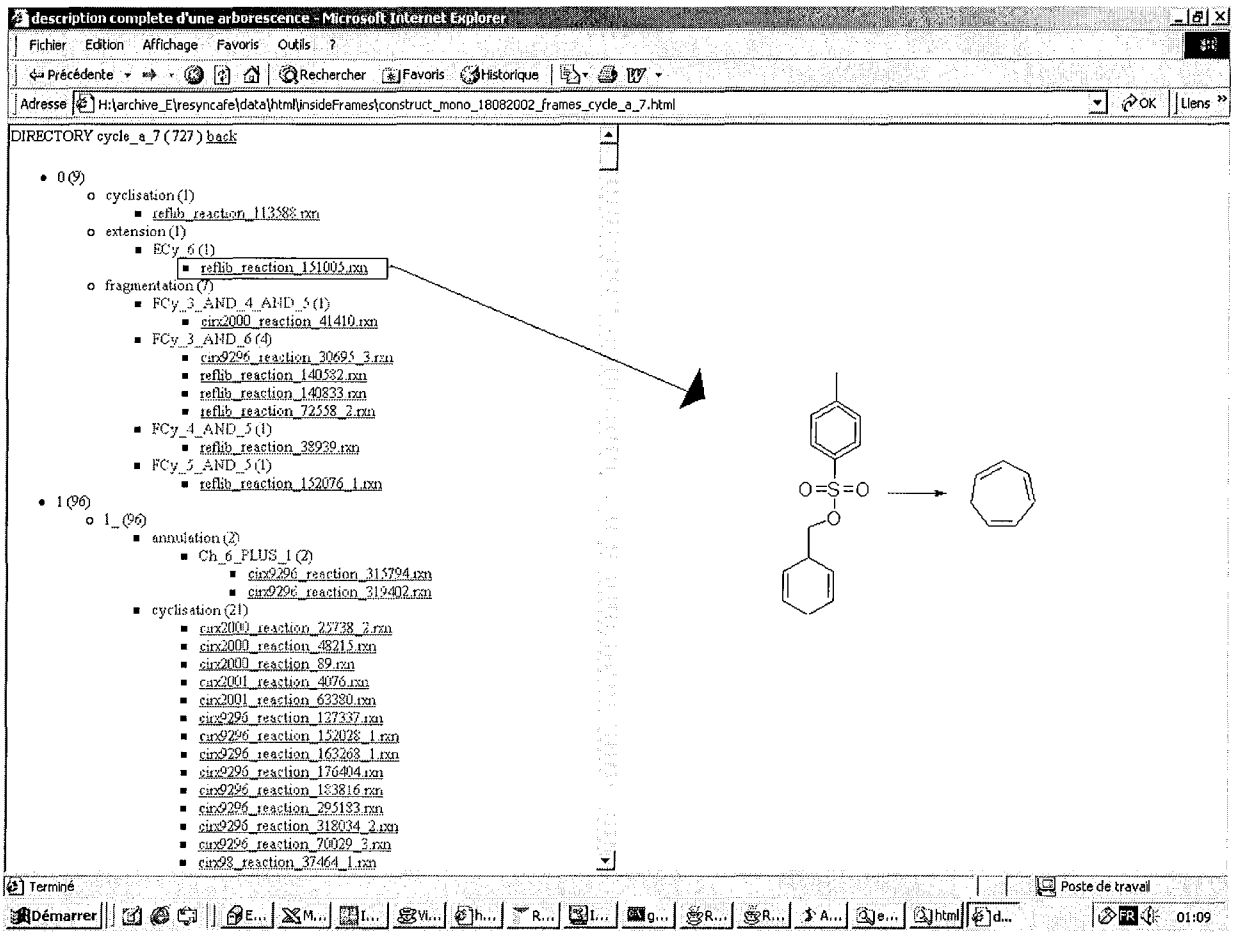


FIG. 8.27 – Visualisation de l'arborescence des méthodes de construction de monocycle de taille 7 dans un navigateur.

méthode de construction de monocycle, on décrit ce cycle (taille, nombre de substituants et positions relatives des substituants) et on détermine les classes topologiques de la méthode (générale puis spécifique). Chacun des éléments de la description de la méthode obtenue permet de placer la méthode dans l'arborescence en créant l'ensemble des répertoires nécessaires pour réaliser le chemin de la source (taille du cycle formé) au dernier répertoire (classe spécifique). On continue à parcourir la liste des exemples d'utilisation de méthodes et on crée les répertoires nécessaires à chaque fois que la méthode examinée appartient à une classe non encore décrite dans l'arborescence. Cette construction incrémentale de l'arborescence permet une mise à jour ne nécessitant pas la reconstruction de celle-ci. Les opérations de mise à jour sont la création et l'ajout. Il n'y a pas de suppression ni de fusion comme dans la classification hiérarchique incrémentale.

Pour faciliter la visualisation de cette arborescence, nous avons développé un module permettant de créer un fichier de type HTML permettant de parcourir l'arborescence rapidement. Dans ce fichier HTML, les feuilles de l'arborescence sont des liens vers des fichiers

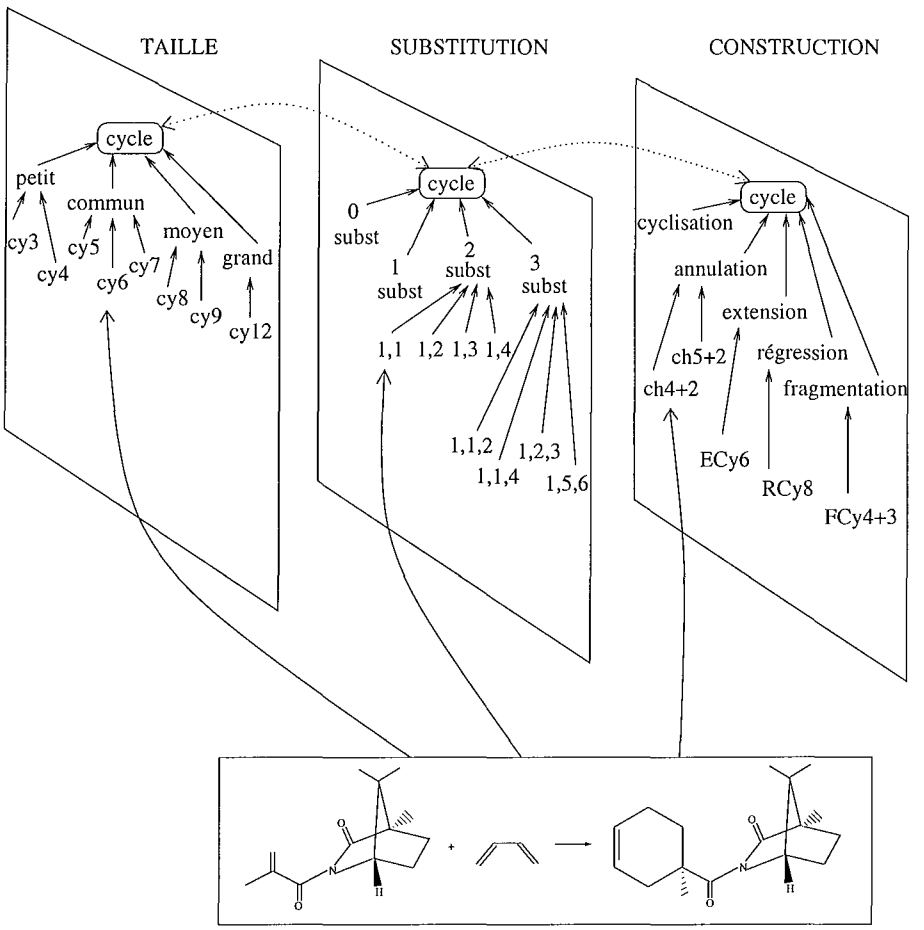


FIG. 8.28 – Le concept Méthode de construction de cycle vu selon différents points de vue.

au format RXN de MDL. Ces fichiers sont visualisables dans un navigateur auquel on a ajouté le *plugin* CHIME [MDLCHIME, url] permettant de visualiser le schéma d'une réaction. Sur la figure 8.27, la capture d'écran montre le début de l'arborescence des méthodes de construction des cycles de taille 7.

Le parcours d'une telle arborescence est liée à la façon dont on l'a construite, il serait préferable de concevoir un système dans lequel les connaissances seraient organisées selon divers points de vue en hiérarchies de classes avec des passerelles entre ces hiérarchies. Ainsi sur le schéma de la figure 8.28, nous avons représenté le concept Méthode de construction de cycle selon les points de vue de la taille ainsi que de la substitution du cycle construit et du type de formation.

Nous avons expérimenté ce type d'approche avec PROTÉGÉ. En suivant le modèle conceptuel décrit dans le chapitre 6, nous avons créé les classes générales d'Action, de Structure, d'Objectif et de Méthode de synthèse. Des classes plus spécifiques ont été ensuite ajoutées. La figure 8.29 montrent les classes qui ont été créées ainsi que la définition de la classe Objectif. Conformément au modèle conceptuel, un Objectif est lié à une Structure et

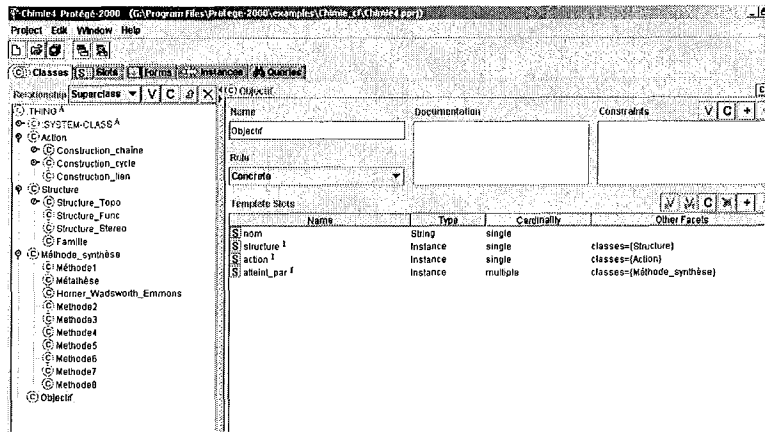


FIG. 8.29 – Le modèle de classes construit dans PROTÉGÉ.

une Action et peut être atteint par plusieurs Méthode de synthèse. Sur la figure 8.30, nous avons sélectionné une instance du concept Métathèse, une classe plus spécifique que celui de Méthode de synthèse. Dans la fenêtre de description de l'instance, nous avons le nom des instances auxquelles est liée l'instance de Méthode de synthèse choisie. En cliquant sur l'un de ces noms, nous obtenons la description de l'instance liée. Ainsi, sur la figure, nous pouvons voir l'objectif lié à la méthode. Cet objectif est atteint par deux méthodes dont celle sélectionnée. La structure construite est celle d'un cycle de taille 7 4-substitué en 1,1,4,5. En plus de la navigation dans l'arborescence des instances, PROTÉGÉ propose un module d'interrogation par mots clés. Celui-ci permet de faire des requêtes sur la base de connaissances qui a été construite.

