THESIS / THÈSE

MASTER EN SCIENCES INFORMATIQUES

Mise au point d'outils en synthèse assistée sur ordinateur

Samyn, Dirk

Award date:
1977

Awarding institution:
Universite de Namur

FACULTÉS UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR
INSTITUT D'INFORMATIQUE

Année accadémique 1976-1977

MISE AU POINT D'OUTILS EN SYNTHÈSE ASSISTÉE SUR ORDINATEUR

Dirk SAMYN

Mémoire présenté en vue de l'obtention
du grade de
Licencié et Maître en Informatique

REMERCIEMENTS

J'exprime mon entière gratitude à Monsieur J.M. ANDRE , qui a bien voulu diriger et orienter ce travail. Ses critiques et conseils en ont facilité l'élaboration et la rédaction.

Qu'il me soit également permis de remercier Monsieur A. KRIEF , directeur du laboratoire de chimie organique, pour l'intérêt porté à ce mémoire, dont la conception a été suggérée par Monsieur R. PICHIOTTINO , chercheur en synthèse assistée à l'université de PARIS VI et VII .

Je tiens, enfin, à remercier les membres du laboratoire de chimie théorique dirigé par Monsieur J.M. ANDRE , pour leur accueil et pour leur collaboration efficace et aide directe dans l'utilisation de leur matériel informatique.

TABLE DES MATIERES

DEUXIEME PARTIE : BASE DE DONNEES TOPOLOGIQUES

TROISIEME PARTIE : BASE DE DONNEES INFORMATIONS

QUATRIEME PARTIE : SYNTHESE ASSISTEE

INTRODUCTION

Deux des principaux champs d'investigation de la chimie organique sont l'analyse et la synthèse de produits organiques. L'analyse identifie les composés chimiques; la synthèse tente de produire de nouvelles molécules en faisant réagir des substances connues et disponibles (et des composés déjà synthétisés). Les deux champs sont complémentaires; le chimiste recourt, en effet, à l'analyse pour vérifier et identifier les substances qu'il a synthétisées.

Les problèmes en synthèse peuvent être divisés en deux classes:

CLASSE I: on donne un produit de départ A et on étudie ses propriétés, notamment les réactions possibles;

$$\boxed{A \longrightarrow ?}$$

Cas particulier: partant d'un produit de départ A et d'une structure cible T, on désire connaître un chemin réactionnel optimal permettant de passer de A à T;

$$\boxed{A \longrightarrow T}$$

CLASSE II: connaissant une structure cible T, on désire trouver les précurseurs possibles de T.

$$\boxed{? \longrightarrow T}$$

En pratique, une synthèse exige:

1: le choix d'une molécule à synthétiser (ce qui implique la description de sa structure);
2: la formulation d'un ensemble de chemins de synthèse "valides";
3: la sélection des étapes individuelles de réaction et leur ordonnancement pour exécution au laboratoire;
4: l'exécution des synthèses et la vérification des résultats;
5: la révision des synthèses si nécessaires.

Le but de la synthèse assistée est de voir l'aide que peut apporter l'ordinateur durant la deuxième phase (formulation des chemins). La figure 1 donne un exemple de chemins trouvés par le programme de Barone (référence 47) pour synthétiser un alkyl thiazole.

Supposons que nous ayons à synthétiser une molécule dénommée "molécule cible" et notée T. Une façon d'aborder le problème est de trouver toutes les réactions qui produisent directement T. Soient C_1, C_2, ..., C_n les prédécesseurs possibles de T (figure 2): ils donnent T en une seule étape. Il s'agit ensuite de créer les molécules C_1, C_2, ..., C_n: C_{11}, C_{12}, ..., C_{1i} sont les prédécesseurs de C_1. Cette procédure est répétée jusqu'à l'obtention d'un ensemble de molécules considérées comme substances disponibles en laboratoire.

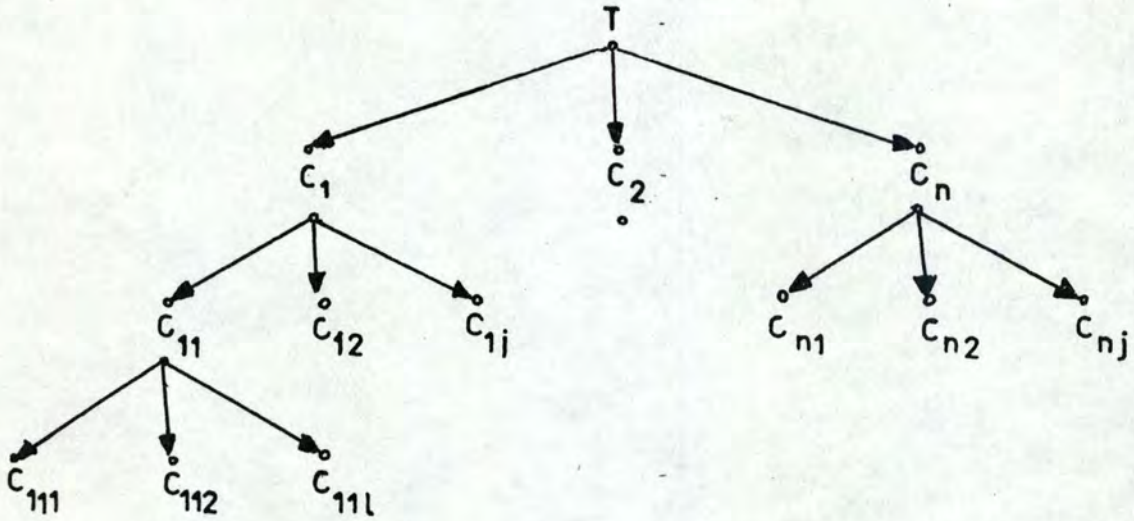


figure 2

Un chemin de synthèse est donc, par exemple: $C_{11} \rightarrow C_1 \rightarrow T$. L'ensemble des chemins de synthèse forme un arbre de synthèse.

Si chaque molécule a, en moyenne, n prédécesseurs et si nous utilisons une synthèse de k étapes, nous avons n^k séquences possibles. Si n égale 40, n^k, pour k=5, vaut 10^8. Certains de ces chemins sont soit absurdes, soit inefficaces. Il existe plusieurs stratégies qui permettent de les éliminer et de choisir les chemins les plus probables et les plus avantageux. Le but de ce mé-

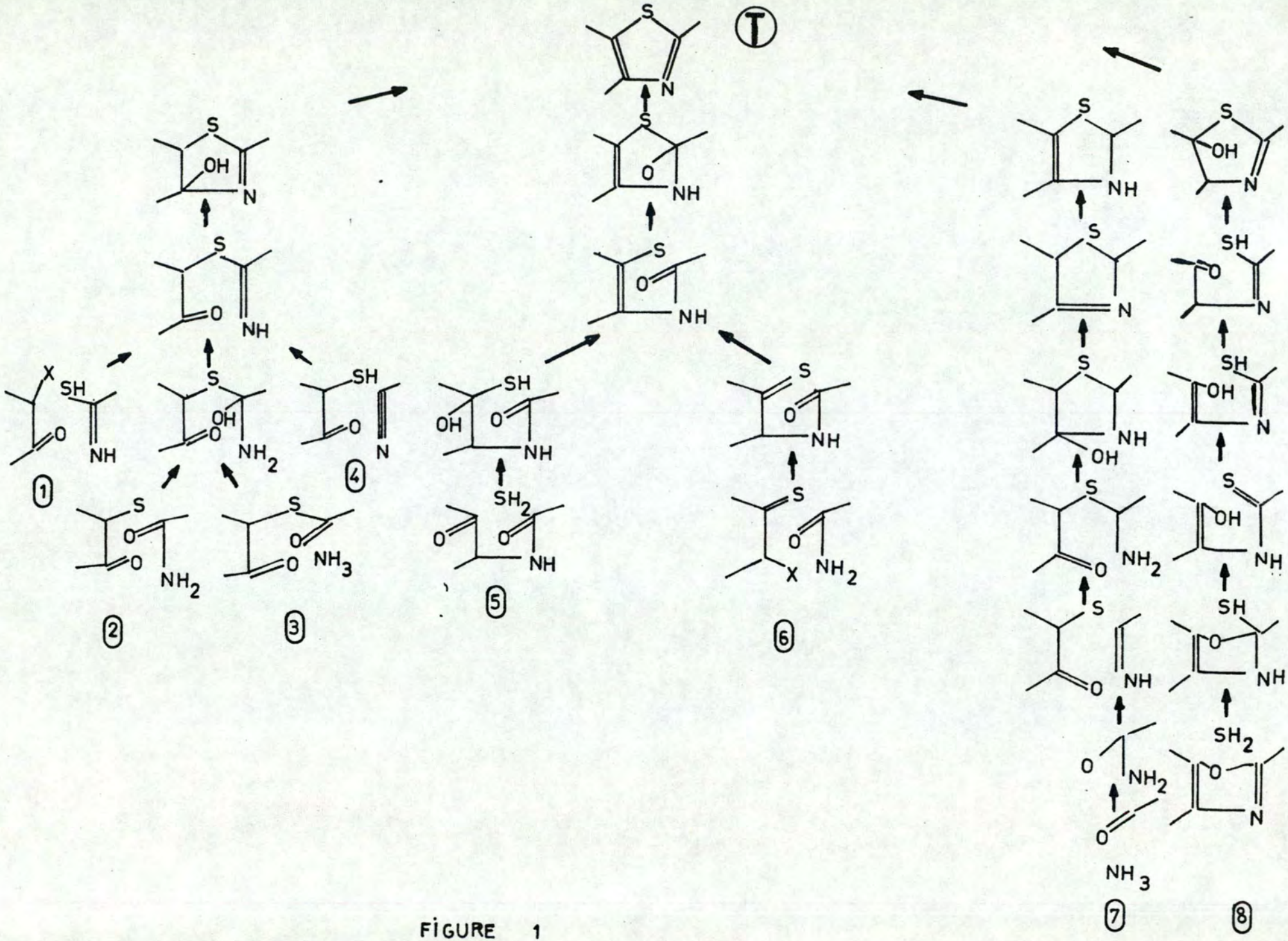


FIGURE 1

moire n'est pas d'imaginer une nouvelle stratégie mais plutôt de concevoir et, éventuellement, d'implémenter des outils de travail en synthèse. Ces outils ne sont valables que pour une synthèse où:

1: les matériaux de départ sont des composés commercialisés;
2: les composés intermédiares peuvent être nouveaux;
3: les différentes étapes de synthèse correspondent à des réactions connues.

Les outils suivants ont été considérés:

- représentation d'une structure moléculaire;
- base de données des composés organiques actuellement connus et répertoriés;
- représentation d'une réaction;
- base de données des réactions existantes;
- système bibliographique permettant de retrouver les références d'un composé ou d'une réaction;
- gestion des produits dans un laboratoire;
- acquisition et restitution de structures.

La majorité des stratégies de synthèse étant basée sur les groupes fonctionnels et les cycles de la structure cible, un dernier outil à considérer sera la reconnaissance de ces deux entités structurales.

Dans la première partie de ce travail, nous donnons une définition d'une base de données (BD) et introduisons les notions de BD informations et topologiques.

Grâce à la théorie des graphes, nous décrivons, dans la deuxième partie, un modèle de BD topologiques après avoir étudié les concepts de structure et de réaction.

Les troisième et quatrième parties contiennent la description d'un modèle de BD informations et un exposé de son implémentation.

L'exploitation de la BD générale étant le résultat de l'exploitation conjuguée des BD topologiques et informations, nous

étudierons dans cette partie une façon de relier des informations provenant de ces BD.

Enfin, la cinquième partie comporte des méthodes de représentation d'informations propres à la synthèse assistée et un algorithme de reconnaissance de cycles dans une structure.

NOTES:

- l'implémentation du système bibliographique, de la gestion des produits et de l'acquisition et restitution de structures a été effectuée sur le matériel disponible au département de chimie: PDP 11/45 (64 K-mots) sous système de multiprogrammation RSX-11M, trois disques 2.400.000 caractères, deux bandes magnétiques 7-9 pistes, un écran graphique GT 42, deux vidéo, une télétype, un lecteur de cartes et une table traçante BENSON;

- en appendice, le lecteur trouvera certaines définitions qui lui permettront de mieux saisir les aspects de chimie organique et de théorie des graphes utilisés dans notre exposé.

PREMIERE PARTIE : BASES DE DONNEES

INTRODUCTION

Historiquement, on a laissé se multipier les fichiers spécifiques à une application (figure 3). Par la suite, on a pris conscience de la nécessité de gérer plus rationnellement cet ensemble de données; une nouvelle approche a été définie: la banque de données (figure 4).

La principale caractéristique de cette approche est le renversement de la hiérarchie traitement-données au profit des données. Les données sont saisies et stockées indépendamment des programmes qui doivent les traiter (programmes statistiques, de gestion, d'édition, etc.).

Les avantages espérés d'une telle approche sont une meilleure cohérence des résultats des différents traitements, un gain au niveau du travail de saisie, de codage et de transmission des données, une réduction du volume des mémoires de stockage, suite à la suppression des redondances et une meilleure disponibilité des données pour l'ensemble des utilisateurs.

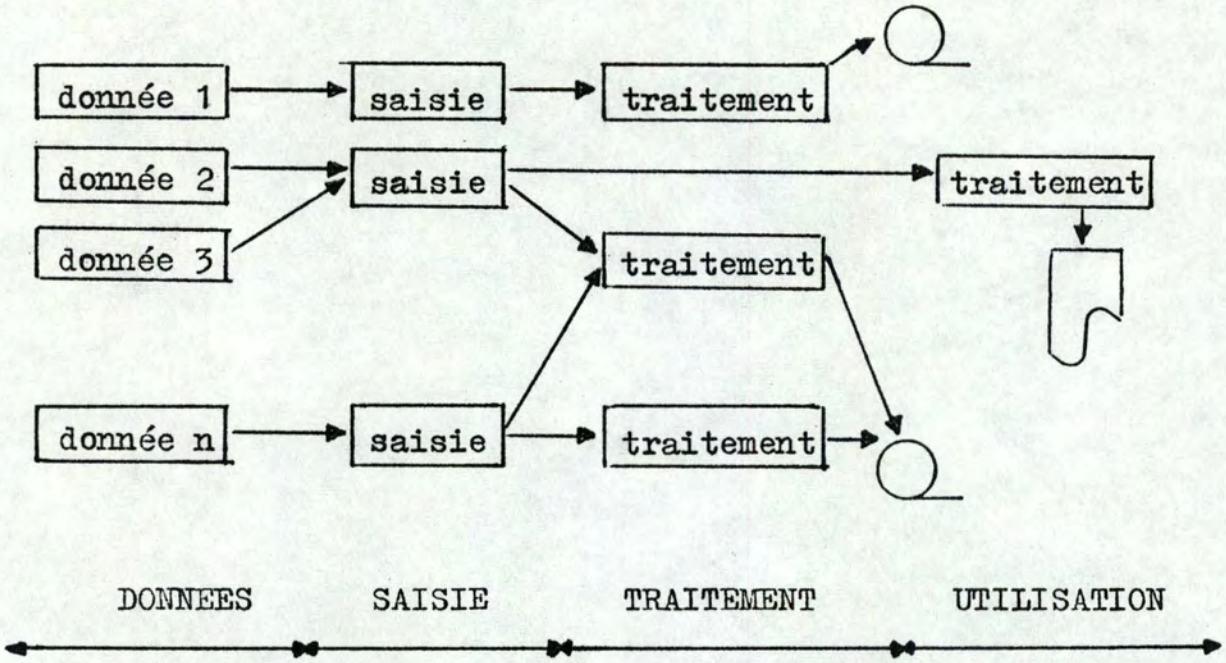


figure 3

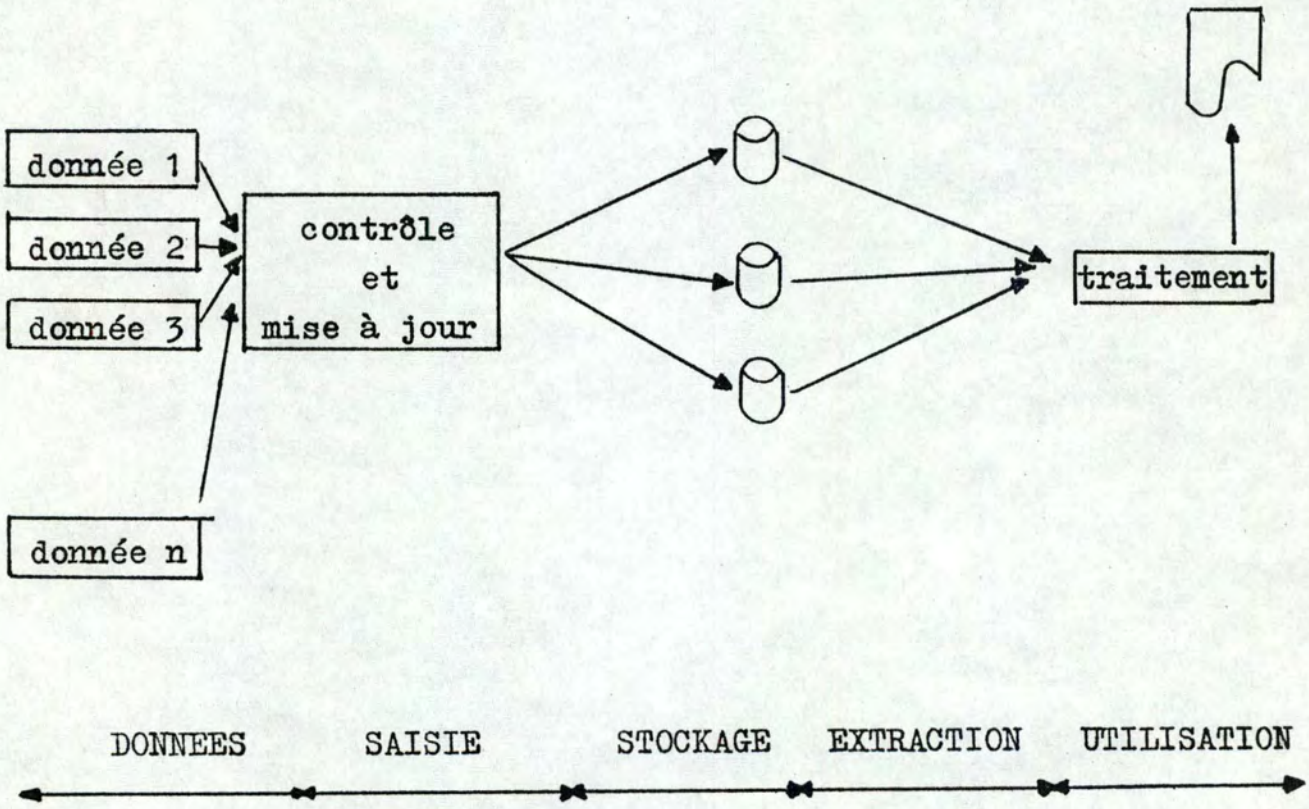


figure 4

1. Principe d'une banque de données

Une banque de données est la réunion de différents ensembles d'informations en un système commun, caractérisé par un ensemble de relations et par l'adoption de règles communes au niveau de la codification et par la mise au point de procédures d'interrogation et d'utilisation suffisamment générales pour être aux besoins de différentes catégories d'utilisateurs.

Une banque de données est donc un ensemble composé des éléments suivants:
- une base de données;
- un système de gestion de la base;
- un système d'exploitation;
- un ensemble de programmes d'exploitation.

2. La base de données

C'est un système physique qui réunit en un ensemble cohérent des données nécessaires à plusieurs fonctions. Son élaboration nécessite de répertorier et de classer ensuite les informations à gérer en considérant les besoins à satisfaire.

Une base peut être constituée par un ensemble cohérent de fichiers physiquement indépendants les uns des autres, mais liés logiquement entre eux (figure 5).

Une base peut être aussi intégrée auquel cas les notions d'enregistrement et de fichier n'ont plus de réalité physique. C'est le système de gestion qui organise la base et crée un ensemble d'informations formant les entités ayant entre elles des relations explicites (figure 6).

La structure "ensemble cohérent de fichiers" facilite les réponses nécessitant le simple balayage des fichiers, tandis que la structure "intégrée" est mieux adaptée aux réponses nécessitant la sélection de groupements d'informations vérifiant certaines propriétés, et aux mises à jour, puisque toutes les redondances

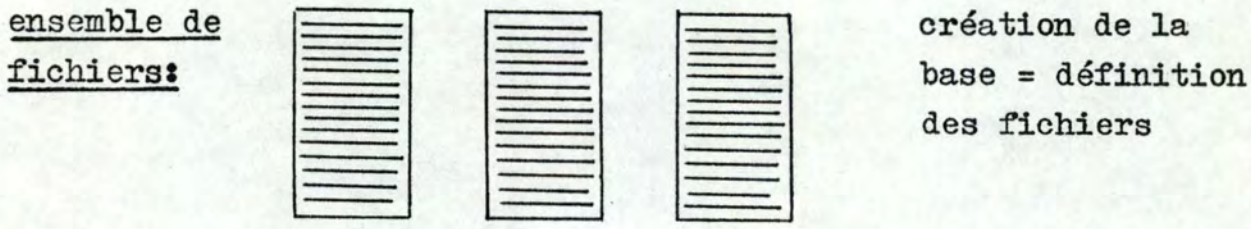


figure 5

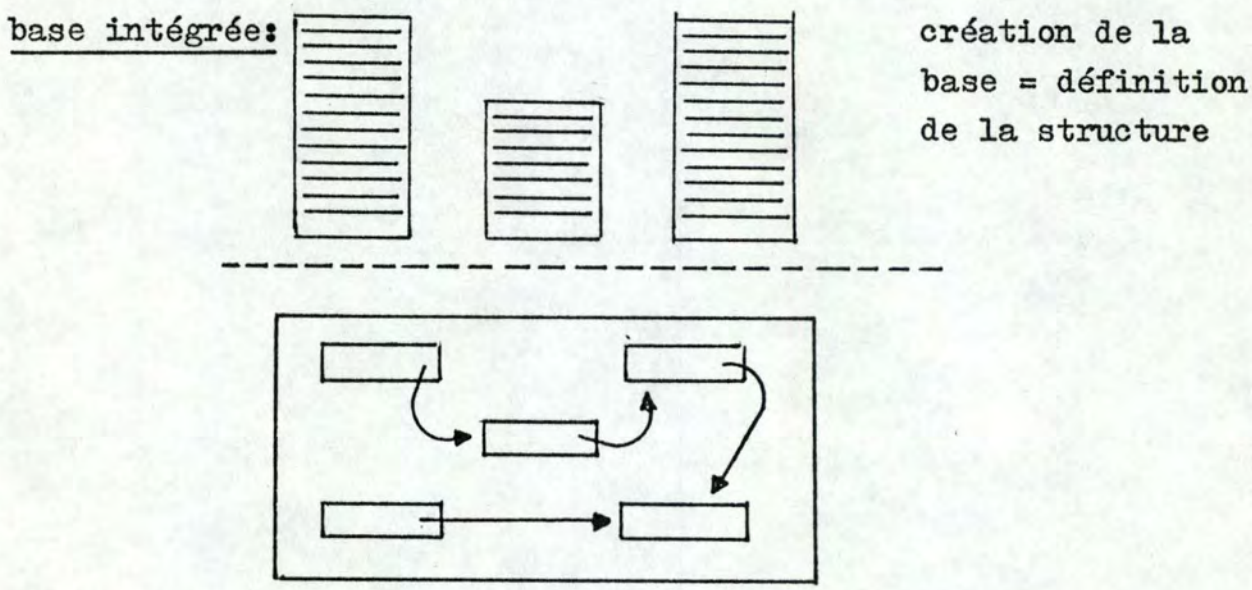


figure 6

sont supprimées et remplacées par des pointeurs.

3. Les bases de données intégrées en chimie

Il n'y a pas d'organisation type pour les bases de données intégrées: la nature des données est un facteur déterminant de l'organisation de la base.

En chimie, les bases de données doivent prendre en compte des informations dont le caractère original est fonction de l'objet propre de la chimie: la molécule. Les bases de données chimiques ne peuvent être organisées qu'autour du concept de molécule: on y trouve, en effet, des données et des informations qui sont toutes associées à des molécules:

- des informations structurales, qui définissent les molécules

(formule développée, par exemple);

- des informations textuelles, qui précisent les données numériques ou bibliographiques, ou les noms systématiques ou triviaux associés aux molécules;
- des informations expérimentales, qui expriment certains comportements (RMN, RX, etc.).

La distinction faite ci-dessus se matérialise par la définition de bases de données distinctes. Chacune de ces bases de données est relative à un type particulier d'informations homogènes (structurales, textuelles ou expérimentales), adapté au traitement automatique.

Nous nous limitons au cas où nous n'avons pas à traiter d'informations expérimentales. La base de données inclut donc:

1: une base de données "informations" (BDI): à chaque composé identifié par son numéro de régistre (NR) correspondent toutes les informations textuelles que l'on juge utiles de mémoriser (origine et disponibilité du produit, références bibliographiques sur une structure, etc.);

2: une base de données "topologique" (BDT): les informations concernant les structures des composés sont mémorisées dans cette base et chaque composé est repéré grâce à son numéro de régistre.

DEUXIEME PARTIE : BASE DE DONNEES TOPOLOGIQUES

INTRODUCTION

La chimie est particulièrement favorisée par rapport à d'autres disciplines car ses concepts fondamentaux sont graphiques dans leur expression:

- les structures (S) ou sous-structures (SS) sont, en effet, représentées par des formules développées qui traduisent intégralement la topologie de la molécule, explicitent les paramètres structuraux de base (nature des liaisons et des atomes, ...) et réfèrent à des informations complémentaires (stéréochimie, longueurs de liaison, ...);
- les hyperstructures (HS) ou ensembles de structures chimiques liées entre elles par des relations formelles ou réactionnelles sont représentées naturellement par des diagrammes de corrélations.

Ainsi, on peut dire que le véritable langage naturel du chimiste est un langage graphique: les lettres sont les atomes ou les liaisons, les mots les structures ou les sous-structures, les phrases les hyperstructures.

Les hyperstructures ne seront pas étudiées dans ce travail (références 30, 31).

CHAPITRE I : REPRESENTATION DES STRUCTURES

1. Graphes en chimie

Le concept de graphe, mathématiquement représenté par G(X,U) où X est l'ensemble des sommets et U l'ensemble des arêtes joue un rôle naturel en chimie.

Si la formule chimique prise dans son ensemble est une représentation naturelle d'une entité, on lui associe toujours une nomenclature dans le but d'arriver à une description linéaire qui peut être manipulée pour diverses opérations. Ce métalangage ne peut être réalisé qu'en identifiant la formule chimique à un graphe approprié. La puissance de la théorie des graphes se trouve dans ses possibilités de description topologique utile pour traiter des combinaisons d'atomes en termes d'entité.

Afin de pouvoir utiliser correctement le concept de graphe en chimie, il est nécessaire de le préciser par la notion de chromatisme.

1.1. Chromatisme

Un graphe chromatique est un graphe dont les sommets X et les arêtes U sont différenciés symboliquement par des "couleurs". Un composé chimique peut donc être assimilé à un graphe chromatique dont les sommets sont les atomes et les arêtes les liaisons, les colorations se référant à des concepts et des relations les unissant. On représente un composé chimique par:

$G_f(X,U,f_X,f_U)$ où
- X est l'ensemble des sommets;
- U est l'ensemble des arêtes;
- f_X est la fonction qui, appliquée à un élément de X, exprime sa couleur;
- f_U est la fonction qui, appliquée à un élément de U, exprime sa couleur.

Le graphe G(X,U) est appelé le graphe topologique associé

au graphe chromatique $G_f(X,U,f_X,f_U)$. En chimie, ce graphe topologique représente le squelette du composé.

Un graphe chromatique partiel est un graphe où certains éléments de X, de U, de f_X, de f_U sont indéterminés. Dans la structure de la figure 7, qui est un graphe chromatique partiel, les atomes X et Y sont indéterminés, de même que les liaisons CH à X et CH à Y.

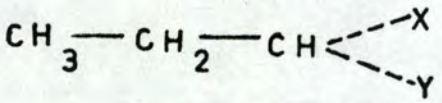


figure 7

1.2. Coloration

L'utilisation du concept de graphe chromatique dans un certain domaine (D) exige un recensement de toutes les couleurs de ce domaine, c'est-à-dire T_D. Ceci implique la détermination des ensembles d'arrivée T_X et T_U des fonctions f_X et f_U dont les ensembles de définition sont X et U.

La fonction f_X d'un graphe chromatique associé à un composé chimique prend ses valeurs dans l'ensemble T_X des 105 éléments de la table préétablie et hiérarchisée conformément au numéro atomique de Mendeleev.

La fonction f_U, quant à elle, prend ses valeurs dans l'ensemble T_U des divers types de liaisons qu'on trouve dans un composé: simple, double, triple, aromatique, tautomérique, dative, ...

1.3. Chromatisme et topologie

En chimie, chromatisme et topologie sont interdépendants: la couleur d'un sommet dépend de la nature d'un atome. Ceci implique la détermination du nombre d'électrons de valence que cet atome échange avec son environnement. En termes topologiques, on dira que la coloration d'un atome impose son degré maximum. Le degré

d'un atome détermine indirectement certains types de coloration des atomes environnants.

NOTE: le graphe chromatique d'une structure peut être simplifié en décidant, par convention, d'exclure les atomes d'hydrogène (degré 1). Le résultat est un graphe "transparent" à partir duquel on peut reconstruire le graphe explicite.

1.4. Terminologie

Le tableau 1 donne une correspondance possible entre certains termes de la théorie des graphes et de la chimie.

Théorie des graphes	Chimie
graphe chromatique	structure chimique
graphe topologique	squelette de la structure
sommet $x \in X$	atome
couleur du sommet	symbole atomique
arête $u \in U$	liaison
couleur de l'arête	valeur de la liaison
arbre	molécule acyclique
cycle	cycle
corde	fermeture de cycle
chaîne de n arêtes	n-polyène
cycle de n arêtes	n-annulène
nombre cyclomatique	nombre de cycles
polynome caractéristique	polynome séculaire
matrice d'incidence	matrice topologique

Tableau 1

2. Représentation des structures

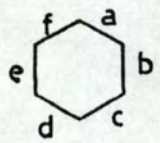
Les représentations sont classées en trois catégories: les représentations matricielles qui rendent aisés les calculs de propriétés physico-chimiques, les tables de connexions dans lesquelles on spécifie tous les atomes de la molécule ainsi que les liaisons, les notations linéaires où l'on désigne les liaisons, les cycles et les groupes fonctionnels par une séquence de symboles moins nombreux que le nombre d'atomes.

2.1. Représentation matricielle

2.1.1. Matrice topologique A

Premier type: matrice topologique des arêtes

$$\boxed{A_{ij}=1 \text{ si l'arête i est adjacente à l'arête j .}}$$



graphe G

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

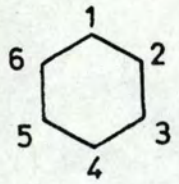
matrice A associée

figure 8

La matrice A correspondant à un graphe G de m arêtes est une matrice mxm symétrique dont le rang est le nombre d'arêtes.

Deuxième type: matrice topologique des sommets

$$\boxed{A_{ij}=1 \text{ si le sommet i est adjacent au sommet j .}}$$



graphe G

$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 0 & 0 & 1 & 0 & 1 \\ 6 & 1 & 0 & 0 & 0 & 1 & 0 \end{array}$$

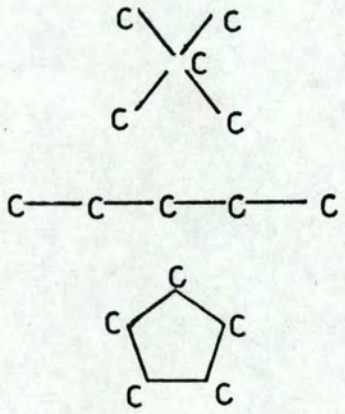
matrice A associée

figure 9

La matrice A correspondant à un graphe G de n sommets est une matrice nxn symétrique dont le rang est le nombre de sommets.

Cette matrice A non seulement décrit l'ensemble de la structure mais permet le calcul de plusieurs propriétés physicochimiques qui dépendent de la topologie:
soit A la matrice topologique associée à un graphe G;
le polynôme $p(x) = \det. \left| A + x E \right| = \sum_{i=1}^{n} k_i x^{n-i}$ où
- E est une matrice unité de même taille que A,
- x est une variable,
- n est le nombre de sommets,
- k_i est le $i^{ème}$ coefficient du polynôme,

est appelé polynôme caractéristique associé à G. La figure 10 donne quelques exemples de polynômes caractéristiques.



$x^5 - 4x^3$

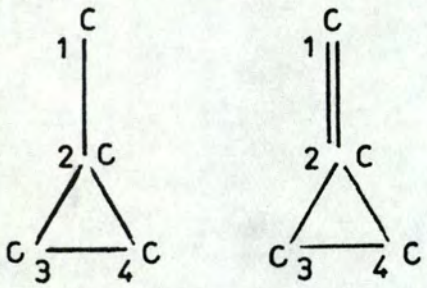
$x^5 - 4x^3 + 3x$

$x^5 - 5x^3 + 5x - 2$

figure 10

En égalant ce polynôme à 0, on obtient l'équation polynomiale $p(x) = 0$ et n valeurs propres (n = le nombre d'atomes).

Il faut noter que $p(x)$ ne définit pas biunivoquement le graphe d'un composé: deux graphes qui ne sont pas équivalents topologiquement (c'est-à-dire non-isomorphiques) peuvent, dans certaines conditions, donner le même polynôme caractéristique et donc le même spectre de valeurs propres (figure 11). Pour cette raison, il

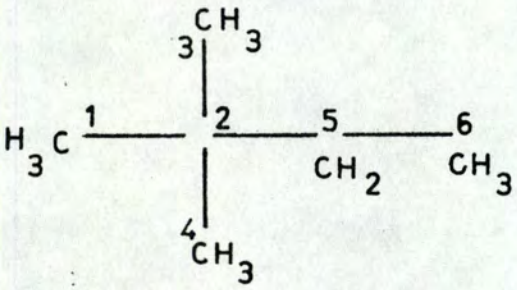


graphe G graphe G'

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 \\ 3 & 0 & 1 & 0 & 1 \\ 4 & 0 & 1 & 1 & 0 \end{array}$$

$p(x) = x^4 - 4x^2 - 2x + 1$

figure 11



$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 & 1 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 5 & 0 & 1 & 0 & 0 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

graphe G matrice topologique

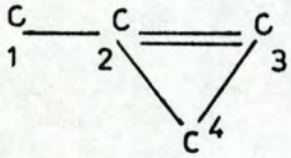
équation caractéristique: $x^6 - 5x^4 + 3x^2$

valeurs propres: $\begin{pmatrix} -2.075 & -0.835 & 0.0 \\ 0.0 & +0.835 & +2.075 \end{pmatrix}$

figure 12

peut être nécessaire de préciser la matrice A:

A_{ii} = symbole atomique de l'atome i,
A_{ij} = valeur de la liaison entre l'atome i et l'atome j .



graphe G

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & C & 1 & 0 & 0 \\ 2 & 1 & C & 2 & 1 \\ 3 & 0 & 2 & C & 1 \\ 4 & 0 & 1 & 1 & C \end{array}$$

matrice A associée

$$C^4 - 7C^2 + 4C + 1$$

polynôme caractéristique associé

figure 13

Cette nouvelle matrice A caractérise biunivoquement le graphe associé G.

La méthode de détermination du polynôme caractéristique repose sur un algorithme issu du théorème de Sachs (référence 2). Soient un graphe G de n sommets et S_n l'ensemble des sous-graphes disjoints contenant n sommets ou un cycle de taille n.

Théorème: les coefficients du polynôme caractéristique sont donnés par:

$$\text{(I)} \qquad \begin{cases} k_o = 1 \\ k_n = \sum_{S_n} (-1)^c 2^r \end{cases}$$

où - i = nombre de composantes simplement connexes de chaque sous-graphe associé,

- n = nombre de sommets considérés pour l'ensemble S_n,
- r = nombre de cycles dans S_n.

