SYNTHÈSE ORGANIQUE ET INFORMATIQUE

Claude LAURENÇO
GDR 1093 du CNRS Traitement Informatique de la Connaissance en Chimie Organique
UPR 9023 du CNRS, CCIPE, 141 rue de la Cardonille, 34094 Montpellier cedex 5
cl@lirmm.fr

1. Introduction

La synthèse multi-étapes de molécules complexes ayant des propriétés désignées à l'avance reste un défi majeur pour les chimistes, tant sur le plan scientifique que sur le plan industriel où les enjeux économiques sont considérables. Son succès est fortement conditionné par l'environnement dans lequel elle est réalisée. Cette réalisation devrait être le travail d'un groupe de spécialistes suffisamment nombreux, disposant à la fois d'un savoir-faire et d'un ensemble de moyens permettant d'accomplir au mieux les différentes tâches qu'elle comporte : (i) choix d'une bonne cible, (ii) conception d'un plan de synthèse, (iii) expérimentation de ce plan, (iv) évaluation de la voie de synthèse suivie. Si la phase expérimentale est primordiale, puisqu'elle doit aboutir à la réalisation matérielle de la synthèse en surmontant tous les imprévus, la phase conceptuelle n'en est pas moins celle qui détermine la qualité de cette synthèse. Dès les années cinquante, Woodward a souligné l'importance de la conception d'un plan dans le processus de la synthèse : *Synthesis must always be carried out by plan, and the synthetic frontier can be defined only in terms of the degree to which realistic planning is possible, utilizing all of the intellectual and physical tools available* [Woodward 56].

Aucune des méthodes actuelles de la chimie théorique ne permet de calculer directement un plan de synthèse pour une molécule donnée, pas plus qu'il n'existe de solution algorithmique à un problème de synthèse. Aussi, pour concevoir un tel plan, les chimistes font-il appel à leur intuition et à leur expérience. Ils mettent en oeuvre des raisonnements divers, le plus souvent de type analogique, qu'ils appliquent à des connaissances chimiques nombreuses et variées : méthodes de synthèse, chemins de synthèse connus, molécules commercialement disponibles, règles de stratégie et de tactique, mécanismes réactionnels, propriétés physico-chimiques, éléments de théorie chimique, etc. Au total, le volume de la connaissance exploitable par les chimistes est considérable ; notamment, plus de 15 millions de molécules et probablement plus de 10 millions de réactions ont déjà été décrites. L'informatique propose différents outils pouvant aider les chimistes à résoudre les problèmes qu'ils rencontrent dans la phase conceptuelle d'une synthèse.

A côté de programmes de modélisation moléculaire ou de prédiction de la réactivité, basés sur des calculs numériques, les chimistes ont à leur disposition divers types de systèmes traitant les données chimiques de manière symbolique. D'une part, des systèmes d'information chimique permettent de stocker et de retrouver dans des bases de données une information sur les molécules et les réactions connues. D'autre part, des systèmes "logiques", dits de Synthèse Assistée par Ordinateur, ont pour vocation d'apporter une aide au raisonnement dans l'élaboration de plans de synthèse ou la simulation de réactions. Ces deux types de systèmes sont fondamentalement différents. Les premiers sont conçus pour retrouver une information ayant été explicitement rangée dans leur base de données et portant sur des *faits* connus. Les seconds, appliquant un mécanisme de raisonnement sur une *connaissance* dont ils disposent, doivent produire dynamiquement de nouvelles connaissances, en l'occurrence des solutions

partielles ou complètes à des problèmes de synthèse. Une telle solution peut avoir, par exemple, la forme d'un arbre de rétrosynthèse. Le terme connaissance recouvre ici différents aspects du savoir (hiérarchies de concepts chimiques, instances connues de ces concepts, règles heuristiques de choix et d'évaluation, etc.) qui se distinguent de l'ensemble de faits non structurés constituant une base de données. Sur le plan informatique, les premiers sont des *systèmes de bases de données* tandis que les seconds s'apparentent, plus ou moins, aux *systèmes experts*. Néanmoins, tous ces systèmes ont en commun de tirer profit de la puissance du langage de la chimie organique, fondé sur les formules structurales, à partir duquel les objets chimiques peuvent être modélisés en termes de graphes. Ces derniers sont particulièrement bien adaptés au traitement informatique.

Préalablement à l'étude de ces deux catégories de systèmes, il convient de s'intéresser aux différents aspects des problèmes de synthèse et à leur résolution.

2. La problématique de la synthèse organique

La synthèse organique est duale. Son objet est soit l'obtention d'une molécule particulière donnée, naturelle ou imaginée, selon une voie originale, soit le développement d'une nouvelle méthode de synthèse, autrement dit d'une réaction ou d'une suite de réactions accomplissant une transformation générique (réduction énantiosélective de la fonction carbonyle, par exemple). Ces deux aspects peuvent être indépendants ou concomitants [Nicolaou & Sorensen 96].

Un problème de synthèse est un *problème de transformation*. Une solution à un tel problème est modélisée par un chemin allant d'un état initial à un état final, en passant le plus souvent par une suite d'états intermédiaires, les transitions entre les états étant assurées par des opérateurs [Kahney 93]. Selon ce modèle, l'état initial d'une voie de synthèse est constitué par un ensemble de molécules disponibles - les produits de départ - l'état final contient la molécule cible et les opérateurs sont des méthodes de synthèse. Un chemin de synthèse doit satisfaire un ensemble de conditions et avoir été validé par l'expérimentation. Les méthodes de synthèse, et plus généralement les réactions, se modélisent de manière semblable.