8.3.3 Discussion

Cette étude basée sur des critères topologiques des méthodes de construction de monocycles a été riche en enseignements.

- Au premier abord, il peut paraître facile de déterminer le type général de construction de cycles si l'on se base sur la représentation simplifiée du schéma 8.20. Cependant, la traitement de données réelles présentes dans les bases de données de réactions nous a conduit à la prise en compte de nombreux cas particuliers. Le développement de l'algorithme a été donc itératif.
- Certains des résultats de la description des objectifs atteints en terme d'action nous ont amené à remettre en cause l'appariement des atomes donné dans les bases de données de réactions. Nous avons mis en évidence un nombre conséquent de cas où cet appariement est erroné. Si l'on reprend l'exemple pris en section 8.2.1.2 (figure 8.8), l'appariement donné dans la base nous conduit à dire que cette méthode correspond à une formation d'un monocycle de taille 5 par annulation alors qu'en réalité cette méthode ne construit pas de cycle. Un effort de réflexion doit être fait sur cet aspect des données des bases de données de réactions. Des pistes de solutions sont proposées dans la thèse de Y. Tognetti qui aborde le problème de l'appariement automatique

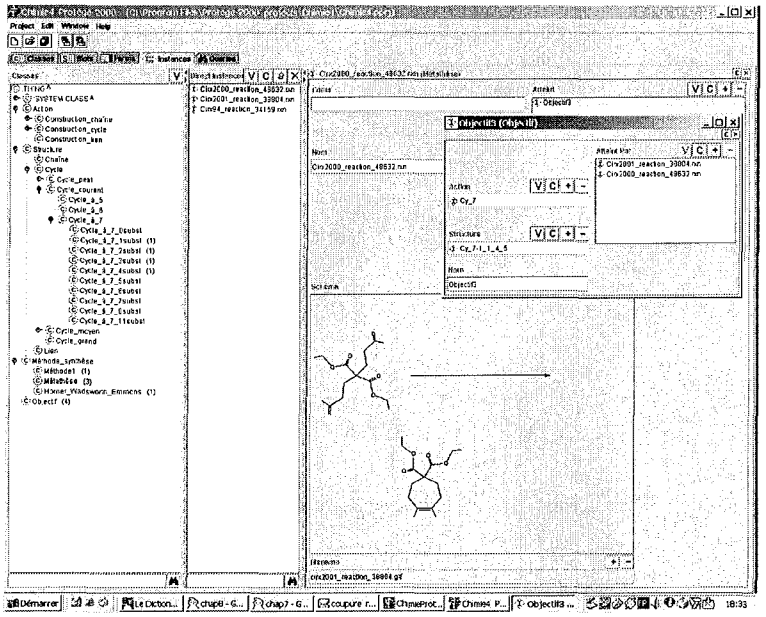


FIG. 8.30 – Une instance du concept de méthode de synthèse créée dans PROTÉGÉ.

[Tognetti, 2002].

- Nous avons identifié un certain nombre de cas particuliers pour lesquels la détermination du type de construction n'est pas possible selon nos critères et avons écarté ces données. Cependant, la majorité (95%) des données ont pu être traitées.
- Il serait intéressant de faire une étude similaire sur les méthodes de construction d'hétérocycles et de comparer les résultats avec l'analyse de la formation des carbocycles.
- L'étude de la construction des polycycles devrait être envisagée sur des principes assez similaires bien que leurs types de construction ne correspondent pas à ceux des monocycles.
- L'étude des conditions réactionnelles doit être également envisagée. Cependant, étant donné la nature de ces données, cela implique un important travail de structuration préalable.

8.4 Conclusion

Nous avons exposé dans ce chapitre les différentes expérimentations d'extraction de connaissances dans les bases de données de réactions que nous avons menées. Nous avons appliqué les techniques de fouille à des données très particulières qui sont à l'origine de type structural. Les transformations des données effectuées induisent la perte de l'information structurale.

La place de l'analyste et des spécialistes du domaine a été très importante dans ce travail. L'intervention de Gilles Niel au niveau de la modélisation des méthodes de construction de cycles a été particulièrement bénéfique pour la compréhension et la prise en compte des

cas particuliers que nous avons rencontrés. Il existe des travaux analogues et, parmi ceux-ci, notons ceux de Sandy Maumus, actuellement en thèse dans l'équipe Orpailleur, qui étudie l'utilisation des méthodes de fouille de données appliquée à l'exploitation de données biologiques et génétiques dans les interactions génotype-phénotype intermédiaire et maladies cardiovasculaires.

L'application à des données réelles nous a montré toute la difficulté de l'exploitation de ce type de données. Malgré les problèmes que nous avons soulevés, les tendances que nous faisons émerger présentent un accord satisfaisant avec les connaissances du domaine. Ces résultats peuvent servir d'heuristiques guidant le choix d'une méthode par rapport à un ensemble. Pour cela, des bases de données répertoriant les règles d'association extraites devraient être couplées au système global d'aide à la planification de synthèse.

Conclusions et perspectives

Cette thèse a eu pour cadre général l'étude des systèmes d'information chimique. Dans l'optique de l'aide à la planification de synthèse, l'objectif principal a été celui de la modélisation des outils indispensables à la résolution de problèmes de synthèse : les méthodes de synthèse. Pour cela, nous avons bénéficié des travaux effectués dans le cadre du GDR 1093 TICCO. Nous nous sommes notamment appuyés sur les résultats de la perception des molécules pour représenter les méthodes de synthèse selon le modèle que nous avons élaboré.

Une méthode de synthèse a été définie comme un moyen d'atteindre un certain nombre d'objectifs de synthèse. Un objectif de synthèse est vu comme une action sur une structure. Le problème a été alors de déterminer les différents types d'objectifs susceptibles d'être atteints par une méthode. La description originale des méthodes de synthèse que nous avons obtenue nous a permis par la suite de poursuivre le processus d'acquisition de connaissances par une extraction de connaissances à partir de bases de données de réactions.

Dans les systèmes d'information chimique, les molécules sont décrites, la plupart du temps, par des graphes moléculaires. Cette représentation permet de nombreux traitements mais a le désavantage de ne présenter qu'un niveau d'analyse. Les travaux du GDR 1093 TICCO ont conduit à multiplier les représentations par des graphes. Celles-ci sont plus ou moins abstraites et se font selon différents points de vue. Au niveau micro, les graphes sont constitués d'atomes et de liaisons. Au niveau méso, les sommets sont des blocs, c'est-à-dire à dire des sous-graphes connexes, et les arêtes des chemins de type méso. Au niveau macro, les graphes sont constitués de systèmes de blocs reliés par des chemins macro. Pour déterminer le bilan d'une méthode de synthèse, il faut disposer d'un appariement entre les atomes des réactants et des produits. Au niveau micro, on en déduit les modifications qui ont eu lieu sur les liaisons (création, destruction, modification). Au niveau méso, nous avons établi les principes d'une correspondance des blocs qui nous permet de déterminer quels blocs ont été transformés (création, destruction) ou inchangés (avec participation ou non) au cours de la méthode. De même, au niveau macro avec les modifications des systèmes de blocs. Cette description des méthodes de synthèse nous donne effectivement les objectifs atteints par la méthode tant en terme de structure (cycle, polycycle, fonction, etc.) que d'action (interchange fonctionnel, construction d'un monocycle, etc.).

L'exploitation de la description des méthodes de synthèse selon les objectifs atteints a été faite dans le cadre de l'extraction de connaissances des bases de données de réactions. Des expérimentations ont été réalisées sur des classes de méthodes importantes du point de

vue de la synthèse : les interchanges fonctionnels, d'une part, les méthodes de construction de monocycles, d'autre part. Les données issues de l'analyse du premier type de méthodes ont pu être traitées en les transformant en données de type textuel. Il s'ensuit une perte d'information d'ordre structurel. Les méthodes de construction de cycles ont fait l'objet d'un traitement particulier. La description adoptée se rapproche de celle des chimistes. Les structures c'est-à-dire les monocycles sont décrits selon leur taille et leur substitution (par exemple, un monocycle de taille 7 trois fois substitué en 1,2,5) tandis que les actions correspondent aux grands types de construction de cycles habituellement utilisés dans le langage de la chimie (une annulation d'une chaîne de taille 4 avec une chaîne de taille 3).

Ces deux expérimentations donnent des résultats satisfaisants par rapport aux connaissances du domaine. Cela nous permet d'envisager une intégration des règles extraites dans un système tel que REsyn ASSISTANT, sous forme d'heuristiques de recherche de solutions connues de problèmes semblables à un problème à résoudre. Ces heuristiques sont des guides potentiels de remémoration dans le raisonnement à partir de cas. De plus, la description des données nous a conduit à développer des méthodes de visualisation et de navigation dans les données au travers d'une nouvelle organisation de celles-ci. Grâce à l'arbre de classification engendré ainsi qu'aux règles extraites, nous pouvons effectuer des requêtes qui ne sont pas réalisables dans les bases de données de réactions existantes. Les nouveaux modes d'interrogation que nous proposons apportent une aide à la planification de synthèse en permettant un accès plus aisé aux méthodes correspondant à un certain objectif. Un des progrès notables est que la méthode développée permet à l'utilisateur de s'affranchir de l'indication de l'appariement atome-atome, la requête pouvant être posée en langage naturel aussi bien que de façon graphique en indiquant uniquement les blocs s'interchangeant. Enfin, nous avons des résultats intéressants sur le contenu des bases qui est resté assez méconnu jusqu'à présent.

Parmi les perspectives que nous envisageons, nous voudrions souligner les suivantes :

- Nous avons étudié les méthodes essentiellement au niveau méso. Il serait intéressant de prolonger ce travail en considérant les systèmes de blocs car ceux-ci sont des objectifs de synthèse plus spécifiques. Ceci implique un nouvel effort de modélisation car les systèmes ne peuvent être considérés exactement de la même façon que les blocs. Il est probable qu'un cycle d'une certaine taille sera construit différemment s'il est isolé ou inclus dans un système polycyclique.
- L'aspect stéréochimique devrait être pris en compte dans les futurs développement, ceci grâce à l'exploitation de la description explicite des stéréogènes. La comparaison des descripteurs de stéréogènes appariés va nous permettre de déterminer si au cours de la réaction on a une rétention ou une inversion de configuration.
- Le choix d'une méthode dépend essentiellement du contexte dans lequel elle doit et peut être appliquée. Le travail de généralisation des contextes doit se poursuivre. En parallèle, une base de connaissances des grandes méthodes de synthèse devrait être élaborée par des spécialistes de la synthèse. Le système sera à terme capable de reconnaître à quelle grande méthode connue un exemple d'utilisation peut être rattaché.
- L'exploitation des conditions réactionnelles mériterait d'être envisagée. Dans ce cadre,

il nous paraît évident que le traitement de ces données particulières nécessite la mise en place de modules de standardisation mais aussi de généralisation comme cela a été effectué dans [Vogel, 2000]. L'ajout des données sur les conditions réactionnelles à la description des méthodes que nous avons développée devrait donner des résultats très intéressants après traitement par les techniques de fouille.