Algorithme: étape 1: construire les ensembles S_n,

étape 2: calculer k_o, k_n au moyen des relations (I).

Exemple:

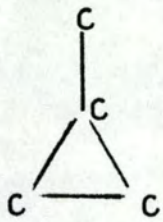
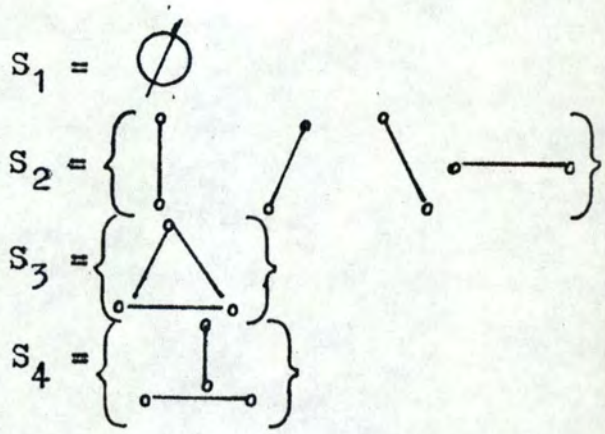


figure 14



$k_o = 1$

$k_1 = 0$

$k_2 = (-1)^1 2^0 + (-1)^1 2^0 + (-1)^1 2^0 + (-1)^1 2^0 = -4$

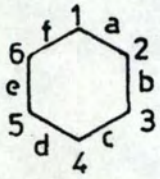
$k_3 = (-1)^1 2^1 = -2$

$k_4 = (-1)^2 2^0 = 1$

polynôme caractéristique: $x^4 - 4x^2 - 2x + 1$

2.1.2. Matrice d'incidence I

$I_{ij} = 1$ si la $j^{ème}$ arête est incidente au $i^{ème}$ sommet.



graphe G

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

matrice I associée

figure 15

La matrice I correspondant à un graphe de m arêtes et de n sommets est une matrice mxn.

Harary (référence 2) a montré qu'il existe une relation

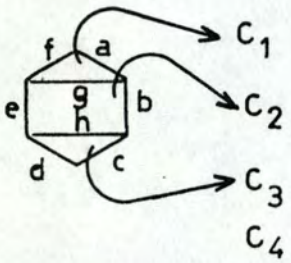
entre A et I:

soient - un graphe G de n sommets et de m arêtes,
- I(G) la matrice d'incidence de G,
- A(G) la matrice topologique de G,
- L(G) la matrice formée en remplaçant les arêtes de G par des sommets tels que deux d'entre eux sont liés si les arêtes correspondantes de G sont adjacentes,

alors $A(L(G)) = I(G).I(G)^t - 2.U$, où "t" signifie la transposée d'une matrice et U la matrice unité mxm.

2.1.3. Matrice des cycles C

C_{ij} = 1 si la $j^{ème}$ arête appartient au $i^{ème}$ cycle.



graphe G

$$\begin{array}{c} \\ C1 \\ C2 \\ C3 \\ C4 \end{array} \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

matrice C associée

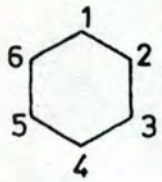
figure 16

Pour une molécule ayant l cycles indépendants, la matrice C correspondant à son graphe G composé de m arêtes est une matrice lxm.

Harary (référence 2) a démontré la relation suivante:
$C.I^t \equiv 0 \pmod 2$ pour tout graphe G.

2.1.4. Matrice des distances D

D_{ij} = nombre minimum d'arêtes nécessaires pour passer du sommet i au sommet j,
= ∞ s'il n'existe pas de chemin allant de i à j.



$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 2 & 3 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 & 3 & 2 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 5 & 2 & 3 & 2 & 1 & 0 & 1 \\ 6 & 1 & 2 & 3 & 2 & 1 & 0 \end{array}$$

figure 17

graphe G

matrice D associée

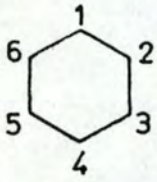
La matrice D correspondant à un graphe G de n sommets est une matrice nxn.

2.1.5. Matrice de Hückel H

C'est une matrice (tridiagonale pour une molécule acyclique) où:
- α = l'intégrale de Coulomb pour deux atomes adjacents,
- β = l'intégrale correspondante de résonance entre ces deux atomes,
- ε' = une valeur propre de l'énergie pour ce système.

Dans une théorie de Hückel simplifiée, tous les termes en α sont égaux, ainsi que les termes en β : les liaisons sont considérées comme équivalentes (figure 18). Le déterminant de la matrice s'appelle le <u>déterminant séculaire</u>.

On peut montrer l'équivalence entre les matrices topologiques et les matrices de Hückel (référence 3). Grâce à ce résultat, nous pouvons tirer de la matrice A des propriétés physico-chimiques, notamment:
- le calcul des orbitales moléculaires pour une valeur propre donnée d'une molécule,
- le calcul de la densité de charge autour d'un noyau atomique,
- etc. .



$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & \alpha-\varepsilon' & \beta & 0 & 0 & 0 & \beta \\ 2 & \beta & \alpha-\varepsilon' & \beta & 0 & 0 & 0 \\ 3 & 0 & \beta & \alpha-\varepsilon' & \beta & 0 & 0 \\ 4 & 0 & 0 & \beta & \alpha-\varepsilon' & \beta & 0 \\ 5 & 0 & 0 & 0 & \beta & \alpha-\varepsilon' & \beta \\ 6 & \beta & 0 & 0 & 0 & \beta & \alpha-\varepsilon' \end{array}$$

graphe G

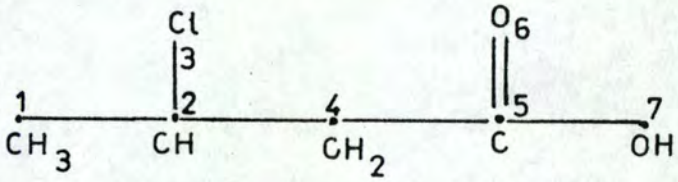
matrice de Hückel associée

figure 18

2.2. Table de connexions

Les atomes de la structure sont d'abord numérotés de façon arbitraire. Une table d'identification est ensuite construite où, pour chaque atome, nous indiquons sa couleur (au sens de la théorie de graphes), celle des liaisons auxquelles il participe, ainsi que le numéro des atomes adjacents. Les figures 19 et 20 donnent les tables de connexions pour une structure acyclique et une structure cyclique.

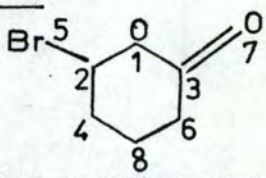
structure acyclique:



numéro de l'atome	couleur de l'atome	numero atome adjacent	couleur de la liaison	numéro atome adjacent	couleur de la liaison	numéro atome adjacent	couleur de la liaison
1	C	2	1	-	-	-	-
2	C	1	1	3	1	4	1
3	Cl	1	1	-	-	-	-
4	C	2	1	5	1	-	-
5	C	4	1	6	2	7	1
6	O	5	2	-	-	-	-
7	O	5	1	-	-	-	-

figure 19

structure cyclique:



numéro de l'atome	couleur de l'atome	numéro atome adjacent	couleur de la liaison	numéro atome adjacent	couleur de la liaison
1	O	2	1	3	1
2	C	1	1	5	1
3	C	1	1	7	2
4	C	2	1	8	1
5	Br	2	1	-	-
6	C	3	1	8	1
7	O	3	2	-	-
8	C	4	1	6	1

figure 20

L'inconvénient de ce genre de table est la présence d'informations redondantes: chaque liaison est notée deux fois. Il est possible d'éliminer cette redondance en prenant un certain nombre de conventions:

convention 1: lors de la numérotation de la structure, dès qu'un atome est numéroté, tous les atomes adjacents à ce dernier sont numérotés en série;

convention 2: dans la table de connexions, seules les liaisons vers des atomes de numéro inférieur sont citées.

La table ainsi construite s'appelle table de connexions compactée. Pour les structures des figure 19 et 20 , les tables de connexions compactées sont:

numéro de l' atome	couleur de l' atome	numero atome adjacent	couleur liaison
1	C	-	-
2	C	1	1
3	Cl	2	1
4	C	2	1
5	C	4	1
6	O	5	2
7	O	5	1

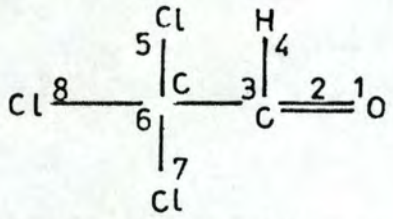
figure 21

numéro de l' atome	couleur de l' atome	numéro atome adjacent	couleur liaison
1	O	-	-
2	C	1	1
3	C	1	1
4	C	2	1
5	Br	2	1
6	C	3	1
7	O	3	2
8	C	4	1
fermeture de cycle:		6-8	1

figure 22

Table de connexions de Ray et Kinsch:

Les atomes et toutes les liaisons multiples sont numérotés; on leur associe les atomes et liaisons adjacents (figure 23).



numéro	couleur	adjacents
1	O	2
2	2	1,3
3	C	2,4,6
4	H	3
5	Cl	6
6	C	3,7,5,8
7	Cl	6
8	Cl	6

figure 23

2.3. Notations linéaires

Le chimiste possède des nomenclatures lui permettant d'identifier les composés chimiques. Il s'agit donc de prévoir, durant la conception d'une base de données chimique, la possibilité de l'interroger au moyen de ces langages. En 1957, IUPAC (International Union of Pure and Applied Chemistry) a standardisé une nomenclature des composés chimiques acceptée par la plupart des chimistes. L'interrogation de la base peut aussi se faire en donnant directement la formule moléculaire. Deux problèmes se posent: comment vérifier la validité d'une notation et comment la traduire en une représentation plus accessible par l'ordinateur (table de connexions) ? Nous introduisons ici quelques éléments de grammaire des langages. Nous donnons ensuite les règles syntaxiques du langage IUPAC et des formules moléculaires, et les règles sémantiques nécessaires pour traduire ces notations en tables de connexions.

2.3.1. Grammaire des langages

Une phrase française "simple" est formée d'un groupe-nom, suivi d'un verbe et d'un groupe-nom. Le groupe-nom est un nom commun précédé d'un article. On dispose alors d'un dictionnaire qui indique les mots que l'on peut utiliser comme nom, article, verbe. Par exemple, un verbe sera "soigner" ou "traiter"; un nom commun sera "médecin" ou "patient", un article "un" ou "le", ... La phrase: "le médecin soigne un patient" sera une phrase correcte, car elle est construite en respectant la structure définie par la grammaire. On peut représenter cette structure par le diagramme suivant:

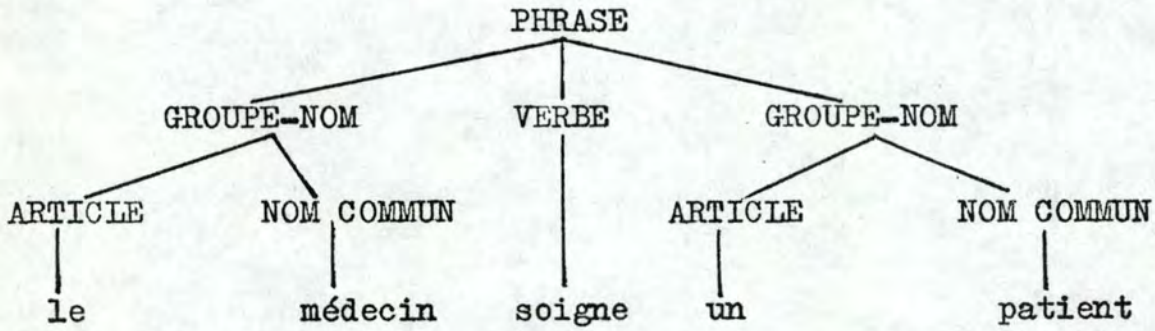


figure 24

La phrase est donc syntaxiquement correcte. Le même diagramme permet d'affirmer que "le patient soigne un médecin" est aussi correct. La syntaxe ne s'intéresse donc qu'à la structure et la construction et non à la signification (la sémantique) de la phrase. Le langage de description de la grammaire, c'est-à-dire les termes qui figurent dans le diagramme, s'appelle métalangage.

Une des notations les plus simples permettant de décrire les règles d'une grammaire est la forme normale de Backus B.N.F. (Backus Normal Form):

::=	signifiant "est défini par",
|	signifiant "ou bien",
< >	entre lesquels seront notées les "notions" (termes

du métalangage),

$[\cdot\cdot\cdot]_1^n$ signifiant que la(es) variable(s) métalinguistique(s) incluse(s) dans les crochets peut (peuvent) être répétée(s) de 1 à n fois.

2.3.2. Langages

On appelle alphabet un ensemble fini non vide d'éléments appelés symboles. Un mot sur cet alphabet est une suite finie de symboles pris dans cet alphabet.

On appelle grammaire formelle G un quadruple G = (V,T,P,A) où:

1: V est un alphabet général ou vocabulaire,
2: T est un alphabet terminal ($T \subset V$),
3: P est un ensemble fini de productions ou règles de production de G,
4: A est un symbole distingué de V - T appelé axiome de G.

L'ensemble V est celui des éléments du métalangage et du langage: groupe-nom, article, médecin, le, ... L'ensemble T regroupe les éléments du langage: médecin, examiner, un, ... P est un ensemble de doublets (u,v) notés u ::= v (u est la partie gauche de la production (u,v), celle qui contient l'axiome, v est la partie droite).

On appelle monoïde sur un alphabet T et noté T^*, l'ensemble de tous les mots sur cet alphabet.

NOTES: - par opposition à T, alphabet terminal, V - T sera souvent appelé alphabet non-terminal de G, ou notions de G;
- l'axiome de G doit apparaître au moins une fois en partie gauche de P. Il peut apparaître en partie droite.

2.3.3. Grammaire des formules moléculaires

Chaque atome est caractérisé par un nom unique (symbole atomique), un poids unique (poids atomique) et un numéro unique (numéro atomique). Les atomes sont combinés entre eux pour former les composés chimiques représentés par une formule moléculaire. A cha-

que atome entrant dans la combinaison est associé un nombre indiquant la proportion de l'atome présent dans le composé (s'il n'est pas présent, par convention, il est égal à 1). Exemples: H_2O (formule moléculaire de l'eau), H_2SO_4 (formule moléculaire de l'acide sulfurique).

Les éléments du métalangage sont: formule (F), chaîne composée (CC), chaîne ionique (CI), chaîne non-ionique (CNI), chaîne (C), chaîne élémentaire (CE), chaîne fermée (CF), facteur de répétition (FR), facteur de répétition intégral (FRI), facteur de répétition non-intégral (FRNI), chiffre (CH), symbole atomique (SA), signe (S); ceux du langage sont: 0,1,2,3,4,5,6,7,8,9,+,-, et les 105 symboles atomiques de la table de Mendeleev. La formule est l'axiome et les productions sont données dans la figure 25 . Les figures 26, 27, 28 donnent les vérifications syntaxiques de quelques formules moléculaires.

2.3.4. Grammaire du système IUPAC (référence 17)

Dans ce système de nomenclature, on exprime:

1: la longueur de la chaîne de carbone (PENT = une chaîne de longueur 5, HEX de longueur 6, ...),
2: un ou plusieurs préfixes indiquant le numéro du carbone auquel est attaché un groupe fonctionnel, suivi par ce dernier,
3: les numéros des carbones et les types de non-saturation qui constituent le suffixe primaire,
4: un suffixe secondaire reprenant les points d'attaches des groupes fonctionnels,
5: le type de fonctionalité (simple ou multiple).

Le composé est l'axiome; les figures 29 et 30 fournissent les productions et un exemple de vérification syntaxique.

2.3.5. Règles sémantiques associées

Les grammaires doivent s'accompagner d'un certain nombre de règles sémantiques afin de transformer un nom de composé en table de connexion plus facilement manipulable.

P11 <formule> ::= <chaîne composée>

P12 <chaîne composée> ::= <chaîne non-ionique>|<chaîne ionique>

P13 <chaîne ionique> ::= [<chaîne non-ionique>]<facteur répétition intégral><signe> |[<chaîne non-ionique>]<signe>

P14 <chaîne non-ionique> ::= $\left[\left\{\begin{array}{l}\text{<chaîne élémentaire>}\\ \text{<chaîne fermée>}\end{array}\right\}\right]_1^n$ |

$[\text{<chaîne>}]_1^n \left[\left\{\begin{array}{l}\text{<chaîne élémentaire>}\\ \text{<chaîne fermée>}\end{array}\right\}\right]_1^n$

P15 <chaîne> ::= $\left[\left\{\begin{array}{l}\text{<chaîne fermée>}\\ \text{<chaîne élémentaire>}\end{array}\right\}\right]_1^n \left\{\begin{array}{c}-\\ =\\ \equiv\end{array}\right\}$

P16 <chaîne fermée> ::= (<chaîne élémentaire>)<facteur répétition>

P17 <chaîne élémentaire> ::= <symbole atomique><facteur répétition>

P18 <facteur répétition> ::= <facteur répétition intégral>|<facteur répétition non-intégral>

P19 <facteur répétition non-intégral> ::= <facteur répétition intégral>.<facteur répétition intégral>

P110 <facteur répétition intégral> ::= $[\text{<chiffre>}]_1^n$

P111 <chiffre> ::= 0|1|2|3|4|5|6|7|8|9|

P112 <symbole atomique> ::= AC|AL|AM|AS|BA|BK| ... |XE|YT|Y|ZN|ZR

P113 <signe> ::= +|-

figure 25

formule ionique

CR207-

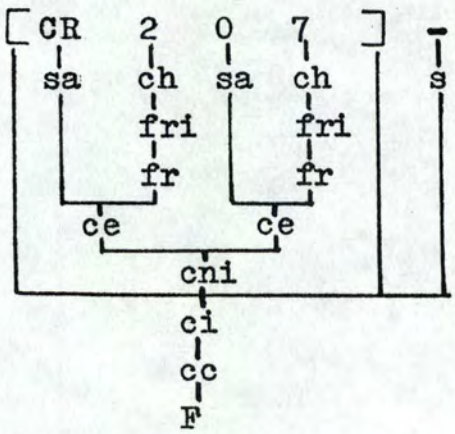


figure 26

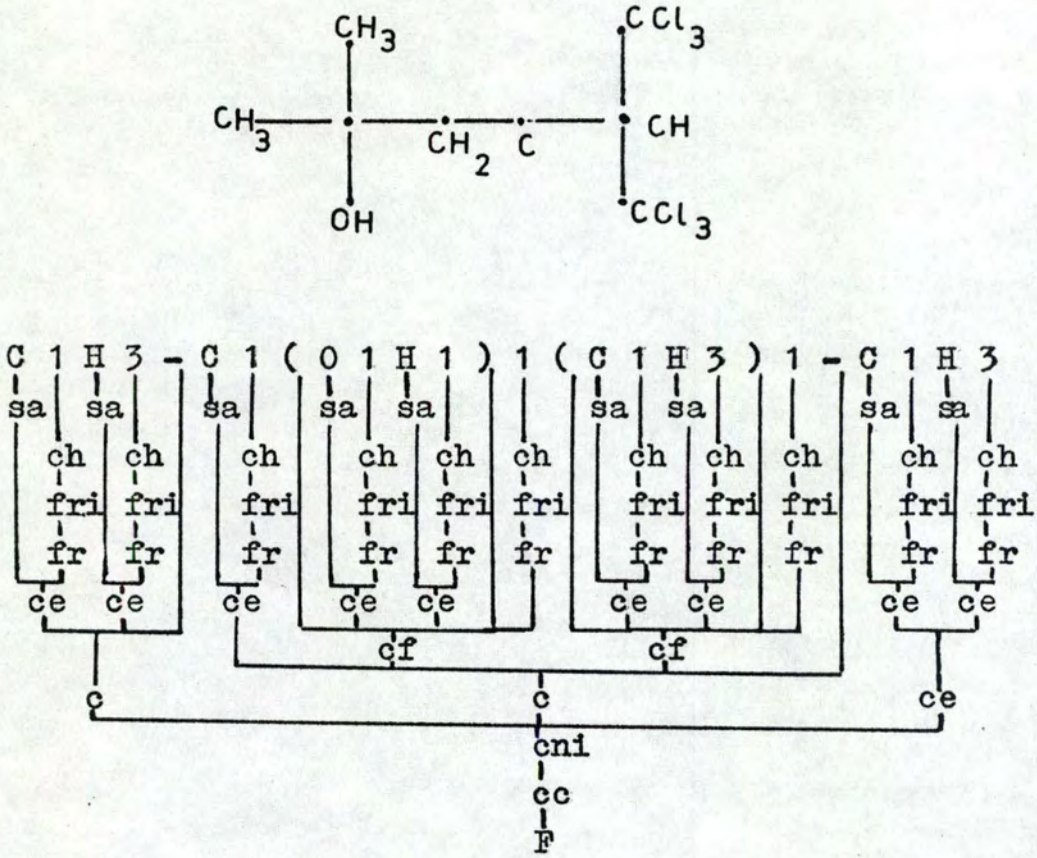


figure 27

composé non-stoichiométrique

ZR1H1.92

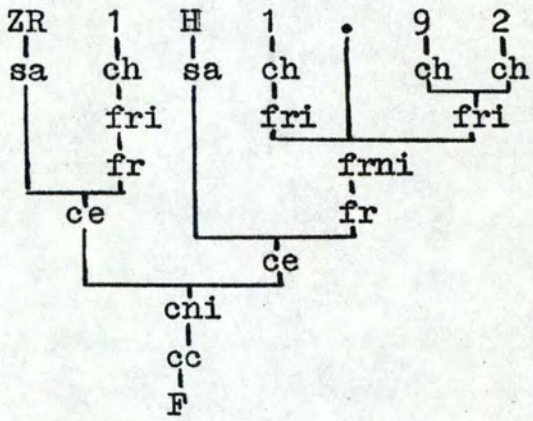


figure 28

```
P21  <composé> ::= <préfixe><carbone><suffixe>|<carbone><suffixe>
P22  <carbone> ::= <alkyl-cyclique>|<non saturation>
P23  <alkyl-cyclique> ::= <alkyl>|CYCLO <alkyl>
P24  <alkyl> ::= METH|ETH|PROP|BUT|PENT|HEX|OCT|NON|DEC
P25  <non-saturation> ::= <deux>-<alkyl-cyclique>ADI|
                          <deux>,<pointeur>-<alkyl-cyclique>ATRI|
                          <pointeur>-<alkyl-cyclique>|
                          <alkyl-cyclique>A |
P26  <préfixe> ::= <préfixe><unité>|<unité>|<unité sans label>
P27  <unité> ::= <pointeur>-<groupe fonct.>-|
                 <groupe de groupes fonct.>-
P28  <unité sans label> ::= <groupe fonct.>-
P29  <suffixe> ::= <suffixe primaire><suffixe secondaire>|
                   <groupe non-saturé><suffixe secondaire>|
                   <suffixe primaire><groupe suffixe><suffixe se-
                   condaire>|<groupe non-saturé><groupe suffixe>
                   <suffixe secondaire>|
P210 <groupe non saturé>::= <groupe non-saturé><groupe suffixe>
                            <suffixe primaire>|<groupe suffixe>
                            <suffixe primaire>|-<suffixe primaire>
P211 <suffixe primaire> ::= AN|EN|YN
P212 <suffixe secondaire> ::= OL|ONE|AL|OIC|ACID|OIC|AMIDE|OATE|
                              ONITRILE|OYL|E
P213 <groupe suffixe> ::= |-<deux fois>|-<trois fois>|-<une fois>
P214 <groupe fonct.> ::= HYDROXY|FLUORO|CHLORO|BROMO|IODO|AMINO|
                         KETO|NITRO|CYANO|METHYL|METHOXY|ETHYL|
                         ETHOXY|PROPYL|BUTYL|BUTOXY|PENTYL|PROPO-
                         XY|
P215 <groupe de groupes fonct.> ::= <deux fois><groupe fonct.>|
                                    <trois fois><groupe fonct.>
P216 <deux fois >::= <deux>-DI
P217 <trois fois> ::= <deux>,<pointeur>-TRI
P218 <une fois> ::= <pointeur>-
P219 <pointeur> ::= entier
P220 <deux> ::= <pointeur>.<pointeur>
```

figure 29

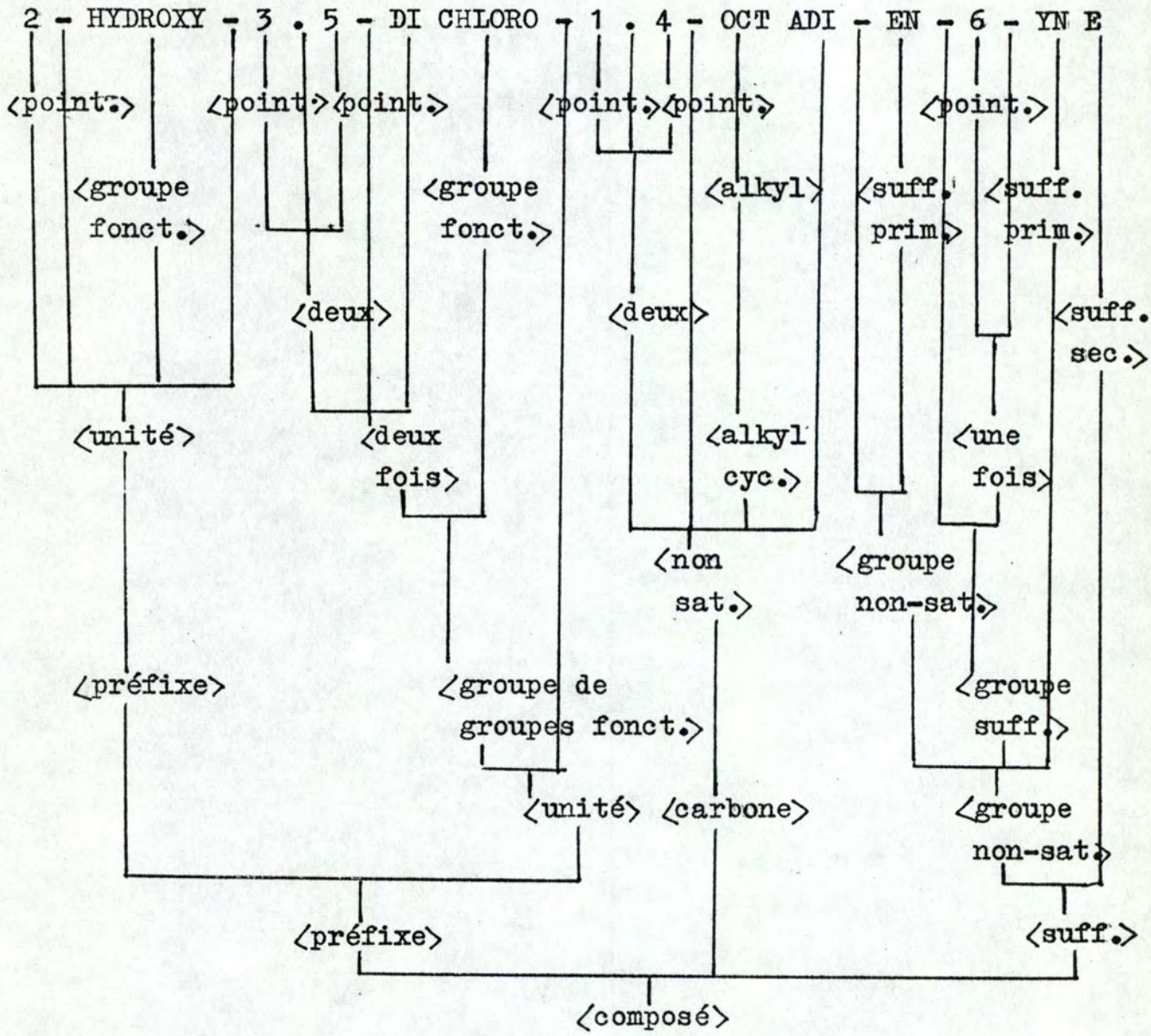

2 - HYDROXY - 3 . 5 - DI CHLORO - 1 . 4 - OCT ADI - EN - 6 - YN E
<point.>
<point.>
<point.>
<point.>
<point.>
<point.>
<groupe fonct.>
<groupe fonct.>
<alkyl>
<suff. prim.>
<suff. prim.>
<deux>
<deux>
<suff. sec.>
<unité>
<deux fois>
<alkyl cyc.>
<une fois>
<non sat.>
<groupe non-sat.>
<préfixe>
<groupe de groupes fonct.>
<groupe suff.>
<unité>
<carbone>
<groupe non-sat.>
<préfixe>
<suff.>
<composé>

figure 30

Grammaire IUPAC: la table de connexions (figure 31) est composée de quatre colonnes: la colonne 1 contient les points d'attache des groupes fonctionnels, la colonne 2 les types de liaisons multiples (double (EN) ou triple (YN)), la colonne 3 les groupes fonctionnels, et la colonne 4 les suffixes secondaires.

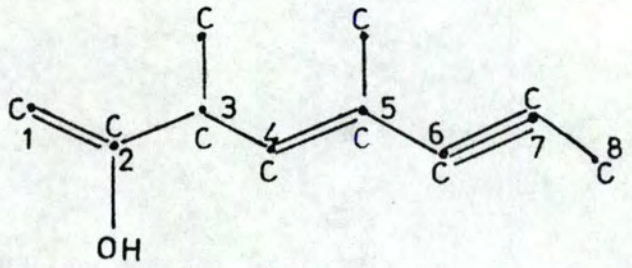
règle syntaxique	règle sémantique associée
P219	créer une nouvelle entrée dans la table,
P214	entrer la fonctionalité en colonne 3. S'il n'existe pas d'entrée correspondante dans la colonne 1, entrer un '1' par défaut (le premier carbone est, par convention, utilisé s'il n'existe pas d'autres spécifications,
P211	entrer l'information en colonne 2,
P24	déterminer la longueur de la chaîne de carbone,
P212	entrer le type de suffixe secondaire en colonne 4.

tableau 2

Grammaire des formules moléculaires: la colonne 4 n'existe plus puisque l'information n'est pas présente (figure 32).

règle syntaxique	règle sémantique associée
P11	déterminer la longeur de la chaîne,
P15	entrer le type de liaison en colonne 2,
P16	créer une nouvelle entrée dans la table, entrer le type de fonctionalité.

tableau 3

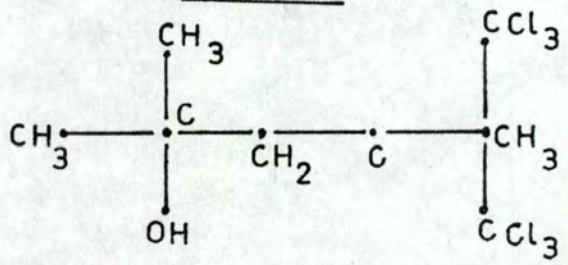


longueur de la chaîne C

2 - HYDROXY - 3 . 5 - DI CHLORO - 1 . 4 - OCT ADI - EN - 6 - YN E

numéro carbone	type liaison	fonctionalité	suffixes secondaires
2	1	OH	-
3	1	Cl	-
5	1	Cl	-
1	2	-	-
4	2	-	-
6	3	-	E

figure 31



C1H3 - C1(O1H1)1(C1H3)1 - C1H2 - C1 - C1H3(C1Cl3)2

numéro carbone	type liaison	fonctionalité
1	1	-
2	1	OH
3	1	-
4	1	-
5	-	CCl,CCl

figure 32

CHAPITRE II : ALGORITHME DE NUMEROTATION DES ATOMES

Si la table de connexions est caractéristique d'une structure de graphe $G_f(X,U,f_X,f_U)$, la numérotation des atomes doit être faite sans ambiguïté: deux molécules identiques doivent avoir la même numérotation. L'algorithme que nous présentons se base sur celui de Morgan (référence 18) dont nous avons modifié certaines règles et que nous avons généralisé au cas de la stéréochimie.

1. Stéréochimie

Un stéréocentre est un centre de la molécule qui, s'il est inversé, produit un stéréoisomère différent. En chimie organique, on considère habituellement deux types de stéréocentres: les atomes de carbone asymétriques et les liaisons C=C capables de cis-trans isomérisme. NOTE: une liaison vers l'avant est représentée par ◀, tandis qu'une liaison vers l'arrière se note - - - -. Le carbone 1 de la figure 33 et la liaison 1=2 de la figure 34 sont des stéréocentres.

figure 33

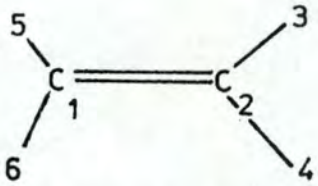
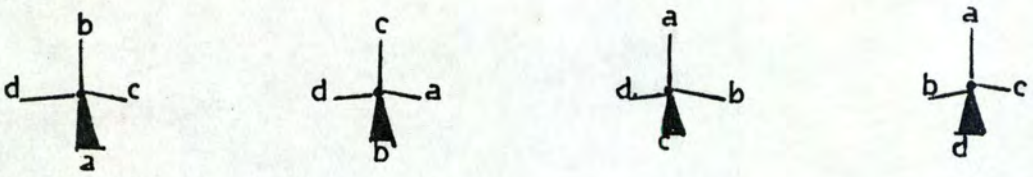


figure 34

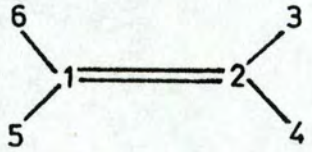
Un stéréocentre du premier type (carbone tétrahédral) peut être représenté par une liste d'atomes: le premier atome est placé en haut, les autres sont notés dans le sens des aiguilles d'une montre. Douze listes sont donc possibles (figure 35). Nous dirons que deux listes sont équivalentes s'il est possible de retrouver l'une à partir de l'autre par un nombre pair d'inversions d'atomes. La liste abcd est équivalente à dacb car abcd ⟶ bacd ⟶ dacb. Si le stéréocentre possède un hydrogène implicite, le nombre de listes possibles est réduit à trois, car, par convention, nous le considérons en quatrième position.



abcd	bcad	cabd	dacb
acdb	badc	cbda	dcba
adbc	bdca	cdab	dbac

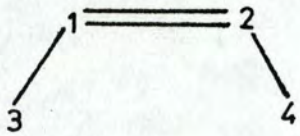
figure 35

Un stéréocentre du deuxième type (C=C) est ordonné de façon similaire: les atomes liés à ceux qui participent à la double liaison sont ordonnés dans le sens des aiguilles d'une montre (figure 36). L'équivalence entre deux listes est définie de la même façon que pour les carbones asymétriques. Les atomes d'hydrogène sont situés en dernière position (figure 37).



1 : 2 5 6
2 : 1 3 4

figure 36



1 : 2 6 (H)
2 : 5 1 (H)

figure 37

2. Algorithme de numérotation

Introduisons, d'abord, quelques définitions et notations:

- le degré d'un sommet i est noté d(i),
- la couleur du sommet i est notée $f_X(i)$,
- celle de l'arête j, $f_U(j)$,
- pour chaque atome, nous définissons la somme des couleurs des atomes voisins, notée SCV, par les relations de récursivité (II),
- partitions sur SCV: il est possible de partitionner l'ensemble des SCV trouvés à l'étape j en classes: deux SCV appartiennent

(II)

soit SCV(i,j) la somme des couleurs des atomes voisins de l'atome i à l'étape j; pour tout atome i, SCV(i,1) = d(i); à l'étape j: $SCV(i,j) = \sum_{k \in K_i} SCV(k,j-1)$ où K_i est l'ensemble des atomes voisins de i.

à la même classe si ces deux valeurs sont égales;

- une <u>séquence ordonnée de SCV</u>, notée SOSCV, est un ensemble de de SCV arrangés par ordre croissant;
- <u>ordre lexicographique</u>:
  soient deux séquences $(x_1, x_2, \ldots, x_n)$ et $(y_1, y_2, \ldots, y_n)$ et considérons la relation L définie par:
  - ou bien $x_1 < y_1$,
  - ou bien il existe un indice $h < n$ tel que $x_i = y_i$ pour tout $i < h$ et $x_h < y_h$,
  - ou bien $x_i = y_i$ pour tout $i < n$ et $x_n < y_n$.