Un problème de transformation est dit *bien-défini* lorsque, dès le début du processus de résolution, l'état initial, l'état final et les opérateurs utilisables - le triplet qui définit l'espace du problème contenant l'ensemble des solutions et celui des voies sans issue - sont eux-mêmes clairement définis. La résolution du problème consiste à parcourir cet espace pour trouver la ou les meilleures solutions. Dans le cas d'un problème de synthèse, on ne connaît initialement que la structure cible - l'état final - et éventuellement de l'information sur ses propriétés et/ou sur des molécules analogues. On a aussi 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. Le problème de la synthèse d'une molécule donnée est donc un problème *mal-défini*. Ceci a plusieurs conséquences : (i) la démarche naturelle de la résolution est analytique, elle consiste à partir de la structure cible et progresser rétrosynthétiquement, par étapes successives, vers les produits de départ, (ii) l'espace du problème est d'une taille considérable car les multiples combinaisons des dizaines de milliers de molécules disponibles constituent autant d'ensembles d'états initiaux possibles et toutes les réactions connues ou à découvrir sont des opérateurs potentiels, (iii) le parcours de l'espace doit être guidé par des heuristiques afin d'éviter l'explosion combinatoire qui résulterait d'une recherche exhaustive des solutions. Un tel problème est trop complexe pour être résolu directement, il doit être décomposé en un ensemble de sous-problèmes plus simples. C'est à cela que vise le processus de planification, chaque étape d'un plan correspondant à la résolution d'un ou plusieurs sous-problèmes et l'ensemble des solutions partielles obtenues composant une solution globale du

problème initial. Une part importante de l'art de la synthèse réside dans la capacité du chimiste à déterminer une décomposition adéquate du problème, autrement dit un bon découpage de la structure cible, une *stratégie de synthèse*.

Une idée de la complexité d'un problème de synthèse est donnée par l'exemple formel suivant : la synthèse totale *stricto sensu* du Taxol (figure 1), à partir de ses éléments constitutifs - carbone, hydrogène, oxygène et azote - nécessiterait de créer successivement les 119 liaisons de cette molécule, ce qui peut être envisagé de plus de 10^{196} façons différentes. Cet exemple montre également les limites actuelles de la synthèse totale. Depuis une vingtaine d'années, plus de 30 équipes ont travaillé à la synthèse de cette molécule qui possède des propriétés anticancéreuses remarquables. Sa première synthèse totale [Holton & al 94], comptant plus de 30 étapes à partir du camphre, a un rendement très faible. Si elle possède un grand intérêt scientifique, elle ne constitue pas pour autant un procédé de fabrication. En fait, le Taxol est produit industriellement par hémisynthèse à partir d'un de ses analogues extrait du milieu naturel en quantité importante.

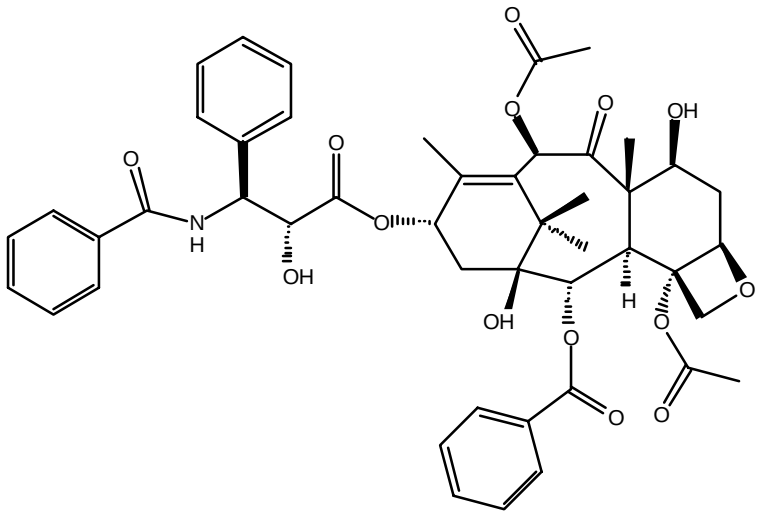

Figure 1 : Formule du Taxol.

La compréhension du problème est la tâche par laquelle commence le processus de sa résolution. Comprendre un problème c'est en construire un modèle et exprimer celui-ci à travers une représentation. Pour cela il faut *percevoir* la structure cible, c'est-à-dire l'analyser et la catégoriser selon différentes perspectives en tenant compte de toute l'information disponible.

3. Les systèmes d'information chimique

La documentation relative à un problème de synthèse fournit des données permettant de mieux comprendre ce problème et de valider des hypothèses formulées pour le résoudre. Les différentes sources d'information chimique appartiennent à trois catégories distinctes : (i) les sources primaires qui décrivent les résultats originaux (articles de journaux, brevets, actes de congrès, etc.), (ii) les sources secondaires qui compilent des références à des documents primaires accompagnées de résumés et qui les indexent afin de faciliter leur recherche (les *Chemical Abstracts* en sont l'exemple type), (iii) les sources tertiaires qui contiennent des informations sélectionnées se rapportant à un certain sujet (mises au point, manuels ou encyclopédies telles que le *Beilstein's Handbuch der Organischen Chemie*).