Vers de nouveaux systèmes d'information

Au delà de l'aide à la structuration et l'interrogation des bases de données de réactions, nous nous intéressons aux systèmes d'information chimique. Il est indéniable que les systèmes d'information vont évoluer dans les années futures. L'augmentation de la quantité d'informations accessibles de façon électronique le montre bien. Selon la vision de B. R. Schatz [Schatz, url], *Internet* sera bientôt supplanté par *Interspace*, un réseau de connaissances interconnectant des collections de données gérées au sein de communautés scientifiques. Ceci suppose que chaque communauté aura pris en charge l'organisation, la constitution et la mise à jour de ses propres ensembles de données. Des projets tels que le projet ENCORE, projet d'encyclopédie multimédia de la chimie organique piloté par le professeur A. Krief, visent cet objectif qui ne peut être atteint qu'à partir d'un important effort de structuration. Cependant, la modélisation d'un domaine tel que la synthèse organique est une tâche difficile. Nous avons été confortés par ces travaux dans la nécessité du dialogue entre experts de la synthèse, modélisateurs et informaticiens. Les résultats que nous avons obtenus montrent que la démarche adoptée est productive et pourra servir de support à d'autres projets.

Annexe A

Quelques mots sur le langage de modélisation UML

UML est devenu un standard dans le domaine des méthodes de modélisation objet. Le but de ce court paragraphe est de présenté les quelques notions nécessaires à la compréhension des diagrammes de classes UML que nous avons construits pour illustrer notre propos.

Tout d'abord, qu'est-ce qu'un objet ? Un objet représente une entité matérielle ou conceptuelle du monde réel. C'est une entité autonome qui regroupe un ensemble de propriétés cohérentes et de traitements associés. Un objet possède une identité qui permet de le différencier des autres entités. Il est caractérisé par des attributs (propriétés) et son comportement est décrit par un certain nombre de méthodes (traitements). Un objet est une instance d'une classe. Une classe est l'abstraction d'un ensemble d'objets qui possèdent une structure commune (c'est-à-dire une liste d'attributs communs) et un même comportement (c'est-à-dire une liste de méthodes commune). L'objet `molécule de Taxol` est une instance de la classe `Molécule`. Construire un diagramme de classes du modèle conceptuel consiste à décrire les classes correspondant aux différents concepts que l'on veut faire apparaître dans le modèle. Les relations entre les classes :

- association entre classes : une association entre classes représente l'abstraction des liens qui peuvent exister entre les objets du monde réel. Une association est définit en lui attribuant les cardinalités et les rôles des classes participant. Par exemple, on sait que la molécule de Taxol agit sur les cancers ovariens. Si l'on généralise : une molécule donnée est susceptible d'agir sur aucune ou plusieurs pathologie. Ceci peut être représenté en créant une association entre la classe `Molécule` et la classe `Pathologie`. Dans cette association `Agit sur`, la classe molécule a le rôle `principe actif` et la classe `Pathologie` celui de `maladie`. Au niveau des cardinalités de l'association `Agit sur`, un objet de type `Molécule` peut agir sur 0 à n objets de type `Pathologie` tandis que l'on connaîtra 0 à n principe actif pour une maladie. Par convention, les notations suivantes sont employées : 0 à n sera noté `0..n` ou `0..*` ou `*`.
- agrégation entre classes : l'agrégation est une forme particulière d'association entre

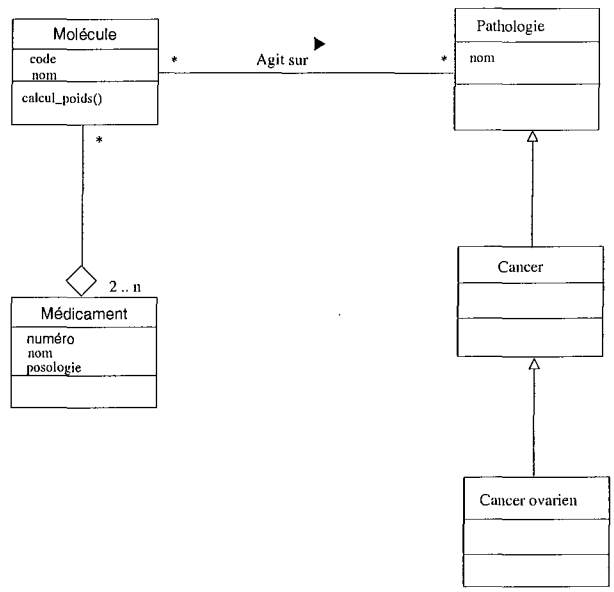


FIG. A.1 – Un exemple de diagramme de classes UML.

classes. Elle exprime le fait qu'une classe peut être composée d'une ou plusieurs autres classes. Par exemple, un objet de la classe `Médicament` est composé d'au moins un principe actif et d'au moins un excipient, qui sont des objets de type `Molécule`.

- généralisation/spécialisation entre classe : la généralisation consiste à factoriser dans une classe, appelée super-classe, les attributs et méthodes communes à un ensemble de classes. La spécialisation représente la démarche inverse qui consiste à créer à partir d'une classe plusieurs classes spécialisées. Ces opérations permettent de construire une hiérarchie de classes. Dans l'exemple, on pourrait par exemple créer une classe `Cancer ovarien` spécialisant la classe `Pathologie`. En fait, il existe une classe intermédiaire entre la classe `Pathologie` et la classe `Cancer ovarien`. On crée donc la classe `Cancer` qui généralise la classe `Cancer ovarien`.

Sur la figure A.1, nous avons représenté le diagramme de classes UML correspondant au modèle que nous avons décrit dans les exemples. Une relation d'association se symbolise par un segment de droite tiré entre les classes associées. Aux extrémités, on indique les cardinalités ainsi que les rôles des classes si besoin. Le nom de l'association donne la sémantique de celle-ci. La relation d'agrégation est représentée par un segment de droite allant de la classe « composant » à la classe « composée » et terminé par un losange. La relation d'héritage est symbolisée par une flèche allant d'une classe à sa super-classe.

Annexe B

Les fichiers de réactions au format MDL

Les données contenues dans les bases de réactions diffusées par la société MDL et accessibles via le système ISIS se présentent sous forme de fiches de réaction. Dans chaque fiche, les données sont organisées selon une hiérarchie de champs dont nous avons représenté une partie sur la figure B.1. Il y a des champs « parents » et des champs « fils ». Sur la figure B.2, on peut voir le résultat visualisable via le système ISIS. La partie de la fiche visualisée permet, entre autres, de connaître le schéma réactionnel (champ `rxnstructure`), le nom de la base d'où provient la fiche (champ `segid`), la référence bibliographique complète (champ `full_citation`), le rendement (champ `yield`) etc. Nous récupérons les données contenues dans les bases réactions par exportation dans des fichiers au format défini par MDL. Un ensemble de fiches de réactions seront stockées dans un fichier au format `RDFILE` (pour Reaction Data File) aussi appelé RDF. La structure de ce type de fichier consiste en une succession d'enregistrements correspondant chacun à une fiche de réaction. Les données sont décrites selon la définition données par la hiérarchie présentée figure B.1. La partie du fichier RDF correspondant à la fiche précédente est donné sur la figure B.3. Pour des raisons de place, nous n'avons conservé que les indications sur les champs que nous récupérons lorsque nous importons les données dans RESYN ASSISTANT. Tout fichiet RDF débute par une entête de deux lignes. Puis les informations sur les fiches se succèdent. Chaque enregistrement de fiche commence par une ligne `$RFMT $RIREG numéro de la fiche dans la base de données`. S'en suit les données correspondant au champ `rxnstructure`. Celles-ci se présentent sous la forme de la succession de tables de connexion. Ensuite, les champs textuels renseignés pour la fiche considérée sont reportés les uns à la suite des autres.

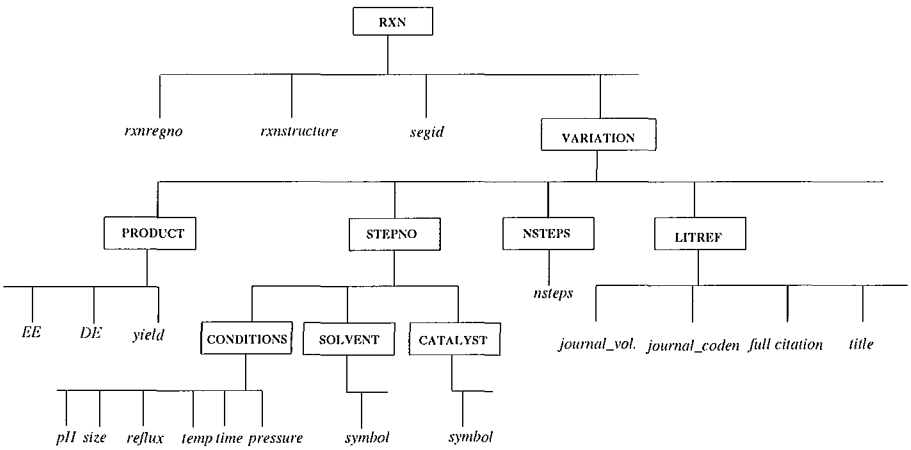


FIG. B.1 – L'organisation hiérarchique des données d'une fiche de réaction dans les bases ISIS.