  La relation L est une relation d'ordre total, que l'on appelle ordre lexicographique, car E étant l'alphabet d'une langue ordonné dans l'ordre normal (a,b, ..., z), l'ordre des mots dans un dictionnaire est un ordre de cette nature.

<u>algorithme:</u>

soit une structure comportant n atomes (les atomes H ne sont pas repris) numérotés arbitrairement au départ de 1 à n;

<u>étape 1:</u> pour tout atome i, SCV(i,1) = d(i); soit C(1) le nombre de classes différentes obtenu en partitionnant les SCV obtenus; k = 2;

<u>étape 2.j :</u> pour tout atome i, $SCV(i,j) = \sum_{k \in K_i} SCV(k,j-1)$; partitionner les SCV obtenus; soit C(j) le nombre de classes différentes; si $C(j) > C(j-1)$, j =j+1 et réitérer; sinon retenir les SCV de l'étape j-1 et aller à l'étape 3;

étape 3 : numéroter les atomes en utilisant les règles RM suivantes (en respectant l'ordre):

RM1 : - choisir l'atome i dont la valeur SCV(i,j-1) est la plus grande,
- lui assigner le numéro 1;

RM2.1 : - assigner la séquence 2, 3, ..., j+1 aux j atomes liés à 1; on assigne les numéros inférieurs aux atomes de valeurs SCV supérieures;

RM2.2 : - assigner de la même façon la séquence j+2, ..., j+2+k aux k atomes liés à l'atome 2;

RM2.3 : - continuer ce processus jusqu'à ce qu'il ne soit plus possible de numéroter d'autres atomes;

RM3 : - si un choix se pose entre plusieurs atomes, prendre celui qui est rattaché par la liaison de couleur supérieure (liaison simple < double < triple);

RM4 : - si un choix se pose entre plusieurs atomes, prendre celui de couleur supérieure (AC < AL < AM < AS < BA < BK < ... < XE < YT < Y < ZN < ZR);

si, après avoir appliqué ces règles, il reste des ambiguïtés, aller à l'étape 4;

étape 4 : essayer toutes les combinaisons possibles: pour chacune d'elles générer le code correspondant (dont la description est donnée plus loin); choisir la numérotation qui correspond au code lexicographiquement inférieur; s'il reste des atomes dont la numérotation est ambiguë, ils sont symétriques et numérotés arbitrairement. Leur numéro est repris dans un descripteur ajouté à la fin du code: le descripteur SYMETRY.

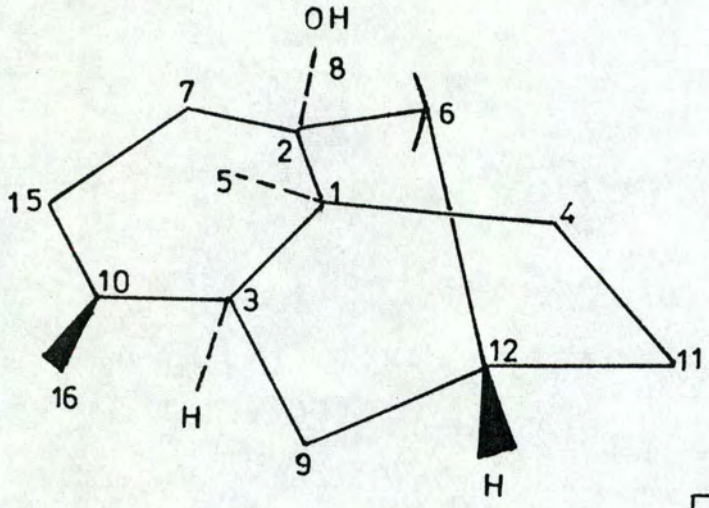
Le code est formé de sept descripteurs juxtaposés:
descripteur FROM: pour chaque atome (par ordre de numéro croissant), on reprend l'atome de plus petit numéro auquel il est relié;
descripteur RING CLOSURE: il comprend toutes les liaisons non reprises dans le descripteur FROM et classées par ordre croissant (la liaison entre les atomes 4 et 9, par exemple, précède celle entre 9 et 12 car 0409 < 0912);
descripteur ATOM TYPE: il donne la couleur des atomes (par ordre de numéro croissant);
descripteur BOND TYPE: il donne la couleur des liaisons (par ordre d'apparition dans les descripteurs FROM et RING CLOSURE);
descripteur MODIFICATION: charges, masses isotopiques, valences anormales y sont spécifiées;
descripteur STEREO1: il donne pour chaque atome du descripteur FROM sa parité;
descripteur STEREO2: il donne pour chaque double liaison (par ordre d'apparition dans le descripteur BOND TYPE) sa parité.

La parité d'un atome ou d'une double liaison peut prendre les valeurs suivantes:

0 pour un non-stéréocentre,
1 pour un stéréocentre impair,
2 pour un stéréocentre pair,
3 pour un stéréocentre de parité inconnue.

Un stéréocentre de type 1 (carbone asymétrique) est pair (respectivement impair) si le nombre d'inversions nécessaires pour amener la séquence des atomes, qui décrivent le stéréocentre, en ordre croissant est pair (respectivement impair). Initialement, l'atome lié par la liaison ◄ (c'est-à-dire celui qui se trouve au-dessus du plan de la feuille) est placé le premier, l'atome lié par la liaison - - - (c'est-à-dire celui qui se trouve en-dessous du plan de la feuille) le dernier; pour les atomes restant, on commence par celui qui possède le numéro le plus petit suivi par les autres parcourus dans le sens des aiguilles d'une montre (figure 38).
Un stéréocentre de type 2 (C=C) est pair (respectivement impair) si le nombre total des inversions nécessaires pour amener les deux séquences en ordre croissant est pair (respectivement impair). Initialement, les séquences sont ordonnées séparément comme pour les

stéréocentres de type 1.



stéréo-centre	séquence initiale
1	2-4-3-5
2	1-7-6-8
3	1-9-10
10	16-3-15
12	H-6-11-9

atomes du plan supérieur
atomes du plan de la feuille
atomes du plan inférieur

permutations:		parité
1	2-4-3-5 $\xrightarrow{4,3}$ 2-3-4-5	impair
2	1-7-6-8 $\xrightarrow{7,6}$ 1-6-7-8	impair
3	1-9-10	pair
10	16-3-15 $\xrightarrow{16,3}$ 3-16-15 $\xrightarrow{16,15}$ 3-15-16	pair
12	H-6-11-9 $\xrightarrow{H,9}$ 9-6-11-H $\xrightarrow{6,9}$ 6-9-11-H	pair

figure 38

Appliquons l'algorithme de numérotation au composé de la figure 39 .

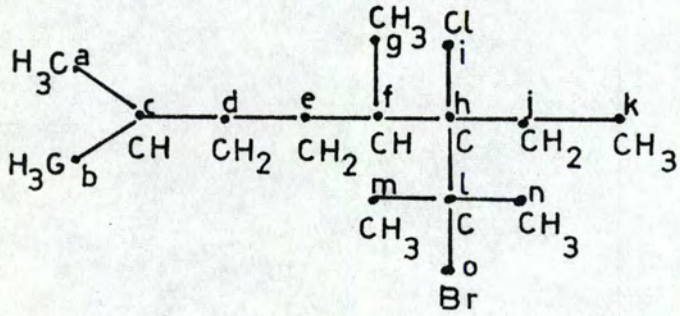


figure 39

étapes 1 et 2:

étapes	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	classes
1.	1	1	3	2	2	3	1	4	1	2	1	4	1	1	1	4
2.1	1	1	4	5	5	7	1	10	1	5	1	7	1	1	1	5
2.2	1	1	7	9	12	16	1	20	1	11	1	13	1	1	1	8
2.3	1	1	11	19	23	33	1	41	1	21	1	23	1	1	1	7

Les SCV retenus sont ceux de l'étape 2.2

étape 3:

numéro de l'atome	SCV	atomes adjacents	numéro (algo.)	règle appliquée
h	20	f,i,j,l	1	RM1
f	16	e,g,h	2	RM2
l	13	h,m,o,n	3	RM2
j	11	h,k	4	RM2
i	1	h	5	RM2
e	12	d,f	6	RM2
g	1	f	7	RM2
m	1	l		
n	1	l	8,9 ?	RM^2_4
o	1	l	10	
k	1	j	11	RM2
d	9	c,e	12	RM2
c	7	a,b	13	RM2
a	1	-	14,15 ?	RM2
b	1	-		RM2

Après l'étape 3, nous obtenons la séquence suivante:

$$h < f < l < j < i < e < g < \{m,n\} < o < k < d < c < \{a,b\}$$

Une ambiguïté subsistant, nous appliquons l'étape 4;

étape 4: toutes les combinaisons possibles sont générées:

combinaison 1: h,f,l,j,i,e,g,m,n,o ,k ,d ,c ,a ,b
numérotation : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

combinaison 2: h,f,l,j,i,e,g,n,m,o ,k ,d ,c ,a ,b
numérotation : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

combinaison 3: h,f,l,j,i,e,g,m,n,o ,k ,d ,c ,b ,a
numérotation : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

combinaison 3: h,f,l,j,i,e,g,n,m,o ,k ,d ,c ,b ,a
numérotation : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Pour chacune des combinaisons, nous générons le code correspondant:

•combinaison	h	f	l	j	i	e	g	m	n	o	k	d	c	a	b
numérotation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FROM	2	1	1	1	1	2	2	3	3	3	4	6	14	—	—
RING CLOSURE	-														
ATOM TYPE	C	C	C	C	CL	C	C	C	C	BR	C	C	C	C	C
BOND TYPE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

•combinaison	h	f	l	j	i	e	g	n	m	o	k	d	c	a	b
numérotation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FROM	2	1	1	1	1	2	2	3	3	3	4	6	14	—	—
RING CLOSURE	-														
ATOM TYPE	C	C	C	C	CL	C	C	C	C	BR	C	C	C	C	C
BOND TYPE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

• combinaison	: h	f	l	j	i	e	g	m	n	o	k	d	c	a	b
numérotation	: 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FROM	: 2	1	1	1	1	2	2	3	3	3	4	6	14	—	—
RING CLOSURE	: -														
ATOM TYPE	: C	C	C	C	CL	C	C	C	C	BR	C	C	C	C	C
BOND TYPE	: 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

• combinaison	: h	f	l	j	i	e	g	n	m	o	k	d	c	b	a
numérotation	: 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FROM	: 2	1	1	1	1	2	2	3	3	3	4	6	14	—	—
RING CLOSURE	: -														
ATOM TYPE	: C	C	C	C	CL	C	C	C	C	BR	C	C	C	C	C
BOND TYPE	: 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Les codes étant lexicographiquement équivalents, il n'est pas possible de discerner les atomes {m,n} et {a,b} : ils sont symétriques. Le descripteur SYMETRY est donc: (8,9 ; 14,15).

En résumé, la numérotation et le code associé à la structure de la figure 39 sont donnés dans la figure 40 , tandis que la figure 41 reprend l'ensemble de la codification.

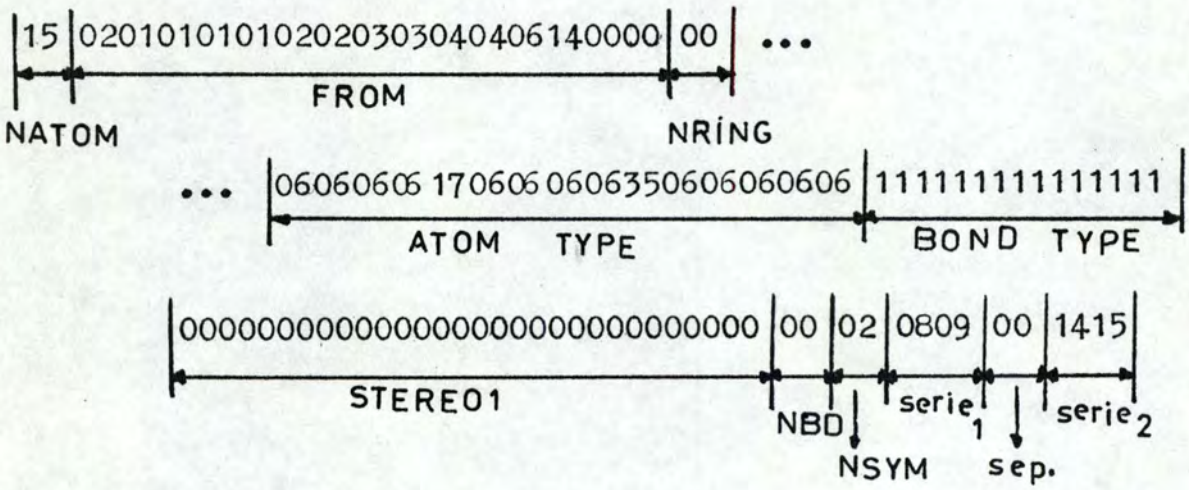


figure 40

code		place mémoire	remarques
NATOM	nombre d'atomes	1 byte	127 atomes max.
FROM	descripteur des liaisons	1 byte par atome	l'ordre est donné par l'algo.
NRING	nombre de cycles	1 byte	
RING CLOSURE	descripteur des cycles	2 bytes par liaison	par ordre croissant
ATOM TYPE	couleur des atomes	1 byte par atome	
BOND TYPE	couleur des liaisons	1 byte par liaison	
STEREO1	parité des atomes	2 bits par atome	00 non-stéréo 01 stéréo impair 10 stéréo pair 11 inconnu le dernier byte est rempli par des 00
NBD	nombre de doubles liaisons	1 byte	
STEREO2	parité des doubles liaisons	2 bits par liaison	idem
NSYM	nombre de séries d'atomes symétriques	1 byte	
SYMETRY	descripteur de symétrie	1 byte par atome	chaque série est séparée par 00

figure 41

La place totale occupée en mémoire est donc:

$$4 \text{ bytes} + NATOM * 3 \text{ bytes} + NRING * 2 \text{ bytes} + \left[NDB * 2 \text{ bits}\right] \text{ bytes} + \left[NATOM * 2 \text{ bits}\right] \text{ bytes} + \sum_{i=1}^{NSYM} SYM_i + (NSYM - 1) * 1 \text{ bytes}$$

où les crochets signifient une conversion du nombre de bits en bytes entiers et SYM_i est le nombre d'atomes appartenant à la série de symétrie i.

Il est possible de définir un certain nombre de critères permettant de juger de la valeur d'un système de codification: simplicité, concision, facilité de manipulation par des méthodes automatiques, pouvoir descriptif, biunivocité. Le code que nous venons d'exposer réalise la synthèse de ces différentes exigences: il permet en un petit nombre de bytes (la structure de la figure 39 utilise 58 bytes) d'appréhender entièrement les graphes chromatiques qui modélisent les structures. Il reste à montrer la biunivocité du code.

3. Biunivocité du code

Il s'agit de prouver (1) que toute représentation correcte d'un même composé donne le même code; (2) que deux composés non identiques sont codés différemment.

3.1. Toute représentation correcte d'un même composé donne le même code

Le problème revient à montrer que le code est indépendant de la position spatiale des sommets et de la numérotation initiale de la structure.

Soient A et B, deux représentations correctes du composé Z; A et B ont le même ensemble de sommets (X) et d'arêtes (U), et à tout sommet de A doit correspondre un sommet de B: A et B sont isomorphiques. Les deux représentations possédant les mêmes composition et configuration, elles ne peuvent différer que par la position spatiale des atomes (sauf pour les stéréocentres) et leur nu-

mérotation. Les coordonnées des atomes n'intervenant pas dans la codification, il reste à montrer que la numérotation finale est indépendante des numéros initiaux attribués aux atomes.

L'algorithme est basé sur les valeurs SCV des atomes (étapes 1 et 2) et n'utilise que les degrés des atomes (et non leur numéro). La numérotation est donc invariante si, durant l'étape 3, aucun choix ne se pose entre deux atomes de même valeur SCV. Par contre, si une ambiguïté apparaît, toutes les combinaisons sont essayées. Soit S l'ensemble des séquences de numérotation correspondant à chaque combinaison: chaque séquence s_i de S ne dépend que des valeurs SCV des atomes et produit un code n_j (il faut noter qu'il n'existe pas de correspondance un à un entre l'ensemble des séquences et l'ensemble des codes N, puisque, si une structure est symétrique, deux séquences différentes peuvent donner le même code). Le code final est le code n_k dont la valeur numérique est minimale. Puisque N et S sont indépendant de la numérotation initiale, nous avons $N_A = N_B$ et $n_A = n_B$. Donc toute représentation correcte du même composé Z conduit au même code.

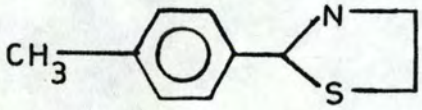
3.2. Les représentations de composés non identiques donnent des codes différents

Prouvons la contraposée: si deux représentations ont le même code, alors elles sont équivalentes et correspondent au même composé. Rappelons que le code est formé de sept descripteurs: FROM, RING CLOSURE, ATOM TYPE, BOND TYPE, MODIFICATION, STEREO1 et STEREO2. Si deux codes sont identiques, tous les descripteurs correspondant sont égaux. Entre autres, les deux représentations ont le même nombre d'atomes ($X = X'$) et de liaisons ($U = U'$). Les squelettes (décrits par les descripteurs FROM et RING CLOSURE) sont identiques: $G(X,U) \equiv G'(X',U')$. De plus, les colorations des deux graphes sont les mêmes ($f_X = f_{X'}$ et $f_U = f_{U'}$), puisque les descripteurs ATOM TYPE et BOND TYPE sont respectivement égaux. L'équivalence entre les descripteurs MODIFICATION indique que les charges, masses isotopiques, ... sont les mêmes dans les deux représentations. Enfin, les stéréocentres y sont identiques puisque les descripteurs STEREO1 et STEREO2 correspondent. Les deux représentations sont donc équivalentes et correspondent au même composé.

CHAPITRE III : INTERROGATION: QUESTIONS DE NATURE TOPOLOGIQUE

Deux types de recherches sont à envisager:

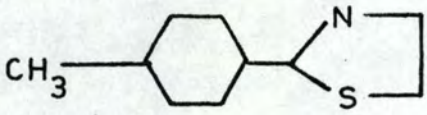
- recherche d'informations topologiques: l'utilisateur s'intéresse à une seule structure complète (figure 42);



" le composé ci-dessus existe-t-il dans le fichier? "

figure 42

- recherche en vue d'établir une corrélation: l'utilisateur peut s'intéresser à une famille de composés et les retrouver dans le fichier en indiquant la sous-structure ou le motif qui les caractérise (figure 43).



" quels sont les composés du fichier qui contiennent la la sous-structure ci-dessus? "

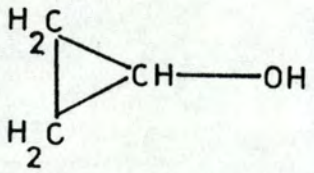
figure 43

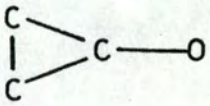
Une structure chimique présente plusieurs ordres possibles (figure 44). La recherche de l'ordre à induire sur un graphe chromatique permet de résoudre ces deux problèmes de façon efficace.

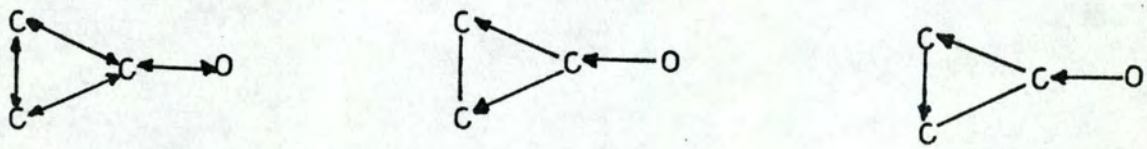
1. Recherche des informations topologiques

1.1. Définitions

Deux graphes chromatiques $G_f(X,U,f_X,f_U)$ et $G_f(X',U',f_{X'},f_{U'})$



le graphe chromatique correspondant est: 



(a) (b) (c)

(a) <u>préordre</u>: un sommet peut être son propre père et grand-père,

(b) <u>ordre partiel</u>: un père peut avoir plusieurs fils et un fils a un seul père,

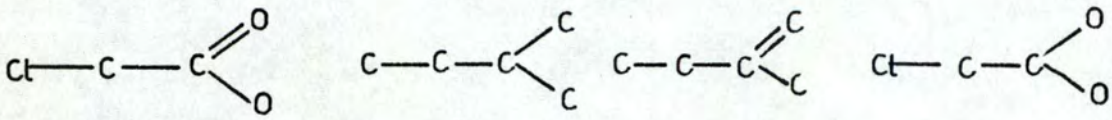
(c) <u>ordre linéaire</u>: un père a un fils et vice-versa.

<u>figure 44</u>

sont - <u>isotopologues</u> s'il existe une bijection entre X et X' compatible avec une bijection entre U et U', c'est-à-dire si leur graphe topologique associé sont isomorphiques;

- <u>isochromatiques</u> sur les sommets ou les arêtes s'ils sont isotopologues et si $f_X = f_{X'}$ ou $f_U = f_{U'}$;
- <u>isomorphiques</u> s'ils sont isochromatiques sur les sommets et arêtes.

Dans la figure 45 , (a), (b), (c), (d) sont isotopologues; (a), (d) et (b), (c) sont isochromatiques sur les sommets; (a) ,(c) et (b), (d) sont isochromatiques sur les arêtes.



(a) (b) (c) (d)

<u>figure 45</u>

Le problème de la recherche d'un composé de graphe G dans un fichier, revient à y trouver un graphe G' qui est isomorphique à G, en vérifiant les trois critères définissant l'isomorphisme de deux graphes. Si l'on induit un préordre sur les graphes, la comparaison devient difficile, car elle est "multidirectionnelle": il s'agit, en effet, de traverser les deux graphes en essayant toutes les combinaisons possibles de comparaison de sommet à sommet en appliquant les trois critères à chaque fois (l'algorithme est défini plus loin). De plus, le parcours de tout le fichier des composés est obligatoire.

L'induction d'un ordre linéaire sur les graphes ramène la comparaison multidirectionnelle à une comparaison linéaire (c'est-à-dire un élément après l'autre); or nous disposons d'un algorithme de numérotation des atomes qui induit un ordre linéaire au niveau du graphe chromatique et au niveau de sa représentation. La détection d'un isomorphisme éventuel peut donc être réalisé par une comparaison de représentations linéaires.

1.2. Ordres dans l'algorithme de numérotation

Durant la numérotation des sommets, qui est une façon d'induire un ordre, le chromatisme des sommets ou des arêtes est utilisé pour régler les cas d'ambiguïtés restant après les premières étapes. Le résultat de la numérotation est une représentation linéaire, numérique et biunivoque. L'ordre partiel est induit sur le graphe par valuation topologique globale: chaque sommet est valué par son degré de connectivité; les sommations de ces valeurs sont itérées jusqu'à stabilisation. Le graphe final est renuméroté suivant les valeurs finales et des décisions chromatiques lexicographiques, ce qui induit l'ordre linéaire.

1.3. Algorithme de recherche

Si le fichier des composés est séquentiel, il s'agit de comparer les codes les uns après les autres avec celui du composé à rechercher. S'il est possible d'utiliser les fichiers en accès direct, le code du composé peut servir de clé d'accès étant donné sa

biunivocité.

2. Recherche en vue d'établir une corrélation

Le problème est de voir si G_f' est un sous-graphe de G_f et, par conséquent, d'identifier certaines parties de G_f (problème de reconnaissance de forme ou d'homomorphisme de graphes chromatiques).

Dans ce cas, l'induction d'un ordre linéaire ne présente pas le même avantage que pour la recherche d'isomorphisme, puisqu'il n'y a aucune raison pour que l'ordre sur la partie de G_f soit le même que celui sur G_f'. Sauf pour de rares exceptions, le graphe G_f et le sous-graphe G_f' n'ont pas la même origine pour l'ordre (figure 46).

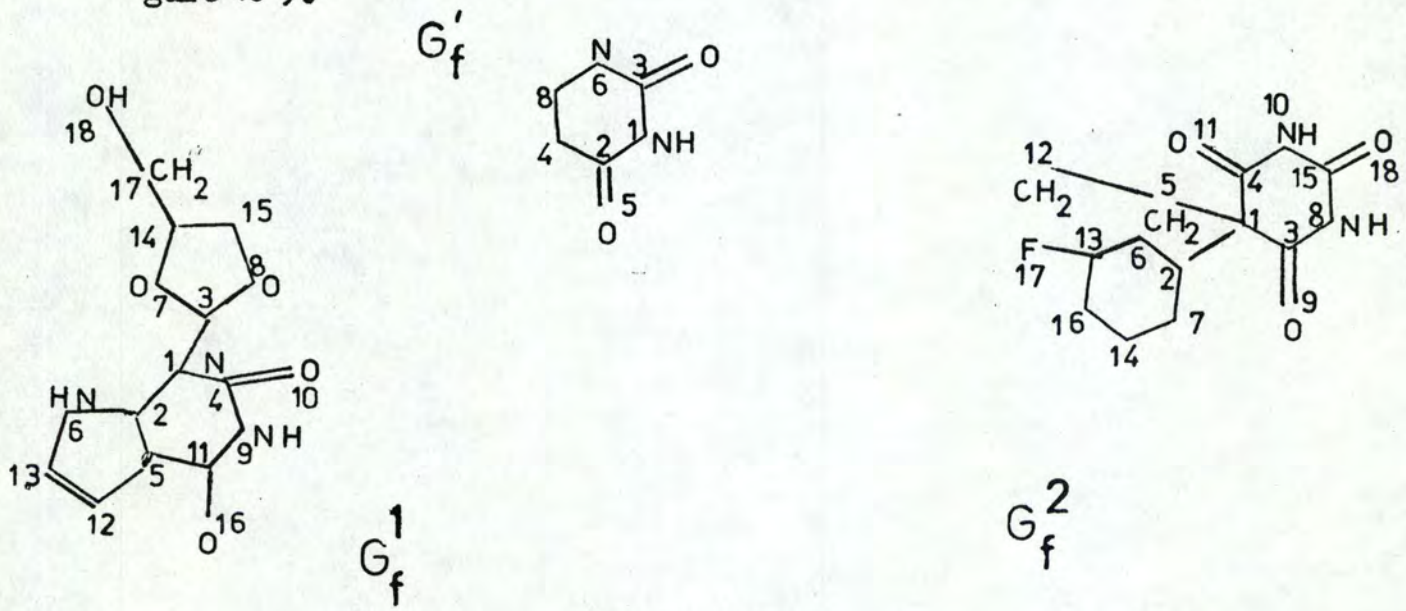


figure 46

2.1. Algorithmes

Plusieurs procédures sont possibles:

- recherche d'homomorphisme par procédé direct: ces procédures s'appliquent directement aux graphes,
- recherche d'homomorphisme par ensembles ordonnés: ces procédures sont appliquées après une étape d'utilisation d' "écrans" qui expriment les analogies structurales entre les structures et la sous-structure-question, et qui sont obtenus par génération automatique.

2.1.1. Procédés directs

2.1.1.1. Comparaison sommet par sommet

définition

L'ordre de Tarry de parcours d'un graphe à partir d'un sommet est celui défini par un promeneur qui, partant du sommet choisi, rencontre pour la première fois chaque sommet en respectant les règles suivantes:

- le promeneur quitte le sommet où il se trouve et chemine dans le sens des arêtes, en prenant le sommet de numéro (assigné par l'algorithme de numérotation) le plus petit qui lui est adjacent;
- lorsque cela n'est plus possible, il fait demi-tour et remonte les arêtes dans le sens contraire jusqu'à ce qu'il atteigne un sommet d'où parte une arête qu'il n'a pas encore utilisée; il applique, ensuite, à nouveau la règle précédente;
- le promeneur retient (ensemble A) les sommets qu'il quitte pour la première fois;
- le promeneur n'emprunte jamais une arête qui aboutit à un sommet de A. Cette arête est une corde.

Le graphe de la figure 47 est parcouru en empruntant successivement les sommets 1,2,5,7,3,6,10,9,4,8. Cela revient à construire un arbre maximal à partir du sommet choisi.

algorithme

Pour essayer de faire correspondre le sous-graphe G'_f avec au moins une partie de G_f, nous appliquons la procédure suivante:

étape 1: générer la table de connexions de G_f;

étape 2: générer la table de connexions de G'_f;

étape 3: prendre chaque sommet de G_f comme racine et générer l'arbre maximal correspondant en parcourant le graphe suivant l'ordre de Tarry;

étape 4: prendre le sommet numéro 1 de G'_f comme racine et générer de la même façon l'arbre maximal;

étape 5: établir une correspondance éventuelle entre l'arbre maxi-

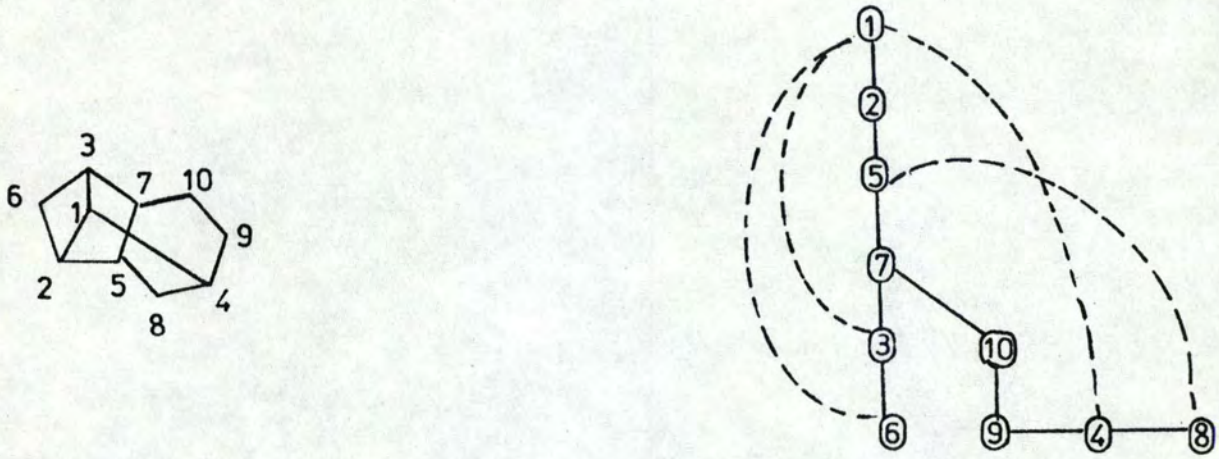


figure 47

mal de G_f' et un sous-arbre des arbres maximaux de G_f par une comparaison niveau par niveau.

La figure 48 donne un exemple de recherche d'un cycle de taille 5.

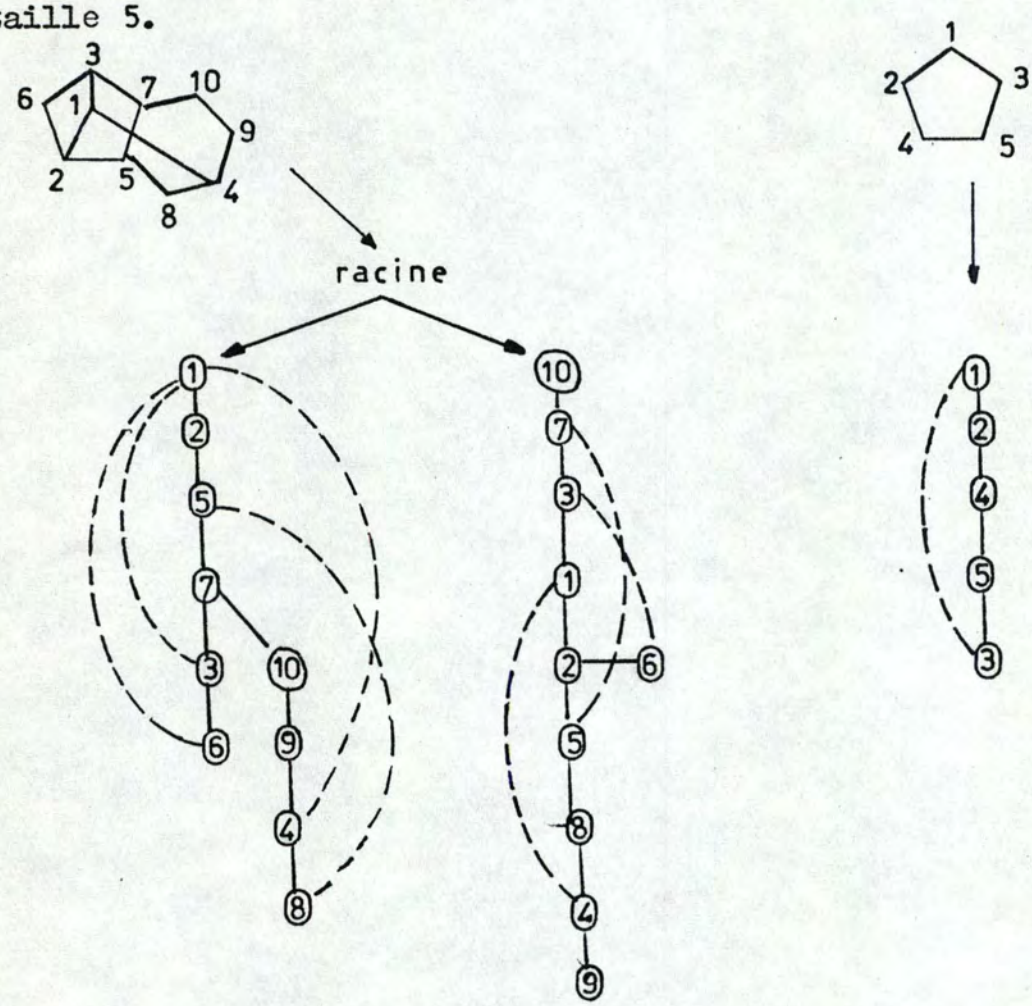


figure 48

NOTES:

- la génération des tables de connexions rend plus aisée celle des arbres maximaux;
- il n'est pas nécessaire de générer d'abord tous les arbres maximaux de G_f: une génération niveau par niveau permet d'éliminer rapidement certains arbres. L'arbre généré à partir de l'atome 10 dans la figure 48 est éliminé dès le niveau 5 car, si les atomes et les liaisons correspondent, les cordes ne correspondent pas (figure 49);
- ce type de recherche ne peut se faire qu'en parcourant tout le fichier des composés qui, de plus, doit être réduit afin qu'elle ne soit pas trop longue.

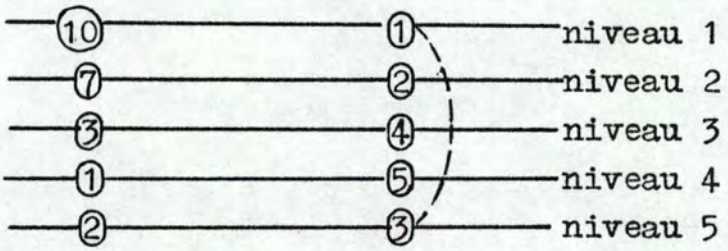


figure 49

2.1.1.2. Recherche par motif

Dans ce type de recherche, la sous-structure ne peut être qu'une entité connue (en général, un groupe fonctionnel). L'ensemble de ces groupes constitue le fichier des motifs. Par une méthode de recherche d'isomorphisme (premier type), on vérifie que la sous-structure existe dans le fichier, d'où on retrouve tous les composés qui la contiennent.

Lors de l'entrée d'un nouveau composé, on recherche tous les motifs qu'il contient en essayant de faire correspondre les graphes des motifs avec une partie du graphe de ce composé (méthode précédente).

L'avantage de cette méthode est que la comparaison atome par

atome est faite une fois pour toute lors de l'entrée du composé (sur base du fichier des motifs qui est assez réduit), et non plus lors de chaque question, ce qui permet un temps de réponse meilleur.