L'accroissement exponentiel du volume de l'information chimique, la difficulté à mener manuellement des recherches bibliographiques et la nécessité pour l'industrie de disposer de

moyens fiables de gestion de l'information se sont conjugués avec les avancées faites dans le domaine de l'informatique pour favoriser, à partir des années soixante, le développement de systèmes d'information chimique. Aujourd'hui, on retrouve sous une forme électronique les trois catégories précédentes de sources d'information chimique. Ainsi, les principaux journaux de l'American Chemical Society, tel que le *Journal of Organic Chemistry*, sont publiés simultanément en version papier et en version électronique [http://www.acs.org]. Certains nouveaux journaux, comme *NetSci*, ne sont accessibles qu'au travers d'Internet [http://www.netsci.org]. Les versions informatisées des *Chemical Abstracts* restent la référence en matière de source secondaire, bien que d'autres produits aient également été développés : *Current Contents*, *Science Citation Index*, etc. *CrossFire+Reactions*, la transcription informatisée du *Beilstein*, est un exemple remarquable de source tertiaire. Dans ce domaine, cependant, il existe d'autres produits très originaux, en particulier le système *ISIS*. Les bases de données les plus importantes comportent des dizaines de milliers, voire des millions, d'éléments d'information et sont continuellement mises à jour. Suivant les cas, elles sont interrogées *online* (par ligne téléphonique) ou par Internet, sur un serveur distant, ou *inhouse*, sur une station de travail ou à travers un réseau local.

Les bases de données intéressant les chimistes spécialisés en synthèse sont de divers types : bases bibliographiques, bases de molécules, bases de réactions. Néanmoins, aucune base de chemins de synthèse n'est actuellement disponible. La recherche d'information peut s'effectuer par termes bibliographiques (auteurs, journal, année de publication), mots clés (nom de substance, nom de réaction, propriété, solvant, rendement ...) et surtout par structures et sous-structures. Ce dernier mode d'interrogation est très efficace et a l'avantage d'utiliser le langage du chimiste, ce dernier pouvant proposer sa requête en dessinant une formule structurale ou un fragment de celle-ci au moyen d'une interface graphique. Une interface de même nature permet de visualiser les résultats de la recherche. Dans le cas des bases de réactions, on peut formuler des requêtes très directement liées aux préoccupations des spécialistes de la synthèse et qui seraient quasiment impossible à mener manuellement. Ainsi, à la figure 2, la requête portant sur la réduction sélective d'un groupement nitro en présence d'un groupement aldéhyde est un exemple typique d'une recherche sous-structurale effectuée dans une base de réactions (ici *CrossFire+Reactions*). Il n'aura fallut au système que 386 secondes de temps cpu pour retrouver les 118 réactions répertoriées par *Beilstein* qui satisfont cette requête.

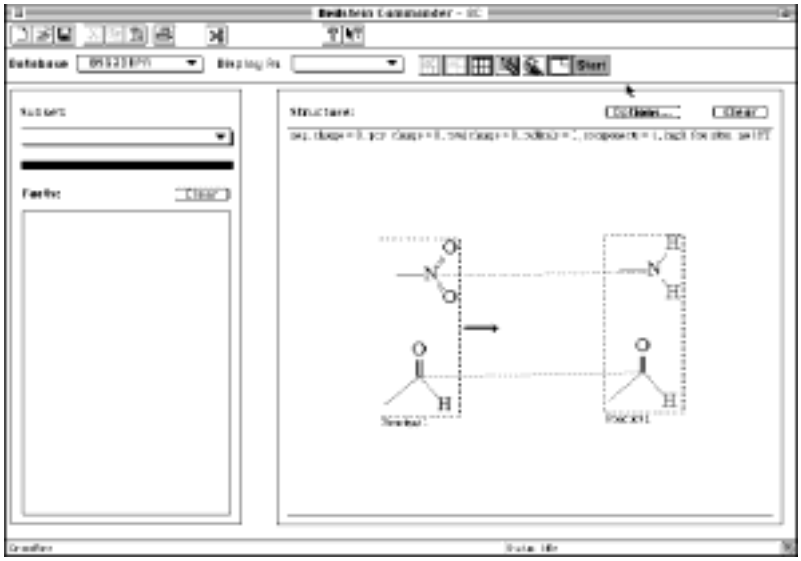

Figure 2a : recherche par sous-structure dans une base de réactions ; visualisation d'une requête.

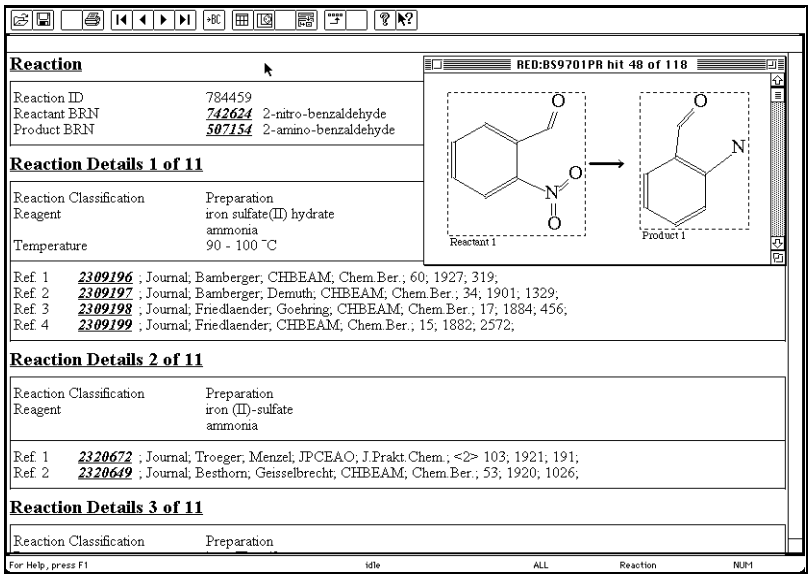

Figure 2b : recherche par sous-structure dans une base de réactions ; visualisation d'une des réponses.