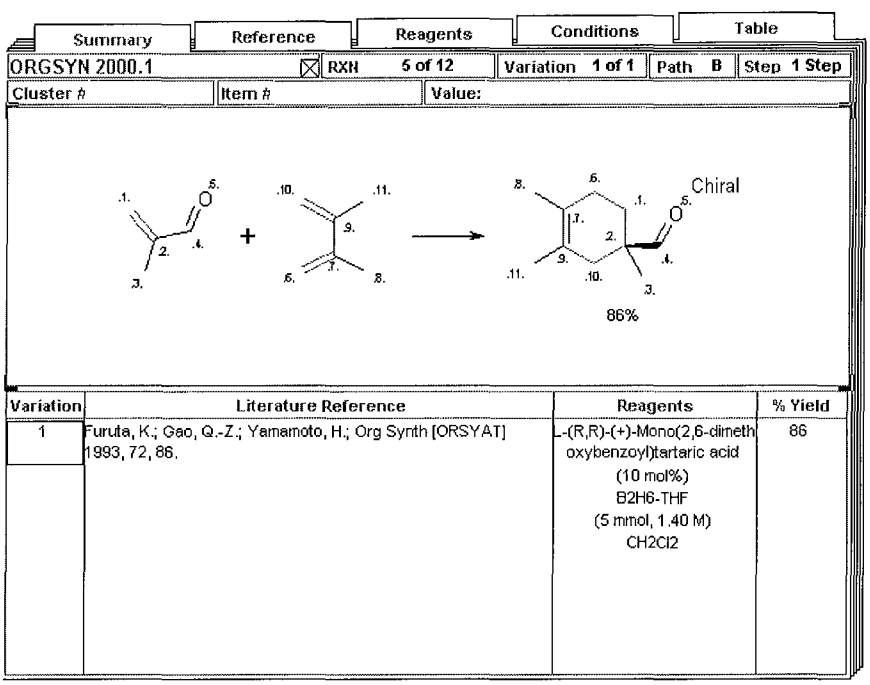


FIG. B.2 – Une fiche d'une méthode de réaction de la base de données ORGSYN.

```
$RDFILE 1
$DATM 11/14/2002 19:0:39
$RFMT $RIREG 61
$RXN
.
.
.
$DTYPE RXN:SEGID
$DATUM orgsyn
$RFMT $RIREG 619
$RXN

      ISIS     111420021900

  2  1
$MOL

  -ISIS-  11140219002D

  5  4  0  0  0  0  0  0  0  0999 V2000
   -9.5628   -2.3388    0.0000 C   0  0  0  0  0  0  0  0  0  2  0  0
   -8.0518   -2.0591    0.0000 C   0  0  0  0  0  0  0  0  0  4  0  0
  -10.0349   -3.8059    0.0000 C   0  0  0  0  0  0  0  0  0  3  0  0
  -10.6609   -1.2160    0.0000 C   0  0  0  0  0  0  0  0  0  1  0  0
   -7.4056   -0.6565    0.0000 O   0  0  0  0  0  0  0  0  0  5  0  0
  1  2  1  0  0  0  2
  1  3  1  0  0  0  2
  1  4  2  0  0  0  8
  2  5  2  0  0  0  2
M  END
$MOL

  -ISIS-  11140219002D

  6  5  0  0  0  0  0  0  0  0999 V2000
   -2.9979   -3.0500    0.0000 C   0  0  0  0  0  0  0  0  0  7  0  0
   -2.9979   -1.5314    0.0000 C   0  0  0  0  0  0  0  0  0  9  0  0
   -1.6400   -3.7657    0.0000 C   0  0  0  0  0  0  0  0  0  8  0  0
   -4.3355   -3.7880    0.0000 C   0  0  0  0  0  0  0  0  0  6  0  0
   -1.6928   -0.7407    0.0000 C   0  0  0  0  0  0  0  0  0 11  0  0
   -4.3965   -0.6960    0.0000 C   0  0  0  0  0  0  0  0  0 10  0  0
  1  2  1  0  0  0  8
  1  3  1  0  0  0  2
  1  4  2  0  0  0  8
  2  5  1  0  0  0  2
  2  6  2  0  0  0  8
M  END
$MOL

  -ISIS-  11140219002D

 11 11  0  0  1  0  0  0  0  0999 V2000
    7.7962   -2.7115    0.0000 C   0  0  1  0  0  0  0  0  0  2  0  0
    6.4533   -3.4853    0.0000 C   0  0  0  0  0  0  0  0  0 10  0  0
    7.7962   -1.1610    0.0000 C   0  0  0  0  0  0  0  0  0  1  0  0
    9.3467   -2.7115    0.0000 C   0  0  0  0  0  0  0  0  0  4  0  0
    8.5730   -4.0502    0.0000 C   0  0  0  0  0  0  0  0  0  3  0  0
    5.1134   -2.7115    0.0000 C   0  0  0  0  0  0  0  0  0  9  0  0
    6.4533   -0.3902    0.0000 C   0  0  0  0  0  0  0  0  0  6  0  0
   10.1205   -1.3686    0.0000 O   0  0  0  0  0  0  0  0  0  5  0  0
    5.1134   -1.1610    0.0000 C   0  0  0  0  0  0  0  0  0  7  0  0
    3.7748   -3.4853    0.0000 C   0  0  0  0  0  0  0  0  0 11  0  0
    3.7748   -0.3902    0.0000 C   0  0  0  0  0  0  0  0  0  8  0  0
  1  2  1  0  0  0  4
  1  3  1  0  0  0  8
  1  4  1  1  0  0  2
  1  5  1  0  0  0  2
  2  6  1  0  0  0  8
  3  7  1  0  0  0  4
  4  8  2  0  0  0  2
  6  9  2  0  0  0  8
  6 10  1  0  0  0  2
  9 11  1  0  0  0  2
  7  9  1  0  0  0  8
M  END
```

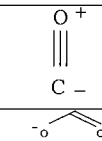
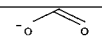
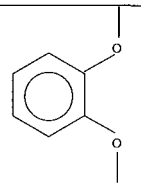
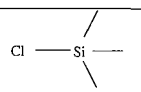
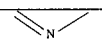
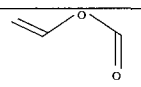
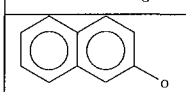
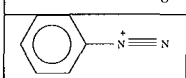
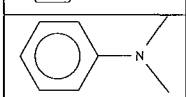
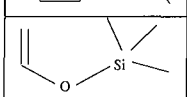
```
$DTYPE RXN:RXNREGNO
$DATUM 619
$DTYPE RXN:VARIATION(1):VARIATION_NO
$DATUM 1
$DTYPE RXN:VARIATION(1):RXNREF(1):EXTREG
$DATUM 720086
$DTYPE RXN:VARIATION(1):RXNREF(1):PATH
$DATUM B
$DTYPE RXN:VARIATION(1):RXNREF(1):STEP
$DATUM 1 Step
.
.
$DTYPE RXN:VARIATION(1):LITREF(1):FULL_CITATION
$DATUM Furuta, K.; Gao, Q.-Z.; Yamamoto, H.; Org Synth [ORSYAT]
1993, 72, 86.
.
.
$DTYPE RXN:VARIATION(1):PRODUCT(1):PRODUCT_NO
$DATUM 1
$DTYPE RXN:VARIATION(1):PRODUCT(1):REGNO
$DATUM 2523
$DTYPE RXN:VARIATION(1):PRODUCT(1):YIELD
$DATUM 8.600000000000000e+001
.
.
$DTYPE RXN:VARIATION(1):STEPNO(1):CATALYST(1):MOL(1):SYMBOL(1):SYMBOL
$DATUM L-(R,R)-(+)-Mono(2,6-dimethoxybenzoyl)tartaric acid
.
.
$DTYPE RXN:VARIATION(1):STEPNO(1):SOLVENT(1):MOL(1):SYMBOL(1):SYMBOL
$DATUM CH2Cl2
.
.
$DTYPE RXN:VARIATION(1):STEPNO(1):CONDITIONS(1):ATMOSPHERE
$DATUM Ar
$DTYPE RXN:VARIATION(1):STEPNO(1):CONDITIONS(1):PRESSURE
$DATUM 1.000000000000000e+000
$DTYPE RXN:VARIATION(1):STEPNO(1):CONDITIONS(1):SIZE
$DATUM 1.000000000000000e-001
$DTYPE RXN:VARIATION(1):WORKUP
$DATUM Aqueous extraction, Drying, Filtration, Evaporation, Distillation
$DTYPE RXN:VARIATION(1):COMMENTS
$DATUM The procedure provides a simple and general method for
the construction of optically active 3-cyclohexene-1-carboxaldehydes.
The present method is characterized as a true catalytic process
with good chemical and high optical yields, simple operation, preparation
of both enantiomers with equal ease, and the ready availability
of the starting materials.For quantitative conversion to the
corresponding acetal, see litref(1).
$DTYPE RXN:VARIATION(1):KEYPHRASES
$DATUM Add_hydrocarbon_group(AHC). Asymmetric_synthesis(ASY).
Chiral (Acyloxy)Borane Complex-Catalyzed Asymmetric Diels-Alder.
Cycloaddition(CYA). Ring_closure(RCL). Stereoselective(SEL).
.
.
$DTYPE RXN:SEGID
$DATUM orgsyn
$RFMT $RIREG 688
$RXN
.
.
.
$DTYPE RXN:SEGID
$DATUM orgsyn
```

FIG. B.3 – Extrait d'un fichier au format RDF de MDL.

Annexe C

Les fonctions « inconnues »

Le tableau ci-dessous présente les fonctions non présentes dans la base de connaissances du prototype ReSyn Assistant. Nous avons sélectionné ici les 10 fonctions qui apparaissent au moins 50 fois dans la base Orgsyn à titre indicatif. Les fonctions « inconnues » ont été nommées par Gilles Niel.

	carbonyl ligand	253
	carbonate ion	137
	dialkylphtalate	109
	chlorosilane	75
	secondary imine	74
	vinyloxycarbonyl	72
	2-naphtol	66
	phenyl diazonium	57
	tertiary aniline	54
	silyl enolether	50

Bibliographie

[Accelrys, url] Accelrys (url). http ://www.accelrys.com/. — *Cité page(s) 28.*

[Al Hulou et al., 1999] R. Al Hulou, A. Napoli, and E. Nauer (1999). Objets semi-structurés, classes polythétiques et classification. In *Actes des Septièmes journées de la Société Francophone de Classification, SFC'99*, pages 379–384, Nancy. F. Le Ber and J.F. Mari and A. Napoli and A. Simon. — *Cité page(s) 71.*

[Arhonovitz, 1999] Y. Arhonovitz (1999). Modélisation de chemins de synthèse. LIRMM. Rapport de Recherche 99091. — *Cité page(s) 122.*

[Aussenac-Gilles et al., 1992] N. Aussenac-Gilles, J. P. Krivine, and J. Sallantin (1992). L'acquisition de connaissances pour les systèmes à base de connaissances. *Revue d'intelligence artificielle*, 6(1-2) :7–18. — *Cité page(s) 63.*

[Baldy, 1999] A. Baldy (1999). Titane : serveur national des bases de données en chimie. *L'actualité chimique.* — *Cité page(s) 27.*

[Barnard, 1991] J. M. Barnard (1991). *Structure representation and searching*, pages 9–56. Chichester. — *Cité page(s) 140.*

[Barnard and Walkowiak, 1998] J. M. Barnard and D. Walkowiak (1998). *Computer systems for substructure searching*, pages 55–72. American Chemical Society, Washington, DC. — *Cité page(s) 35.*

[Bastide, 2000] Y. Bastide (2000). *Data mining : algorithmes par niveaux, techniques d'implantation et applications.* Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand. — *Cité page(s) 76, 83, 84, 86, 87, 190.*

[Bastide et al., 2000] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal (2000). Mining frequent patterns with counting inference. *ACM SIGKDD Explorations*, 2(2) :66–75. — *Cité page(s) 83.*

[Bastide et al., 2002] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal (2002). PASCAL : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1) :65–95. — *Cité page(s) 83.*