2.1.2. Ensemble ordonné

Dans la méthode de recherche par motifs, on extrait du graphe G_f un ensemble de sous-graphes figés. La méthode par ensemble ordonné permet une fragmentation dynamique et flexible de la structure: on extrait des graphes G_f et G'_f un ensemble de sous-graphes standards, non nécessairement disjoints. Ensuite, par une méthode de recherche d'isomorphisme, on vérifie que les sous-graphes définis dans G'_f sont présents dans G_f. Si ce n'est pas le cas, il n'y a pas homomorphisme.

définition

Un écran canonique (c'est-à-dire défini par un règle) est l'environnement retenu à partir de chaque atome de degré supérieur ou égal à trois et constitué de deux zones concentriques (couches A et B). Chaque écran dans une structure est décrit indépendamment des autres atomes (figure 50).

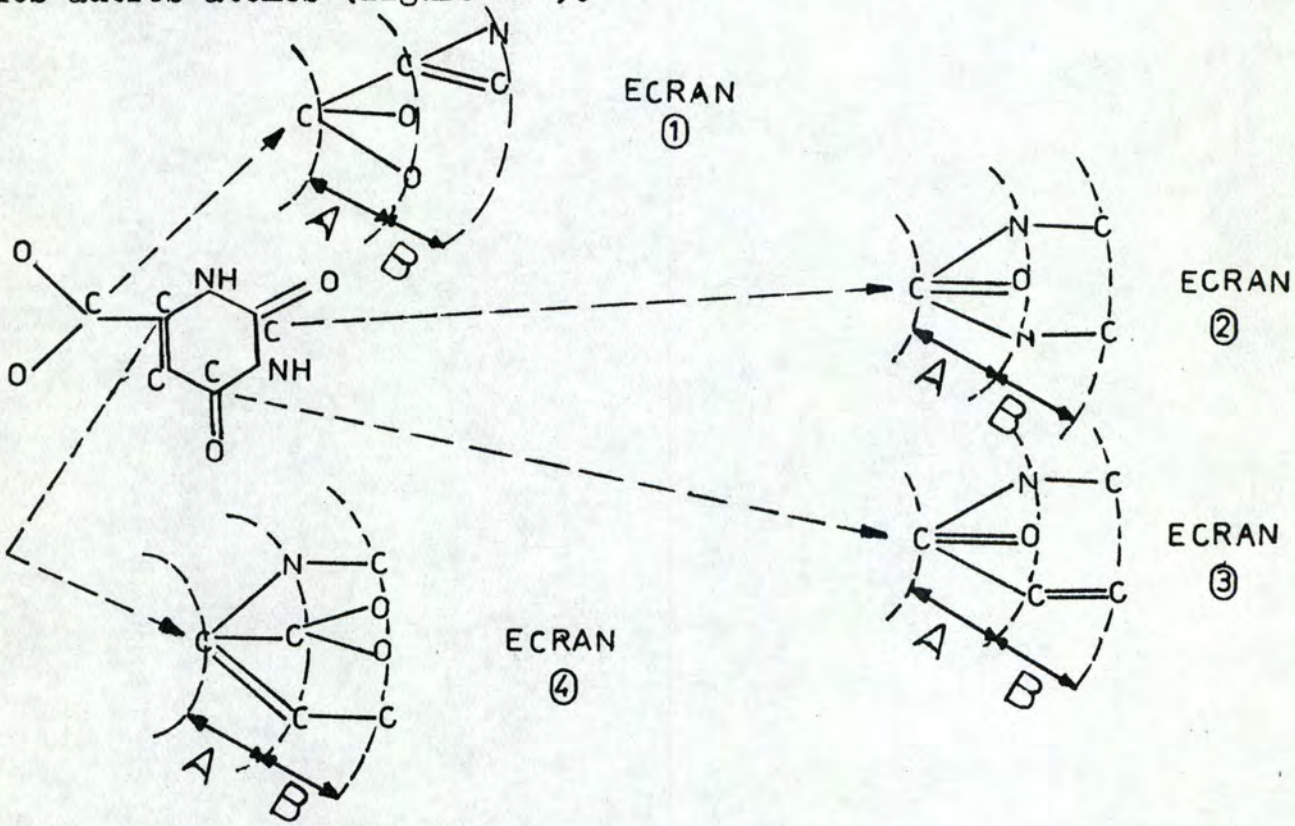


figure 50

Les écrans associés à une structure ne la décrivent pas exhaustivement (il existe une perte d'information, particulièrement pour les chaînes linéaires puisque les sommets de degré inférieur ou égal à deux ne sont pas pris en considération). Cependant la collection des écrans d'un composé décrit de façon très spécifique les aspects topologiques essentiels et le chromatisme intéressant de ce composé:

- les motifs sont centrés sur des atomes de haut degré et donc décrivent les parties de la molécule riches en chromatisme et topologie;
- la même zone de la molécule est souvent couverte par plusieurs écrans. L'overlapping illustre les caractéristiques topologiques importantes de la molécule (figure 51).

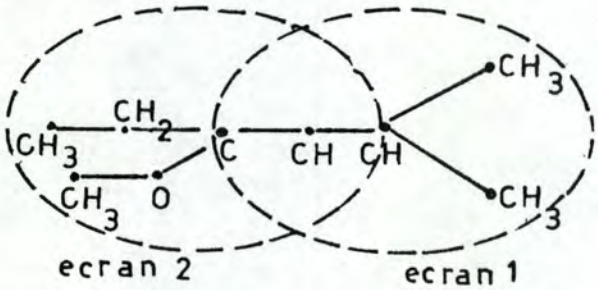


figure 51

Lors d'une recherche par sous-structure, il suffit de comparer les écrans de la question aux écrans associés aux composés du fichier pour décider de la pertinence d'un composé. Soient C l'ensemble des composés que l'on désire accéder par la recherche et $\widetilde{C}$ l'ensemble des composés qui contienne les écrans de la sous-structure; C est contenu dans $\widetilde{C}$ mais C devrait être très proche de $\widetilde{C}$: la comparaison des écrans doit permettre de sélectionner rapidement avec le moins de bruit possible (une réponse est accompagnée de bruit si cette réponse inclut des éléments non pertinents pour la question posée) et peu de silence (une réponse est accompagnée de silence si des éléments pertinents pour la question posée n'apparaissent pas dans la réponse). Le bruit peut être complètement éliminé en terminant la recherche par une phase de comparaison atome par atome entre la sous-structure et les composés résultant de la comparaison des écrans qui n'a alors pour objectif qu'un criblage primaire du fichier des composés.

CHAPITRE IV : ACQUISITION ET RESTITUTION: GTDES

L'originalité de la documentation chimique automatique réside dans le fait qu'elle manipule simultanément des informations textuelles et structurelles; l'intérêt des systèmes graphiques se situe au niveau des fonctions d'acquisition et de restitution:

- acquisition: les systèmes graphiques permettent à l'utilisateur de s'exprimer dans son langage familier (formule développée associée d'une structure) par simple dessin sur une console graphique pour formuler sa question ou pour créer la base de données structurales;
- restitution: les systèmes graphiques permettent à l'utilisateur de prendre connaissance, par l'intermédiaire d'une console, des structures qui satisfont à ses paramètres d'interrogation et éventuellement d'affiner ces derniers et de conduire ainsi sa recherche en mode conversationnel.

Nous avons mis au point un programme (GTDES) d'acquisition et de restitution de structures sur une console graphique DIGITAL GT42 (figure 52).

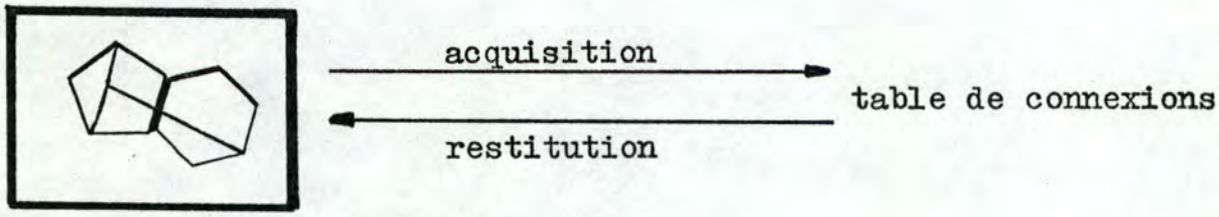


figure 52

La structure est dessinée en trois étapes successives:

- dessin des atomes et de leur couleur,
- dessin des charges atomiques,
- dessin des liaisons et de leur couleur.

Une quatrième phase permet de corriger, éventuellement, le dessin. A chaque étape correspond un "schéma" différent: un schéma comporte des mots-contrôle que l'on peut activer par l'intermédiaire d'un crayon optique. L'ensemble des mots-contrôle forme le "menu" du schéma.

1. Dessin des atomes: schéma 1

Pour dessiner un atome, on place d'abord correctement le repère (un losange) qui indiquera sa position sur l'écran. On active ensuite un des mots-contrôle du menu:

(1) CARBON: atome de carbone,
(2) HYDROGEN: atome d'hydrogène,
(3) GROUPE: groupe fonctionnel,
(4) ERASE: effacer tous les atomes,
(5) EXIT: passer à l'étape suivante,
(6) ERATOM: effacer un atome en particulier,
(7) HETERO: hétéro-atome.

En activant le dernier mot-contrôle, un schéma intermédiaire apparaît:

(1) NITROGEN: atome N,
(2) PHOSPHORUS: atome P,
(3) OXYGEN: atome O,
(4) SULFUR: atome S,
(5) IODINE: atome I,
(6) FLUORINE: atome F,
(7) BORON: atome B,
(8) CHLORINE: atome CL,
(9) SELENIUM: atome SE,
(10) EXIT: retourner au schéma 1.

2. Dessin des charges: schéma 2

Pour désigner des atomes chargés positivement ou négativement, on active le mot-contrôle correspondant. On pointe ensuite l'atome concerné. La présence d'une charge sur un atome lui permet d'être connecté par un nombre de liaisons différent de sa valence:

(1) CHARGE +,
(2) CHARGE -,
(3) CHARGE +2,
(4) CHARGE -2,
(5) CHARGE +3,
(6) CHARGE -3,

(7) EXIT: passer à l'étape suivante.

3. Dessin des liaisons: schéma 3

On pointe d'abord un des mots-contrôle puis les deux atomes concernés par la liaison. Quand le nombre maximum de liaisons permis pour un atome (valence + sa charge) est dépassé, un message d'erreur apparaît et la liaison est détruite:

(1) SINGLE: liaison simple C.——.C
(2) DATIVE: liaison dative C.—·—·.C
(3) FORWARD: liaison en avant C.▬▬.C
(4) BACKWARD: liaison en arrière C.— — —.C
(5) DOUBLE: liaison double C.═══.C
(6) TRIPLE: liaison triple C.≡≡≡.C

4. Correction

Une fois la strcuture dessinée, l'utilisateur peut modifier la position d'un ou de plusieurs atomes afin de rendre son dessin plus clair: il suffit de pointer l'atome à déplacer et d'indiquer, au moyen du repère, la nouvelle position de l'atome. Ce dernier et toutes les liaisons qui y sont attachées sont déplacées automatiquement.

5. Exit

Les informations (contenues dans la table de connexions générée par le programme) nécessaires pour redessiner la structure sont mémorisées dans un fichier.

6. Représentation des données

Les données (figure 53) sont représentées par une série de tableaux:

(1) ATPOS: les coordonnées des atomes sur l'écran,
(2) ATOM: les couleurs des atomes,
(3) GRPOS: position des groupes fonctionnels (pointeur vers un atome de ATPOS),
(4) GROUPE: les groupes fonctionnels,
(5) BOND: - pour chaque liaison, ce tableau contient la position

des atomes qui y participent (pointeurs vers ATPOS),
- les couleurs des liaisons,

(6) CHARGE: - position de l'atome (pointeur vers ATPOS),
- charge de l'atome ($-3 \leq c \leq +3$).

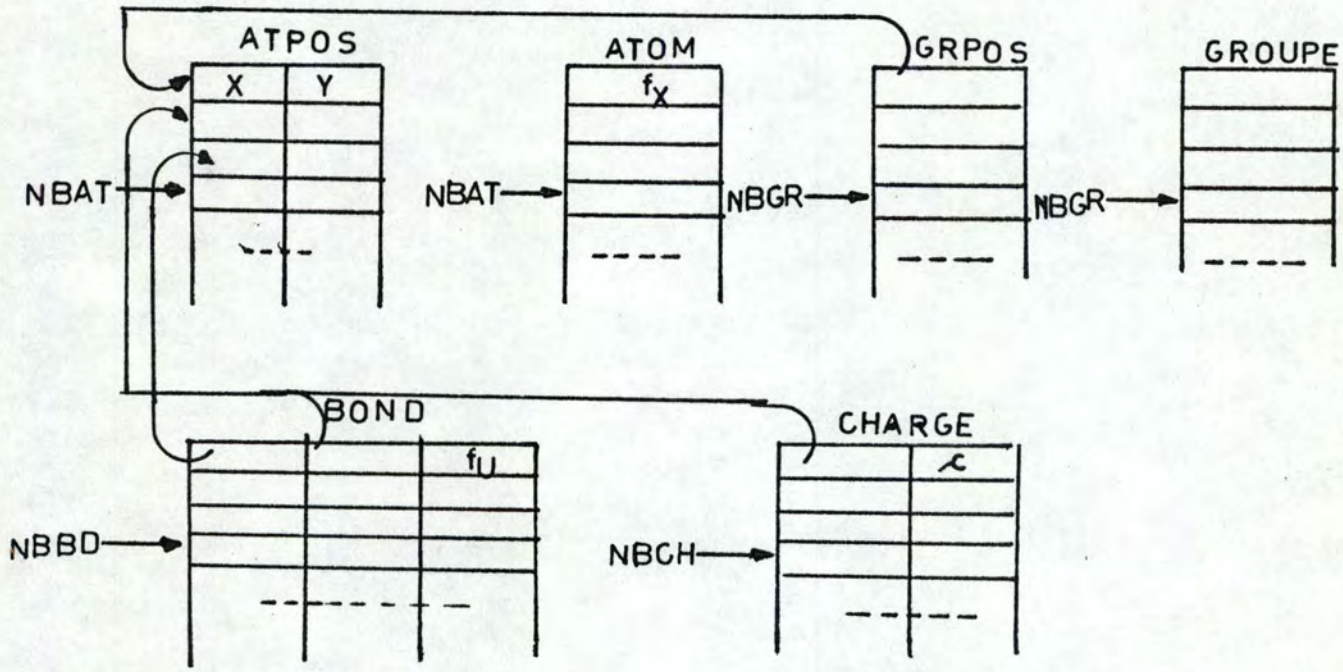


figure 53

7. Améliorations possibles de GTDES

7.1. Amélioration de l'édition graphique

La molécule peut être représentée par le modèle de Dreiding. Dans ce cas, la molécule est matérialisée par un agencement de boules et de batonnets figurant les atomes et leurs liaisons. Les boules sont caractérisées dans le plan d'observation par des cercles identiques ou proportionnels au rayon de covalence. Les liaisons sont représentées par une série de droites parallèles étroitement rapprochées.

Des options permettent d'améliorer cette représentation en

précisant les points d'attaches des liaisons sur les boules et en faisant varier l'échelle de l'image en fonction de la distance observateur-objet. L'effet de perspective produit peut être représenté par un effet de perspective locale obtenu en représentant la liaison par un batonnet conique dont l'épaisseur varie avec la différence de niveau des atomes qui la constituent (figure 54).

7.2. Amélioration de l'interaction

Il serait souhaitable d'avoir la possibilité d'agrandir et de regarder la molécule sous tous les angles et d'avoir simultanément plusieurs représentations de la même vue suivant l'option d'édition choisie.

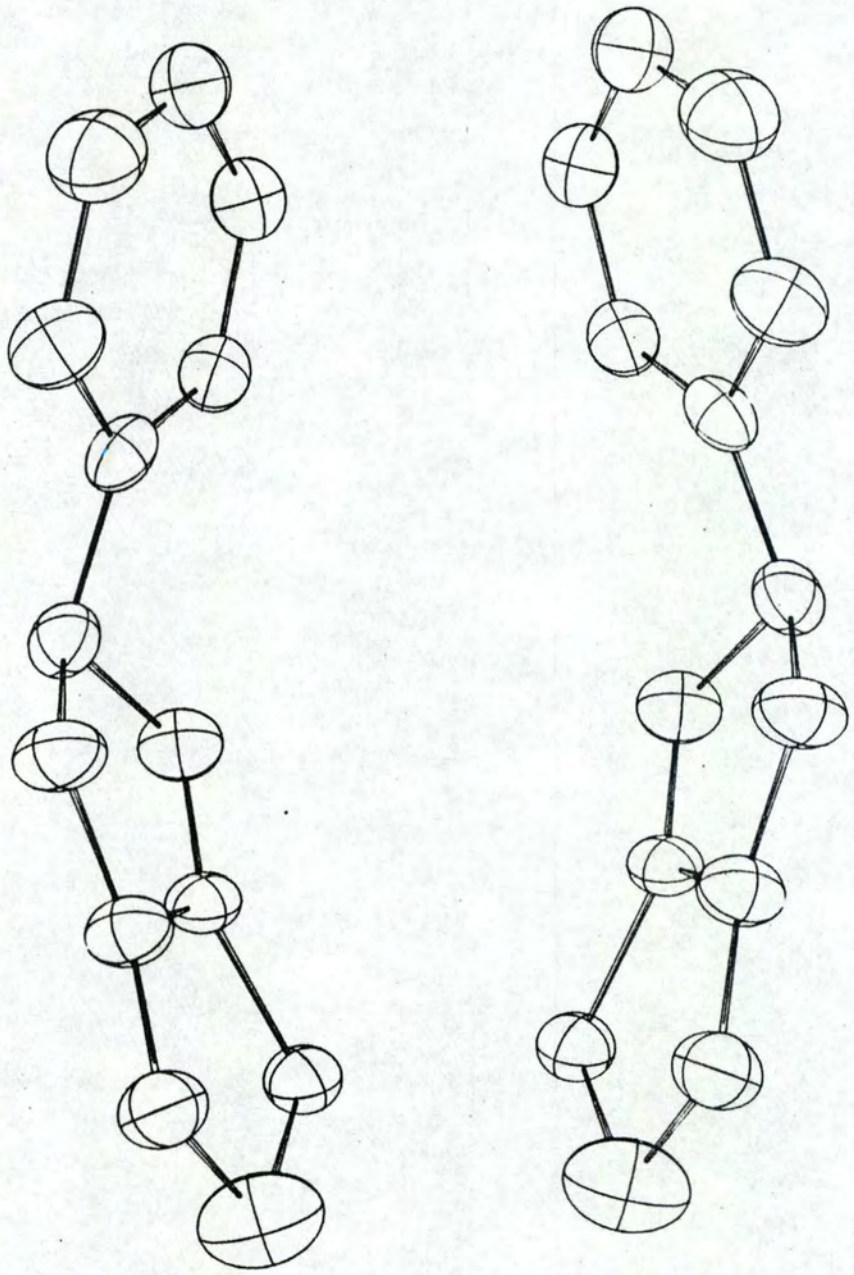


figure 54

CHAPITRE V : REPRESENTATION DES REACTIONS

Une réaction élémentaire se décrit par une relation entre deux membres (figure 55). Le produit COOMe s'appelle le produit P1 de

$$\text{COOMe} + \text{LAH} \underset{\text{AR}}{\overset{\text{R}}{\rightleftharpoons}} \text{CH}_2\text{OH}$$

figure 55

départ, CH_2OH le produit P2 d'arrivée, LAH le réactif (RCT). R est le sens réactionnel, AR le sens antiréactionnel. Une réaction est donc un triplet (P1,RCT,P2). Dans le sens R, pour P1 et RCT donnés, il existe un et un seul triplet (P1,RCT,P2); dans le sens AR, pour RCT et P2 donnés, il existe un ensemble ($i \in I$) de triplets (Pi,RCT, P2) (figure 56). Une réaction est un ensemble de triplets (P1,RCT1,

$$\text{COOMe} + \text{LAH} \longrightarrow \text{CH}_2\text{OH}$$

$$\text{CHO} + \text{LAH} \longrightarrow \text{CH}_2\text{OH}$$

figure 56

P2), (P2,RCT2,P3), (P3,RCT3,P4), ..., (Pn-1,RCTn-1,Pn), c'est-à-dire (P1,PR,Pn) où PR = (RCT1,RCT2, ..., RCTn-1). Nous supposerons qu' une réaction élémentaire ne modifie qu'un ensemble d'atomes et de liaisons adjacents.

1. Exigences du chimiste

Dans la documentation des réactions, le chimiste peut fixer son attention sur différents points:

- produit d'arrivée: il doit "préparer" un produit et désire savoir si ce produit a déjà été préparé et par quelle(s) réaction(s);
- produit de départ: le chimiste possède un produit et désire connaître les réactions possibles sur ce produit;
- produit de départ et d'arrivée ou produit de départ et réactif ou

réactif et produit d'arrivée: quelles sont les réactions possibles?

- type de réactions: les réactions, en chimie organique, sont réties en classes d'après la nature des transformations qui s'y produisent. On peut, en effet, observer que les changements produits par les réactions peuvent se ramener à un petit nombre de types, tels que: constitution d'un édifice moléculaire à partir d'éléments d'abord séparés (réactions de synthèse), ou l'inverse (dissociation), échange d'éléments, ou de groupes d'éléments (qu'on nomme radicaux), entre des espèces chimiques (réactions de substitution ou d'échange), union de deux espèces chimiques pour donner un seul composé (réactions d'addition), échange de protons(réactions acide-base), échange d'électrons (réactions d'oxydoréduction), etc. Parmi l'ensemble des réactions-réponses qui lui sont proposées, le chimiste peut ne s'attacher qu'à une classe particulière;
- réaction directe ou indirecte: la réaction doit se faire en une ou plusieurs étapes;
- mécanisme de réaction: le chimiste ne s'attache plus au produit de départ et d'arrivée mais aux sites. Un site (réactionnel ou antiréactionnel) est un ensemble d'atomes, de liaisons ou cycles modifiés par la réaction (dans le produit de départ ou d'arrivée).

2. Représentation

L'analyse de ces exigences fait apparaître la nécessité de représenter une réaction de différentes façons:

- méthode orientée composée: chaque produit du triplet (P1,RCT,P2) est codé et accédé par les méthodes vues précédemment;
- méthode orientée groupe fonctionnel: dans ce cas, une réaction opère une transformation entre les groupes GF1 du produit P1 et GF2 du produit P2;
- méthode orientée site: la réaction est décrite indépendamment des produits de départ et d'arrivée. En fait ceux-ci ne sont que les états de la réaction. Si nous considérons les produits P1 et P2 des réactions de la figure 57 , R1 et R2 sont du même type: formation d'un amine à partir d'un nitro ($AR-NO_2 \longrightarrow AR-NH_2$). Par contre, si nous regardons le type de transformation, nous avons: dans R1 et R2, une rupture N, O et une formation N, H, une rupture C, N et une formation C, N.

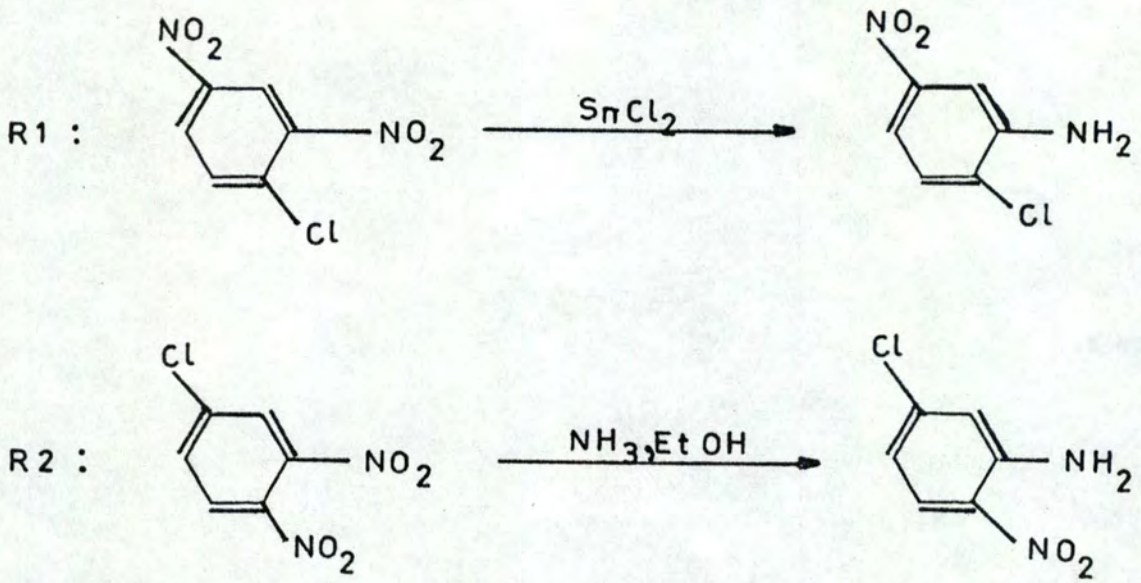


figure 57

3. Recherche du site d'une réaction

Le site d'une réaction est formé par le rassemblement de petites structures partielles appelées fragments (figure 58).

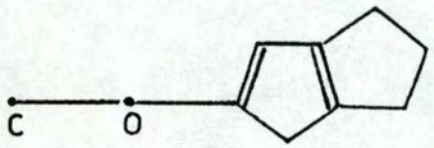


figure 58

- fragments centrés sur un atome:

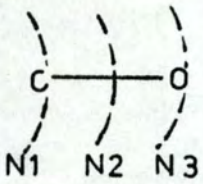


figure 59

- fragments centrés sur une liaison:

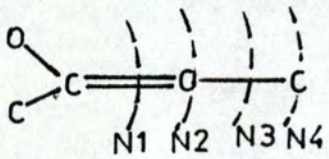


figure 60

- fragments centrés sur un cycle:

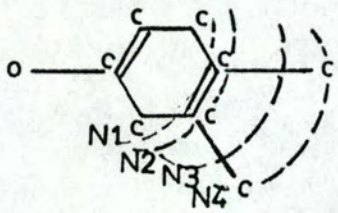


figure 61

Le code des fragments est formé par la couleur de l'élément central (atome = numéro atomique, liaison = valeur de la liaison, cycle = taille du cycle) suivi, par ordre croissant, des couleurs des éléments d'un même niveau. La figure 62 donne les codes de quelques fragments.

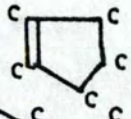
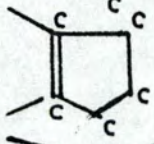
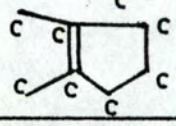
type de fragment	exemple	code
atome niveau 1	C	06
niveau 2	C—	0601
niveau 3	C—O	060116
liaison niveau 1	=	02
niveau 2	C=C	020606
niveau 3	>C=C—	020601060101
niveau 4	O>C=C—C	0206010606010601 6
cycle niveau 1		0101010102
niveau 2		0101010102060606060 6
niveau 3		0101010102060606060601 01
niveau 4		0101010102060606060601 01 0606

figure 62

Pour former le site d'une réaction, les produits de départ et d'arrivée sont analysés en termes de fragments. Les deux ensembles de fragments sont ensuite comparés:

- certains fragments sont communs. Parmi ceux-là, certains sont communs en nombre et en conformation (fragments COM); d'autres sont communs en conformation et non en type (fragments EXTRA);
- certains ne sont pas communs (fragments NONCOM).

EXTRA ∪ NONCOM forme l'ensemble des fragments sur lequel sera basé le réassemblage du site. NOTE: les fragments doivent être du même type. Cherchons, par exemple, le site liaison niveau 3 de la réaction de la figure 63 .

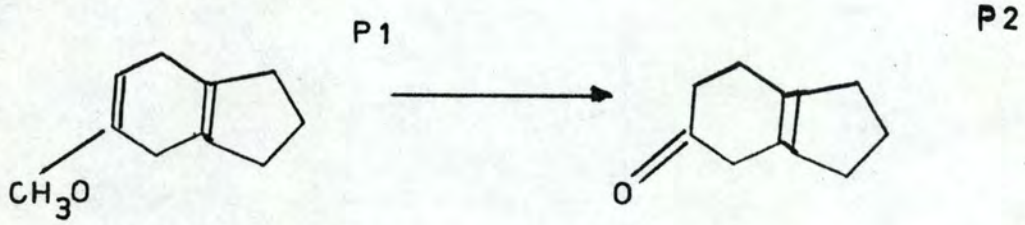


figure 63

L'ensemble des fragments liaison niveau 3 pour P1 et P2 sont:

P1	nombre	P2	nombre
C—O—	(1)	O=C<	(1)
=C—C—	(1)	>C—C—	(6)
>C—O—	(1)	—C—C—	(3)
—C—C— ‖	(5)	>C=C<	(1)
—C—C—	(2)		
>C=C—	(1)		
>C=C<	(1)		

figure 64

NOTE: les chiffres entre parenthèses donnent la fréquence du fragment dans le composé.

EXTRA = { —C—C(=)— , —C—C— }

NONCOM = { C—O— , =C—C— , >C—O— , >C=C— , O=C< }

En parcourant la structure de départ et d'arrivée dans l'ordre assigné par l'algorithme de numérotation biunivoque aux atomes, on assemble les fragments appartenant soit à EXTRA, soit à NONCOM (figure 65).

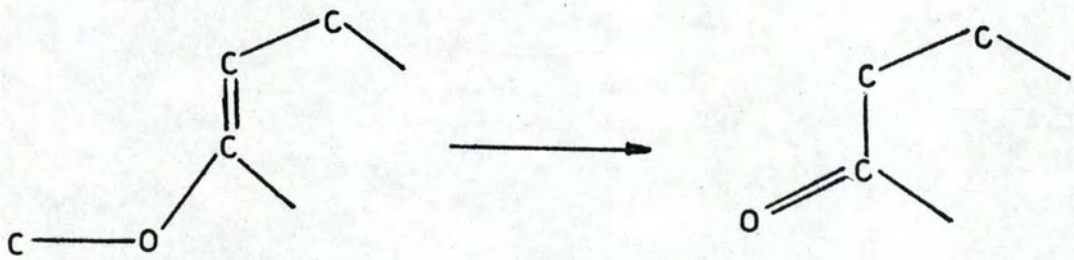


figure 65

Le code d'un site de type donné est obtenu par juxtaposition (par ordre croissant) des codes des fragments, qui le forment (figure 66).

01061601 011601060102 020601060101 0106010602
F1 F2 F3 F4

----> 0216060101 01060106010216 0106010601
F1 F2 F3

figure 66

CONCLUSION

NOTE: nous conseillons au lecteur de consulter d'abord le chapitre concernant le système bibliographique qui a été implémenté et qui définit les principaux concepts informatiques utilisés.

Le code, qui sert de type d'indicatif pour plusieurs types d'entité, est composé de pointeurs vers les descripteurs, ce qui permet de classer et d'accéder les composés par descripteur:

- deux composés qui ont, par exemple, le même descripteur FROM, comportent le même pointeur vers ce descripteur;
- à partir d'un descripteur, on peut accéder aux composés qui le comportent.

Les trois types de représentations d'une réaction sont présents (figure 67):
- transformation de produits,
- transformation de sites,
- transformation de groupes fonctionnels.

Après avoir accédé à un réactif, à un site, à un groupe fonction nel, ou à un composé (par l'intermédiaire du code, du nom ou de la formule développée), on peut retrouver toutes les réactions concernées. L'utilisation de listes inversées permet la recherche de réactions satisfaisant à plusieurs conditions. Par exemple: " Quelles sont les réactions travaillant sur le groupe fonctionnel UN-TEL et par l'intermédiaire de tel réactif ?"

Enfin, en associant à un composé l'ensemble des écrans qu'il comporte, on permet la recherche de structures par écrans canoniques.

La figure 67 donne un essai de diagramme structural pour une implémentation future de la base de données.

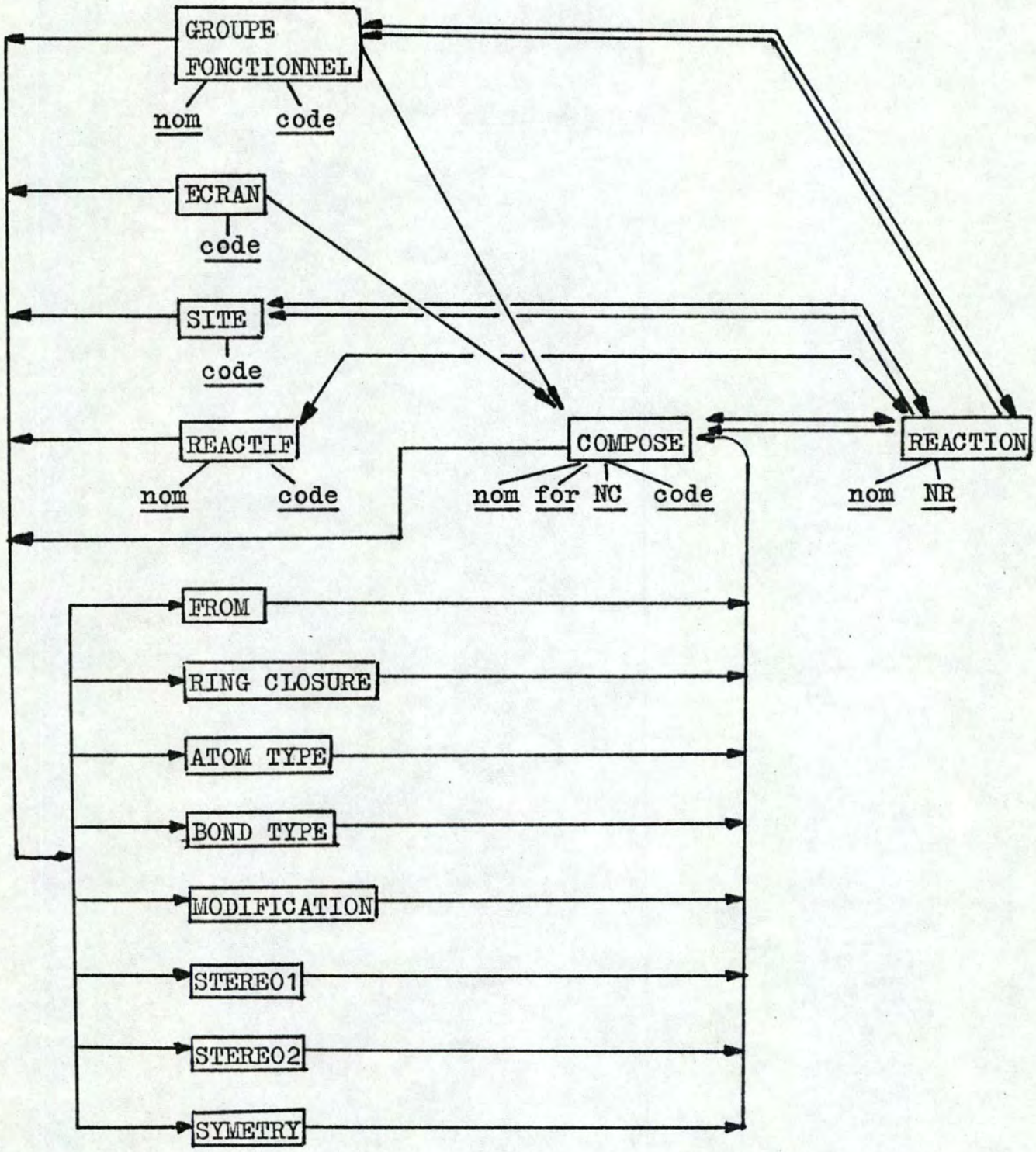


diagramme du système topologique

figure 67

TROISIEME PARTIE : BASE DE DONNEES INFORMATIONS

INTRODUCTION

Les exigences à satisfaire lors de la création d'une base de données informations (BDI) sont différentes. Les informations à traiter sont des suites de données textuelles ou de valeurs numériques qu'il est difficile de coder et qui doivent être restituées sans déformation.

Les champs d'applications de cette BDI, dans notre problème, sont limités. Ils sont de deux types:

- système bibliographique: il s'agit de fournir au chercheur la bibliographie existante sur un composé ou une réaction;
- système de gestion des composés commerciaux: "tel composé existe-t-il sous forme commercialisée ? Si oui, sous quelle forme et quelle est la quantité en stock ?"