L'emploi de l'ordinateur permet de croiser des requêtes sur une base ou même sur plusieurs bases. La combinaison, à l'aide d'opérateurs logiques, d'une recherche par sous-structure avec une recherche par mots clés précise une question et augmente la pertinence des réponses.

La figure 3 rassemble les différentes interrogations qui peuvent être effectuées sur une base de données chimiques telle que celle de *CrossFire+Reaction*.

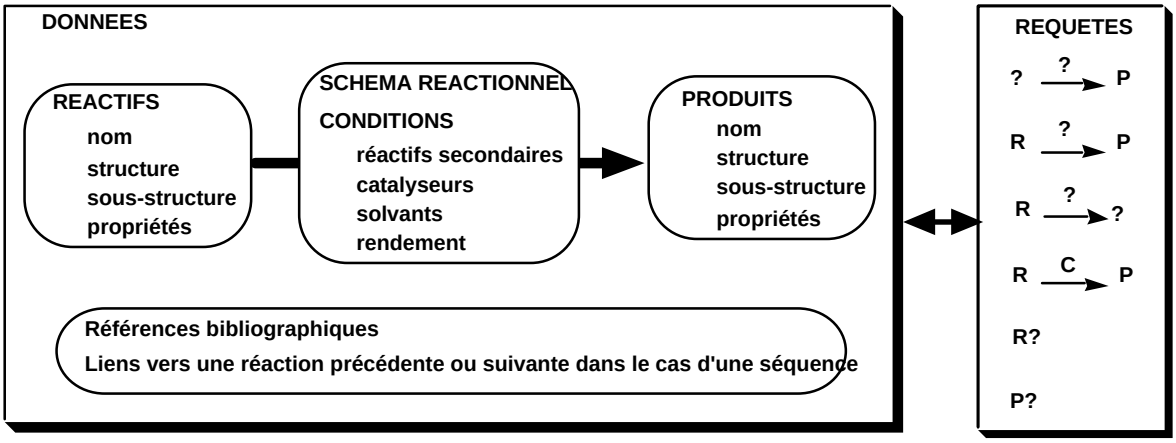

Figure 3 : Requêtes sur une base de données chimiques.

Quelques bases de données chimiques intéressant les spécialistes de la synthèse organique sont accessibles dans les établissements d'Enseignement Supérieur ou dans les laboratoires du CNRS au tarif réduit Education :

• ***Bases de Chemical Abstracts Service*** [http://www.cas.org]

Les bases de données de CAS, une division de l'American Chemical Society, sont les plus complètes existant dans le domaine de l'information chimique. Couvrant tous les secteurs de la chimie, elles contiennent plus de 14 millions de résumés de publications de la littérature, issus de 8000 journaux, ainsi que de brevets et elles répertorient 17 millions de substances dont 80% sont des molécules organiques. Leur utilisation s'impose lorsqu'on souhaite faire des

recherches bibliographiques exhaustives. Leur inconvénient principal est de fournir des résultats souvent trop nombreux, limités aux résumés et aux références qui renvoient aux articles originaux. Par exemple, on peut obtenir la référence à un article décrivant des propriétés physiques d'une molécule mais on n'accède pas directement à leur description. Cependant, un nouveau produit, *ChemPort*, devrait permettre le passage entre les références CAS et le texte complet des articles publiés dans les principaux journaux. Les bases de données CAS sont accessibles sur le Web à l'aide du logiciel *SciFinder* ou par l'intermédiaire du réseau Scientific & Technical Information Network [http://www.fiz-karlsrhue.de], *online* à l'aide du logiciel *STN Express* ou sur le Web par *STN Easy*. Le réseau STN donne aussi accès à plus de 200 bases de données dans de nombreux domaines scientifiques, techniques ou commerciaux. En chimie, outre plusieurs bases de réactions, on peut interroger des bases de données spectroscopiques, physicochimiques, biologiques, sur la sécurité chimique, sur la disponibilité commerciale des molécules, etc. Il existe également des versions des bases de données de CAS sur support CD-ROM qui permettent leur interrogation *inhouse*.

• ***Base de Beilstein*** [http://www.beilstein.com]

La base de données *Beilstein* appartient au groupe Information Handling Services (IHS) et à l'Institut Beilstein (lesquels ne devraient pas tarder à la céder à Elsevier Science). Initialement orientée vers les molécules, elle couvre la littérature depuis 1779 et contient des structures chimiques avec leurs propriétés, préparations, réactions et références respectives. Elle est accessible par STN mais il est plus pratique de l'utiliser *inhouse* à l'aide du système *CrossFire+Reactions* qui donne accès à 8 millions de molécules associées à 10 millions de réactions et à de multiples propriétés chimiques, physiques, biologiques, etc. (350 champs sont associés aux structures). Le système est très performant grâce à son moteur de recherche et à des liens hypertexte mais il fournit aussi de très nombreux résultats et son traitement des réactions n'est pas vraiment orienté vers la synthèse, c'est-à-dire qu'elles sont moins vues comme des méthodes de synthèse que comme des propriétés des molécules.