[Beach et al., 1979] A. J. Beach, H. F. Dabek, J. R. Hosansky, and N. L. Hosansky (1979). Chemical reactions information retrieval from chemical abstracts service publications and services. *J. Chem. Inf. Comput. Sci.*, 19(3) :149–155. — *Cité page(s) 26.*

[Beilstein, url] Beilstein (url). http ://www.beilstein.com. — *Cité page(s) 28, 32.*

[Benitez-Guerrero et al., 2001] E. Benitez-Guerrero, C. Collet, and M. Adiba (2001). Entrepôts de données : caractéristiques et problématique. *Technique et science informatiques*, 20(2) :145–178. — *Cité page(s) 68.*

[Berasaluce et al., 2002] S. Berasaluce, C. Laurenço, and A. Napoli (2002). Extraction de connaissances à partir de bases de données de réactions en chimie organique. In *13èmes journées francophones d'Ingénierie des Connaissances (IC'2002)*, pages 151–162, Rouen. — *Cité page(s) 173.*

[Berge, 1973] C. Berge (1973). *Graphes et Hypergraphes.* Dunod, Paris. — *Cité page(s) 128.*

[Bersohn and Esack, 1976] N. Bersohn and A. Esack (1976). Computer and organic synthesis. *Chemical reviews*, 76(2) :269–282. — *Cité page(s) 21.*

[Blake and Dana, 1990] J. E. Blake and R. C. Dana (1990). CASREACT : more than a million reactions. *J. Chem. Inf. Comput. Sci.*, 30 :394–399. — *Cité page(s) 28, 39.*

[Blower et al., 1997] P. E. Blower, G. J. Myatt, and M. W. Petras (1997). Exploring functional group transformations on CASREACT. *J. Chem. Inf. Comput. Sci.*, 37 :54–58. — *Cité page(s) 181, 183.*

[Booch et al., 1999] G. Booch, J. Rumbauch, and I. Jacobson (1999). *The Unified Modeling Langage user guide.* Oject Technology Series. Addison-Wesley. — *Cité page(s) 62.*

[Boulicaut et al., 2000] J. F. Boulicaut, A. Bykowski, and C. Rigotti (2000). Approximation of frequency queries by means of free-sets. In D. A. Zighed, H. J. Komorowski, and J. M. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings*, volume 1910 of *Lecture Notes in Computer Science*, pages 75–85. Springer. — *Cité page(s) 84.*

[Brachman and Anand, 1996] R. J. Brachman and T. Anand (1996). The process of knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 37–57. AAAI Press / MIT Press, Menlo Park, California. — *Cité page(s) 72.*

[Brakel, 1997] J. Van Brakel (1997). Chemistry as the science of the transformation of substances. *Synthese*, 111 :253–282. — *Cité page(s) 98.*

[Bush, 1945] V. Bush (1945). As we may think. *The Atlantic Monthly.* — *Cité page(s) 3.*

[Cahn et al., 1966] R. S. Cahn, C. K. Ingold, and V. Prelog (1966). Specification of molecular chirality. *Angew. Chem. Int. Ed. Engl.*, 5 :385–415. — *Cité page(s) 143.*

[CAS, url] CAS (url). http ://www.cas.org/. — *Cité page(s) 28.*

[CASREACT, url] CASREACT (url). http ://www.cas.org/casfiles/casreact.html. — *Cité page(s) 28, 32.*

[Cayley, 1874] A. Cayley (1874). On the mathematical theory of isomers. *Philosophical Magazine*, 47 :444–447. — *Cité page(s) 25.*

[Chandrasekaran and Johnson, 1993] B. Chandrasekaran and T. R. Johnson (1993). *Generic tasks and task structures : history, critique and new directions*, pages 232–272. Springer-Verlag. — *Cité page(s) 59.*

[Chandrasekaran et al., 1999] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins (1999). What are ontologies, and why do we need then. *IEEE Intelligent Systems*, pages 20–26. — *Cité page(s) 58.*

[Charlet et al., 2000] J. Charlet, M. Zacklad, G. Kassel, and D. Bourigault (2000). Ingénierie des connaissances : recherches et perspectives. In J. Charlet, M. Zaclad, G. Kassel, and D. Bourigault, editors, *Ingénierie des connaissances. Evolutions récentes et nouveaux défis*, pages 1–22. Eyrolles. — *Cité page(s) 54, 55, 56.*

[Chatterjee et al., 1957] A. Chatterjee, C. Chatterjee, and B. K. Bhattacharyya (1957). Synthesis of 5-keto-9-methoxy-i :2 :3 :4 :4a :5 :6 :6a :iia :iib-decahydroxychrysofluorene. *Jour. Indian Chem. Soc.*, 34(12) :855–858. — *Cité page(s) 13.*

[Cherfi and Toussaint, 2002] H. Cherfi and Y. Toussaint (2002). Interprétation des règles d'association extraites par un processus de fouille de texte. In *RFIA 2002.* — *Cité page(s) 85, 86.*

[Chevreau et al., 2001] H. Chevreau, F. Fuster, and B. Silvi (2001). La liaison chimique : mythe ou réalité ? les méthodes topologiques de description de la liaison. *L'Actualité Chimique*, pages 15–22. — *Cité page(s) 98.*

[Chittimoori et al., 1999] R. Chittimoori, L. Holder, and D. Cook (1999). Applying the subdue substructure discovery system to the chemical toxicity domain. In *Proceedings of the Twelfth International Florida AI Research Society Conference*, pages 90–94. — *Cité page(s) 190.*

[Corey, 1967] E. J. Corey (1967). General methods for the construction of complex molecules. *Pure & Appl. Chem.*, 14 :19–37. — *Cité page(s) 41.*

[Corey, 1971] E. J. Corey (1971). Computer-assisted analysis of complex synthetic problems. *Quart. Rev. Chem. Soc.*, 25 :455–482. — *Cité page(s) 19.*

[Corey and Cheng, 1989] E. J. Corey and X.-M. Cheng (1989). *The logic of chemical synthesis.* Wiley-InterScience, New York. — *Cité page(s) 8.*

[Corey et al., 1985] E. J. Corey, K. A. Long, and S. D. Rubenstein (1985). Computer-assisted analysis in organic synthesis. *Science*, 228 :408–418. — *Cité page(s) 20, 41.*

[Coste, 2002] J. Coste (2002). L'information en chimie. notes de cours sur les systèmes d'information chimique disponibles dans le système académique français. — *Cité page(s) 26, 30, 32.*

[Coste et al., 1999] J. Coste, O. Gien, A. Dietz, and C. Laurenço (1999). A propos de l'utilisation des bases de données de réactions. *L'actualité chimique*, pages 27–32. — *Cité page(s) 27, 37, 39.*

[David, 1995] J. M. David (1995). Les systèmes experts de seconde génération ou de l'importance de la modélisation dans la construction de systèmes à base de connaissances.

Techniques et Sciences Informatiques, 14(4) :435–471. — *Cité page(s) 54, 59, 61, 64, 65.*

[Daylight, url] Daylight (url). http ://www.daylight.com. — *Cité page(s) 28.*

[Dehaspe et al., 1998] L. Dehaspe, H. Toivonen, and R. D. King (1998). Finding frequent substructures in chemical compounds. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *4th International Conference on Knowledge Discovery and Data Mining*, pages 30–36. AAAI Press. — *Cité page(s) 190.*

[Deslongchamps, 1984] P. Deslongchamps (1984). Le concept de stratégie en synthèse organique. *Bull. Soc. Chim. Fr.*, II :349–361. — *Cité page(s) 11.*

[Dietz, 1995] A. Dietz (1995). Yet another representation of molecular structure. *J. Chem. Inf. Comput. Sci.*, 35(5) :787–802. — *Cité page(s) 145.*

[Dietz et al., 1997] A. Dietz, J. Coste, and C. Laurenço (1997). Computer-oriented representation of stereoselective synthetic methods. *Bull. Soc. Chim. Fr.*, 134 :669–676. — *Cité page(s) 164.*

[Dietz et al., 2000] A. Dietz, C. Fiorio, M. Habib, and C. Laurenço (2000). Representation of stereochemistry using combinatorial maps. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 51 :117–128. — *Cité page(s) 145.*

[Ducournau, 1989] R. Ducournau (1989). *Y3. Langage à objets, Manuel de référence.* Sema Group, Montrouge. — *Cité page(s) 47.*

[Duribreux-Cocquebert and Houriez, 2000] M. Duribreux-Cocquebert and B. Houriez (2000). Application industrielle d'une approche mixte de modélisation des connaissances. In J. Charlet, M. Zacklad, G. Kassel, and D. Bourigault, editors, *Ingénierie des connaissances — Évolutions récentes et nouveaux défis*, pages 357–369. Eyrolles, Paris. — *Cité page(s) 62.*

[Dury et al., 2001] L. Dury, T. Latour, L. Leherte, F. Barberis, and D. P. Vercauteren (2001). A new graph descriptor for molecules containing cycles. application as screening criterion for searching molecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.*, 41 :1437–1445. — *Cité page(s) 35.*

[Eiben and Schoenauer, 2002] A. E. Eiben and M. Schoenauer (2002). Evolutionary computing. *Information Processing Letters*, 82 :1–6. — *Cité page(s) 89.*

[Ermine, 2000] J. L. Ermine (2000). La gestion des connaissances, un levier stratégique pour les entreprises. In *IC'2000.* http ://www.irit.fr/IC2000/ACTES/texte-ermine.html. — *Cité page(s) 65.*

[Euzenat, 1998] J. Euzenat (1998). Représentation de connaissance par objet. In R. Ducournau, J. Euzenat, G. Masini, and A. Napoli, editors, *Langages et modèles à objets. État des recherches et perspectives.*, Collection didactique, chapter 10, pages 293–319. INRIA. — *Cité page(s) 113.*

[Failed Reactions, url] Failed Reactions (url). http ://www.accelrys.com/chem_db/failedreact.hti — *Cité page(s) 39.*

[Famili et al., 1997] A. Famili, W. M. Shen, R. Weber, and E. Simoudis (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1 :3–23. — *Cité page(s) 72.*

[Fayyad et al., 1996] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smith (1996). From data mining to knowledge discovery : An overview. In *Advances in knowledge discovery and data mining*. AAIT/MIT Press. — *Cité page(s) 73.*

[Fujita, 1988] S. Fujita (1988). Canonical numbering abd coding reaction center graphs and reduced reaction center graphs. a novel approach to the linear coding of individual organic reactions types. *J. Chem. Inf. Comput. Sci.*, 28 :137–142. — *Cité page(s) 154.*

[Funatsu et al., 1988] K. Funatsu, T. Endo, N. Kotera, and S. I. Sasaki (1988). Automatic recognition of reaction sites in organic chemical reactions. *Tetrahedron Computer Methodology*, 1(1) :53–69. — *Cité page(s) 30.*