Cependant, nous avons voulu généraliser ces deux systèmes afin de les rendre opérationnels indépendamment du problème et de répondre à des questions du type: "quels sont les articles écrits par l'auteur UNTEL, en telle année et sur tel sujet ?" ou "quels sont les composés commerciaux de telle marque utilisés dans tel domaine d'application ?".

CHAPITRE I : BASES DE DONNEES

Nous définissons dans ce chapître quelques notions concernant l'organisation d'une base de données, notamment celles d'entité et de relation.

1. Définitions

- type d'item: nom donné à un ensemble A non vide représentant les caractéristiques d'un fait ou d'un objet sur lequel on désire enrégistrer des informations;
- item: un élément de l'ensemble A;
- type d'entité: nom donné à une association B de types d'items;
- entité: un élément (qui est une association d'items) de l'ensemble B;
- indicatif: item qui sert à repérer l'entité dans laquelle il est contenu;
- type d'indicatif: type d'item dont les éléments sont utilisés comme indicatifs.

2. Relations

On appelle relation R entre x et y décrivant respectivement deux ensembles A et B, toute propriété définie sur A x B, c'est-à-dire une propriété caractéristique des éléments d'une partie G de A x B. G s'appelle le graphe de la relation R. On dit que (x,y) appartient à G ou que x et y vérifient la relation binaire R:

$R(x,y) \Longleftrightarrow (x,y) \in G$

ou $G = \left\{ (x,y) \in A \times B \mid R(x,y) \right\}$

Supposons A = B = E. Une relation entre x et y de E est appelée relation simple entre éléments de E ou relation simple définie sur E; elle est caractérisée par son graphe G qui est une partie de E x E.

A partir d'un ensemble fondamental R, appelé référentiel, dont les éléments sont susceptibles de vérifier une certaine propriété caractéristique p, on peut définir un autre ensemble E dont les é-

léments vérifient cette propriété. On écrit parfois E(p) pour rappeler que les éléments de E sont caractérisés par la propriété p.
NOTE: une propriété est caractéristique si tout élément de R qui la vérifie appartient à l'ensemble défini E, et si tout élément qui n'appartient pas à E ne la vérifie pas. Au sous-ensemble des éléments qui conviennent correspond une relation unaire, car, pour chaque élément pris individuellement, il est possible de dire s'il est ou non dans la relation. Une relation unaire peut donc être considérée comme une partie d'un référentiel.

CHAPITRE II : SYSTEME BIBLIOGRAPHIQUE

La base de données a été structurée en fonction des questions susceptibles de lui être posées:

- quels sont les articles correspondant à un mot clé ?
- quels sont les articles écrits par un auteur ?
- quels sont les articles parus en telle année ?
- quels sont les articles possédés par un chercheur ?
- quels sont les articles traitant d'un composé ou d'une réaction ?
- quels sont les articles correspondant à un mot clé et parus en telle année ?
- quels sont les sujets traités par un auteur ?
- etc.

1. Types d'entités et d'items

types d'entités	types d'items
les mots clés (CLE)	mot clé (MOT)
les auteurs (AUTEUR)	nom (NOM), initiales (INIT)
les propriétaires (PROPRIETAIRE)	nom (NOM)
les composés ou réactions (STRUCTURE)	nom (NOM), formule (FOR)
les références (REFERENCE)	numéro de bibliographie (NB) titre (TITRE) revue (REVUE) année de parution (ANNEE) numéro du volume (VOL) première page (PG1) dernière page (PG2)

tableau 6

Les figures 69 et 78 donnent deux autres formes de représentations possibles des types d'entités.

2. Relations, opérateurs et chemins d'accès

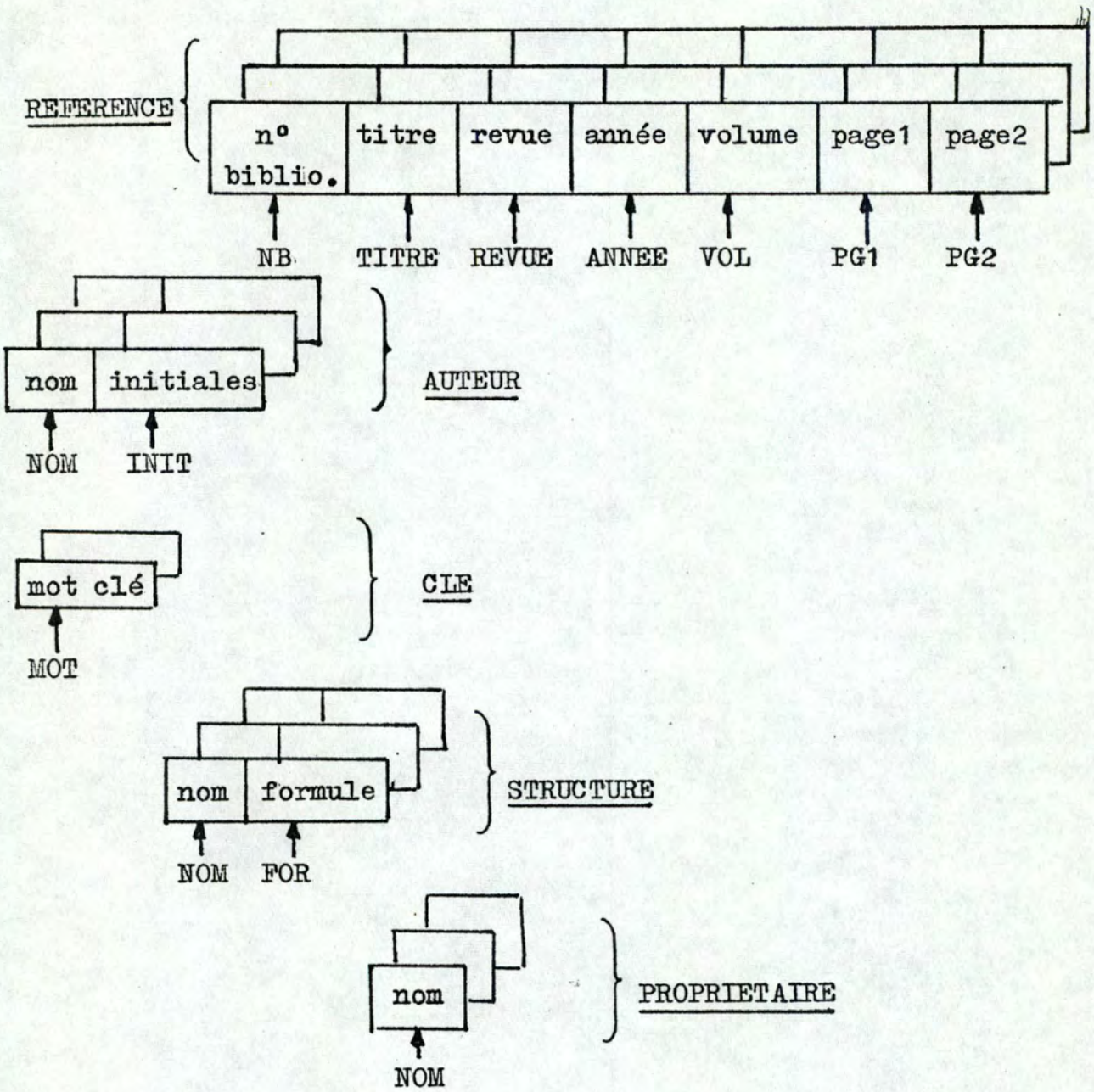


figure 69

sous forme de diagramme:

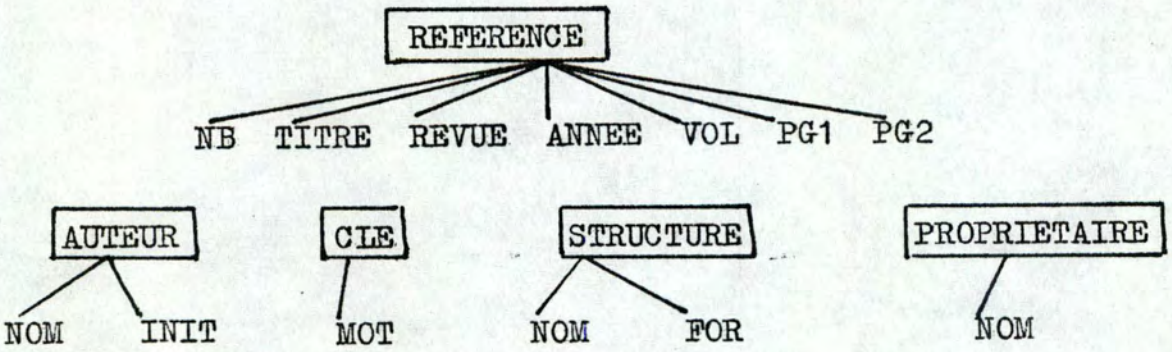


figure 70

Si nous analysons les questions, nous remarquons qu'elles sont de deux types: les questions élémentaires et les questions composées.

2.1. Question élémentaire

- quels sont les articles écrits par: ----
- quels sont les articles parus durant: ----
- etc.

Il s'agit de retrouver tous les articles qui satisfont à une condition p.

2.2. Question composée

Imaginons la question suivante: "quels sont tous les articles écrits par UNTEL et AUTRETEL en 1972 ?"; elle peut se décomposer en trois questions élémentaires:

(1) quels sont les articles écrits par UNTEL ?
(2) parmi ceux-ci, quels sont les articles écrits par AUTRETEL ?
(3) parmi ceux-ci, quels sont les articles parus en 1972 ?

La condition composée P peut s'exprimer avec quelques conditions élémentaires p_1, p_2, ..., p_n liées entre elles par un opérateur logique. Nous n'avons retenu que l'opérateur .AND. qui est le plus fréquent.

Le tableau 4 donne les relations que nous avons considérées entre les types d'entités, les opérateurs et les chemins d'accès associés. Le chemin d'accès associé à une relation R est le chemin qui implémente physiquement la relation; l'opérateur associé est la procédure qui, en parcourant le chemin d'accès, permet de retrouver les éléments qui satisfont à la relation.

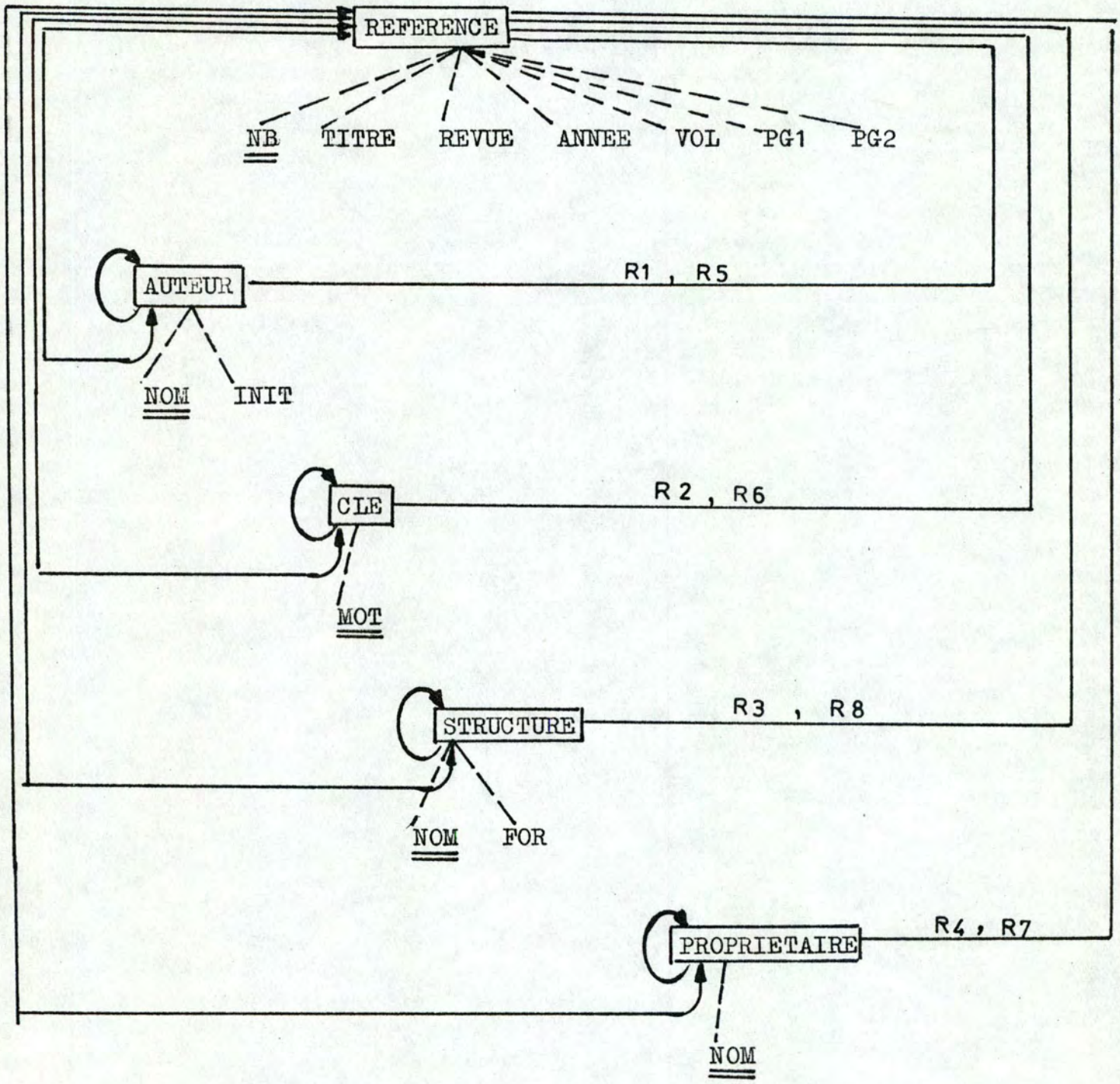
3. Implémentation des chemins d'accès

A une relation R(A,B,a) entre les types d'entités A et B, qui à une entité a de A associe toutes les entités de B satisfaisant à la relation R, nous avons associé deux chemins d'accès:

Relation	types d'entités concernés		opérateur associé	chemin d'accès
I) Relations binaires				
R1 être auteur de	AUTEUR	REF.	OAUT(AUTEUR,REF,auteur)	- à "auteur" - de "auteur" à tous les articles dont il est l'auteur
R2 être mot clé de	CLE	REF.	OCLE(CLE,REF,clé)	- à "clé" - de "clé" à tous les articles dont il est le mot clé
R3 être référencé dans	STRUCTURE	REF.	OSTR(STRUCTURE,REF,structure)	- à "structure" - de "structure" à tous les art. qui la référencent
R4 posséder	PROPRIETAIRE	REF.	OPROP(PROP,REF,prop)	- à "prop" - de "prop" à tous les articles qu'il possède
R5 être écrit par	REF.	AUTEUR	OREF1(REF,AUTEUR,nb)	- à l'article de n° "nb" - de cet article à tous ses auteurs
R6 traiter de	REF.	CLE	OREF2(REF,CLE,nb)	- à l'article de n° "nb" - de cet article à tous ses mots clé
R7 être possédé par	REF.	PROPRIE-TAIRE	OREF3(REF,PROP,nb)	- à l'article de n° "nb" - de cet article à son propriétaire
R8 référencer	REF.	STRUCTU-RE	OREF4(REF,STRUCTURE,nb)	- à l'article de n° "nb" - de cet article à toutes les structures référencées

II) Relations simples				
R9 précéder lexicographiquement	AUTEUR	—	OLAUT(AUTEUR)	ordre lexicographique sur les auteurs
R10 précéder lexicographiquement	CLE	—	OLCLE(CLE)	ordre lexicographique sur les mots clé
R11 précéder lexicographiquement	STRUCTURE	—	OLSTRU(STRUCTURE)	ordre lexicographique sur les noms de structures
R12 précéder lexicographiquement	PROP	—	OLPROP(PROPRIETAIRE)	ordre lexicographique sur les propriétaires
R13 précéder lexicographiquement	REF	—	OLREF(REFERENCE)	ordre lexicographique sur tous les articles
III) Relations unaires				
R14 être paru en	REF	année	OANN(nb)	à l'article de n° "nb"
R15 être paru dans	REF	revue	OREV(nb)	à l'article de n° "nb"

tableau 4



Graphe des relations et des chemins d'accès

figure 71

- chemin d'accès direct: qui permet de retrouver l'entité a;
- chemin d'accès par itinéraire: qui, à a, associe toutes les entités de B liées par R.

3.1. Chemins d'accès par itinéraire associés aux relations binaires

Nous avons choisi de les implémenter sous forme de listes inversées. Par cette méthode, la composition d'opérateurs (pour répondre aux questions composées) est facilitée: il suffit de faire l'intersection des listes inversées correspondantes. Nous appelons liste inversée un ensemble d'éléments, non nécessairement adjacents, chaque élément étant divisé en deux champs:
- le premier champ contient un pointeur sur l'information définie par cet élément;
- le deuxième champ contient un pointeur sur l'élément suivant de la liste.

L'accès à la liste peut se faire à l'aide d'un pointeur P sur le premier élément; le dernier contient une entrée spéciale, notée *, pour signifier un pointeur nul dans le second champ.

La relation R1 qui, à un auteur, associe tous les articles dont il est l'auteur, est implémentée par la liste inversée LI1 (figure 72).

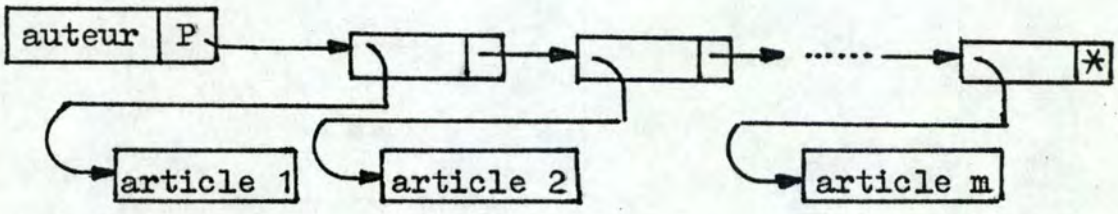


figure 72

En groupant les éléments de la liste par bloc de n éléments, nous réalisons un gain de place, car certains pointeurs deviennent inutiles (figure 73).

NOTES:

(1) Si une liste inversée facilite la recherche d'intersection, el-

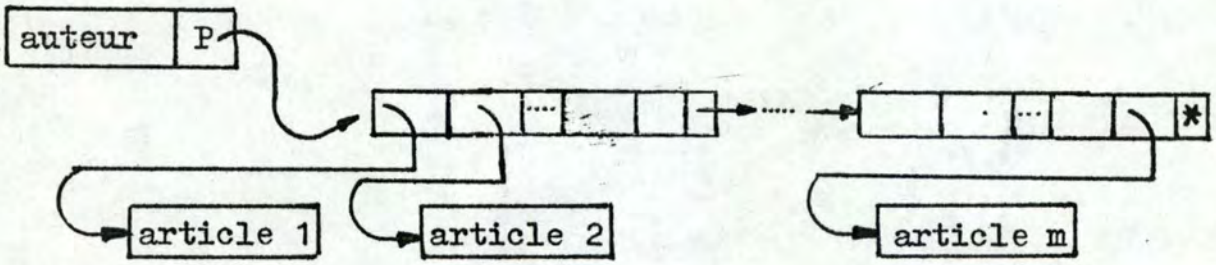


figure 73

le est aussi caractéristique de relations que nous appellerons "non-disjointes".

définition

- soit une relation R(A,B);
- pour tout $a_i \in A$, construisons $B_i = \{b_i \in B \mid R(a_i,b_i)\}$;
- nous dirons qu'une relation est disjointe si $B_i \cap B_j = \emptyset$, pour tout i,j. Elle est non-disjointe dans le cas contraire.

La relation R1 (être auteur de), par exemple, est non-disjointe, car plusieurs auteurs peuvent avoir écrit le même article.

Une relation non-disjointe ne peut pas être implémentée sous forme de liste simple: des difficultés de parcours se poseraient (figure 74). Ces dernières n'existent plus dans le cas de relations disjointes que l'on peut donc créer au moyen de listes simples si le problème de recherche d'intersection ne se pose pas (figure 75).

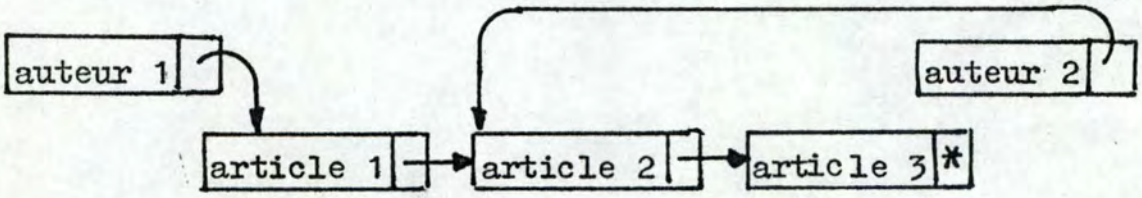


figure 74 : 'relation être auteur de'; supposons que l'auteur 1 ait écrit trois articles et l'auteur 2 un seul. Comment savoir (sans artifice) que l'auteur 2 n'a pas écrit l'article 3 ?

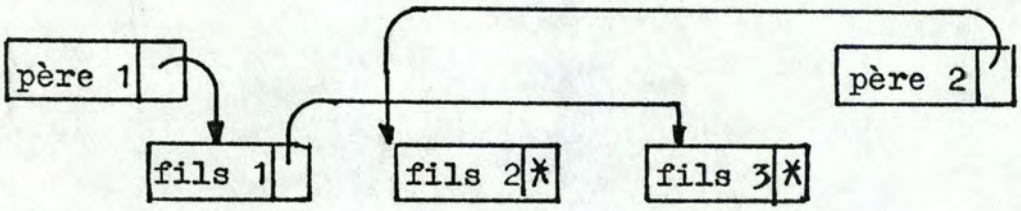


figure 75 : 'relation être père de'; deux pères différents ne peuvent avoir le même fils. Le parcours des listes simples n'est donc pas ambigu.

(2) Une relation R(A,B) qui à une entité a de A associe une et une seule entité b de B (relation biunivoque) est une liste inversée d'un type particulier (figure 76); on peut remplacer cette liste par le pointeur P (figure 77).

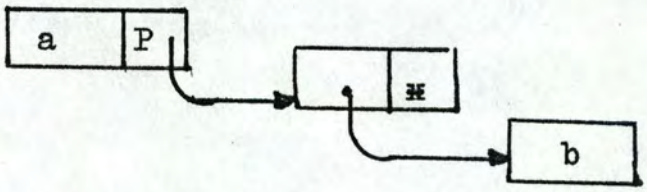


figure 76

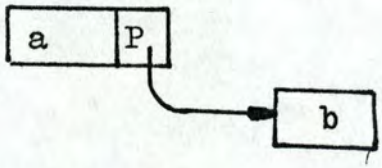


figure 77

3.2. Chemins d'accès par itinéraire associés aux relations simples

Ils sont représentés par des chaînes. Nous appellons chaîne un ensemble d'entités liées les unes aux autres. L'accès à une entité ne peut se faire que par balayage de la chaîne à partir de l'une ou l'autre extrémité.

Une chaîne associée à une relation simple sur un ensemble d'entités d'un type donné parcourt ceux-là par ordre alphabétique (figure 78). L'accès aux chaînes triées (CT) se fait par l'intermédiaire d'une table de pointeurs (figure 79). L'utilisation de ces chaînes facilitent l'édition de classements par mot clé, structure, auteur, propriétaire ou titre.

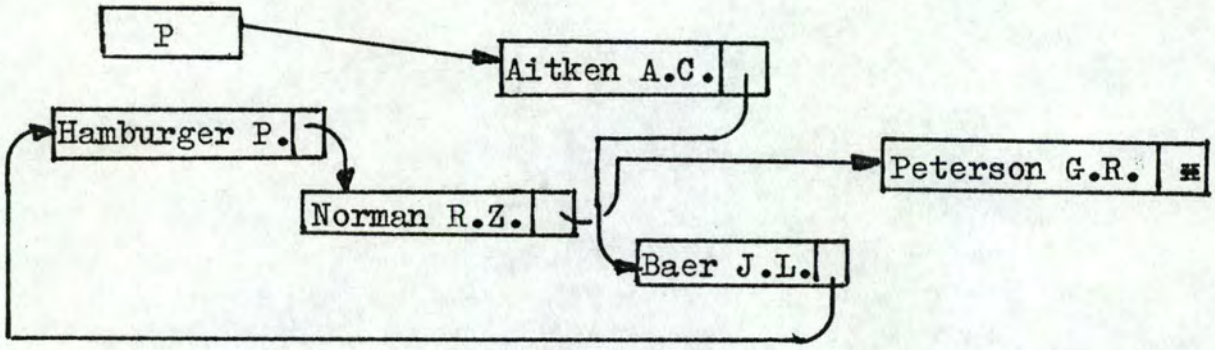
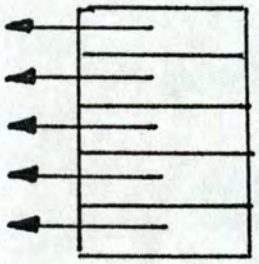


figure 78



pointeur vers la CT1 des mots clé
pointeur vers la CT2 des noms de structures
pointeur vers la CT3 des auteurs
pointeur vers la CT4 des propriétaires
pointeur vers la CT5 des titres

figure 79

3.3. Chemins d'accès directs

Ces chemins découlent de la méthode adoptée pour le rangement des entités, méthode qui dépend de ce que l'on veut en faire et notamment de la nécessité de procéder à des recherches d'entités en ordre aléatoire.

Rangement par "adresses calculées"

Cette méthode consiste à ranger les entités à une adresse qui découle de l'indicatif par un calcul plus ou moins complexe. Quand on veut ranger une entité, on calcule l'adresse qu'elle doit occuper. Cette adresse pourra être aisément recalculée quand il s'agira de retrouver l'entité pour une consultation ou mise à jour.

Le rangement en adresses calculées demande d'examiner:
- le choix d'une méthode de calcul,
- le rangement des "synonymes",
- le groupement des entités,
- le traitement des entités en longueur variable.

Le coefficient de remplissage et le temps de recherche dépendent de ce choix.

3.3.1. Rangement des synonymes

Les méthodes de calcul d'adresse présentent l'inconvénient d'introduire des "synonymes", ce qui revient à affecter à deux ou plusieurs entités la même adresse. On résoud ce problème en créant des zones de débordement: à côté de la zone normale de rangement, on prévoit une zone où ranger l'entité qui ne trouve pas place à son lieu normal de stockage, du fait qu'il est déjà occupé.

Le premier problème est donc de trouver, à partir de la zone normale de stockage, l'endroit où est stockée l'entité rejetée, pour une recherche éventuelle.

méthode des listes

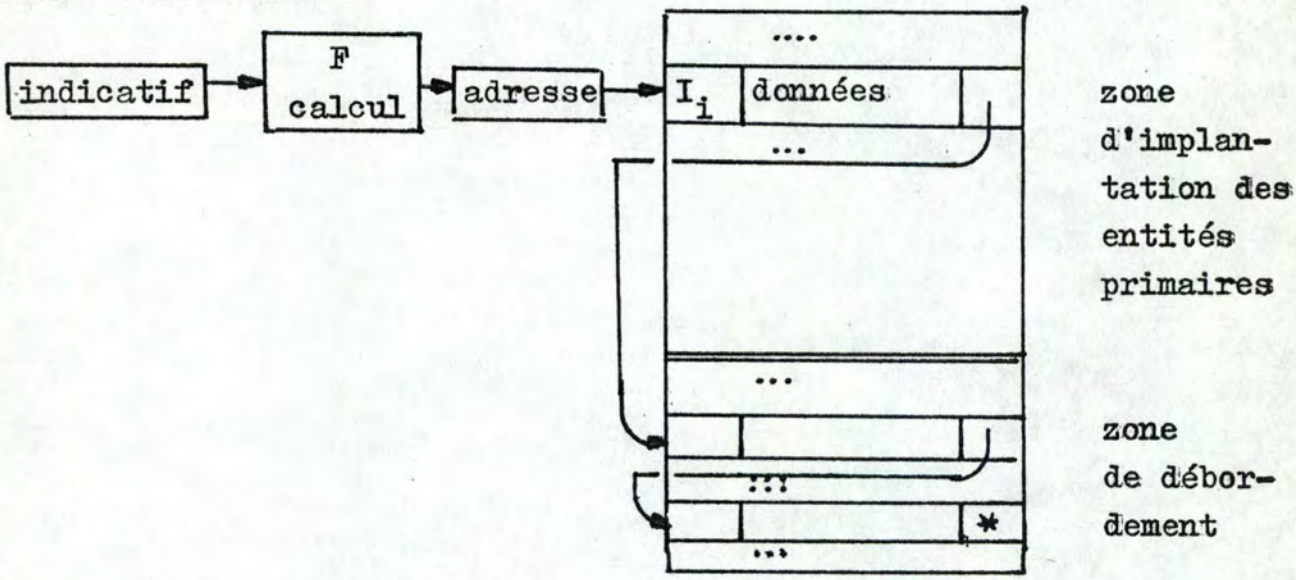


figure 80

A côté de la zone réservée à l'entité, se trouve un emplacement réservé où sera rangé l'adresse de l'entité déplacée (figure 80).

méthode des concaténations

entité primaire	indicatif	adresse	indicatif	adresse	...

figure 81

Il s'agit de prévoir à côté de l'entité primaire, l'emplacement d'une zone de chaînage où l'on range l'adresse de renvoi et l'indicatif des entités déplacées (figure 81). Cette méthode évite, en cas de rejets multiples, de faire des extractions successives; mais elle oblige à la prévision anticipée du nombre de rejets et entraîne une perte de place assez importante.

Le deuxième problème est de définir la taille des zones d'implantation des entités primaires et de débordement. La première zone est définie une fois pour toutes. La deuxième peut être agrandie au fur et à mesure des besoins par un système d'allocation dynamique de mémoire.

- zone d'implantation supérieure à la zone de débordement: le nombre de rejets sera peu élevé et la longueur moyenne des listes également, ce qui entraîne une diminution du temps de recherche d'une entité déplacée. Cette méthode représente cependant une perte de place importante en raison du nombre initial élevé de "trous" dans la zone d'implantation;
- zone d'implantation inférieure à la zone de débordement: les listes de synonymes s'allongent mais la perte de place diminue: la zone d'implantation est minime et, dès le départ, entièrement remplie, tandis que la zone de débordement pourra augmenter au fur et à mesure des ajoutes.

Notre objectif étant un gain de place, la première méthode de chacun des problèmes a été retenue.

3.3.2. Calcul d'adresse

Le premier objectif du calcul d'adresse est d'établir une correspondance entre les nombres représentant les adresses de la zone mémoire affectée à un fichier, et les indicatifs des entités. Le deuxième but est d'obtenir une distribution uniforme des entités dans les différentes listes.

Les indicatifs (mot clé, nom d'auteur, nom de propriétaire, nom de structure) possèdent un certain nombre de caractéristiques:
- ils ne comportent que des caractères alphanumériques,

- tout l'indicatif est représentatif: il n'est pas possible de ne retenir, dans l'indicatif, que la partie la plus représentative et d'écarter le reste qui ne jouerait que peu sur l'individualité de l'indicatif,
- l'indicatif ne peut pas être découpé en zones ayant une signification de nature, de région, de catégorie.

méthode de calcul d'adresse en Fortran:

soit un indicatif de m caractères;

(1) tronçonnage: nous ne considérons que les n premiers caractères de l'indicatif afin d'abréger le calcul;
(2) pliage: somme des valeurs internes de ces n caractères;
(3) division: nous divisons la somme par X, le reste R ($0 \leq R \leq X-1$) donne l'adresse de rangement dans la zone d'implantation, relative à l'adresse origine: X = taille de la zone d'implantation - 1.

La division peut produire des restes égaux pour deux indicatifs différents: ils déterminent des "synonymes" rangés dans la zone de débordement.

3.3.3. Groupement d'articles

En Fortran, la portion de mémoire unitairement adressable est le record logique défini dans le DEFINE FILE n(m,p) où n est le numéro logique du fichier, m la taille en mots (de deux bytes) d'un enregistrement, p le nombre d'enregistrements logiques.

Il n'est donc pas possible de ranger plusieurs synonymes dans une même portion de mémoire afin de diminuer le nombre d'accès au disque.

3.3.4. Rangement d'entités de longueur variable

Toutes les entités traitées dans ce système sont de longueur variable. Le titre, par exemple, peut varier de 50 à 200 caractères. Réserver une place maximale pour ces entités entraînerait une

perte de place très importante. Si on réserve, par exemple, 250 bytes pour le titre qui fait en moyenne 100 caractères, on perd 150 bytes.

Le Fortran ne permet malheureusement pas la manipulation d'enregistrements de longueur variable. Nous avons procédé de la manière suivante: les items de longueur fixe sont groupés au début de l'entité; les items de longueur variable sont découpés en segments de longueur fixe. A l'adresse que l'entité doit normalement occuper d'après le calcul effectué, nous rangeons un pointeur vers une zone tertiaire (zone des découpes) où une chaîne permet de retrouver tous les segments de l'item (figure 82).