• ***Bases de Molecular Design Ltd*** [http://www.mdli.com]

MDL est une division du groupe Elsevier Science. Cette société développe le système *ISIS* (Integrated Scientific Information System) servant à l'interrogation des bases de données de provenances diverses qu'elle diffuse. Une part importante de ces bases sont ciblées vers la synthèse organique. A l'inverse des précédentes, ces bases de réactions n'ont pas la vocation d'être exhaustives. Les réactions introduites sont sélectionnées en fonction de leur caractère prototypique : une classe de réactions ayant valeur de méthode de synthèse est décrite par un ou quelques exemples remarquables. Une réaction peut être incluse dans une séquence. Les résultats obtenus sont moins nombreux qu'avec la base de *Beilstein* mais sont le plus souvent pertinents. Néanmoins, lorsqu'une requête engendre un ensemble trop grand de réponses celui-ci peut-être soumis à un processus de *clustérisation* qui le partitionne en sous-ensembles selon des critères définis par l'utilisateur ou prédéfinis. Parmi les bases exploitables par *ISIS*, certaines sont très générales, comme *ChemInform*, tandis que d'autres sont très spécialisées, comme *Comprehensive Heterocyclic Chemistry* ou *ORGSYN* dont la reproductibilité des quelques milliers de réactions qu'elle comporte a été expérimentalement vérifiée. Au total, *ISIS* permet d'accéder à près d'un million de réactions. Par ailleurs, *Available Chemicals Directory* (ACD) est une base de molécules très intéressante pour la synthèse organique puisqu'elle rassemble 420 catalogues de fournisseurs de produits commerciaux (255 000 composés) avec toute l'information utile sur le coût, la pureté, les quantités disponibles, etc.

ISIS est utilisable uniquement *inhouse*. Il autorise la création et la maintenance de bases de données privées, de molécules et de réactions, pour répondre aux besoins des sociétés

industrielles qui doivent archiver leur résultats et conserver leur savoir faire ainsi qu'à ceux des universitaires qui souhaitent développer des bases de données spécialisées.

4. Les systèmes d'aide au raisonnement

Dès 1963, Vléduts avait imaginé simuler la démarche des chimistes sur un ordinateur afin de les aider à résoudre leurs problèmes de synthèse [Vléduts 63]. Corey et Wipke ont décrit en 1969 *OCSS*, le premier système de Synthèse Assistée par Ordinateur [Corey &Wipke 69]. Celui-ci, pour une molécule cible qui lui était soumise, procédait à une analyse rétrosynthétique selon le principe défini par Vléduts et formalisé par Corey [Corey & Cheng 89]. Le résultat d'une telle analyse est un arbre de rétrosynthèse représentant un ou plusieurs plans de synthèse. A la racine et aux feuilles correspondent respectivement la structure cible et des produits de départ ou des précurseurs de la cible à partir desquels l'exploration de l'espace du problème n'a pas été poursuivie (figure 4).

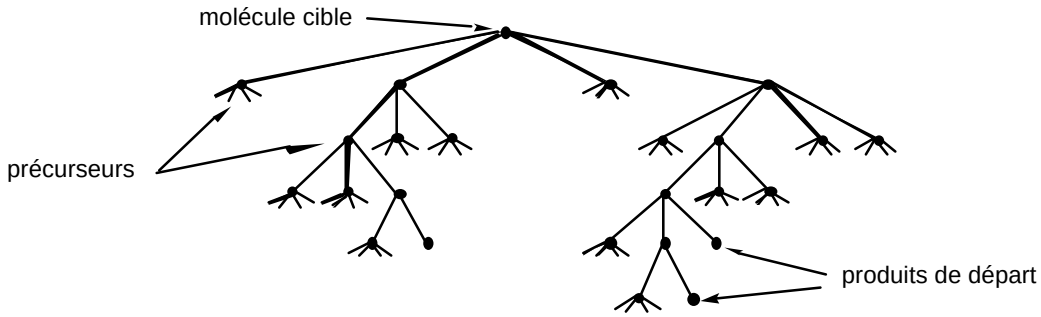

Figure 4 : Arbre de rétrosynthèse.

L'arbre est construit de manière récursive. Un précurseur est engendré à partir du graphe moléculaire de la cible par application d'une *transformation*, opération effectuant dans le sens rétrosynthétique les modifications structurales réalisées par une méthode de synthèse (casser ou créer une liaison, inverser un stéréocentre, etc.). Lorsqu'un précurseur formé n'est pas reconnu comme un produit de départ, il peut être pris comme nouvelle cible et le processus est renouvelé jusqu'à l'obtention d'au moins une solution satisfaisante.

A la suite d'*OCSS* beaucoup d'autres systèmes ont été conçus et réalisés, le plus souvent sans dépasser le stade du prototype, et différentes approches ont été explorées [Ott & Noordik 92]. Deux conceptions s'opposent : celle de Corey qui part de l'idée que la résolution des problèmes complexes de synthèse doit faire appel à l'expérience des chimistes [Corey & Cheng 89], c'est une approche *empirique*, et celle de Ugi qui repose sur des bases logiques et mathématiques en respectant les principes fondamentaux de la théorie structurale [Ugi & al 94], c'est une approche *formelle* qui n'utilise aucune connaissance d'expert. Les systèmes connus ont été conçus selon l'une ou l'autre de ces approches, ou selon des approches intermédiaires. Certains systèmes sont interactifs tandis que d'autres sont automatiques. Certains fonctionnent dans le sens synthétique, d'autres dans le sens rétrosynthétique et d'autres encore dans les deux sens. Tous communiquent avec l'utilisateur au moyen d'interfaces graphiques. Quelques uns d'entre eux sont représentatifs de l'ensemble :