[Ganter and Rudolph, 2001] B. Ganter and S. Rudolph (2001). 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. Springer-Verlag. — *Cité page(s) 190.*

[Garey and Johnson, 1979] M. R. Garey and D. S. Johnson (1979). *Computers and Intractability : A guide to the theory of NP-completness*. W. H. Freeman, San Francisco. — *Cité page(s) 35.*

[Gasteiger et al., 1992] J. Gasteiger, W. D. Ihlenfeldt, and P. Röse (1992). A collection of computer methods for synthesis design and reaction prediction. *Recl. Trav. Chim., Pays-Bas*, 111 :270–290. — *Cité page(s) 45.*

[Gelernter et al., 1990] H. Gelernter, J. R. Rose, and C. Chen (1990). 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(4) :492–504. — *Cité page(s) 42.*

[Gennari et al., 1989] J. H. Gennari, P. Langley, and D. Fisher (1989). Models of incremental concept formation. *Artificial Intelligence*, 40 :11–61. — *Cité page(s) 89.*

[Gien, 1998] O. Gien (1998). *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. — *Cité page(s) 21, 40, 48, 97, 118.*

[Gilleron and Tommasi, 2000] R. Gilleron and M. Tommasi (2000). Découverte de connaissances à partir de données. Notes de cours correspondant à une partie d'un cours donné en IUP MIAGE troisième année à Lille 1 et en Maîtrise MASS à Lille 3. — *Cité page(s) 68.*

[Goodman, 2000] J. M. Goodman (2000). Solutions for chemistry : synthesis of experiment and calculation. *Phil. Trans. R. Soc. Lond. A.*, pages 387–398. — *Cité page(s) 8.*

[Greeno, 1978] J. G. Greeno (1978). Handbook of learning and cognition processes. Lawrence Erlbaum Associates. — *Cité page(s) 12.*

[Grethe, 1995] G. Grethe (1995). From hardcopy to electronic reaction classification and beyond : increasing the effectiveness of reaction retrieval for the synthetic chemist. In *The 1995 International Chemical Information Conference*, pages 40–47, England. — *Cité page(s) 36.*

[Grethe and Moock, 1990] G. Grethe and T. E. Moock (1990). Similarity searching in REACCS. a new tool for the synthetic chemist. *J. Chem.inf. Comput. Sci.*, pages 511–520. — *Cité page(s) 36.*

[Guillaume, 2000] S. Guillaume (2000). *Traitement de données volumineuses. Meures et algorithmes d'extraction de règles d'association et règles ordinales.* Thèse de doctorat, Université de Nantes. — *Cité page(s) 86.*

[Hanessian, 1983] S. Hanessian (1983). *Total synthesis of natural products : the 'Chiron' Approach.* Organic chemistry series. Pergamon press, London. — *Cité page(s) 8, 22.*

[Hanser et al., 1995] T. Hanser, P. Jauffret, E. Gruber, J. F. Marchaland, L. Ellermann, T. Fang, and G. Kaufmann (1995). CGS : condensed graph of synthesis, a new computer representation for molecules, reactions and synthesis. In *AIP Conference Proceedings, Computational Chemistry*, pages 575–581. — *Cité page(s) 155.*

[Hassner and Stumer, 1994] A. Hassner and C. Stumer (1994). *Organic Syntheses Based on Name Reactions and Unnamed Reaction.* Elsevier Science Ltd. — *Cité page(s) 112.*

[Heller, 1998] S. R. Heller, editor (1998). *The Beilstein System. Strategy for effective searching.* American Chemical Society, Washington, DC. — *Cité page(s) 28.*

[Hendrickson and Sander, url] J. B. Hendrickson and T. L. Sander (url). http ://www.webreactions.com. — *Cité page(s) 27.*

[Hendrickson and Parks, 1992] J. B. Hendrickson and C. A. Parks (1992). A program for the forward generation of synthetic routes. *J. Chem. Inf. Comput. Sci.*, 32 :209–215. — *Cité page(s) 22.*

[Hereth et al., 2000] J. Hereth, G.Stumme, R. Wille, and U. Wille (2000). Conceptual knowledge discovery and data analysis. In *Proceedings of the 8th International Conference of Conceptual Structures (ICCS'2000)*, pages 421–437, Montpellier, France. — *Cité page(s) 80.*

[Herges, 1988] R. Herges (1988). Reaction planning : computer-aided reaction design. *Tetrahedron Computer Methodology*, 1(1) :15–25. — *Cité page(s) 44.*

[Hopkinson et al., 1990] G. A. Hopkinson, T. P. Cook, and I. P. Buchan (1990). Computer treatment of chemical reactions and synthetic problems. In D. Bawden andE. Mitchell, editor, *Chemical Information Systems Beyond the Structure Diagram*, pages 83–91. Ellis Horwood. — *Cité page(s) 141.*

[Inokuchi et al., 2000] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery*, pages 13–23. — *Cité page(s) 190.*

[Ireland, 1969] R. E. Ireland (1969). *Organic Synthesis.* Prentice-Hall, Inc., N. J. — *Cité page(s) 97.*

[IUPAC, 1989] IUPAC (1989). Terminology for organic chemical transformations. *Pure Appl. Chem.*, 61 :725–768. — *Cité page(s) 19.*

[IUPAC, 1994] IUPAC (1994). Recommendation 1994, Glossary of terms used in physical organic chemistry. — *Cité page(s) 19, 193.*

[Jambaud, 1996] P. Jambaud (1996). *Le dialogue comme processus de résolution de problème. Une application en chimie organique.* Thèse de doctorat, Université Montpellier II. — *Cité page(s) 41, 48.*

[Johnson, 1988] A. P. Johnson (1988). Reaction indexing : an overview of current approaches. In W. A. Warr, editor, *Chemical structures. The international langage of chemistry.*, pages 297–301. Springer-Verlag. — *Cité page(s) 27.*

[REACCS-JSM, url] REACCS-JSM (url). http ://www.derwent.com/. — *Cité page(s) 173.*

[Kahn et al., 1985] G. Kahn, S. Nowlan, and J. McDermott (1985). Strategies for knowledge acquisition. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-7(5) :511–522. — *Cité page(s) 59, 60.*

[Kahney, 1986] H. Kahney (1986). *Problem solving : a cognitive approach.* Open Guides to Psychology. Open University Press, Philadelphia. — *Cité page(s) 10, 15.*

[Kayser, 1997] D. Kayser (1997). *La représentation des connaissances.* Hermes, Paris. — *Cité page(s) 54.*

[Kocovský et al., 1990] P. Kocovský, F. Turecek, and J. Hájícek (1990). *Synthesis of natural products : problems of stereoselectivity*, volume 1. CRC Press, Boca Raton, Florida, USA. — *Cité page(s) 191.*

[Kodratoff, 1991] Y. Kodratoff (1991). Faut-il choisir entre science des explications et science des nombres ? In Y. Kodratoff and E. Diday, editors, *Induction symbolique et numérique à partir de données*, volume 1, pages 193–239. Cépaduès-Éditions. — *Cité page(s) 76.*

[Kodratoff, 2001] Y. Kodratoff (2001). Comparing machine learning and knowledge discovery in databases : an application to knowledge discovery in texts. In *Machine learning and its applications*, Lecture Notes in Artificial Intelligence 2049, pages 1–22. Springer-Verlag. — *Cité page(s) 86.*

[Kuramochi and Karypis, 2002] M. Kuramochi and G. Karypis (2002). An efficient algorithm for discovering frequent subgraphs. Technical report, Department of computer science/Army HPC Research Center, University of Minnesota. Technical Report 02-026. — *Cité page(s) 190.*

[Laurenço, 1998] C. Laurenço (1998). Traitement informatique de la connaissance en chimie organique. LIRMM. rapport d'activité. — *Cité page(s) 60, 123.*

[Laurenço, 1985] C. Laurenço (1985). *Synthèse Organique Assistée par Ordinateur.* Thèse de doctorat d'État, Université Louis Pasteur (Strasbourg I). — *Cité page(s) 40.*

[Laurenço et al., 1990] C. Laurenço, M. Py, A. Napoli, J. Quinqueton, and B. Castro (1990). Représentation de connaissance en synthèse organique à l'aide d'un langage à objets. *New J. Chem.*, 14(12) :921–931. — *Cité page(s) 47, 94, 155, 160.*

[Laurenço, 1998] C. Laurenço (1998). Synthèse organique et informatique. In *8_mes Journées Informatique et Pédagogie des Siences Physiques*, Montpellier. http ://www.inrp.fr/Tecne/Rencontre/Jipsp8/Pdf/Lauren1.pdf. — *Cité page(s) 30, 32.*

[Laurenço et al., 2002] C. Laurenço, S. Berasaluce, A. Napoli, and G. Niel (2002). Knowledge extraction from organic reaction databases. Sino-french workshop, Molecular informatics and drug design. Shangai, P. R. China. — *Cité page(s) 191.*

[Lawson, 1998] A. J. Lawson (1998). *CrossFireplusReactions*, pages 73–97. American Chemical Society, Washington, DC. — *Cité page(s) 28.*

[Lebbe, 1998] J. Lebbe (1998). Représentations par objets et classifications biologiques. In R. Ducournau, J. Euzenat, G. Masini, and A. Napoli, editors, *Langages et modèles à objets. État des recherches et perspectives.*, Collection didactique, chapter 15, pages 421–447. INRIA. — *Cité page(s) 87.*

[Lebbe and Vignes, 1991] J. Lebbe and R. Vignes (1991). Génération de graphes d'identification à partir de description de concepts. In Y. Kodratoff and E. Diday, editors, *Induction symbolique et numérique à partir de données*, volume 1, pages 193–239. Cépaduès-Éditions. — *Cité page(s) 87, 88.*

[Leblanc, 2000] H. Leblanc (2000). *Sous-hiérarchie de Galois : un modèle pour la construction et l'évolution des hiérarchies d'objets.* Thèse de doctorat, Université de Montpellier II. — *Cité page(s) 80.*

[Lehn, 1996] J. M. Lehn (1996). L'enseignement de la chimie. *Pour la science*, (222) :25–26. — *Cité page(s) 5, 8.*

[Lehn, 1998] J. M. Lehn (1998). L'intelligence du chimiste. *Pour la science*, (254) :187–189. — *Cité page(s) 8.*

[Lieber, 1997] J. Lieber (1997). *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). — *Cité page(s) 48, 124, 125.*

[Lieber and Napoli, 2000] J. Lieber and A. Napoli (2000). Planification à partir de cas et classification. In J. Charlet, M. Zacklad, G. Kassel, and D. Bourigault, editors, *Ingénierie des connaissances — Évolutions récentes et nouveaux défis*, pages 357–369. Eyrolles, Paris. — *Cité page(s) 48.*

[Luckenbach, 1998] R. Luckenbach (1998). *The Beilstein Handbook*, pages 13–28. American Chemical Society, Washington, DC. — *Cité page(s) 26.*