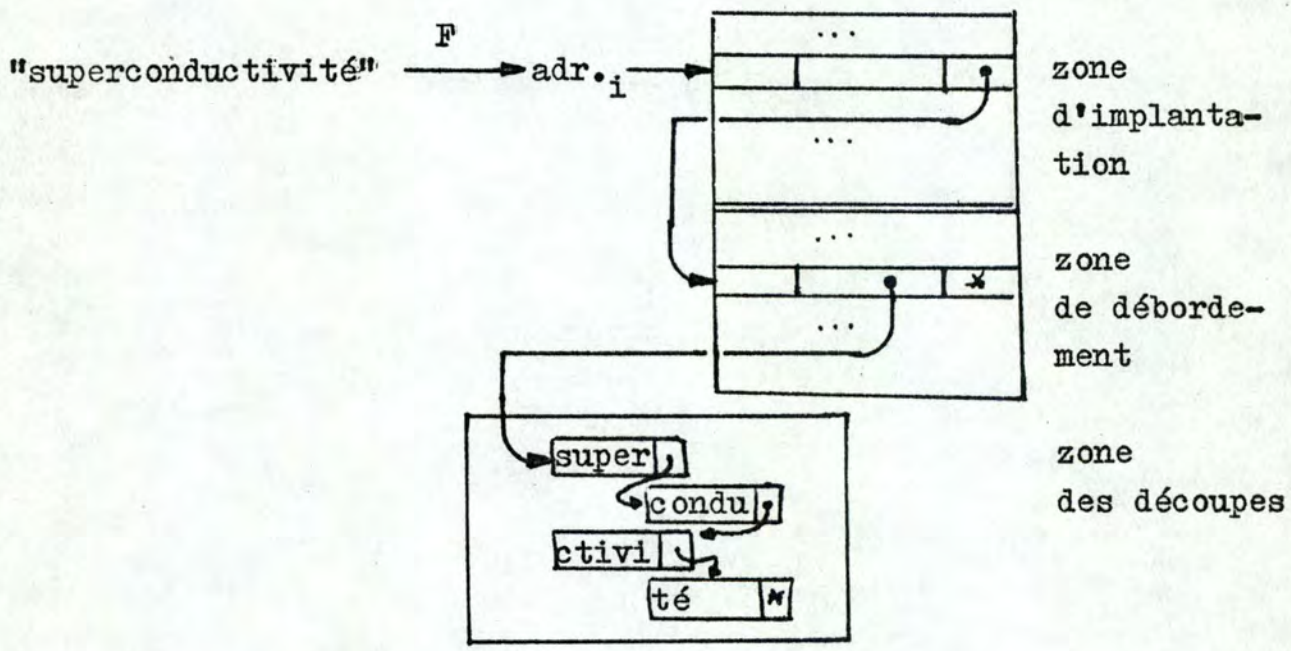


figure 82

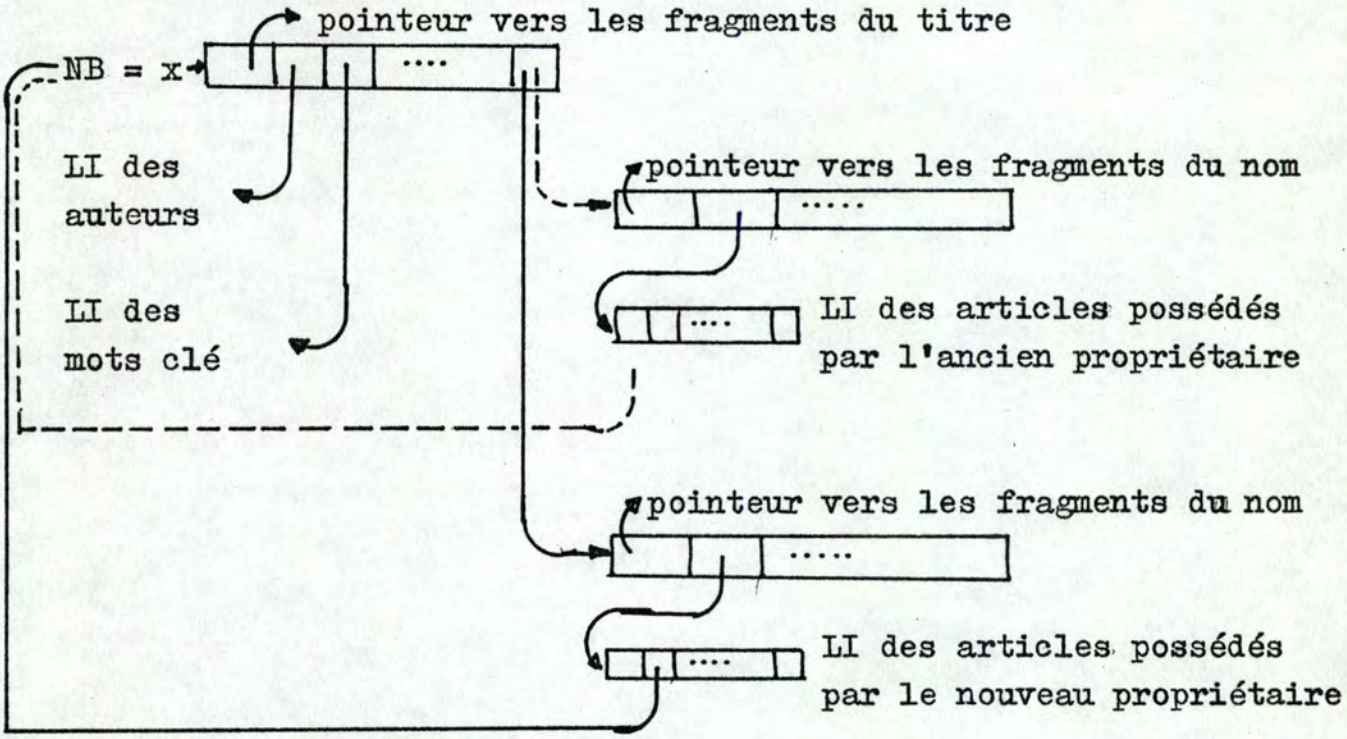
4. Mise à jour

Les seules mises à jour à effectuer dans notre système sont des changements de propriétaire d'un article et des ajoutes d'articles.

Le mode de rangement par adresses calculées est spécialement conçu en vue de la consultation et de la mise à jour instantanée. En conséquence, l'ajoute d'articles fait appel aux méthodes décrites pour le rangement: calcul de l'adresse de l'entité, éventua-

lité de synonymes, recherche et rangement en zone de débordement.

En ce qui concerne le changement de propriétaire, à chaque article est associé un numéro de bibliographie qui n'est autre qu'une numérotation séquentielle suivant leur entrée. Lors d'une demande de changement, l'utilisateur repère l'article par son numéro, entre le nom de l'ancien propriétaire (en vue d'une vérification) et celui du nouveau (figure 83).



------ avant la modification

———— après la modification

figure 83

5. Système d'interrogation

Pour l'utilisateur, les opérateurs apparaissent sous forme de questions élémentaires, une composition d'opérateurs sous forme de question composée qui n'est en fait qu'une série de questions élémentaires.

"Quels sont tous les articles écrits par ---- sur le sujet ---- parus en ---- ?". Les opérateurs correspondant sont:
OAUT .AND. OCLE .AND. OANN .
(1) quels sont les articles écrits par ---- ?
(2) quels sont les articles traitant du sujet ---- ?
(3) quels sont les articles parus durant ---- ?

Il faut noter que, si les questions sont posées séquentiellement, les opérateurs ne sont pas appliqués de la même façon: un opérateur simple parcourt une liste inversée; une composition d'opérateurs réalise une intersection de listes inversées. Cette intersection (un ensemble de numéros de bibliographie) est stockée dans un fichier auxiliaire qui permet à l'utilisateur de voir évoluer ses réponses; il est ainsi mieux à même de préciser les questions lorsque les réponses ne lui paraissent pas satisfaisantes. On évite ainsi à l'utilisateur de devoir conditionner complètement sa question avant l'interrogation: la mise au point de la question devient ainsi tout-à-fait conversationnelle. La figure 86 donne un exemple de dialogue en vue d'une recherche bibliographique.

6. Edition des résultats

Une fois la recherche terminée, il est évidemment nécessaire de pouvoir procéder à une vérification. A cet effet, il est prévu une commande permettant l'affichage sur vidéo ou sur écran graphique (figure 87), de toutes les informations relatives à un numéro de bibliographie donné. On peut ainsi s'assurer de la pertinence de chaque résultat.

Par ailleurs, pour pallier le caractère temporaire d'un affichage sur l'écran, il est possible d'obtenir un listing sur papier des bibliographies.

SYSTEME BIBLIOGRAPHIQUE: diagramme d'accès

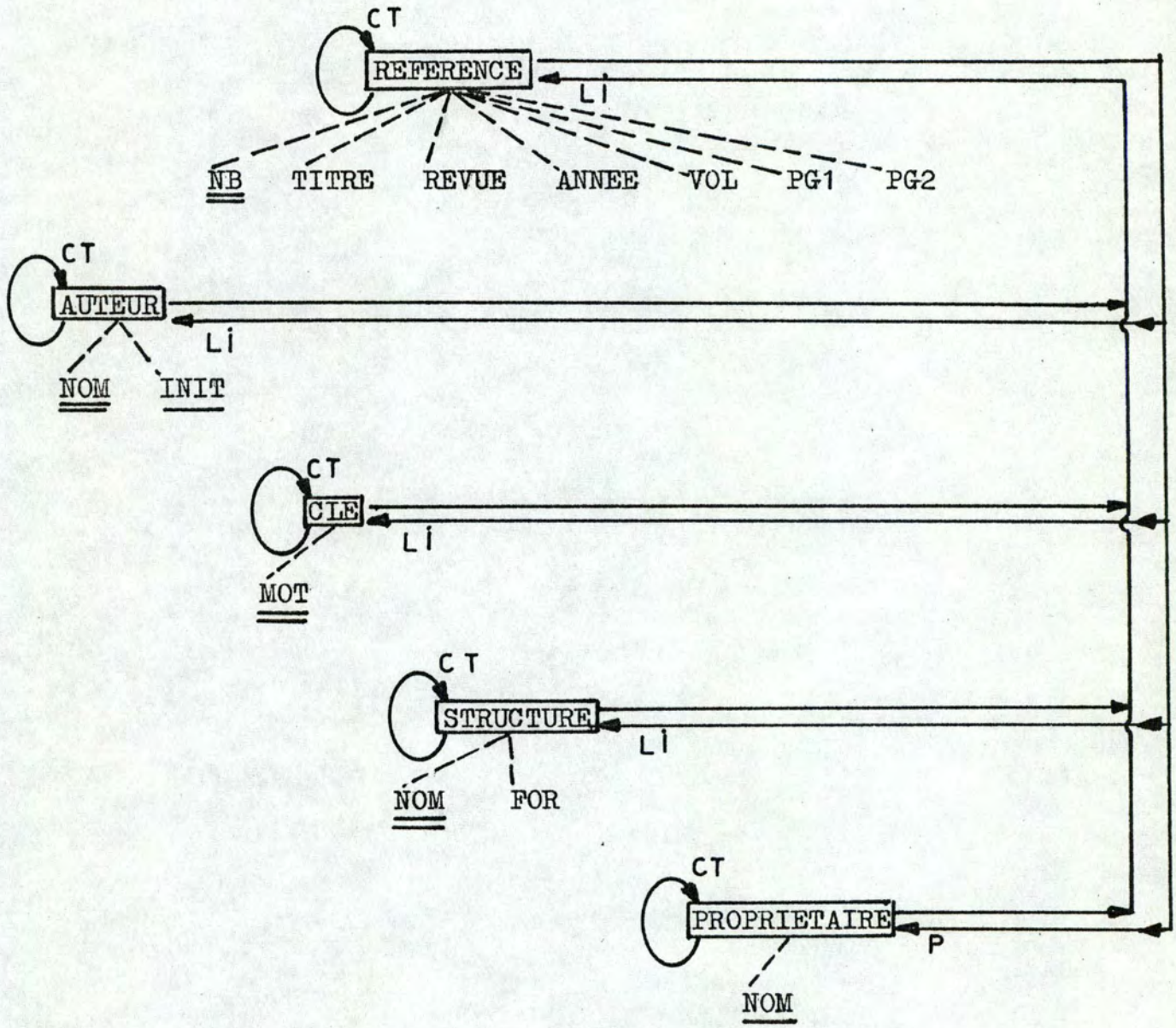


figure 84

LI : liste inversée

CT : chaîne triée

P : pointeur

NOTE: les types d'entités soulignés sont les types d'indicatifs

SYSTEME BIBLIOGRAPHIQUE: enchaînement des modules programmés

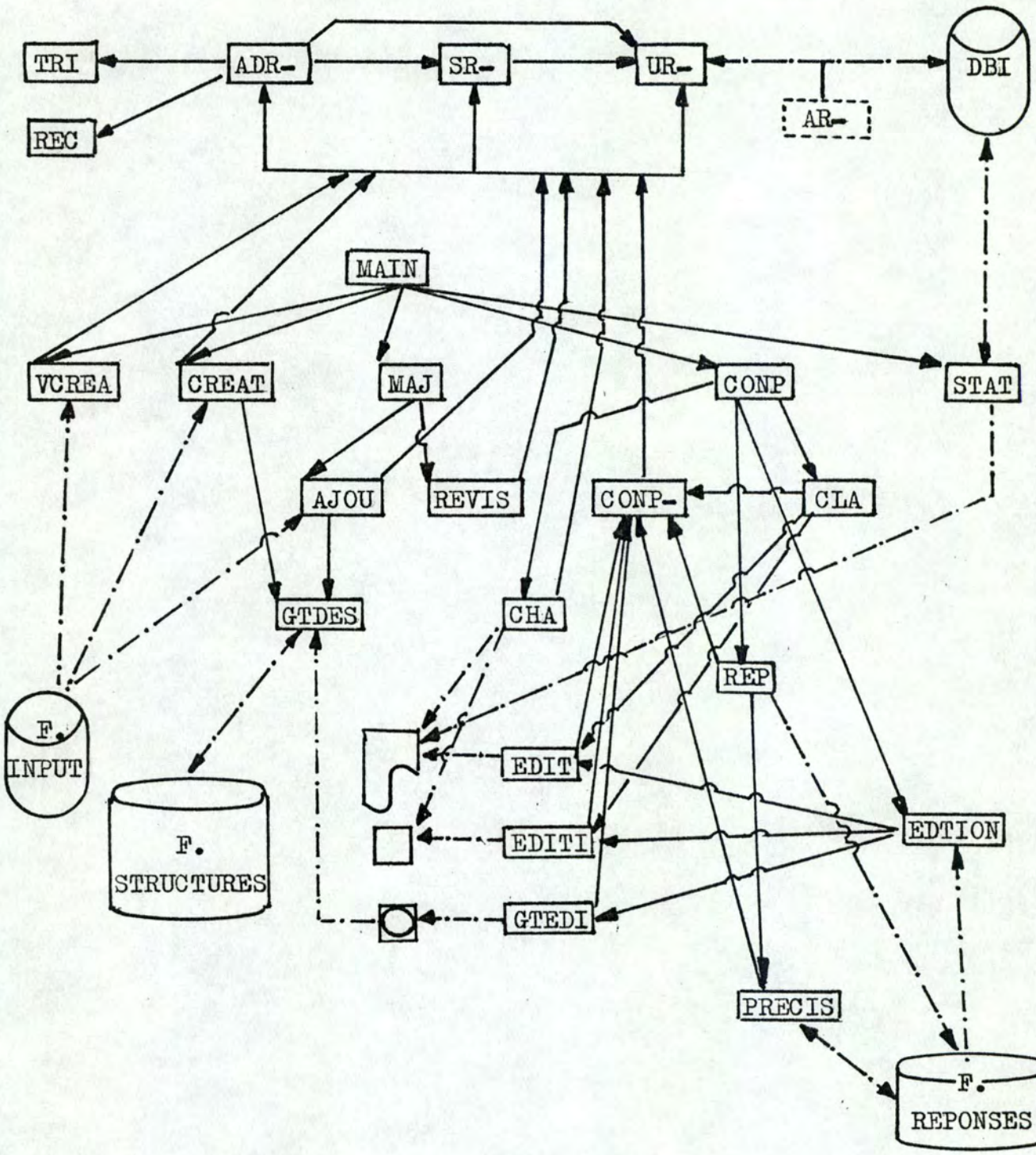


figure 85

modules

appel de module

fichiers

transfet de données

zone de communication entre modules

série de modules

nom du module	fonction du module	
	données en entrée	données en sortie et fonction
UR-	adresse d'une entité	items fixes de cette entité
SR-	adresse d'une entité	items fixes de cette entité recomposition des items (de longueur variable)
ADR-	indicatif d'une entité	adresse de l'entité
TRI	numéro de chaîne triée indicatif	insertion de l'entité (correspondant à l'indicatif) dans la chaîne triée
STAT	BDI	statistiques sur l'utilisation de la BDI
CREAT	fichier input	création de la BDI
VCREA	fichier input	vérification des données
MAJ	-	coordination des modules de mises à jour
CONP	-	coordination des modules de consultation
GTDES	écran graphique	dessin de(s) structure(s) référencée(s) dans un article
AJOU	fichier input	ajouter des articles dans la BDI
REVIS	n° de bibliographie nouveau propriétaire	changement de propriétaire d'un article
CONP-	-	ensemble des opérateurs

CLA	n° de chaîne triée	enchaînement des opérateurs en vue d'un classement des articles suivant un des ordres définis
REP	-	- enchaînement des opérateurs pour répondre à la question posée - les nb appartenant à l'$\cap$ sont stockés dans le F. réponse
PRECIS	-	enchaînement des opérateurs en vue de préciser une question posée
EDTION	fichier réponses	édition du fichier réponse
EDIT	n° de bibliographie	édition sur imprimante des renseignements sur un article
EDITI	n° de bibliographie	édition sur terminal des renseignements sur un article
GTEDI	n° de bibliographie	édition sur écran graphique des renseignements sur un article et des structures référencées par cet article
CHA	n° de chaîne triée	parcours d'une de ces chaînes
REC	indicatif	calcul de l'adresse de l'entité
SQUES	-	module d'interrogation
LIST	-	gestion des listes

tableau 5

```
>
>SET /UIC=[11,13]
>RUN DK1:BIBLI
LABORATOIRE DE CHIMIE THEORIQUE APPLIQUEE
FACULTES NOTRE-DAME DE LA PAIX  BELGIUM
BIBLI  PROGRAM FOR BIBLIOGRAPHICAL RESEARCH


-IDENTIFICATION DE L'UTILISATEUR?
 (1=THEORIQUE,Ø=ORGANIQUE)
TAPEZ LE CHIFFRE CORRESPONDANT
Ø
-QUELLE OPERATION DESIREZ-VOUS?
*CONSULTATION (1)
*MISE A JOUR (2)
*CREATION (3)
*VERIFICATION (4)
*STATISTIQUE (5)

TAPEZ LE CHIFFRE CORRESPONDANT
1
-QUELLE EDITION DESIREZ-VOUS?
*LES ARTICLES CORRESPONDANT A UN MOT CLE (1)
*LES ARTICLES ECRITS PAR UN AUTEUR (2)
*LES ARTICLES PARUS EN TELLE ANNEE (3)
*LES ARTICLES POSSEDES PAR UN CHERCHEUR (4)
*LES ARTICLES TRAITANT D'UN COMPOSE OU D'UNE REACTION (8)
*LES SUJETS TRAITES PAR UN AUTEUR (5)
*CLASSEMENT (6)
*LISTE  (7)
TAPEZ LE CHIFFRE CORRESPONDANT
3
-QUE DESIREZ-VOUS?
*LES ARTICLES PARUS EN TELLE ANNEE (1)
*LES ARTICLES PARUS ENTRE TELLE ET TELLE ANNEE (2)
*LES ARTICLES PARUS APRES TELLE ANNEE (3)
TAPEZ LE CHIFFRE CORRESPONDANT
3
-INTRODUISEZ L'ANNEE (FORMAT:I2)
69
-VOULEZ-VOUS SPECIFIER UNE REVUE?
 (1=OUI,Ø=NON)
TAPEZ LE CHIFFRE CORRESPONDANT

-VOULEZ-VOUS SPECIFIER UN AUTEUR?
 (1=OUI,Ø=NON)
TAPEZ LE CHIFFRE CORRESPONDANT

-VOULEZ-VOUS SPECIFIER UN MOT CLE?
 (1=OUI,Ø=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
```

```
-NOMBRE DE REPONSE:     21
-VOULEZ-VOUS PRECISER VOTRE QUESTION?
  (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-VOULEZ-VOUS SPECIFIER UNE REVUE?
 (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-INTRODUISEZ LA REVUE
 (80 CARACTERES MAXIMUM)
ANGEW. CHEM. INTERNAT. EDIT
-VOULEZ-VOUS SPECIFIER UN AUTEUR?
 (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-INTRODUISEZ LE NOM
 (80 CARACTERES MAXIMUM)
UGI
-INTRODUISEZ LES INITIALES
 (4 CARACTERES MAXIMUM)
 (SI VOUS NE DESIREZ PAS SPECIFIER LES INITIALES,
  TAPEZ LE CARACTERE .)
I
-VOULEZ-VOUS SPECIFIER UN MOT CLE?
 (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-INTRODUISEZ LE MOT CLE
 (80 CARACTERES MAXIMUM)
STRUCTURE
-VOULEZ-VOUS SPECIFIER UNE ANNEE PARTICULIERE?
 (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-QUE DESIREZ-VOUS?
*LES ARTICLES PARUS EN TELLE ANNEE (1)
*LES ARTICLES PARUS ENTRE TELLE ET TELLE ANNEE (2)
*LES ARTICLES PARUS APRES TELLE ANNEE (3)
TAPEZ LE CHIFFRE CORRESPONDANT
1
-INTRODUISEZ L'ANNEE (FORMAT:I2)
70
```

```
-NOMBRE DE REPONSE:      1
-VOULEZ-VOUS PRECISER VOTRE QUESTION?
  (1=OUI,0=NON)
TAPEZ LE CHIFFRE CORRESPONDANT


-QUE DESIREZ-VOUS ?
*SORTIE SUR TERMINAL (1)
*SORTIE SUR IMPRIMANTE (2)
*SORTIE SUR ECRAN GRAPHIQUE (3)

TAPEZ LE CHIFFRE CORRESPONDANT
1
```

```
*  LABORATOIRE DE CHIMIE THEORIQUE APPLIQUEE  *
*  FACULTES NOTRE-DAME DE LA PAIX  BELGIUM    *
*  BIBLI  PROGRAM FOR BIBLIOGRAPHICAL RESEARCH*
*                                             *
*  DATE:17-MAY-77                             *
-TITRE:            (NO DE BIBLIOGRAPHIE:      1)
CHEMISTRY AND LOGICAL STRUCTURES

-REVUE:
ANGEW. CHEM. INTERNAT. EDIT
 VOLUME:     9      PAGE:    703   A   730      ANNEE:  1970
-LISTE DES MOTS CLES:
STRUCTURE
SET THEORY
TOPOLOGY
GROUP THEORY
ANALOGY
CLASSIFICATION
NOMENCLATURE
REACTION
-LISTE DES AUTEURS:
(I    )  UGI
(D    )  MARQUARDING
(H    )  KLUSACEK
(G    )  GOKEL
(P    )  GILLESPIE
-PROPRIETAIRE:
J.M. ANDRE
TAPEZ 99 SI VOUS DESIREZ ARRETER. SINON TAPEZ          <CR>

TT3  --  STOP
>
```

figure 86

```
*************************************************
*  LABORATOIRE DE CHIMIE THEORIQUE APPLIQUEE    *
*  FACULTES NOTRE-DAME DE LA PAIX   BELGIUM     *
*  BIBLI  PROGRAM FOR BIBLIOGRAPHICAL RESEARCH*
*                                               *
*  DATE:17-MAY-77                               *
*************************************************

$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
$NUMERO DE BIBLIOGRAPHIE:          1        $
$                                             $
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$


-TITRE:
-------
CHEMISTRY AND LOGICAL STRUCTURES


-REVUE:
-------
ANGEW. CHEM. INTERNAT. EDIT
 VOLUME:      9
 PAGE:    703    A      730
 ANNEE:  1970


-LISTE DES MOTS CLES:
---------------------
STRUCTURE
SET THEORY
TOPOLOGY
GROUP THEORY
ANALOGY
CLASSIFICATION
NOMENCLATURE
REACTION


-LISTE DES AUTEURS:
-------------------
UGI
INITIALES: I

MARQUARDING
INITIALES: D

KLUSACEK
INITIALES: H

GOKEL
INITIALES: G

GILLESPIE
INITIALES: P


-PROPRIETAIRE:
--------------
J.M. ANDRE
```

figure 87

CHAPITRE III : SYSTEME DE GESTION DES PRODUITS COMMERCIAUX

Lorsqu'on recherche, par exemple, le MERCURE OXYDE (MERCURIQUE) dans le catalogue des produits de la marque BILLAULT, on trouve les renseignements suivants:

nom du composé: mercure oxyde (mercurique),
formule: HgO,
extension: rouge pur pour analyses FB,
numéro de catalogue: 625788,
unité de vente: 1 kg,
prix unitaire: 200 fb,
degré de pureté: supérieur à 99%,
liste des impuretés:

Pb	0.0005%
Cu	0.0005%
Cd	0.0005%
Fe	0.0010%
Zn	0.0005%
N	0.005%
SO_4	0.005%

domaine d'application: réactif pour analyse,
liste des synonymes: mercure peroxyde, mercure (II) oxyde.

NOTES:

(1) signification de la notion d'extension:
exemple: ANHYDRIDE IODIQUE pur pour analyses FB
nom du composé: ANHYDRIDE IODIQUE
extension: pur pour analyses FB;

(2) pour un même nom de composé, il peut exister des produits correspondant à des extensions différentes:
exemple: FER CHLORURE (FERRIQUE) en solution à 30° Bé
FER CHLORURE (FERRIQUE) en solution à 45° Bé
FER CHLORURE (FERRIQUE) sec
FER CHLORURE (FERRIQUE) hexahydraté pur pour analyses FB
etc.;

(3) pour un même composé, il peut exister des produits correspondant à des marques différentes:
exemple: ANHYDRIDE SELENIEUX sublimé

commandé chez BILLAULT
ANHYDRIDE SELENIEUX pur sublimé
commandé chez LABOSI;

(4) pour un même numéro de catalogue, il peut exister des produits de versions différentes:
exemple: mercure oxyde (mercurique) rouge pur pour analyses FB
numéro de catalogue: 625788
première version: unité de vente: 1 kg,
prix unitaire: 200 fb,
deuxième version: unité de vente: 5 kg,
prix unitaire: 800 fb.

En prenant ces informations comme base, nous avons réalisé un système qui permet de répondre à des questions du type:

- quels sont les produits appartenant à tel domaine d'application ?
- quelles sont les références d'un produit de tel nom ou de numéro de catalogue untel ?
- quels sont les produits utilisables dans tel domaine et de telle marque ?

Ce système doit aussi être capable de gérer les stocks des produits dans un laboratoire et de vérifier la circulation des produits:

- qui possède tel produit et par quel projet cette personne est-elle concernée ?
- quels sont les produits actuellement en possession d'un chercheur travaillant sur tel projet ?

1. Types d'entités et d'items

types d'entités	types d'items
les domaines d'application (DOMAINE)	domaine d'application (DOM)
les noms des produits (STRUCTURE)	nom (NOM) formule (FOR)
les références des produits (REFERENCE)	numéro de catalogue (NOCA) marque (MAR)

	quantité de vente (QVEN)
	unité de vente (UVEN)
	prix unitaire (PUNI)
	unité monétaire (UNIM)
	degré pureté (DPUR)
les extensions (EXTENSION)	extension (EXT)
les synonymes (SYNONYME)	synonyme (SYN)
les impuretés (IMPURETE)	impureté (IMP)
les stocks (DISPONIBILITE)	quantité en stock (QSTO)
	date d'ouverture du produit (DOUV)
	date de validité du produit (DVAL)
les chercheurs (CHERCHEUR)	nom (NOM)
	initiales (INIT)
	poste (POST)
les projets (PROJET)	nom (NOM)

tableau 7

Sous forme de diagramme:

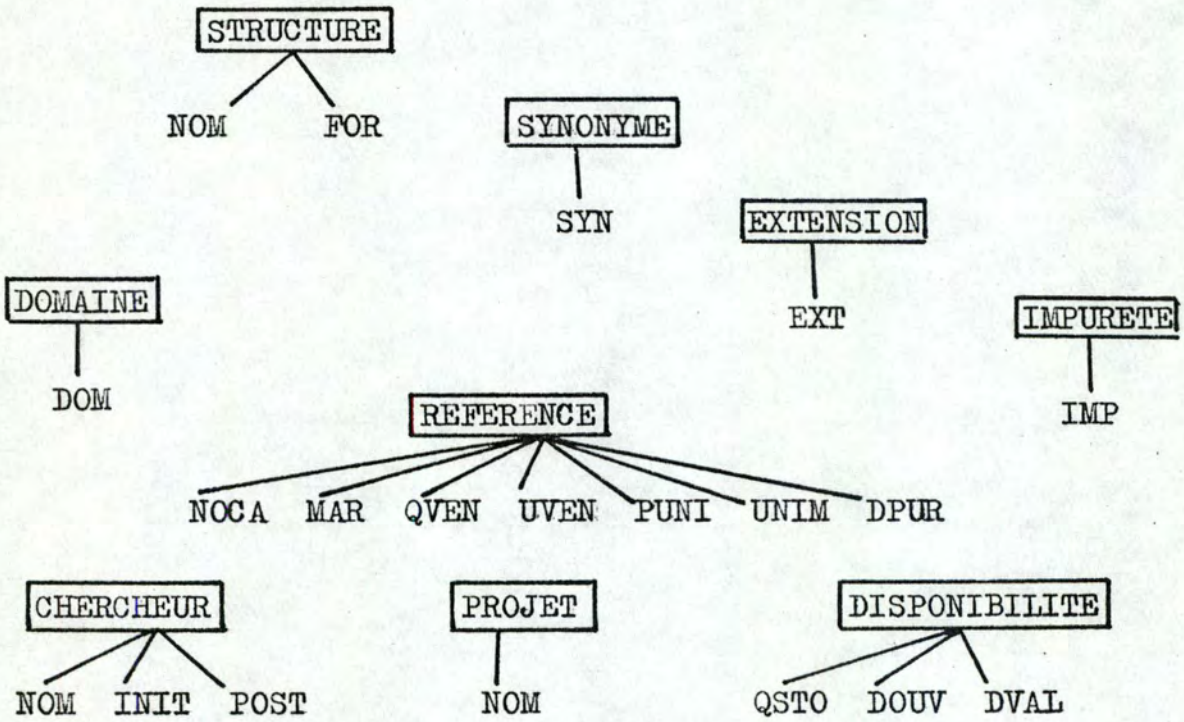


figure 88

Relation	types d'entités concernés		opérateur associé	chemin d'accès
I) Relations binaires				
R1 être domaine de	DOMAINE	REF.	ODOM(DOMAINE,REF,dom)	- à "domaine" - de "domaine" à tous les produits qui y sont utilisables
R2 être nom de	STRUCTURE	REF.	OSTR(STRUCTURE,REF,structure)	- à "structure" - de "structure" à tous les produits de ce nom
R3 avoir pour synonymes	STRUCTURE	SYNONYME	OSYN(STRUCTURE,SYNONYME, structure)	- à "structure" - de "structure" à tous ses synonymes
R4 être extenxion de	EXTENSION	REF.	OEXT(EXTENSION,REF,extension)	- à "extension" - de "extension" à tous les produits de cette extension
R5 posséder	CHERCHEUR	REF.	OCHE(CHERCHEUR,REF,chercheur)	- à "chercheur" - de "chercheur" à tous les produits qu'il possède
R6 utiliser	PROJET	REF.	OPRO(PROJET,REF,projet)	- à "projet" - de "projet" à tous les produits utilisés
R7 être utilisé dans	REF.	DOMAINE	OREF1(REF,DOMAINE,np)	- au produit de n° np - du produit à tous ses domaines d'application
R8 avoir pour nom	REF.	STRUCTURE	OREF2(REF,STRUCTURE,np)	- au produit de n° np - du produit à son nom

R9 avoir pour extension	REF.	EXTENSION	OREF3(REF,EXTENSION,np)	- au produit de n° np - du produit à son extension
R10 avoir pour impuretés	REF.	IMPURETE	OREF4(REF,IMPURETE,np)	- au produit de n° np - du produit à toutes ses impuretés
R11 avoir pour stock	REF.	DISPONIBILITE	OREF5(REF,DISPONIBILITE,np)	- au produit de n° np - du produit à tous ses stocks
R12 être possédé par	REF.	CHERCHEUR	OREF6(REF,CHERCHEUR,np)	- au produit de n° np - du produit au chercheur qui le possède
R13 être utilisé par	REF.	PROJET	OREF7(REF,PROJET,np)	- au produit de n° np - du produit au projet qui l' utilise
R14 être synonyme de	SYNONYME	STRUCTURE	OSSY(SYNONYME,STRUCTURE, synonyme)	- au "synonyme" - au nom dont il est le synonyme
II) Relations simples				
R15 précéder lexico.	REF.		OLREF(REF)	- ordre lexico. sur les numéros
R16 précéder lexico.	DOMAINE		OLDOM(DOMAINE)	- ordre lexico. sur les domaines
R17 précéder lexico.	STRUCTURE		OLSTR(STRUCTURE)	- ordre lexico. sur les noms

R18 précéder lexico.	CHERCHEUR		OLCHE(CHERCHEUR)	- ordre lexico. sur les noms
R19 précéder lexico.	PROJET		OLPRO(PROJET)	- ordre lexico. sur les noms
R20 exister en d'autres versions	REF.		OVREF(REF)	- au produit de n° np - du produit à toutes ses versions
III) Relations unaires				
R21 être de la marque	REF.	MAR	OMAR(np)	- au produit de n° np
R22 stock positif ou nul	DISPONIBILITE	QSTO	OSTO(ns)	- au stock de n° ns

tableau 8

2. Implémentation des chemins d'accès

Les chemins d'accès par itinéraire associés aux relations simples sont des chaînes triées.

Les chemins d'accès par itinéraire associés aux relations binaires sont implémentés sous forme de listes inversées rendant ainsi la composition d'opérateurs plus aisée. Cas particuliers:

- la relation R10 est créée au moyen d'une liste inversée car, même si elle n'est pas impliquée dans une recherche d'intersection, elle est non-disjointe: de nombreux composés ont les mêmes impuretés;
- la relation R11 est disjointe et n'intervient pas dans une recherche d'intersection: elle est implémentée sous forme de liste simple;
- les relations R8, R9, R12 et R13 sont biunivoques: elles sont implémentées par des pointeurs.

Les chemins d'accès directs sont implémentés par "adresses calculées".