• ***RAIN & IGOR***

RAIN (Reaction And Intermediates Networks) & *IGOR* (Interactive Generation of Organic Reactions) sont deux systèmes développés par le groupe de Ugi à l'Université de Munich [Ugi & al 94]. Il sont basés sur le modèle Dugundji-Ugi dont l'équation fondamentale **B + R = E** représente les réactions chimiques. **B**, **E** et **R** sont des matrices décrivant respectivement l'ensemble des graphes moléculaires des réactifs, celui des graphes

moléculaires des produits et le schéma de la redistribution électronique permettant d'*isomériser* **B** en **E**, ou **E** en **B** si on remplace l'opérateur **+** par l'opérateur **-** . Etant donné une paire {**B**,**E**}, en appliquant en séquence des générateurs de réactions **R**, *RAIN* peut construire des réseaux de réactions qui relient les deux ensembles **B** et **E** et qui respectent le *principe de distance chimique minimale* (redistributions électroniques minimales). *RAIN* apporte une aide à l'élucidation de mécanismes réactionnels et à la prédiction de réactions. Il peut aussi avoir un intérêt en chimie combinatoire.

IGOR a été conçu pour découvrir de nouvelles réactions : à partir d'un générateur de réactions (spécifiant les redistributions électroniques sans préciser les types des atomes) il propose des paires {**B**,**E**} possibles.

L'approche formelle adoptée pour réaliser ces systèmes ne nécessite pas de créer et de maintenir des bases de connaissances chimiques, tâches évidemment très lourdes. Il suffit d'un petit nombre de schémas généraux pour décrire l'ensemble des réactions de la chimie organique. S'ils sont innovants du point de vue méthodologie de la synthèse, ils n'ont pratiquement pas d'intérêt en planification de synthèse car ils ne disposent d'aucune connaissance stratégique ou tactique.

• ***SYNGEN*** [http://syngen2.chem.brandeis.edu]

SYNGEN (SYNthesis GENerator) est réalisé par l'équipe de Hendrickson à l'Université Brandeis (Massachusetts) [Hendrickson 90]. Son but est d'engendrer automatiquement, pour une molécule cible donnée, les voies de synthèse les plus courtes et les plus économiques. Basé sur la *théorie des demi-réactions*, il ne requiert pas de base de connaissance. Pour faire une réaction, il faut combiner une demi-réaction nucléophile avec une demi-réaction électrophile, les combinaisons chimiquement valides de 25 demi-réactions fournissant une centaine de réactions.

Le système procède en deux étapes. Au cours d'une étape stratégique, il établit des *plans d'assemblage* de la structure cible. Pour cela, il considère le squelette de la cible (abstraction faite de la fonctionnalité) et le fragmente en deux parties de toutes les façons possibles en déconnectant à chaque fois une liaison (si un cycle est présent, deux liaisons peuvent être déconnectées). Pour respecter le *principe de convergence* favorisant les synthèses les plus courtes, une fragmentation est retenue lorsque le plus petit des deux fragments contient au moins le quart du nombre total des atomes du squelette. Chaque fragment est alors comparé aux squelettes de molécules présentes dans une base de 6000 produits commerciaux. Un fragment est conservé s'il correspond à une molécule de cette base, sinon il est à son tour fragmenté en deux parties de toutes les façons possibles et les fragments formés sont comparés à ceux de la base de molécules. Les ensembles dans lesquels tous les fragments correspondent à des squelettes de produits de départ potentiels sont retenus. Pour chacun de ces ensembles est créé un ensemble ordonné des liaisons déconnectées qui constitue un plan d'assemblage de la structure cible. Dans l'étape tactique, *SYNGEN* considère dans l'ordre les liaisons stratégiques du plan d'assemblage et recherche des réactions susceptibles de les créer en utilisant la fonctionnalité initiale de la cible.

L'approche suivie par *SYNGEN* est séduisante dans son principe, cependant ce système produit de très nombreuses solutions. Par exemple, pas moins de 1432 plans d'assemblage sont proposés pour une molécule de complexité moyenne telle que l'estrone. Par ailleurs, alors que la plupart des synthèses décrites consacrent au moins la moitié des étapes à l'aménagement fonctionnel, afin de faciliter la construction de la structure, *SYNGEN* n'autorise qu'un minimum de modifications de la fonctionnalité, écartant ainsi de nombreuses solutions pertinentes.

• ***CAMEO*** [http://zarbi.chem.yale.edu/programs/cameo.html]

CAMEO (Computer Assisted Mechanistic Evaluation of Organic Reactions) est un système prédisant les produits de réactions organiques étant donné un ensemble de réactifs et des conditions réactionnelles. Il est développé à l'Université Yale (Connecticut) par le groupe de Jorgensen [Jorgensen 90]. Ce système emploie différents modèles chimiques en fonction de la classe des réactions et des conditions réactionnelles choisies par l'utilisateur. Certains modèles reposent sur des théories qui permettent de faire des calculs numériques, d'autres sont purement empiriques. *CAMEO* traite actuellement les réactions nucléophiles et en milieu basique, les réactions électrophiles et en milieu acide, les réactions péricycliques, les oxydoréductions, les réactions radicalaires, les réactions des carbènes, la formation des hétérocycles, les réactions photochimiques, les réactions des composés organométalliques du Palladium et les substitutions électrophiles aromatiques. Le programme utilise aussi des données de thermochimie pour estimer la stabilité des produits formés. La prédiction des réactions est problème très difficile et la qualité des résultats obtenus avec *CAMEO* varie en fonction des classes de réactions et des modèles utilisés. Ce système trouve son utilité en synthèse multi-étapes lorsqu'on veut vérifier la pertinence des méthodes de synthèse sélectionnées au cours de la rétrosynthèse, détecter les éventuelles compétitions, envisager des protections de groupements fonctionnels, etc.