[Lynch and Willett, 1978] M. F. Lynch and P. Willett (1978). The automatic detection of chemical reaction sites. *J. Chem. Inf. Comput. Sci.*, 18(3). — *Cité page(s) 30.*

[March, 1992] J. March (1992). *Advanced organic chemistry. Reactions, mechanisms, and structure.* Wiley-Interscience, New York, USA, 4th edition. — *Cité page(s) 184, 188.*

[Martin-Mattei and Ricard, 1992] C. Martin-Mattei and B. Ricard (1992). Modéliser des connaissances pour les applications d'aide au diagnostic. *Revue d'Intelligence Artificielle*, 6 :157–194. — *Cité page(s) 62.*

[Masseglia, 2002] F. Masseglia (2002). *Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données.* PhD thesis, Université de Versailles St-Quentin en Yvelines. — *Cité page(s) 84.*

[McDermott, 1988] J. McDermott (1988). Preliminary Steps Towards a Taxonomy of Problem Solving Methods. In S. Marcus, editor, *Automatic Knowledge Acquisition for Experts*, pages 225–295. Kluwer Academic Publishers, Boston. — *Cité page(s) 59.*

[McGraw and Harbison-Briggs, 1989] K. McGraw and K. Harbison-Briggs (1989). Knowledge acquiqition for expert systems. In *Knowledge acquisition, principles and guidelines*, pages 1–27. Prentice-Hall International Editions. http ://www.scism.sbu.ac.uk/inmandw/review/knowacq/review/rev15561.html. — *Cité page(s) 65.*

[McGregor and Willett, 1981] J. J. McGregor and P. Willett (1981). Use of a maximal common subgraph algorithm in the automatic identification of the ostensible bond changes occuring in chemical reactions. *J. Chem. Inf. Comput. Sci.*, 21 :137–140. — *Cité page(s) 30.*

[MDL, url] MDL (url). http ://www.mdli.com/. — *Cité page(s) 28.*

[MDL CHIME, url] MDL CHIME (url). http ://www.mdli.com/chime/. — *Cité page(s) 212.*

[Medler, 1998] D. A. Medler (1998). A brief history of connectionism. *Neural Computing Survey*, 1 :61–101. disponible à l'adresse http ://www.icsi.berkeley.edu/ jagota/NCS/vol1.html. — *Cité page(s) 89.*

[Métivier, 1987] P. Métivier (1987). *Synthèse Organique Assistée par Ordinateur : L'approche synthétique.* thèse de doctorat, Université Louis Pasteur (Strasbourg I). — *Cité page(s) 45, 139.*

[Mockus and Stobaugh, 1980] J. Mockus and R. E. Stobaugh (1980). The chemical abstracts service chemical registry system. vii tautomerism and alternating bonds. *J. Chem. Inf. Comput. Sci.*, 20 :18–22. — *Cité page(s) 139, 140, 141.*

[Moock et al., 1988a] T. E. Moock, D. L. Grier, W. D. Hounshell, G. Grethe, K. Gronin, and J. G. Nourse J. Theodosiou (1988a). Similarity searching in the organic reaction domain. *Tetrahedron Computer Methodology*, 1(2) :117–128. — *Cité page(s) 35.*

[Moock et al., 1988b] T. E. Moock, D. L. Grier, W. D. Hounshell, and J. G. Nourse (1988b). The implementation of atom-atom mapping and related features in the reactions access system (REACCS). In W. A. Warr, editor, *Chemical structures. The international langage of chemistry.*, volume 1, pages 303–313. — *Cité page(s) 30.*

[Motta et al., 1990] E. Motta, T. Rajan, and M. Eisenstadt (1990). Knowledge acquisition as a process of model refinement. *Knowledge Acquisition*, 2 :21–49. — *Cité page(s) 66.*

[Musen, 1993] M. A. Musen (1993). *An overview of knowledge acquisition*, pages 405–427. Springer-Verlag. — *Cité page(s) 55, 57.*

[Napoli, 1992] A. Napoli (1992). *Représentations à objets et raisonnement par classification en intelligence artificielle.* Thèse de doctorat d'État, Université Nancy I. — *Cité page(s) 16, 60, 79, 132.*

[Napoli and Laurenço, 1993] A. Napoli and C. Laurenço (1993). 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. — *Cité page(s) 47.*

[Nauer, 2001] E. Nauer (2001). *Principes de conception de systèmes hypertextes pour la fouille de données bibliographiques multibases.* Thèse de doctorat, Université Nancy I. — *Cité page(s) 80.*

[Newell, 1982] A. Newell (1982). The knowledge level. *Artificial Intelligence*, 18(1) :87–127. — *Cité page(s) 57, 58.*

[Newell and Simon, 1972] A. Newell and H. A. Simon (1972). *Human Problem Solving.* Prentice-Hall, Londres. — *Cité page(s) 13.*

[Nicolaou et al., 1994] K. C. Nicolaou, Z. Yang, J. J. Liu, H. Ueno, P. G. Nantermet, R. K. Guy, C. F. Claiborne, J. Renaud, E. A. Couladouros, K. Paulvannan, and E. J. Sorensen (1994). Total synthesis of Taxol. *Nature*, 367 :630–634. — *Cité page(s) xii, 95.*

[Niel et al., 2002a] G. Niel, S. Berasaluce, S. Mouné, and C. Laurenço (2002a). Extraction de connaissances à partir de bases de données chimiques. 7^{eme} Symposium ICSN-CNRS - De la chimie à la biologie. Poster. — *Cité page(s) 191.*

[Niel et al., 2002b] G. Niel, S. Berasaluce, S. Mouné, and C. Laurenço (2002b). Knowledge extraction from chemical reaction databases. 16^{eme} Symposium franco-japonais de chimie fine et médicinale. Poster. — *Cité page(s) 191.*

[Noy et al., 2000] N. Fridman Noy, R. W. Fergerson, and M. A. Musen (2000). The knowledge model of PROTÉGÉ-2000 : combining interoperability and flexibility. In *2th International Conference on Knowledge Management (EKAW'2000)*, Juan-les-Pins, France. — *Cité page(s) 61.*

[Object Management Group, url] Object Management Group (url). http ://www.omg.org. — *Cité page(s) 62.*

[ORGSYN, url] ORGSYN (url). http ://www.orgsyn.org/. — *Cité page(s) 173.*

[Ott and Noordik, 1992] M. A. Ott and J. H. Noordik (1992). 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. — *Cité page(s) 26, 27.*

[Parkar and Parkin, 1999] F. A. Parkar and D. Parkin (1999). Comparaison of beilstein cross fire plus reactions and the selective reaction databases under isis. *J. Chem. Inf. Comput. Sci.*, 39 :281–288. — *Cité page(s) 27, 37, 38.*

[Petrarca et al., 1967] A. E. Petrarca, M. F. Lynch, and J. E. Rush (1967). A method for generating unique computer structural representations of stereoisomers. *Journal of Chemical Documentation*, 7(3) :154–165. — *Cité page(s) 148.*

[Polaillon, 1998] G. Polaillon (1998). *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme.* Thèse de doctorat, Université Paris IX-Dauphine. — *Cité page(s) 76, 80.*

[Polya, 1965] G. Polya (1965). *Comment poser et résoudre un problème.* Jacques Gabay. — *Cité page(s) 10, 14.*

[Poulovassilis and Levene, 1994] A. Poulovassilis and M. Levene (1994). A nested-graph model for the representation and manipulation of complex objects. *ACM Transactions on information systems*, 12 :35–68. — *Cité page(s) 134.*

[PROTÉGÉ, url] PROTÉGÉ (url). http ://protege.stanford.edu. — *Cité page(s) 61.*

[Py, 1992] M. Py (1992). *Un agent rationnel pour raisonner par analogie.* Thèse de doctorat, Université Montpellier II. — *Cité page(s) 48, 124.*

[Rational, url] Rational (url). http ://www.rational.com. — *Cité page(s) 62.*

[Régin, 1995] J.-C. Régin (1995). *Développement d'outils algorithmiques pour l'Intelligence Artificielle et application à la chimie organique.* Thèse de doctorat, Université Montpellier II. — *Cité page(s) 48, 133, 171.*

[Rich and Knight, 1991] E. Rich and K. Knight (1991). *Artificial intelligence.* International Edition. — *Cité page(s) 16, 17, 53, 89, 116.*

[Ridley, 2002] D. D. Ridley (2002). *Information retrieval : SciFinder and SciFinder Scholar.* John Wiley & Sons, LTD, England. — *Cité page(s) 28.*

[Rivail, 1994] J. L. Rivail (1994). *Éléments de chimie quantique à l'usage des chimistes.* InterÉditions / CNRS Éditions, Paris, seconde édition edition. — *Cité page(s) 8.*

[Rose and Gasteiger, 1994] J. R. Rose and J. Gasteiger (1994). Horace : an automatic system for the hierarchical classification of chemical reactions. *J. Chem. Inf. Comput. Sci.*, 34 :74–90. — *Cité page(s) 36.*

[Rouvray, 1986] D. H. Rouvray (1986). *Scientific American*, pages 36–43. — *Cité page(s) 101.*

[Rouvray, 1995] D. H. Rouvray (1995). Combinatorics in chemistry. In R. Graham, M. Grötschel, and L. Lovasz, editors, *Hanbook of combinatorics*, chapter 38. Elsevier Science. — *Cité page(s) 128.*

[Schatz, url] B. R. Schatz (url). http ://ai.bpa.arizona.edu/interspace/. — *Cité page(s) 219.*

[Schildknecht, 2001] S. Schildknecht (2001). *Le projet GRAMS. Construction automatique et exploitation de bases de connaissances réactionnelles.* PhD thesis, Université Louis Pasteur, Strasbourg I. — *Cité page(s) 40.*

[Schreiber et al., 2000] A. Th. Schreiber, M. Cruzbézy, and M. Musen (2000). A case study in using PROTÉGÉ-2000 as a tool for CommonKADS. In R. Dieng and O. Corby, editors, *Knowledge Engineering and Knowledge Management : 12th International Conference EKAW2000*, volume 1937 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag. — *Cité page(s) 60, 61.*

[Simon, 2000] A. Simon (2000). *Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données.* Thèse de doctorat, Université Nancy 1. — *Cité page(s) 70, 73, 88, 89.*

[Speel et al., 2001] P-H. Speel, A. Th. Schreiber, W. van Joolingen, G. van Heijst, and G. J. Beijer (2001). *Conceptual modelling for knowledged-based systems.* Marcel Dekker Inc., New York. — *Cité page(s) 56, 57, 65.*

[Srikant and Agrawal, 1994] R. Srikant and R. Agrawal (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, Santiago, Chile. — *Cité page(s) 80, 83.*

[Stefik, 1995] M. Stefik (1995). *Introduction to Knowledge Systems.* Morgan Kaufmann Publishers, Inc., San Francisco, California. — *Cité page(s) 54.*