GESTION DES PRODUITS COMMERCIAUX: diagramme d'accès

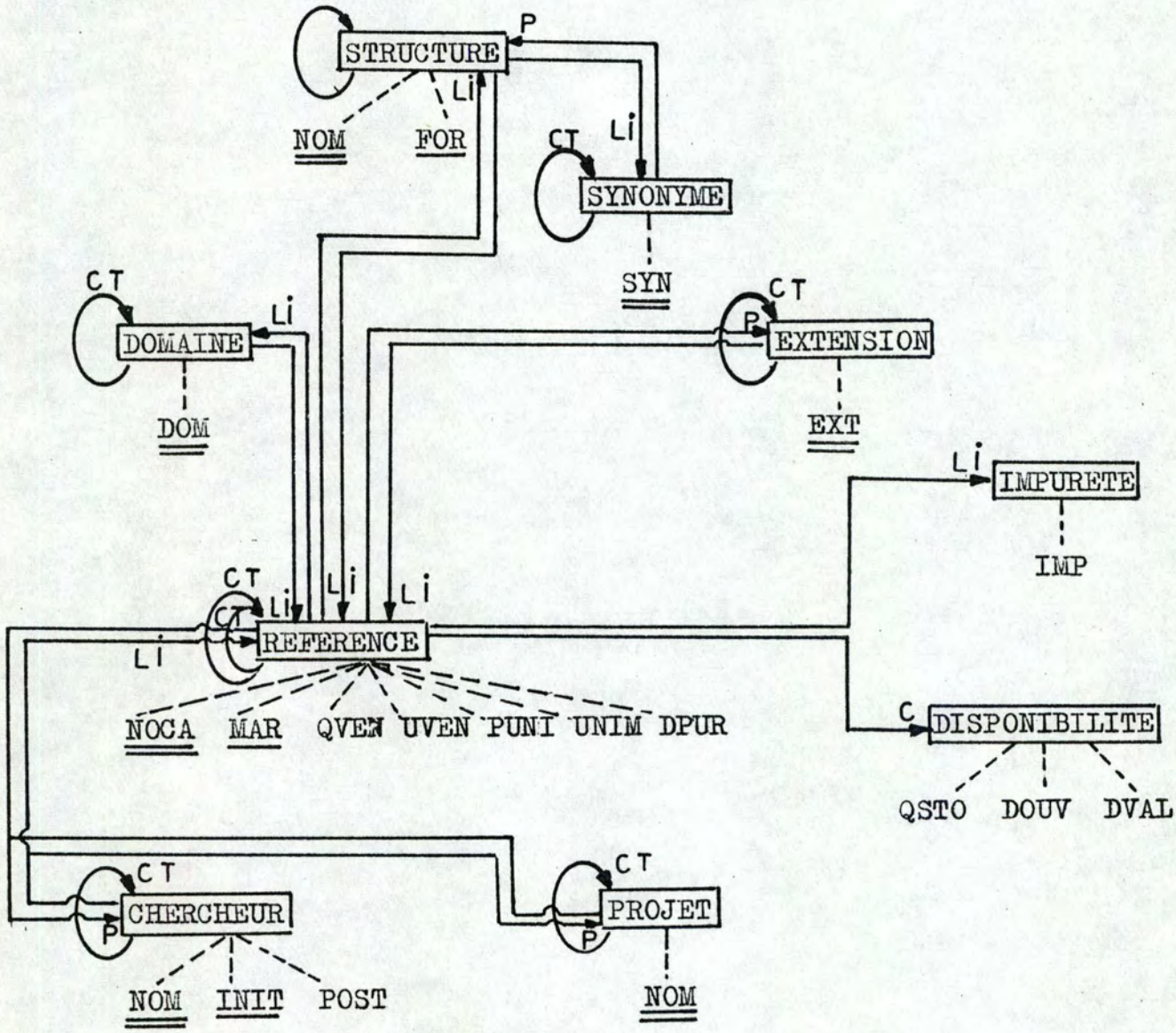


figure 89

LI: liste inversée

CT: chaîne triée

P : pointeur

NOTE: les types d'entités soulignés sont les types d'indicatifs

GESTION DES PRODUITS COMMERCIAUX: enchaînement des modules

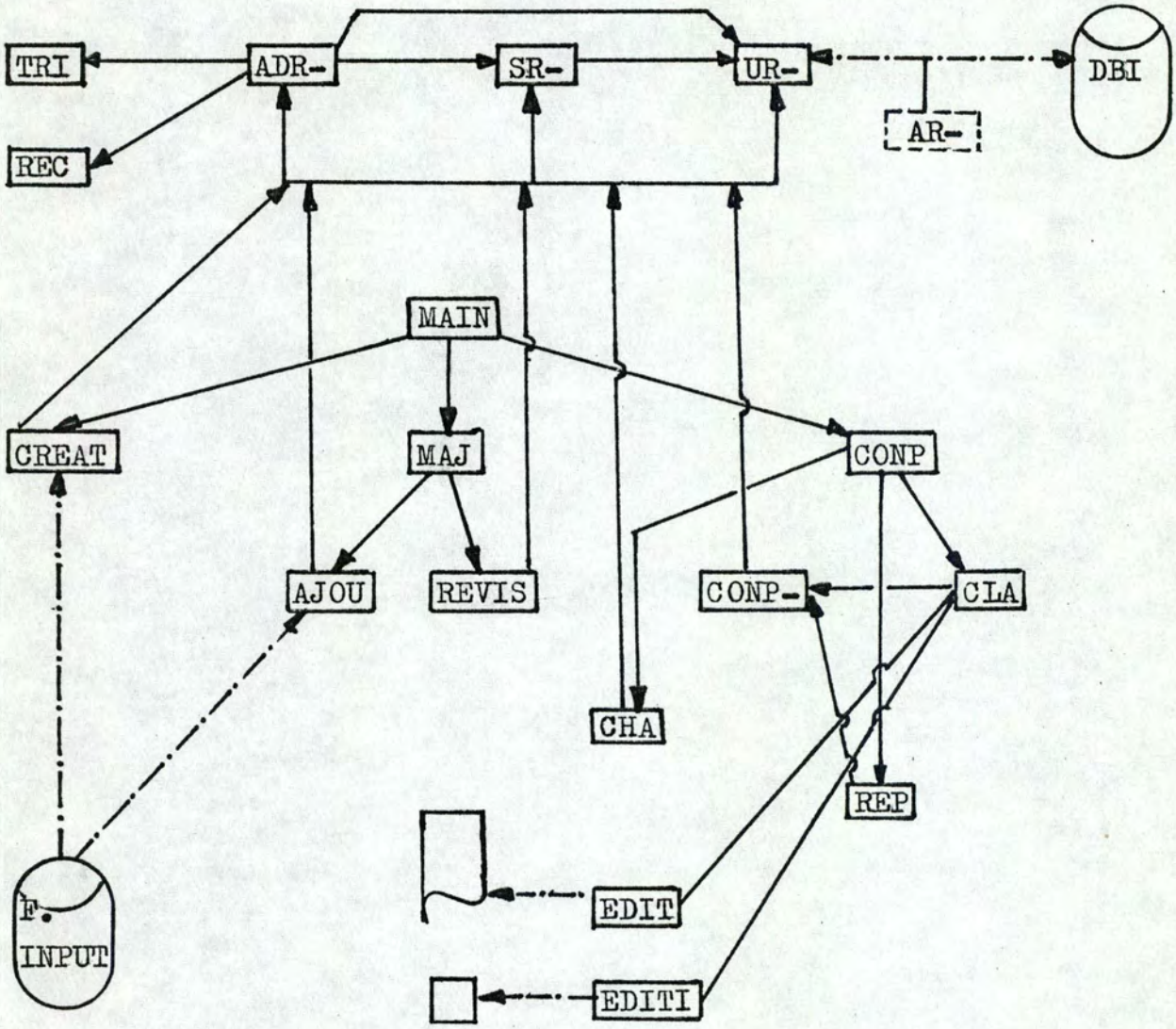


figure 90

modules

appel de module

fichiers

transfert de données

zone de communication entre modules

série de modules

nom du module	fonction du module	
	données en entrée	données en sortie et fonction
UR-	adresse d'une entité	items fixes de cette entité
SR-	adresse d'une entité	items fixes de cette entité recomposition des items (de longueur variable)
ADR-	indicatif d'une entité	adresse de l'entité
TRI	numéro de chaîne triée indicatif	insertion de l'entité (correspondant à l'indicatif) dans la chaîne triée
CREAT	fichier input	création de la BDI
MAJ	-	coordination des modules de mises à jour
CONP	-	coordination des modules de consultation
AJOU	fichier input	ajouter des produits dans la BDI
REVIS	n° de stock n° de produit	- mise à jour d'un stock - changement de propriétaire (chercheur et/ou projet) d'un produit
CONP-	-	ensemble des opérateurs
CLA	n° de chaîne triée	enchaînement des opérateurs en vue d'un classement des produits suivant un ordre défini
REP	-	enchaînement des opérateurs en vue de répondre à la question posée

EDIT	n° de produit	édition sur imprimante des renseignements sur un produit
EDITI	n° de produit	édition sur terminal des renseignements sur un produit
CHA	n° d'une chaîne triée	parcours d'une de ces chaînes
REC	indicatif	calcul de l'adresse de l'entité
SQUES	-	module d'interrogation
LIST	-	gestion des listes

tableau 9

CHAPITRE IV : INTEGRATION

L'utilisateur doit pouvoir accéder aux deux bases informations et topologiques conjointement afin d'obtenir des réponses à des questions de nature mixte du type:

- quels sont les composés comportant telle sous-structure et quelles en sont les références bibliographiques ?
- sous quelle(s) forme(s) existe le composé référencé dans tel article ?
- quel est le dessin (sur écran graphique) de la structure de tel composé ?

Des questions de nature triple sont aussi à envisager:

- quelles sont les références bibliographiques et les produits disponibles concernant telle structure dessinée sur l'écran ?

L'entité STRUCTURE étant commune à toutes les BD, il est possible de ne la représenter qu'une seule fois (figure 91). Cette façon de procéder entraîne une complication due au grand nombre d'informations qu'il faudra pouvoir accéder. Pour répondre à une question de type mixte, deux BD doivent être présentes: soit BDT et BDI1 (système bibliographique), soit BDT et BDT2 (le système de gestion des produits). Les réponses à des questions triples nécessitent l'accès aux trois BD.

Puisque le nombre d'armoires à disques est généralement limité et donc le nombre d'informations accessibles en même temps insuffisant sur un mini-ordinateur, il faudra accéder séparément aux BD. L'interface est facilement réalisable: chaque structure topologique est répérée par son numéro de composé (NC), chaque réaction par son numéro de réaction (NR). Le numéro de bibliographie (NB) caractérise un article et un numéro de produit (NP) un produit commercial (figure 92).

Une présélection est établie dans une des BD; dans l'autre une recherche est effectuée à partir du fichier réponse: les numéros de sortie servent d'entrée dans cette BD. La question "quels sont les composés comportant telle sous-structure et quelles en

sont les références bibliographiques ?" comporte deux sélections:

(1) recherche dans la BDT des composés comportant la sous-structure spécifiée; les réponses sont des NB;

(2) recherche dans la BDT1 des références concernant les NB du fichier réponse.

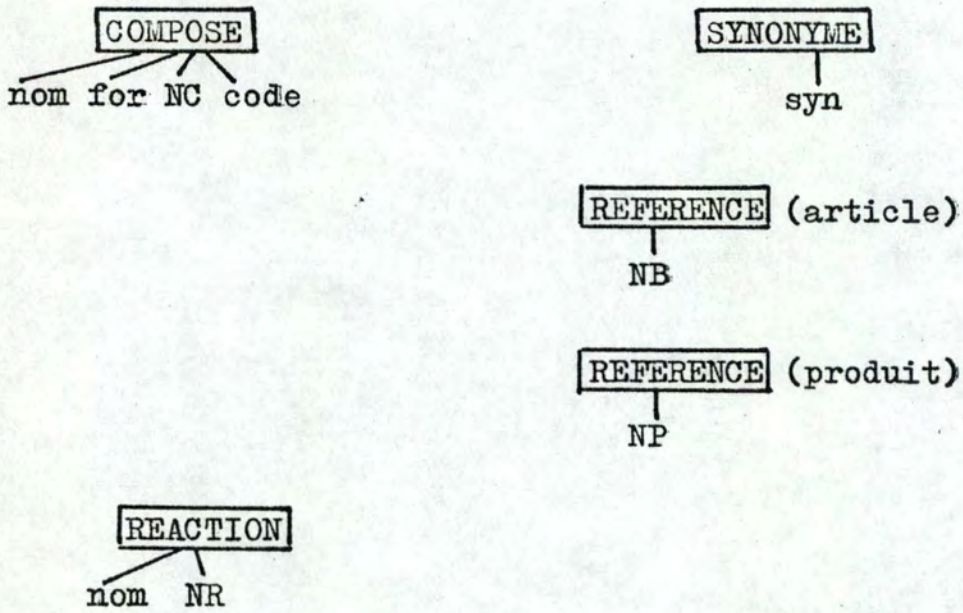


figure 91

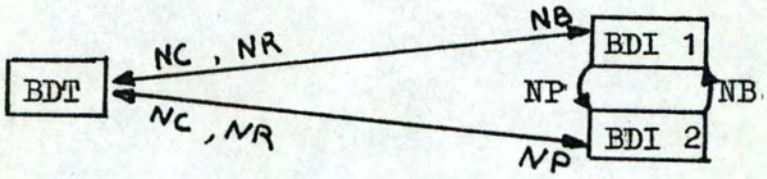


figure 92

QUATRIEME PARTIE : SYNTHESE ASSISTEE

INTRODUCTION

Tous les outils de représentation d'entités en synthèse assistée ont été étudiés dans les chapîtres précédents. Il reste à définir des outils facilitant l'énumération automatique de toutes les voies d'accès possibles et plausibles au produit à synthétiser.

Nous présentons d'abord brièvement les "heuristiques" de Corey qui fut le premier à définir un ensemble de règles et de stratégies permettant de choisir les chemins de synthèse optimaux. Rappellons qu'une heuristique est un procédé empirique, une stratégie, un tour qui améliore l'efficacité d'un système qui tente de découvrir les solutions d'un problème complexe. Un programme heuristique est un programme automatique qui utilise ces heuristiques.

CHAPITRE I : HEURISTIQUES DE COREY

En chimie organique (B⟶?), on essaye
- de former des liaisons stratégiques,
- d'ajouter des groupes fonctionnels très sensibles aux acides, bases le plus tard possible,
- de créer le plus de fonctionalités possibles,
- d'insérer des groupes d'interférences après les autres étapes (exemple: création de liaisons stratégiques),
- de favoriser la rupture de ponts chimiques,
- d'utiliser des réactions puissantes et utiles.

Corey se base sur ces caractéristiques pour définir un ensemble de règles de stratégies:

1. Liaisons stratégiques C--C

Une liaison stratégique C--C doit satisfaire aux règles suivantes:
(1) appartenir à un cycle de taille 5, 6 ou 7,
(2) ne pas appartenir à un cycle de taille inférieure à 3,
(3) ne pas appartenir à l'intersection de deux cycles dont l'enveloppe possède plus de sept atomes,
(4) ne pas appartenir à un stéréocentre.

L'heuristique accompagnant ce concept est de briser ces liaisons le plus vite possible: il s'agit, en effet, de simplifier la structure par des ruptures de liaisons qui conduiront à un précurseur plus accessible et plus simple et qui minimiseront le nombre de cycles pontés et de cycles de tailles moyennes ou grandes.

2. Entités complexes

Toujours dans un but de simplification, on tente de supprimer le plus de groupes fonctionnels, de centres stéréochimiques, de groupes d'interférences possibles. Par exemple, si la rupture d'une liaison stratégique interfère avec une sous-structure, on applique

une réaction qui supprimera cette sous-structure.

3. Choix des réactions

Les synthèses les plus puissantes sont celles qui utilisent des réactions puissantes: les réactions Diels-Alder, Robinson annulation, Birch reduction, cation-olefin cyclization.

4. Utilisation d'"annexes"

Une annexe de cycle est un groupe d'atomes attachés au cycle par une liaison n'appartenant pas au cycle lui-même.

Un atome peut définir une annexe si:
(1) la liaison entre l'atome et le cycle n'appartient pas au cycle,
(2) soit l'atome est un atome de carbone, soit l'atome est un hétéroatome avec au moins deux atomes de carbone.
NOTE: les annexes de cycles à trois liaisons ne sont pas pris en considération.

Une branche annexe doit contenir au moins trois atomes de carbone et doit prendre son origine soit en un atome n'appartenant pas au cycle et qui a trois ou plus de liaisons vers des atomes autres qu'hydrogène, soit en des liaisons doubles ou triples non terminales.

On choisira l'annexe de telle façon que sa rupture entraîne une simplification significative de la molécule cible et réduise le nombre de branchements dans la structure. Différentes stratégies sont possibles:
(1) stratégie RA-RA (annexe de cycle vers annexe de cycle):
deux annexes appartenant au même cycle ou à des cycles différents sont reliés l'une à l'autre;
(2) stratégie RA-R (annexe de cycle vers cycle):
un atome sur une annexe est relié à un cycle ou à un atome fonctionnel directement rattaché à un cycle;
(3) reconnection acyclique:
cette stratégie connecte au moins deux stéréocentres en un

un cycle;

5. Interconversion de la fonctionalité

Une interconversion de fonctionalité (FGI) est la transformation, dans la molécule, d'un groupe fonctionnel en un autre (figure 93). Le but de la stratégie est de produire une séquence (dont la profondeur est au maximun 4) de FGI qui conduira à une simplification de la molécule par des ruptures de liaisons stratégiques ou la formation de cycles.

Dans la séquence de FGI de la figure 94 A, B, C, D, E sont des groupes fonctionnels, A est le groupe cible ou sujet, B le groupe but ou objet; B, C, D sont les groupes intermédiaires et P1, P2, P3, P4 les transformations.

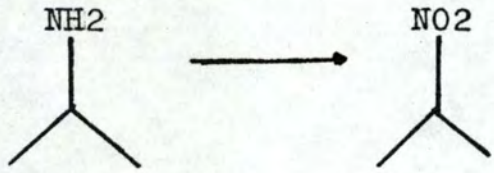


figure 93

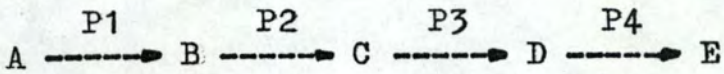


figure 94

CHAPITRE II : GROUPEMENT DES INFORMATIONS

De nombreux chercheurs ont proposés d'autres heuristiques (références 34 à 52) mais toutes se basent sur les informations contenues dans la table de connexions de la molécule à synthétiser.

Dans ce chapître, nous essayons de grouper ces informations de telle façon qu'elles soient plus facilement manipulables par ces stratégies.

1. Table de connexions

Initialement, les informations concernant la molécule cible sont contenues dans deux tables: celle des atomes et celle des liaisons (figure 95).

1.1. Table des atomes

1	C	1	3	3	2 6,7	1,6,7
2	C	1	3	3	1,3,12	1,2,13
3	C	1	3	3	2,4,9	2,3,10
4	C	1	4	4	3,5,8,13	3,4,9,14
5	C	0	2	2	4,6	4,5
6	C	1	4	4	5,1,10,11	5,6,11,12
7	C	0	2	2	1,8	7,8
8	C	2	3	3	7,4,14	8,9,15
9	N	0	1	1	3	10
10	O	0	1	1	6	11
11	C	0	1	1	6	12
12	C	0	1	1	2	13
13	H	0	1	1	4	14
14	O	0	1	1	8	15

colonne 1: numérotation des atomes
colonne 2: couleur des atomes
colonne 3: stéréocentre (0 = non-stéréo, 1 = stéréo, 2 = non spécifié
colonne 4: nombre d'électrons de valence

colonne 5: nombre d'atomes adjacents
colonne 6: numéro des atomes adjacents
colonne 7: numéro des liaisons adjacentes

1.2. Table des liaisons

1	1	0	1	2
2	1	0	2	3
3	1	0	3	4
4	1	0	4	5
5	1	0	5	6
6	1	0	6	1
7	1	0	1	7
8	1	0	7	8
9	1	0	8	4
10	1	1	3	9
11	1	6	6	10
12	1	1	6	11
13	1	0	2	12
14	1	1	4	13
15	1	4	8	14

colonne 1: numérotation des liaisons
colonne 2: couleur des liaisons
colonne 3: (0 = non-stéréo, 1 = liaison vers l'avant, 6 = liaison vers l'arrière, 4 = position indéfinie)
colonne 4: n° du premier atome
colonne 5: n° du deuxième atome

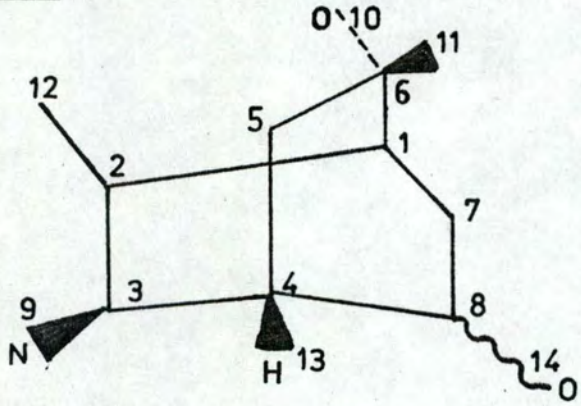


figure 95

2. Groupement d'informations en "set"

Certaines informations contenues dans les tables sont groupées en "set". Un set est une chaîne de bits référencée par le nom d'une propriété, dans laquelle le $i^{ème}$ bit est 1 si la "$i^{ème}$ chose" possède cette propriété.

2.1. SETs d'atomes

A1- atomes partageant une liaison simple
A2- atomes partageant une liaison double
A3- atomes partageant une liaison triple
A4- atomes C
A5- atomes H
A6- atomes O
A7- atomes P
A8- atomes S
A9- atomes N

A10- atomes portant 1 atome non-hydrogène
A11- atomes portant 2 atomes non-hydrogène
A12- atomes portant 3 atomes non-hydrogène
A13- atomes portant 4 atomes non-hydrogène
A14- atomes ne portant aucune charge
A15- atomes portant une charge +
A16- atomes portant des électrons non appareillés
A17- atomes portant une charge négative
A18- atomes N, O, S, P (hétéroatome)
A19- atomes portant un atome H explicite ou implicite
A20- atomes portant un atome H explicite
A21- atomes terminaux
A22- atomes têtes de pont
A23- stéréocentres
A24- atomes origines d'un groupe fonctionnel
A25- atomes appartenant à un cycle aromatique

A26(i)- atomes appartenant au cycle i
A27(a)- atomes adjacents à l'atome a

A28(b)- atomes adjacents à la liaison b
A29(a,n)- atomes n fois plus éloignés de l'atome a
A30(a1,a2,...,ai)- atomes adjacents à un ensemble d'atomes a1,...,ai
A31(b1,b2,...,bj)- atomes adjacents à un ensemble de liaisons b1,...,bj

2.2. SETs de liaisons

L1- liaisons simples
L2- liaisons doubles
L3- liaisons triples
L4- liaisons vers un hétéroatome
L5- liaisons vers un atome H
L6- liaisons multiples conjuguées
L7- liaisons appartenant à un cycle
L8- liaisons doubles appartenant à un stéréocentre
L9- liaisons appartenant à un cycle aromatique

L10(i)- liaisons appartenant au cycle i
L11(a)- liaisons adjacentes à l'atome a
L12(b)- liaisons adjacentes à la liaison b
L13(b,n)- liaisons n fois plus éloignées de b
L14(b1,b2,...,bj)- liaisons adjacentes à l'ensemble de liaisons b1,...,bj
L15(a1,a2,...,ai)- liaisons adjacentes à l'ensemble d'atomes a1,...,ai

2.3. Opérations sur les SETs

L'avantage des SETs est qu'ils sont facilement manipulables par des opérations et fonctions logiques.

2.3.1. Opérations logiques

(1) OR(S1,S2) : OU inclusif entre les deux ensembles S1 et S2, c'est-à-dire $S1 \cup S2 = \{s \text{ tq } s \in S1 \text{ ou } s \in S2\}$;

(2) AND(S1,S2) : intersection des deux ensembles S1 et S2, c'est-à-dire $S1 \cap S2 = \{s \text{ tq } s \in S1 \text{ et } s \in S2\}$;

(3) XOR(S1,S2) : OU exclusif entre les deux ensembles S1 et S2, c'est-à-dire $S1 \setminus S2 = \{ s \mid s \in S1,\ s \in S2 \}$;

(4) NOT(S1) : complément de l'ensemble S1, c'est-à-dire $\complement S1 = \{ s \mid s \in U,\ s \notin S1 \}$.

2.3.2. Fonctions logiques

NOTE: deux ensembles sont équivalents $(S1 \equiv S2)$ si pour tout i, $v_1(i) = v_2(i)$ où v = 1 ou 0 .

```
(1) EQV(S1,S2)   : if S1 ≡ S2 then EQV = .TRUE. else EQV = .FALSE.
(2) MEM(S,a)     : if atome a ∈ S then MEM = .TRUE.
                                  else MEM = .FALSE.
(3) ON(S,a)      : if atome a ∉ S then begin
                                           ON = .TRUE.
                                           ajouter a dans S
                                       end
                                  else ON = .FALSE.
(4) OFF(S,a)     : if atome a ∈ S then begin
                                           OFF = .TRUE.
                                           enlever a de S
                                       end
                                  else OFF = .FALSE.
(5) SUB(S1,S2)   : if S1 ⊂ S2 then SUB = .TRUE. else SUB = .FALSE.
```

2.3.3. Fonctions integer

(1) COUNT(S) : charger le nombre d'éléments dans S dans la variable COUNT

(2) NEXT(S,a) : mettre l'élément suivant a dans la variable NEXT

3. Groupement d'informations en liste

Une liste permet le stockage d'informations en nombre variable et plus complexes que des appartenances à une classe.

(1) <u>liste des cycles</u> (figure 96)

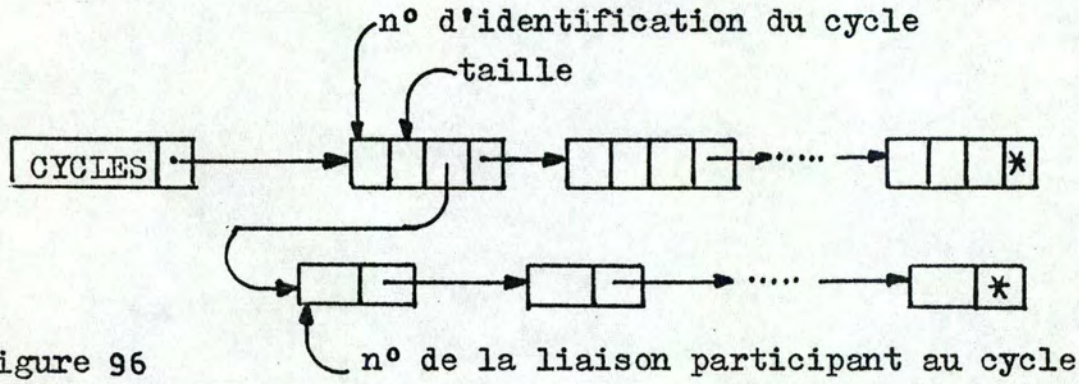


figure 96

(2) <u>liste des groupes fonctionnels</u> (figure 97)

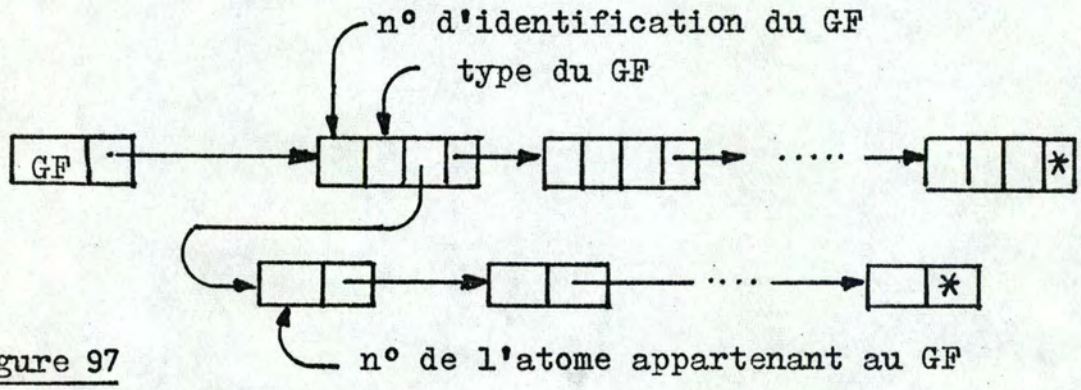


figure 97

<u>NOTES</u>:

(1) la reconnaissance des groupes fonctionnels peut se faire par l'une des méthodes vues précédemment;

(2) un algorithme de recherche de cycles sera décrit dans le chapître suivant.

CHAPITRE III : RECHERCHE DES CYCLES

La recherche des cycles dans une structure présente plusieurs difficultés. Si des structures tri-dimensionnelles sont manipulées, il est nécessaire de les projeter dans un plan. La projection peut varier suivant le plan choisi (figure 98): la projection A fait apparaître un cycle de taille 5 contenant deux cycles de taille 5 et 6; la projection B présente deux cycles 5 dans un cycle 6. Afin d'éliminer l'arbitraire de la projection choisie, il est nécessaire de considérer des cycles d'importance égale (cycles fondamentaux).

La deuxième difficulté provient des conventions utilisées dans les systèmes de documentation pour reconnaître un cycle et qui ne correspondent pas toujours aux vues du chimiste. Supposons la question: "Quelles sont les structures contenant un N et un S sur un cycle de taille 6 ?" (figure 99). Si les conventions sont telles que seules les structures A et D sont retenues, soit le chercheur sera orienté vers des structures qui ne l'intéressent pas, soit certaines structures seront oubliées.

figure 98

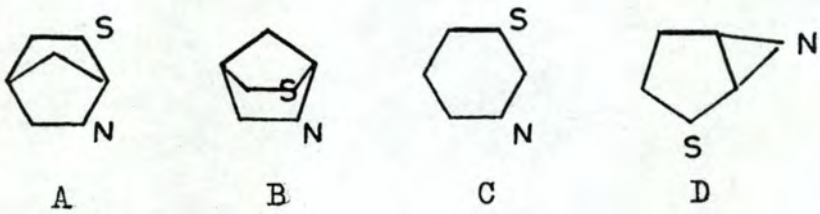


figure 99

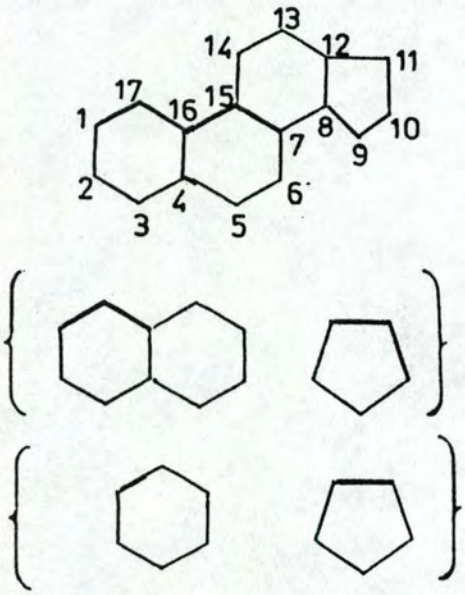
1. Définitions

- un cycle simple et élémentaire est un cycle qui ne passe qu'une et une seule fois par chacun de ses sommets et chacune de ses arêtes utilisés.

Les cycles simples et élémentaires du composé de la figure 100 sont:
- 1,2,3,4,16,17,1 (taille 6),
- 16,4,5,6,7,15,16 (taille 6),
- 15,7,8,12,13,14,15 (taille 6),
- 12,8,9,10,11,12 (taille 5),
- 1,2,3,4,5,6,7,15,16,17 (taille 11),
- 1,2,3,4,5,6,7,8,12,13,14,15,16,17,1 (taille 16),
- etc.

Il est évident que cette définition ne correspond pas au point de vue du chimiste, pour qui il existe trois cycles 6 et un cycle 5.

- une union disjointe de cycles est l'ensemble des liaisons appartenant à deux ou plusieurs cycles disjoints deux à deux. Deux cycles sont disjoints s'ils ne sont pas fusionnés et s'ils n'ont pas de liaisons communes (figure 100).



n'est pas une union disjointe de cycles

est une union disjointe de cycles

figure 100

- <u>somme de liaisons</u>: $S1 \oplus S2$. Soient deux ensembles de liaisons S1 et S2; $S1 \oplus S2$ (ou exclusif) est l'ensemble des liaisons appartenant à S1 ou S2, mais non à S1 et S2.

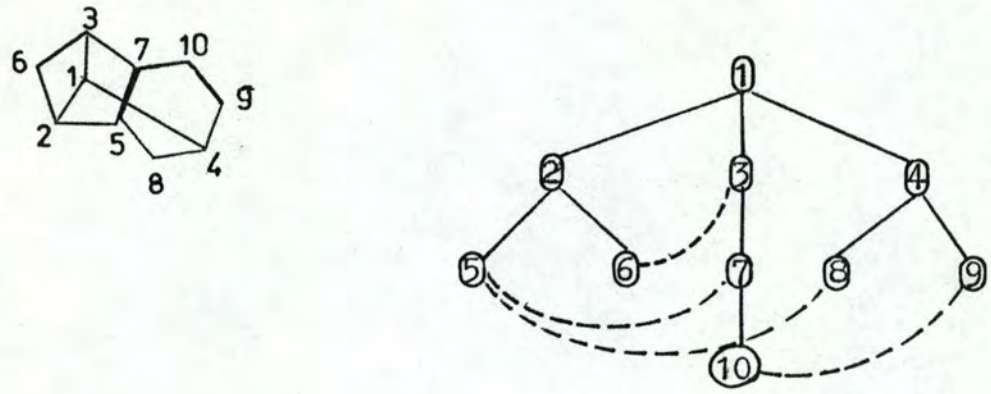
L'ensemble X de tous les cycles et de toutes les unions disjointes de cycles forme un groupe muni de la loi interne $\oplus$:
- la loi $\oplus$ est associative: $(R1 \oplus R2) \oplus R3 = R1 \oplus (R2 \oplus R3)$,
- le cycle de taille 0 est l'élément neutre,
- tout élément est symétrisable: $R1 \oplus R2 = R2 \oplus R1$.

Chaque cycle ou union disjointe de cycles forme un vecteur dans cet espace.

- <u>cycle fondamental</u>: soient n_a le nombre d'atomes , n_l le nombre de liaisons dans une structure G et T un arbre maximal de G; alors $n_a - 1$ est le nombre de liaisons dans T et $n_l - n_a + 1$ est le nombre de cordes dans G. Soit, de plus, une arête joignant a_i et a_j: T contient a_i et a_j (par définition d'un arbre maximal), et un chemin entre a_i et a_j (par définition d'un graphe simplement connexe) qui est unique (T est acyclique). Si on ajoute une corde à T, on forme un cycle appele cycle fondamental. Il existe une correspondance un à un entre les $n_l - n_a + 1$ cordes et l'ensemble des cycles fondamentaux de G.

Soient $R = (b_1, b_2, \ldots, b_i, b_{i+1}, b_{i+2}, \ldots, b_j)$ un cycle, où $b_1, b_2, \ldots, b_i$ sont des cordes et $b_{i+1}, \ldots, b_j$ des liaisons de T, et R_k le cycle fondamental correspondant à la corde b_k; formons $R^1 = R_1 \oplus R_2 \ldots \oplus R_i$: R et R^1 contiennent $b_1, b_2, \ldots, b_i$.
De ce qui précède, deux propositions sont déductibles:
(1) $R \oplus R^1$ contient uniquement des liaisons de T;
(2) $R \oplus R^1$ est un cycle ou une union disjointe de cycles.
Ces deux propositions sont contradictoires: il faut donc que $R \oplus R^1 = \emptyset$, ou que $R = R^1$. Tout cycle peut donc être exprimé comme une combinaison linéaire des cycles fondamentaux (figure 101).

R_k étant le seul cycle fondamental contenant b_k et R_k ne pouvant pas être exprimé comme $\oplus$ d'autres cycles fondamentaux, ces derniers sont donc linéairement indépendants et forment une base



- l'ensemble des cordes est: ((6,3), (5,7), (5,8), (10,9))

- R1 = ((6,2), (2,1), (1,3), (3,6))
  R2 = ((5,2), (2,1), (1,3), (3,7), (7,5))
  R3 = ((5,2), (2,1), (1,4), (4,8), (8,5))
  R4 = ((10,7), (7,3), (3,1), (1,4), (4,9), (9,10))

- R = ((6,3), (3,7), (7,5), (5,2), (2,6))

- R^1 = R1 $\oplus$ R2 = ((6,2), (5,2), (3,6), (3,7), (7,5))

- donc R^1 = R

figure 101

pour l'espace des cycles, qui est de dimension $n_l - n_a + 1$.

- un cycle minimal: pour décrire complètement un réseau de cycle, il est nécessaire de considérer des cycles autres que ceux contenus dans la base. Le nouvel ensemble A formé doit être tel que tout cycle qui n'y appartient pas puisse être généré par une $\oplus$ de cycles de l'ensemble (il est à noter que les cycles de A ne sont pas nécessairement linéairement indépendants). L'ensemble A s'appelle ensemble de cycles minimaux. Un cycle minimal est un cycle qui ne peut pas être exprimé comme une $\oplus$ de cycles plus petits. Si pour une liaison b donnée, il n'existe pas de cycle plus petit que le cycle R contenant b, alors R est un cycle minimal, car il ne peut pas être exprimé comme $\oplus$ de cycles plus

petits. L'ensemble A contient la base de l'espace des cycles.

2. Algorithme pour trouver l'ensemble des cycles minimaux

première étape: génération d'un ensemble A de cycles fondamentaux R_i en construisant un arbre maximal;

(1) choix de la racine de l'arbre maximal: l'atome de numéro 1;
(2) construire l'arbre en parcourant pour chaque sommet ceux qui lui sont adjacents (par ordre croissant); si, en empruntant une liaison, on aboutit à un sommet déjà noté , cette liaison est une corde et un cycle fondamental est formé.

deuxième étape: génération de l'ensemble des cycles fondamentaux. Supposons un ensemble de cycles S_r, qui contient l'ensemble des cycles minimaux. Soit S_{ms} un ensemble vide: si on y inclue les cycles les plus petits de S_r qui sont linéairement indépendants avec les cycles de S_{ms}, alors S_{ms} contient l'ensemble des cycles minimaux.

Soit $R = (b_1, b_2, \ldots, b_i, b_{i+1}, \ldots, b_j)$ un cycle où $b_1, \ldots, b_i$ sont des cordes; supposons R plus grand que chacun des R_k (pour tout $k = 1$ à i, R_k est le cycle fondamental associé à la corde b_k). R n'est pas un cycle minimal, car il peut être exprimé comme combinaison linéaire de cycles plus petits. Si, pour toute corde b_l, on rassemble les cycles qui ne sont pas plus grands que R_l, on formera l'ensemble des cycles minimaux.

(1) Etant donné un ensemble de cycles fondamentaux R_i ($i = 1$ à $E-N+1$), on forme $R_{ij} = R_i \oplus R_j$ ($i < j \leqslant E-N+1$);
(2) on ordonne l'ensemble R_i R_{ij} en utilisant la hiérarchie suivante:
   - taille minimale
   - somme arithmétique des numéros maximale
   - séquence des numéros lexicographiquement inférieur;
(3) à partir de cet ensemble ordonné, on sort les $(E-N+1)$ premiers éléments linéairement indépendants.