• ***LHASA*** [http://www.chem.leeds.ac.uk/LUK]

LHASA (Logic and Heuristics Applied to Synthesis Analysis) est, toutes catégories confondues, le programme d'aide à la synthèse le plus performant. Il est développé depuis le début des années 70 par le groupe de Corey à l'Université Harvard (Massachusetts) [Corey & al 85] ainsi qu'en Angleterre, à l'Université de Leeds, dans le cadre du club LHASA UK. Fondamentalement, c'est un système de rétrosynthèse interactif. L'utilisateur fait des choix stratégiques ou tactiques que le programme cherche à satisfaire (*LHASA* peut suggérer certains de ces choix). Le système fonctionne pas à pas mais peut, dans certaines circonstances, enchaîner plusieurs étapes. *LHASA* exploite une base de connaissance comportant environ 2100 transformations et 500 combinaisons tactiques de transformations. Ces combinaisons représentent des séquences où des méthodes de synthèse s'enchaînent logiquement pour atteindre un objectif important. Les transformations sont représentées au moyen du langage spécialisé *CHMTRN*. En plus du *retron* de la transformation, c'est-à-dire la sous-structure nécessairement présente dans la cible pour que cette transformation soit sélectionnée, et des actions à effectuer pour engendrer les précurseurs, la description indique au système les contextes structurels et fonctionnels et les conditions dans lesquels la transformation est généralement applicable ou non applicable. Cette connaissance est établie à partir de l'expérience d'experts de la synthèse. Les stratégies rétrosynthétiques implantées dans *LHASA* sont basées sur : (i) la fonctionnalité (certaines associations de groupements fonctionnels sont le retron de transformations pouvant nettement simplifier la structure cible), (ii) la topologie (les liaisons dont la déconnexion réduit la complexité des systèmes polycycliques sont considérées comme stratégiques), (iii) l'application prioritaire des transformations relatives aux méthodes de synthèse les plus performantes, (iv) la stéréochimie (v) la reconnaissance dans la cible de sous-structures correspondant à des produits de départ potentiels.

L'analyse rétrosynthétique d'une molécule aussi complexe que le Taxol effectuée avec l'aide de *LHASA* donne un aperçu des performances actuelles du programme [Van Rozendaal 94]. Celui-ci est capable de proposer des tactiques pertinentes (transformations et séquences de transformations) pour atteindre un objectif fixé mais le grand nombre de résultats obtenus oblige l'utilisateur à des tris fastidieux. Ceux-ci sont rendus difficiles parce qu'aucune

explication des résultats n'est donnée. Le niveau stratégique est faible, *LHASA* ne proposant que des options générales à partir desquelles l'utilisateur va faire les choix lui permettant de contrôler la construction de l'arbre de rétrosynthèse. Etant limité par le contenu de sa base de connaissance, *LHASA* ne fournit pas toutes les solutions possibles et ne peut pas inventer de nouvelles réactions. Il peut néanmoins employer, dans de nouveaux contextes, de nombreuses réactions connues et suggérer des idées conduisant à une résolution satisfaisante du problème posé. Dans tous les cas, le chimiste doit exercer son sens critique pour évaluer les solutions proposées.

5. Une nouvelle génération de systèmes d'aide à la synthèse

Bien que certains d'entre eux aient des performances honorables, les systèmes de Synthèse Assistée par Ordinateur restent peu employés. A l'inverse, l'utilité et l'efficacité des systèmes d'information chimique ont été reconnues par les chimistes. *CrossFire+Reactions* ou *ISIS* sont couramment employés en synthèse, au-delà même de leur fonction documentaire initiale. Pour élaborer un plan de synthèse, les chimistes préfèrent souvent combiner l'analyse rétrosynthétique faite par eux-mêmes avec une interrogation de bases de réactions. Celle-ci leur fournit des exemples de réactions ayant permis de résoudre des problèmes similaires à ceux posés par le franchissement des différentes étapes de leur plan, lesquels sont alors résolus par analogie avec ces exemples. Si les systèmes d'information chimique sont aujourd'hui considérés comme des outils indispensables, c'est qu'ils sont suffisamment performants et que leur utilisation apporte un progrès considérable par rapport à la situation antérieure. En fait, ces systèmes sont beaucoup plus faciles à concevoir et à rendre opérationnels, jusqu'au niveau de produits commerciaux, que de vrais systèmes d'aide à la résolution des problèmes de synthèse.

A l'analyse, on constate que les systèmes de Synthèse Assistée par Ordinateur n'ont jamais atteint un stade de développement suffisant et qu'ils suscitent la plupart des critiques faites habituellement aux systèmes experts : conception monolithique, difficulté d'acquisition des connaissances, production limitée d'explications, incapacité à apprécier les limites de leur compétence, etc. Afin de corriger ces défauts, les spécialistes de l'intelligence artificielle ont conçu des méthodologies et des techniques nouvelles faisant évoluer le concept de système expert vers celui, plus général, de *système à base de connaissance* qui combine différents types de connaissances, de représentations et de méthodes de résolution [David & al 93]. Ces méthodologies et techniques doivent être employées pour progresser dans la réalisation de nouveaux systèmes d'aide à la synthèse. C'est l'objectif poursuivi par le GDR 1093 du CNRS, structure réunissant des informaticiens et des chimistes - des experts de la synthèse et des "modélisateurs" de la chimie - tant du secteur public que de l'industrie. *RESYN*, le prototype d'un système d'aide à la conception de plans de synthèse est utilisé comme plate-forme pour expérimenter de nouveaux modèles chimiques et des techniques avancées d'informatique. Les connaissances sont modélisées selon une approche *orientée objets*. La plupart des concepts sont décrits en termes de graphes organisés en hiérarchies, selon des relations de *subsomption*, et sur lesquelles *RESYN* applique un *raisonnement par classification* [Vismara 95], [Régin 95]. Cette approche facilite l'acquisition des connaissances en s'affranchissant des modes de description procéduraux des transformations et permet d'envisager une construction automatisée de bases de connaissance à partir de bases de données de réactions. Des fonctions de perception des molécules ont été développées pour donner différents points de vue d'une cible (topologie, stéréochimie, chimie, ...) et des modes de *raisonnement distribué* [Jambaud 96] et de *raisonnement à partir de cas* [Lieber 97] ont été étudiés et implantés.