[Stumme et al., 2002] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal (2002). Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, 42 :189–222. — *Cité page(s) 83.*

[SYNGEN, url] SYNGEN (url). http ://syngen2.chem.brandeis.edu/syngen.html. — *Cité page(s) 44.*

[Tognetti, 2002] Yannic Tognetti (2002). *Contribution à la modélisation des systèmes d'information chimique par la théorie et l'algorithmique de graphes.* PhD thesis, LIRMM, Université Montpellier II. — *Cité page(s) 125, 133, 169, 177, 214.*

[Tognetti and Vismara, 2000] Y. Tognetti and P. Vismara (2000). Un max-var-csp pour la recherche de décompositions symétriques. *Actes RFIA2000.* — *Cité page(s) 48, 133.*

[Tonnelier, 1990] C. Tonnelier (1990). *Outils informatiques pour l'acquisition des connaissances en synthèse assistée par ordinateur.* Thèse de doctorat, Université Strasbourg 1. — *Cité page(s) 20.*

[Tripos, url] Tripos (url). http ://www.tripos.com. — *Cité page(s) 28.*

[Trombini, 1987] C. Trombini (1987). Planning organic synthesis. I – IV. *La Chim. e l'Ind.*, 6 (5) :82–86, (6) :82–86, (9) :99–104, (12) :129–133. — *Cité page(s) 117.*

[Two Crows Corporation, 1999] Two Crows Corporation (1999). Introduction to data minig and knowledge discovery. Two Crows Corporation. http ://www.twocrows.com/intro-dm.pdf. — *Cité page(s) 75.*

[Ugi et al., 1994] I. Ugi, J. Bauer, C. Blomberger, J. Brandt, A. Dietz, E. Fontain, B. Gruber, A. Von Scholley-Pfab, A. Senff, and N. Stein (1994). Models, concepts, theories and formal languages in chemistry and their use as a basis for computer assistance in chemistry. *J. Chem. Inf. Comput. Sci.*, 34 :3–16. — *Cité page(s) 43, 44.*

[Valdés-Pérez, 1999] R. E. Valdés-Pérez (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence*, 107 :335–346. — *Cité page(s) 72.*

[Valente, 1995] A. Valente (1995). Knowledge-level analysis of planning systems. *SIGART Bulletin*, 6. — *Cité page(s) 94.*

[Valtchev, 1999] P. Valtchev (1999). *Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets.* PhD thesis, Université Joseph Fourier, Grenoble 1. — *Cité page(s) 80.*

[Velluz et al., 1971] L. Velluz, J. Valls, and G. Nominé (1971). Recent advances in the total synthesis of steroids. *Angew. Chem.*, 4 :181–200. — *Cité page(s) 13.*

[Vilain et al., 2000a] P. Vilain, D. Schwabe, and C.Sieckenius de Souza (2000a). A diagrammatic tool for representation user interaction in UML. In *Lecture Notes in Computer Science 1939 - «UML» 2000 - The Unified Modeling Language advancing the standard. Third International conference York*, pages 1–12, UK. — *Cité page(s) 62.*

[Vilain et al., 2000b] P. Vilain, D. Schwabe, and C.Sieckenius de Souza (2000b). Use cases and scenarios in the conceptuel design of web applications. Technical Report Inf.MCC12/00, PUC, Rio. — *Cité page(s) 62.*

[Vismara, 1995] P. Vismara (1995). *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). — *Cité page(s) 48, 128, 135, 137.*

[Vismara et al., 1998] P. Vismara, P. Jambaud, C. Laurenço, and J. Quinqueton (1998). RESYN : objets, classification et raisonnement distribué en chimie organique. In R. Ducournau, J. Euzenat, G. Masini, and A. Napoli, editors, *Langages et modèles à objets : États des recherches et perspectives*, volume 19 of *Collection didactique*, chapter 14, pages 397–419. INRIA. — *Cité page(s) 47.*

[Vismara and Laurenço, 2000] P. Vismara and C. Laurenço (2000). An abstract representation for molecular graphs. *DIMACS Series In Discrete Mathematics and Theoretical Computer Science*, 51 :343–366. — *Cité page(s) 48.*

[Vladutz, 1986] G. Vladutz (1986). Do we still need a classification of reactions ? In *Modern approach to chemical searching*, pages 202–220. Grower, York. — *Cité page(s) 154.*

[Vleduts, 1963] G. E. Vleduts (1963). Concerning one system of classification and codification of organic reactions. *Inform. Stor. Retr.*, 1 :117–146. — *Cité page(s) 18, 26, 154.*

[Vogel, 1988] C. Vogel (1988). *Génie cognitif.* siences cognitives. Masson, Paris. — *Cité page(s) 60, 62.*

[Vogel, 2000] H. Vogel (2000). *Apprentissage automatique de connaissances réactionnelles. Acquisition d'exemples de réactions à partir de bases de données et prise en compte des conditions réactionnelles.* Thèse de doctorat, Université Louis Pasteur, Strasbourg 1. — *Cité page(s) 38, 219.*

[Warren, 1986] S. Warren (1986). *Organic Synthesis. The disconnection approach.* John Wiley and Sons, Chichester, Great Britain. — *Cité page(s) 17, 207.*

[Welford, 1988] S. M. Welford (1988). Tautomer processing in the Beilstein Registry System. — *Cité page(s) 141.*

[Wender and Miller, 1993] P. A. Wender and B. L. Miller (1993). Toward the ideal synthesis : connectivity analysis and multibond-forming processes. *Organic Synthesis : Theory and applications*, 2 :27–66. — *Cité page(s) 8.*

[Wilcox and Levinson, 1986] C. S. Wilcox and R. A. Levinson (1986). A self-organized knowledge base for recall, desing, and discovery in organic chemistry. In T. H. Pearce and B. A. Holme, editors, *Artificial Intelligence Applications in Chemistry*, pages 209–230. American Chemical Society. — *Cité page(s) 154.*

[Willett, 1986] P. Willett (1986). The reaction indexing problem : a historical viewpoint. In *Modern approaches to chemical reaction searching. Proceedings of a conference organized by the Chemical Structure Association at the University of Yor*, pages 1–17, England. — *Cité page(s) 27.*

[Willett, 1998] P. Willett (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38(6) :983–996. — *Cité page(s) 35, 36.*

[Wipke et al., 1977] W. T. Wipke, H. Braun, G. Smith, F. Choplin, and W. Sieber (1977). SECS-similation and evaluation of chemical synthesis : Strategy and planning. In W.T. Wipke and W.J. Howe, editors, *Computer-Assisted Organic Synthesis*, 61, pages 97–127. ACS Symposium. — *Cité page(s) 41.*

[Wipke and Dyott, 1974] W. T. Wipke and T. M. Dyott (1974). Simulation and evaluation of chemical synthesis computer representation and manipulation of stereochemistry. *J. Am. Chem. Soc.*, 96 :4825–4834. — *Cité page(s) 148, 164.*

[Yen and Chen, 2001] S. J. Yen and A. L. P. Chen (2001). A graph-based approach for discovering various types of association rule. *IEEE Transactions on knowledge and data engineering*, 13(5). — *Cité page(s) 200.*

[Ziegler, 1979] H. J. Ziegler (1979). Roche Integrated Reaction System (RIRS). A new documentation system for organic reaction. *J. Chem. Inf. Comput. Sci.*, 19(3) :141–149. — *Cité page(s) 36.*

Index

Université HENRI POINCARÉ
NANCY 1

Mademoiselle **BERASALUCE Sandra**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY 1

en CHIMIE INFORMATIQUE & THEORIQUE

VU, APPROUVÉ ET PERMIS D'IMPRIMER

n° 776

Nancy, le 16 janvier 2003

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ, NANCY 1 ★ LE PRÉSIDENT

Cl. BURLET

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél. : 03 83 68 20 00 - Fax : 03 83 68 21 00

Résumé : Les systèmes d'information chimique sont des outils indispensables pour les chimistes organiciens. Ceux-ci disposent notamment de grandes bases de données décrivant des millions de réactions chimiques. Bien que constituant un progrès non négligeable, ces bases de données ne sont pas exemptes de défauts. Dans cette thèse, nous avons essayé de dépasser les limites de ces bases en ajoutant des connaissances structurant les données. Ceci nous permet d'envisager de nouveaux modes d'interrogation, plus performants, de ces bases. In fine, l'objectif est de concevoir des systèmes possédant à la fois des fonctionnalités de bases de données et de systèmes à base de connaissances. Dans le processus d'acquisition de connaissances, nous avons mis l'accent sur la modélisation des objets chimiques. Nous nous sommes intéressés, plus particulièrement, aux méthodes de synthèse que nous avons décrites en terme d'objectifs de synthèse atteints et de contextes structuraux d'application. Nous nous sommes ensuite appuyés sur le modèle élaboré pour appliquer des techniques de fouille de données et faire émerger des connaissances des bases de données de réactions. Les expérimentations que nous avons effectuées ont porté sur les méthodes de construction de monocycles, d'une part, et sur les interchanges fonctionnels, d'autre part. Nous avons utilisé pour nos divers développements la plate-forme RESYN ASSISTANT, un système d'aide à la planification de synthèse permettant de représenter les objets chimiques selon le modèle élaboré. Les résultats obtenus donnent des tendances en accord avec les connaissances du domaine et serviront d'heuristiques de choix des méthodes de synthèse.

Title : Data mining and knowledge acquisition from chemical reaction databases.

Abstract : Chemical information systems are indispensable tools for synthetic chemists. These ones can use, in particular, very large databases which describe millions of chemical reactions. Also they mark an advance in the field, these ones are not free from flaws. In this thesis, we have tried to overcome the databases limits by adding knowledge which structures data. This allows us to consider new efficient modes for query these databases. In the end, the goal is to design systems having both functionalities of databases and knowledge based systems. In the knowledge acquisition process, we emphasized on the modelling of chemical objects. More precisely, we were interested in synthetic methods which we have described in terms of synthetic objectives and structural contexts of application. Afterward, we based ourselves on the elaborated model to apply data mining techniques and to extract knowledge from chemical reaction databases. The experiments we have done concern synthetic methods which construct monocycles, on one side, and the functional interchanges, on the other side. For our various developments, we use the RESYN ASSISTANT platform, a system for synthesis planning assistance which provides representations of chemicals objects according to the elaborated model. The results give trends in good agreement with domain knowledge and will be used as heuristics in the choice of synthetic methods.

Keywords : chemical information systems, chemical reaction databases, knowledge based systems, knowledge acquisition, knowledge discovery from chemical reaction databases, data mining

---

Discipline : Chimie Informatique et Théorique

Mots-clés : systèmes d'information chimique, bases de données de réactions chimiques, systèmes à base de connaissances, acquisition de connaissances, extraction de connaissances à partir de bases de données, fouille de données

---

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503