Recherchons les cycles minimaux de la structure de la figure 101:

- <u>cycles fondamentaux</u>: R1 = 6-2-1-3
  R2 = 5-2-1-3-7
  R3 = 5-2-1-4-8
  R4 = 10-7-3-1-4-9

- <u>les combinaisons</u>: R12 = R1 $\oplus$ R2 = 5-2-6-3-7
  R13 = R1 $\oplus$ R3 = 5-2-6-3-1-4-8
  R14 = R1 $\oplus$ R4 = 10-7-3-6-2-1-4-9
  R23 = R2 $\oplus$ R3 = 7-3-1-4-8-5
  R24 = R2 $\oplus$ R4 = 9-4-1-2-5-7-10
  R34 = R3 $\oplus$ R4 = 10-7-3-1-2-5-8-4-9

- les cycles pris en compte suivant les règles hiérarchiques:
  R1, R12, R3, R2;

- R2 étant combinaison linéaire de R1 et R12, on le remplace par le cycle R4 sélectionné d'après les règles. L'ensemble des cycles minimaux est donc formé par R1, R12, R3, R4 qui sont linéairement indépendants et forment encore une base de l'espace des cycles.

CHAPITRE IV : SYNTHESE ASSISTEE

1. Représentation de l'arbre de synthèse

Les données concernant un précurseur sont représentées de la façon suivante:

pointeur vers le PARENT
pointeur vers le FRERE
pointeur vers le 1er FILS
n° d'ordre de création
descripteurs de codification
pointeur vers la REACTION utilisée
taux r1 du chimiste
taux r2 du programme
changement par rapport au PARENT
coordonnées des atomes

figure 102

Ces données sont stockées en mémoire auxiliaire. Chaque fois qu'un précurseur est choisi comme molécule à synthétiser, sa table de connexions, ses SETs et listes sont générées.

Les taux r1 et r2 sont des estimations, soit calculées, soit données par le chimiste, de la probabilité de réussite de l'étape en cours.

Les changements de structure que présente un précurseur par rapport au parent sont décrits au moyen des opérations élémentaires suivantes:

(1) DELA n° atome : supprimer l'atome repéré par son numéro et les liaisons qui y aboutissent;

(2) ADDA n° atome, c : ajouter un atome de couleur c;
(3) BREB n° liaison : briser une liaison;
(4) MAKB n° at1, n° at2, c : créer une liaison de couleur c entre deux atomes;
(5) MOVC n° at1, n° at2 : déplacer la charge d'un atome vers un autre;
(6) DELC n° atome : supprimer la charge d'un atome;
(7) ADDC n° atome : ajouter une charge sur un atome.

L'intérêt de ces opérations élémentaires est qu'il n'est pas nécessaire de mémoriser la table de connexions des prédécesseurs: elle peut être retrouvée à partir de la table de la molécule cible T.

Grâce aux pointeurs PARENT, FILS, FRERE on pourra à tout moment soit sortir un chemin particulier, soit sortir tous les chemins déjà générés (figure 103).

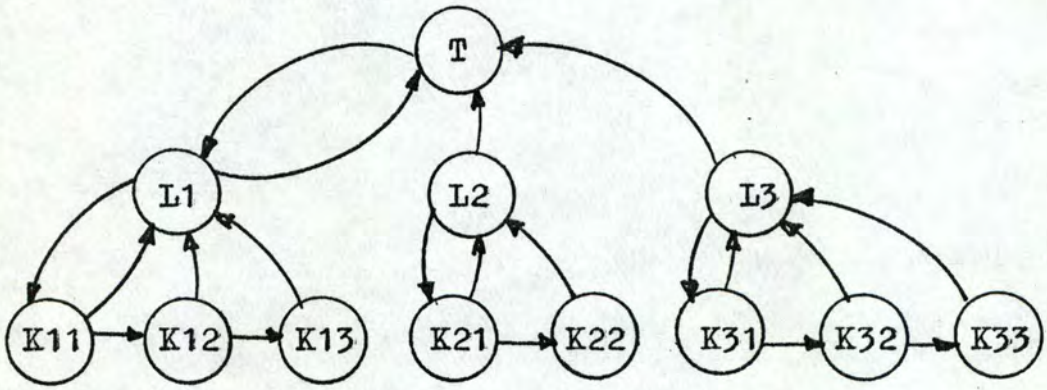


figure 103

2. Organigramme d'une session de travail en synthèse

La figure 104 donne un organigramme général des différentes phases d'une synthèse assistée. L'utilisateur développe la structure de la molécule à synthétiser, sur écran graphique. Le programme génère, ensuite, la table de connexions et perçoit, dans la mo-

lécule, les informations pertinentes pour l'analyse notamment, les groupes fonctionnels, les cycles, les SETs, etc.. Si plusieurs stratégies sont disponibles, l'utilisateur peut en choisir une en particulier sinon elles seront utilisées les unes après les autres. Les précurseurs sont générés niveau par niveau et présentés au chercheur pour évaluation. Celui-ci peut demander certaines références afin de repérer dans l'ensemble des précurseurs proposés ceux qui lui semblent les plus pertinents, les autres étant supprimés de l'arbre de synthèse. L'utilisateur peut choisir un précurseur particulier comme nouvelle molécule à synthétiser (et éventuellement une nouvelle stratégie) sinon tous les précurseurs de ce niveau seront traités les uns après les autres.

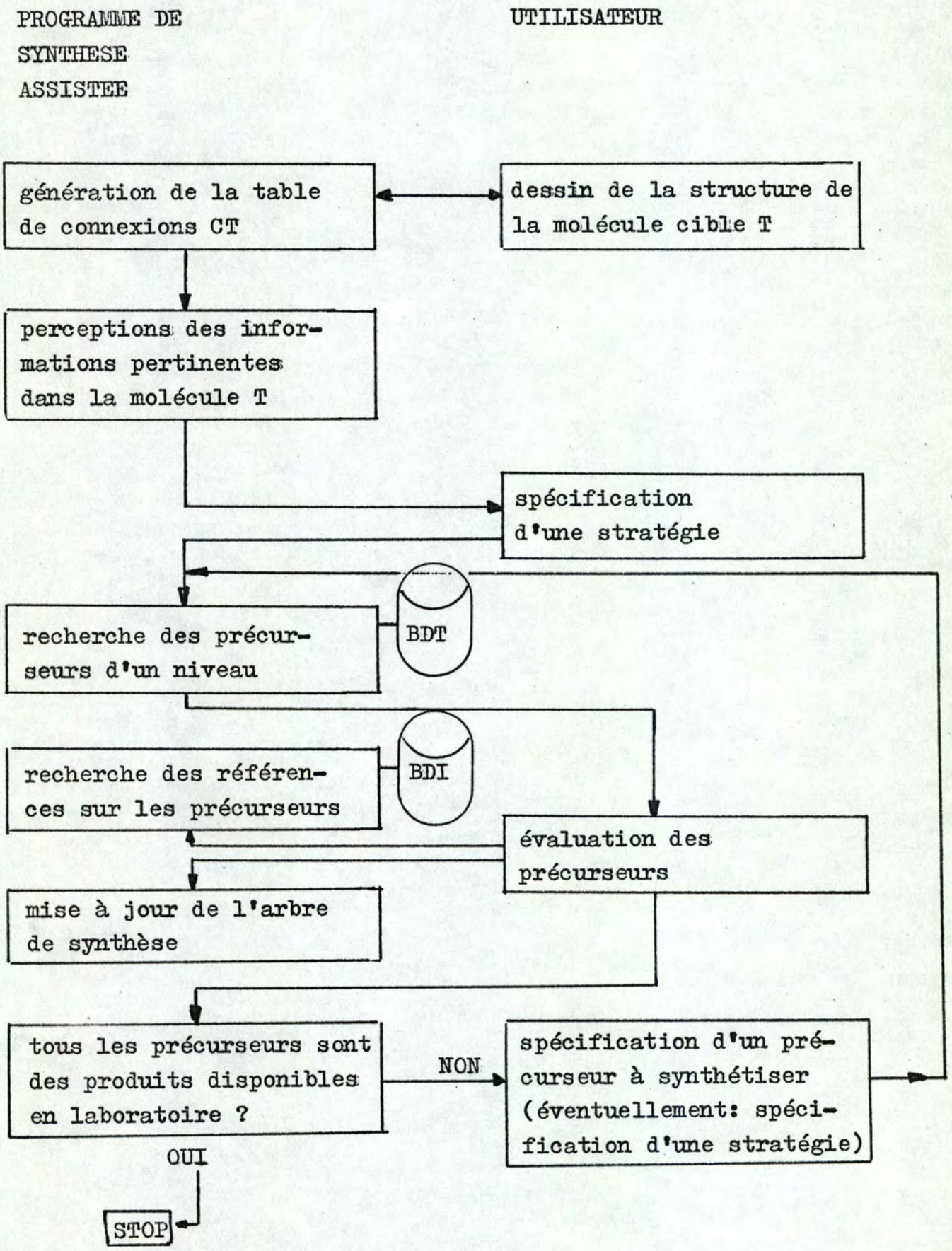

PROGRAMME DE SYNTHESE ASSISTEE
UTILISATEUR
génération de la table de connexions CT
dessin de la structure de la molécule cible T
perceptions des informations pertinentes dans la molécule T
spécification d'une stratégie
recherche des précurseurs d'un niveau
BDT
recherche des références sur les précurseurs
BDI
évaluation des précurseurs
mise à jour de l'arbre de synthèse
tous les précurseurs sont des produits disponibles en laboratoire ?
NON
spécification d'un précurseur à synthétiser (éventuellement: spécification d'une stratégie)
OUI
STOP

figure 104

CONCLUSION ET PROLONGEMENTS POSSIBLES

Actuellement, plusieurs systèmes de codification des composés et réactions chimiques sont opérationnels. Citons, par exemple, les systèmes GREMAS, CCBF (Computer Processing of Chemical and Biological Facts), IDC (Internationale Dokumentationsgesellschaft für Chemie mbH, Frankfurt).

Ces systèmes présentent tous un inconvénient majeur: la translation de la formule chimique en son code exige un exercice intellectuel dont seuls sont capables des spécialistes doués d'une connaissance approfondie des règles parfois compliquées de codage.

En considérant une structure chimique comme un graphe simple et non orienté, nous disposons d'un langage de description et d'un outil de manipulation: la théorie des graphes. La codification est plus aisée et non ambiguë; de plus, nous l'avons rendue transparente à l'utilisateur qui peut s'exprimer dans son langage naturel de visualisation au moyen d'un écran graphique.

En synthèse assitée, les outils que nous avons conçus et, pour certains d'entre eux, implémentés permettent:
- une meilleure représentation des informations manipulées,
- à partir de références bibliographiques et commerciales, une meilleure évaluation, par l'utilisateur, de l'arbre de synthèse,
- des recherches de stratégies plus rapides et plus générales grâce à l'utilisation d'une base de données topologiques.

Le champ d'application de ces outils n'est d'ailleurs pas li-

mité à la synthèse assistée: les BDI et BDT peuvent être utilisées en tant que systèmes d'information et de documentation indépendants.

A partir ce ce travail, plusieurs prolongements sont souhaités:

Tout d'abord, il nous semble intéressant d'implémenter la BDT.

Une stratégie de synthèse basée sur ces outils de représentation et de manipulation pourrait être mise au point.

La mise en place de la base orientée vers la gestion des produits devrait permettre d'évaluer les arcs d'un chemin de synthèse au niveau des coûts (prix des produits intervenant dans les réactions, coûts des expériences, etc.) et au niveau des rendements des réactions.

ANNEXE 1 : RAPPELS DE THEORIE DES GRAPHES

Un graphe G est un couple G = (X,U) constitué par:

- un ensemble X = x1, x2, ..., xn fini ou dénombrable, dont les éléments sont les sommets du graphe,
- une famille U = (u1, u2, ..., un) d'éléments dans X x X, appelés les arcs du graphe.

Intuitivement, un graphe est un schéma constitué par un ensemble de points x1, x2, ..., xn en nombre fini ou dénombrable et que l'on appelle sommets, reliés entre eux par des branches orientées u1, u2, ..., un que l'on appelle arcs (figure 105).

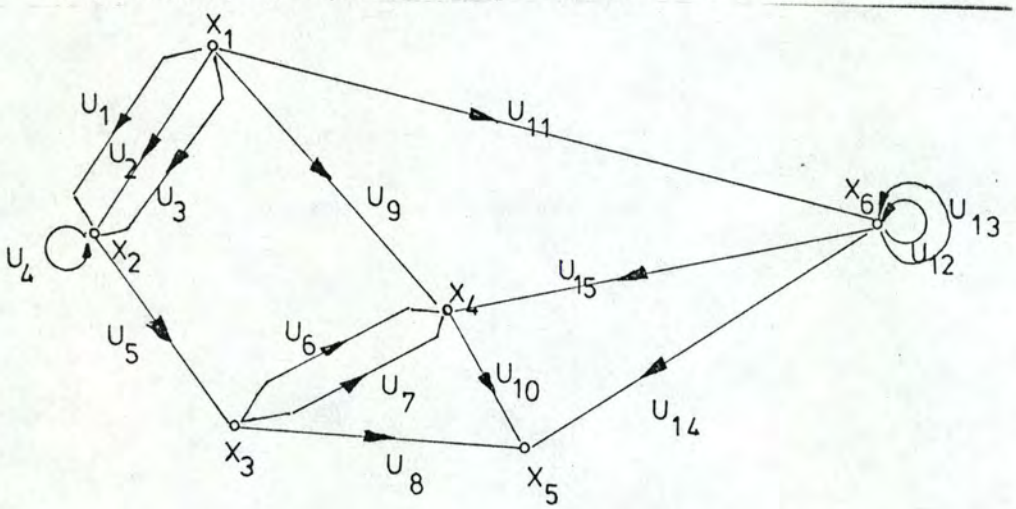


figure 105

Lorsque, pour des raisons conceptuelles, on néglige l'orientation des arcs d'un graphe, on dit que l'on obtient un multigraphe. Les branches non orientées sont alors appelées les arêtes du multigraphe. Un multigraphe G est donc un couple G = (X,$\bar{U}$) où X est l'ensemble des sommets et $\bar{U}$ la famille des arêtes (figure 106).

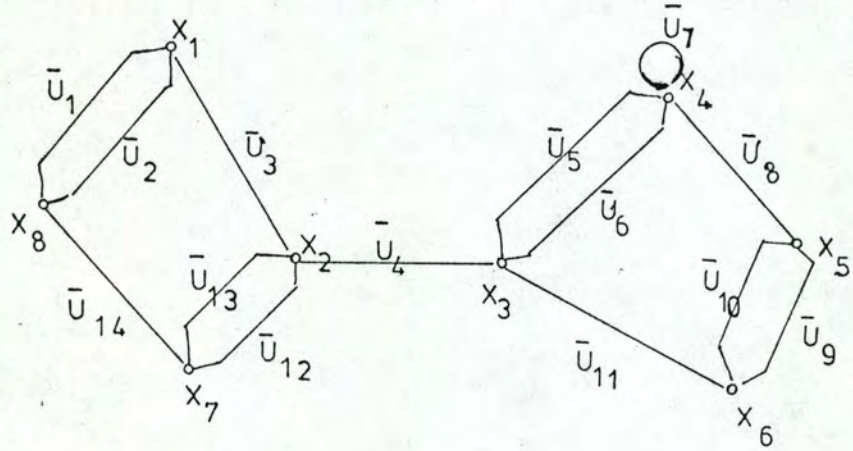


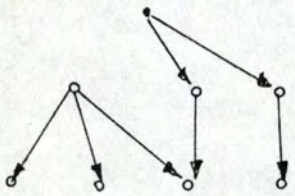
figure 106

CONCEPTS NON ORIENTES

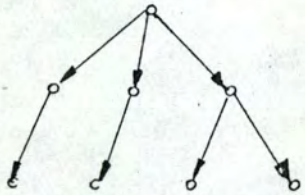
soit $G = (X,\bar{U})$ un multigraphe;

- l'ordre du multigraphe est le nombre de ses sommets;
- si $\bar{u} = (\overline{x_i,x_j})$ est une arête du multigraphe, x_i est l'extrémité initiale de $\bar{u}$ et x_j l'extrémité terminale de $\bar{u}$;
- deux sommets sont adjacents s'ils sont les extrémités d'une arête les reliant;
- deux arêtes sont adjacentes si elles ont au moins une extrémité commune;
- une arête est incidente à un sommet si ce sommet est l'une de ses extrémités;
- le degré d'un sommet est le nombre d'arêtes qui lui sont incidentes;
- une arête est incidente à un sous-ensemble A de sommets si l'une de ses extrémités (et une seule) se trouve dans A. Le degré de A est donc aussi le nombre d'arêtes incidentes à A.
- Un pseudo-chaîne est une suite $\bar{u} = (\bar{u}_1, \ldots, \bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}, \ldots, \bar{u}_q)$ d'arêtes (pas nécessairement distinctes) telle que chaque arête $\bar{u}_i$ $(1 \leq i \leq q)$ est reliée à $\bar{u}_{i-1}$ par une de ses extrémités et à $\bar{u}_{i+1}$ par l'autre;
- le nombre q d'arêtes composant la suite est la longueur de la pseudo-chaîne;
- un pseudo-cycle est une pseudo-chaîne qui part d'un sommet x_i et aboutit au même sommet x_i;
- une chaîne est une pseudo-chaîne dans laquelle, s'il y a répétition d'arêtes, chaque arête répétée est parcourue dans le même sens;
- un cycle est un pseudo-cycle dans lequel, s'il y a répétition d'arêtes, chaque arête répétée est parcourue dans le même sens;
- une chaîne est simple lorsqu'elle n'utilise pas plus d'une fois la même arête;
- un cycle est simple lorsqu'il n'utilise pas plus d'une fois la même arête;
- une chaîne est élémentaire lorsqu'elle ne passe pas plus d'une fois par chacun de ses sommets;
- un cycle est élémentaire lorsqu'il ne passe pas plus d'une fois par chacun de ses sommets;

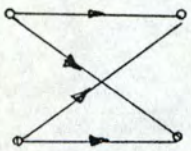
- G est un multigraphe sans cycle s'il est impossible d'y trouver une suite d'arêtes définissant un cycle;
- G est un graphe simplement connexe s'il existe une chaîne entre tout couple de sommets distincts;
- un arbre est un graphe simplement connexe et sans cycle;
- une arborescence est un arbre qui possède une racine (un sommet qui est ascendant à tout sommet du graphe est appelé une racine);
- G est un graphe fortement connexe s'il existe deux chemins reliant en sens opposés tout couple de sommets distincts;
- un arbre maximal d'un graphe G simplement connexe est un sous-graphe de G contenant tous les sommets de G.



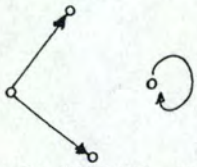
ARBRE



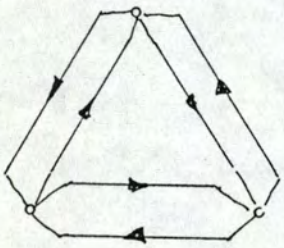
ARBORESCENCE



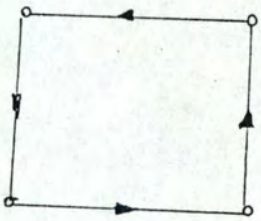
SIMPLEMENT CONNEXE



NON SIMPLEMENT CONNEXE



FORTEMENT CONNEXE



NON FORTEMENT CONNEXE

ANNEXE 2 : ELEMENTS DE CHIMIE ORGANIQUE

L'expérience montre que, si le nombre des différentes substances que l'on peut rencontrer est très élevé, toutes ces substances peuvent être obtenues à partir d'un petit nombre (une centaine) de corps que l'on nomme "simples", tous les autres étant des "composés". Il était naturel, puisque toute matière est formée de particules, de penser qu'à chaque corps simple correspond une particule ultime, la plus petite parcelle susceptible d'entrer en combinaison, et qu'on nomme atome; de telle sorte qu'un échantillon de corps simple est formé d'atomes tous identiques, alors qu'un échantillon, et même une particule ultime d'un corps composé sont formés par l'union d'atomes différents, provenant de différents corps simples.

Classification naturelle périodique des éléments

Les éléments y sont rangés par numéro atomique croissant et disposés: horizontalement suivant des périodes correspondant au remplissage progressif des couches K, L, M, ... ; verticalement, en groupes d'éléments analogues chimiquement.

Notation chimique

Chaque élément est représenté par son symbole, formé de la lettre initiale majuscule du nom de l'élément, suivie, en cas de confusion possible, d'une deuxième lettre minuscule; la liste des symboles se trouve dans le tableau des masses atomiques. Chaque corps pur, simple ou composé, est représenté par sa formule, constituée à l'aide des symboles des éléments qui entrent dans la composition du corps; en outre, le symbole de chaque élément représente par définition une masse déterminée de cet élément.

Liaisons chimiques

Tous les corps sont formés d'atomes; il est cependant peu fréquent que, dans les conditions ordinaires, un échantillon de matière soit constitué d'atomes séparés et pratiquement indépendants

(gaz inertes); presque toujours, des liaisons chimiques unissent entre eux des atomes, semblables ou différents; le résultat de cette union interatomique peut être la formation de molécules, particules pluriatomiques et électriquement neutres dont les dimensions sont généralement de l'ordre de quelques angströms; le corps pur est dans ce cas formé de molécules toutes identiques, il est de constitution moléculaire; des forces d'attraction intermoléculaire assurent, dans les états condensés, c'est-à-dire solide et liquide, la cohésion de l'ensemble; elles deviennent beaucoup plus faibles dans les états dilués (gaz, solutions) en raison des distances plus grandes qui séparent en moyenne, les molécules. Les corps purs de constitution moléculaire sont très nombreux; citons l'hydrogène, l'oxygène, l'azote, le chlore, l'eau, ..., et surtout la presque totalité des corps organiques: méthane, acéthylène, alcool, ...

Réaction chimique

On dit qu'un système de corps a été le siège d'une réaction chimique lorsque l'analyse immédiate permet de retirer de ce système des espèces chimiques qui ne se trouvaient pas dans le mélange initial, à tout le moins lorsque les proportions relatives des espèces chimiques ont varié. Exemple: $H_2 + CuO \longrightarrow Cu + H_2O$. Une réaction chimique est donc une transformation qui fait passer un système de corps d'un état initial à un état final.

Isomérie

La formule brute ne suffit pas à caractériser une substance. Par exemple, l'alcool éthylique et l'oxyde de méthyle ont tous deux la formule brute C_2H_6O . Ils sont appelés isomères. L'expérience montre que le nombre d'isomères croît très vite dès que la formule brute se complique; on connaît 13 composés différents pour $C_5H_{12}O$, des millions sont prévus pour $C_{30}H_{50}O_4N_2$.

Formules développées planes

Si les isomères sont constitués par les mêmes atomes en même nombre, ils se distinguent l'un de l'autre par un arrangement dif-

férent des atomes au sein de la molécule. On a supposé les atomes reliés les uns aux autres, et cette dépendance a été symbolisée par un tiret appelé liaison. La formule développée plane représente l'assemblage deux à deux des atomes; elle est dite plane du fait que cet enchaînement peut être traduit sur une feuille de papier.

Valence

L'hydrogène ne peut réunir deux atomes: il ne dispose que d'une seule liaison; il est dit univalent. Les halogènes, à de très rares exceptions, sont également univalents. Le problème est moins simple pour les autres éléments. Dans le méthane, le carbone est lié à quatre atomes d'hydrogène; il est donc, ici, quadrivalent; dans l'éthane, le carbone est quadrivalent. De même si l'oxygène est bivalent dans l'alcool méthylique H_3COH , il est univalent dans l'aldéhyde $H_3C - C \begin{smallmatrix} H \\ O \end{smallmatrix}$.

Liaisons multiples

En remarquant que le carbone de valence apparente inférieure à 4, l'oxygène de valence inférieure à 2, l'azote de valence inférieure à 3 étaient toujours voisins d'un atome présentant la même anomalie, les chimistes ont convenu d'unir les atomes à valence incomplète par plusieurs liaisons: l'éthylène est écrit $CH_2 = CH_2$, l'aldéhyde $CH_3 - CH = O$. Les composés dont la formule renferme des liaisons multiples sont dits non saturés; en effet, ils peuvent additionner 2 ou 4 atomes d'hydrogène pour conduire à de nouvelles molécules dans lesquelles toutes les valences redeviennent normales: $CH_3 - C \equiv N + 2H_2 \longrightarrow CH_3 - CH_2 - NH_2$. Moyennant cet artifice, les valences respectives 4, 2, 3 du carbone, de l'oxygène, de l'azote deviennent constantes, ce qui confère à la notion de valence une grande simplicité.

Formules semi-développées

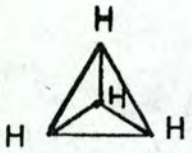
Les formules développées planes tiennent beaucoup de place. On convient de les condenser en groupant certains atomes dont la disposition n'offre aucune ambiguïté. C'est ainsi qu'on écrira l'éthane $CH_3 - CH_3$.

Stéréochimie

Nous avons considéré jusqu'ici que toutes les formules écrites dans le plan sont équivalentes pourvu qu'elles respectent l'enchaînement successif des divers atomes. Or l'expérience a montré que certaines formules planes représentent plusieurs isomères rencontrés; par exemple, aux formules planes $CH_3 - CHOH - CO_2H$, $CH_3 - CHCl - CHOH - CH_3$, correspondent respectivement 2 et 4 isomères connus. Les formules planes sont donc insuffisantes pour les distinguer et il convient d'étudier la forme de la molécule dans l'espace; c'est le but de la stéréochimie.

Carbone tétraédrique

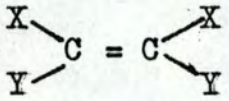
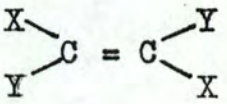
Les 4 valences émanant d'un atome de carbone ne sont pas dans un même plan; si les 4 substituants du carbone sont monoatomiques et identiques (CH_4), les 4 atomes identiques occupent les sommets d'un tétraèdre régulier dont le carbone occupe le centre:



Si les 4 substituants ne sont pas identiques, le tétraèdre est déformé mais ils ne sont jamais dans le même plan. Un carbone lié à 4 radicaux tous différents s'appelle carbone asymétrique.
Notation: les substituants Z et T sont vers l'observateur, X et Y en arrière.

Isomérie géométrique

Soit un composé CXY = CXY , X et Y étant définis comme ci-dessus. Les deux atomes de carbone et les quatre atomes attachant X et Y à ces carbones sont dans un même plan; il y a donc deux dispositions:

cis  trans 

On trouve donc une nouvelle isomérie, dite isomérie cis-trans ou géométrique.

BIBLIOGRAPHIE

- théorie des graphes et codification

(1) NARSINGH DEO, Graph theory with applications to engineering and computer science, Prentice-Hall, Inc., Englewwood Cliffs, N.J. (1974).

(2) D.H. ROUVRAY, The search for useful topological indices in chemistry, American Scientist, Vol. 61, No. 6, Nov.-Dec. 1973, pp. 729-735.

(3) D.H. ROUVRAY, Uses of graph theory, Chemistry in Britain, Vol. 10, No. 1, Jan. 1974, pp. 11-15.

(4) R.W. JOTHAM, A topological subject, Chemical Society Reviews, Vol. 2, No. 4, 1973, pp. 457-474.

(5) E.A. FEIGENBAUM, J. LEDERBERG, Mechanization of inductive inference in organic chemistry, John Wiley and Sons, Inc., New York, 1968, pp. 187-218.

(6) D.J. GLUCK, A chemical structure storage and search system developed at Du Pont, J. Chem. Doc., Vol. 5, No. 1, Feb. 1965, pp. 43-51.

(7) J. LEDERBERG, Topology of molecules, The Mathematical Sciences, The M.I.T. Press, Cambridge, Mass., 1969, pp. 37-51.

(8) M.F. LYNCH, J.M. HARRISON, W.G. TOWN, J.E. ASH, Computer Handling of chemical structure information, Elsevier Publishing Company, Amsterdam, 1971.

(9) A.R. MEETHAM, Partial isomorphisms in graphs and structural similarities in tree-line molecules, Proc. IFIP Congress, 1968.

(10) R.H. PENNY, A connectivity code for use in describing chemical structures, J. Chem. Doc., Vol. 5, No. 2, May 1965, pp. 113-117.

(11) F.A. TATE, Handling chemical compounds in information systems, Ann. Rev. Inf., Sci. Tech. Vol. 2 (C.A. Cuadra, ed.), John Wiley and Sons, Inc., New York, 1967, pp. 285-309.

(12) J.M. HARRISON, M.F. LYNCH, Computer analysis of chemical reactions for storage and retrieval, J. Chem. Soc., Vol. C, 1970, pp. 2082-2087.

(14) J. FICHEFET, Théorie des graphes et applications, Document interne de l'Institut d'Informatique, F.N.D.P., Namur (1973).

(15) M. BERSOHN, A. ESACK, A computer representation of synthetic organic reactions, Computers and Chemistry, Vol. 1, 1976, pp. 103-107.

(16) R. FUGMANN, U. DÖLLING, H. NICKELSEN, A topological approach to the problem of ring structures, Angew. Chem. internat. Edit., Vol. 6, No. 9, Sept. 1967, pp. 723-818.

(17) N. CARPENTIER, Syntax-directed translation of organic chemical formulas into their two-dimensional representation, Computers and Chemistry, Vol. 1, 1976, pp. 25-28.

(18) H.L. MORGAN, The generation of a unique machine description for chemical structures - A technique developed at chemical abstracts service, J. Chem. Doc., Vol. 5, 1965, pp. 107-113.

(19) A.M. DUFFIELD, A.V. ROBERTSON, C. DJERASSI, B.G. BUCHANAN, G.L. SUTHERLAND, E.A. FEIGENBAUM, J. LEDERBERG, Applications of artificial intelligence for chemical inference: I. The number of possible organic compounds (acyclic structures containing C,H,O,N); II. Interpretation of low-resolution mass spectra of ketones, J.A.C.S., Vol. 91, No. 11, May 1969, pp. 2973-2981.

(20) G. KRESZE, Present-day communication in chemistry - problems and possibilities, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 545-550.

(21) C. WEISKE, Chemical information services, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 550-555.

(22) R. FUGMANN, Theoretical aspects of communication in chemistry, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 555-576.

(23) M.A. LOBECK, Use of IDC system, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 576-583.

(24) E. MEYER, Versatile computer techniques for searching by structural formulas, partial structures, and classes of compounds, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 583-589.

(25) R. FUGMANN, H. NICKELSEN, I. NICKELSEN, J.H. WINTER, TOSAR - a topological method for the representation of synthetic and analytical relations of concepts, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 589-595.

(26) W. NUBLING, W. STEIDLE, The dockumentationsring der chemisch-pharmazeutischen industrie: aims and methods, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 596-598.

(27) O. SCHIER, W. NUBLING, W. STEIDLE, J. VALLS, A system for the documentation of chemival reactions, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 599-604.

(28) G. OHNACKER, W. KALBFLEISCH, CCBF - a system for the computer processing of chemical and biogical facts, Angew. Chem. internat. Edit., Vol. 9, No. 8, 1970, pp. 605-610.

(29) J.E. DUBOIS, H. VIELLARD, Système DARC, Bulletin de la Société Chimique de France,
1968, No. 3, pp. 900-904;
1968, No. 3, pp. 905-912;
1968, No. 3, pp. 913-919;
1971, No. 3, pp. 839-848;
1973, No. 6, pp. 1988-1996;
1973, No. 6, pp. 1996-2002;
1975, No. 5-6, pp. 1390-1400;
1975, No. 5-6, pp. 1401-1406;
1975, No. 9-10, pp. 2100-2110.

(30) J.E. DUBOIS, D. LAURENT, A. PANAYE, Y. SOBEL, Système DARC: hyperstructures formelles d'antériorité, C. R. Acad. Sc. Paris, t. 281, Nov. 1975, pp. 687-690.

(31) J.E. DUBOIS, D. LAURENT, A. PANAYE, Y. SOBEL, Système DARC: concept d'hyperstructure formelle, C. R. Acad. Sc. Paris, t. 280, Avril 1975, pp. 851-854.

- structure de données et banques de données

(32) P. SIEWIOREK, Introduction to computer organization and data structures: PDP-11 Edition, McGraw-Hill, Inc., 1975.

(33) M.N. CHEMAN, Organisation et interrogation d'une banque de données biomédicales, Thèse de doctorat, Université de TOURS, 1974.

- stratégies de synthèse

(34) E.J. COREY, Choosing chemical routes, MOSAIC Fall, 1974, pp. 21-25.

(35) E.J. COREY, W. TODD WIPKE, RICHARD D. CRAMER III, W. JEFFREY HOWE, Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive graphics, J.A.C.S., Vol. 94, No. 2, Jan. 1972, pp. 421-429.

(36) E.J. COREY, RICHARD D. CRAMER III, W. JEFFREY HOWE, Computer-assisted analysis for complex molecules methods and procedures for machine generation of synthetic intermediates, J.A.C.S., Vol. 94, No. 2, 1972, pp. 440-459.

(37) E.J. COREY, W. TODD WIPKE, RICHARD D. CRAMER III, W. JEFFREY HOWE, Techniques for perception by a computer of synthetically significant structural features in complex molecules, J.A.C.S., Vol. 94, No. 2, 1972, pp. 431-439.

(38) E.J. COREY, GEORGE A. PETERSSON, An algorithm for machine perception of synthetically significant rings in complex cyclic organic structures, J.A.C.S., Vol. 94, No. 2, 1972, pp. 460-465.

(39) E.J. COREY, WILLIAM L. JORGENSEN, Computer-assisted synthetic analysis. Generation of synthetic sequences involving sequential functional group interchanges, J.A.C.S., Vol. 98, No. 1, 1976, pp. 203-209.

(40) E.J. COREY, W. JEFFREY HOWE, H.W. ORF, DAVID A. PENSAK, GEORGE PETERSSON, General methods of synthetic analysis. Strategic bond disconnections for bridged polycyclic structures, J.A.C.S., Vol. 97, No. 21, 1975, pp. 6116-6124.

(41) E.J. COREY, W. JEFFREY HOWE, DAVID A. PENSAK, Computer-assisted analysis. Methods for machine generation of synthetic intermediates involving multistep look-ahead, J.A.C.S., Vol. 96, No. 25, 1974, pp. 7724-7737.

(42) E.J. COREY, WILLIAM L. JORGENSEN, Computer-assisted synthetic analysis. Synthetic strategies based on appendages and use of reconnective transforms, J.A.C.S., Vol. 98, No. 1, 1976, pp. 189-203.

(43) E.J. COREY, H.W. ORF, DAVID A. PENSAK, Computer-assisted synthetic analysis. The identification and protection of interfering functionality in machine-generated synthetic intermediates, J.A.C.S., Vol. 98, No. 1, 1976, pp. 210-221.

(44) JAMES B. HENDRICKSON, A systematic characterisation of structures and reactions for use in organic synthesis, J.A.C.S., Vol. 93, No. 25, 1971, pp. 6847-6862.

(45) JAMES B. HENDRICKSON, Systematic synthesis design III. The scope of the problem, J.A.C.S., Vol. 97, No. 20, 1975, pp. 5763-5783.

(46) JAMES B. HENDRICKSON, Systematic synthesis design IV. Numerical codification of construction reactions, J.A.C.S., Vol. 97, No. 20, 1975, pp. 5784-5800.

(47) R. BARONE, M. CHANON, J. METZGER, Ordinateur et synthèse organique:utilisation des mécanismes réactionnels, Tetrahedron Letters, No. 32, 1974, pp. 2761-2764.

(48) M. BERSOHN, Automatic problem solving applied to synthetic chemistry, Bulletin of the chemical society of Japan, Vol. 45, 1972, pp. 1897-1903.

(49) H.W. WHITLOCK, A heuristic solution to functional group switching problem in organic synthesis, J.A.C.S., Vol. 98, No. 11, 1976, pp.3225-3233.

(50) PAUL E. BLOWER, HOWARD W. WHITLOCK, An application of artificial intelligence to organic synthesis, J.A.C.S., Vol. 98, No. 6, 1976, pp. 1499-1510.

(51) N.S. SRIDHARAN, A heuristic program to discover syntheses for complex organic molecules, Information Processing 74, North-Holland Publishing Company, 1974, pp. 828-833.

(52) B.G. BUCHANAN, G.L. SUTHERLAND, E.A. FEIGENBAUM, Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry, Machine Intelligence, Vol. 4, Edinburgh University Press, Edinburgh, 1969.