6. Conclusion

La synthèse organique par la créativité qu'elle requiert est un art tout autant qu'une discipline scientifique. Il est improbable qu'avant longtemps l'ordinateur puisse concevoir seul de *belles* synthèses. Par contre, il a déjà montré qu'il savait manipuler des données chimiques, organiser des connaissances, appliquer des raisonnements, trouver des régularités, etc. et être ainsi un outil efficace susceptible de stimuler l'esprit d'invention des chimistes. Par ailleurs, les systèmes d'aide à la synthèse sont aussi des outils pédagogiques. Certains d'entre eux sont notamment utilisés dans le cours de *logique de la synthèse organique* donné par Jacques Coste à l'Ecole Nationale Supérieure de Chimie de Montpellier.

7. Remerciements

Merci à J. Coste, A. Dietz, O. Gien, P. Vismara et, d'une manière générale aux membres du GDR 1093 du CNRS, dont le travail m'a beaucoup aidé dans la rédaction de cet article.

8. Bibliographie

[Corey &Wipke 69] Corey E.-J., Wipke W.-T., Computer-Assisted Design of Complex Organic Syntheses, *Science*, 166, 1969, 178-192.

[Corey & al 85] Corey E.-J., Long A.-K., Rubenstein S.-D., Computer-assisted Analysis in Organic Synthesis, *Science*, 228, 1985, 408-418.

[Corey & Cheng 89] Corey E.-J., Cheng X.-M., *The Logic of Chemical Synthesis*, John Wiley & Sons, New York, 1989.

[David & al 93] *Second Generation Expert Systems*, David J.-M., Krivine J.-P., Simmons, R. (Eds), Springer Verlag, 1993.

[Hendrickson 90] Hendrickson J.-B., Organic Synthesis in the Age of Computers, *Angew. Chem. Int. Ed. Engl.*, 29, 1990, 1286-1295.

[Holton & al 94] Holton R.-A., Somoza C., Kim H.-B., Liang F., Biediger R.-J., Boatman P.-D., Shindo M., Smith C.-C., Kim S., Nadizadeh H., Suzuki Y., Tao C., Vu P., Tang S., Zhang, P., Murthi K.-K., Gentile L.-N. et Liu J.-H., First Total Synthesis of Taxol. 1. Functionalization of the B Ring, *J. Am. Chem. Soc.*, 1994, 116, 1597-1598 ; 2. Completion of the C and D Rings, *idem*, 1599-1600.

[Jambaud 96] Jambaud P., *Le dialogue comme processus de résolution de problème*, Thèse de l'Université Montpellier II, 1996.

[Jorgensen 90] Jorgensen W.-L., Laird E.-R., Gushurst A.-J., Fleisher J.-M., Gothe S.-A., Helson H.-E., Paderes G.-D., Sinclair S., CAMEO: a program for the logical prediction of the products of organic reactions, *Pure & Appl. Chem.*, 62, 1990, 1921-1932.

[Kahney 93] Kahney H., *Problem Solving* (2nd ed.), Open University Press, 1993.

[Lieber 97] Lieber J., *Raisonnement à partir de cas et classification hiérarchique*, Thèse de l'Université Nancy I, 1997.

[Nicholaou & Sorensen 96] Nicolaou K.-C., Sorensen E.-J., *Classics in Total Synthesis*, VCH, 1996.

[Ott & Noordik 92] Ott M.-A, Noordik J.-H., Computer tools for reaction retrieval and synthesis planning in organic chemistry, *Recl. Trav. Chim. Pays-Bas*, 111, 1992, 239-246.

[Régin 95] Régin J.-C., *Développement d'outils algorithmiques pour l'intelligence artificielle*, Thèse de l'Université Montpellier II, 1995.

[Ugi & al 94] Ugi I., Bauer J., Blomberger C., Brandt J., Dietz A., Fontain E., Gruber B., von Scholley-Pfab A., Senff A., Stein N., 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, 1994, 3-16.

[Vléduts 63] Vléduts G.-E., Concerning one System of Classification and Codification of Organic Reactions, *Inf. Stor. Retr.*, 1, 1963, 117-146.

[van Rozendaal 94] van Rozendaal E.-L.-M., Ott M.-A., Scheeren, H.-W., A LHASA analysis of Taxol, *Recl. Trav. Chim. Pays-Bas*, 113, 1994, 297-303.

[Vismara 95] Vismara P., *Reconnaissance et représentation d'éléments structuraux pour la description d'objets complexes*, Thèse de l'Université Montpellier II, 1995.

[Woodward 56] Woodward R.-B., in *Perspectives in Organic Chemistry*, Todd, A. R. (Ed.), New York, Interscience, 1956.