# Facteurs électroniques et reconnaissance de formes en CAO de molécules pharmacologiquement actives : application à la famille des benzodiazépines

Roger Rozot

UNIVERSITE DE NANCY I
U.E.R. Sciences de la Matière

**THESE**

Présentée à
l'UNIVERSITE DE NANCY I

Pour obtenir
le titre de Docteur de l'Université de Nancy I
en Chimie Informatique et Théorique



par
Roger ROZOT

# Facteurs électroniques et reconnaissance de formes en C.A.O. de molécules pharmacologiquement actives. Application à la famille des benzodiazépines

Soutenue publiquement devant la Commission d'examen
le 13 Juillet 1988

Membres du Jury : MM.

| | |
|---|---|
| G. SIEST | Président |
| A. CALVET | Rapporteur |
| D. CANET | Rapporteur |
| R. MOHR | Examinateur |
| A. CARTIER | Examinateur |
| J.L. RIVAIL | Examinateur |

UNIVERSITE DE NANCY I
U.E.R. Sciences de la Matière

**THESE**

Présentée à
l'UNIVERSITE DE NANCY I

Pour obtenir
le titre de Docteur de l'Université de Nancy I
en Chimie Informatique et Théorique

par
Roger ROZOT

# Facteurs électroniques et reconnaissance de formes en C.A.O. de molécules pharmacologiquement actives. Application à la famille des benzodiazépines

Soutenue publiquement devant la Commission d'examen
le 13 Juillet 1988

| | | |
|---|---|---|
| Membres du Jury : MM. | G. SIEST | Président |
| | A. CALVET | Rapporteur |
| | D. CANET | Rapporteur |
| | R. MOHR | Examinateur |
| | A. CARTIER | Examinateur |
| | J.L. RIVAIL | Examinateur |

*à SARAH ...*

J'exprime ma profonde gratitude à Monsieur le Professeur Jean-Louis RIVAIL qui m'a accueilli au Laboratoire de Chimie Théorique, de l'Université de Nancy I, qu'il dirige et qui m'a permis de réaliser ces travaux.

Je le remercie également pour ses précieux conseils et la constante attention qu'il a portée à la rédaction de ce mémoire.

Je remercie également tous les chercheurs du Laboratoire de Chimie Théorique pour leur accueil et leur aide, en particulier:

- Monsieur Alain CARTIER, qui s'intéresse de près aux problèmes de relations structure-activité et qui a réalisé l'essentiel des analyses statistiques classiques et des calculs de structure moléculaire nécessaires pour étayer mes travaux, pour ses excellents conseils et ses très bonnes remarques sur la rédaction de ce mémoire;

- Monsieur Daniel RINALDI pour son aide dans le domaine des calculs de chimie quantique et de leur programmation;

- Mademoiselle Marilia T. C. MARTINS COSTA pour son logiciel d'infographie moléculaire.

Je remercie vivement les membres du Jury:

- Monsieur le Professeur Gérard SIEST, Directeur du Centre du Médicament à la Faculté de Pharmacie de l'Université de Nancy I, qui a bien voulu s'intéresser à mes travaux et qui a accepté de présider le Jury malgré ses activités très prenantes;

- Monsieur Alain CALVET, Ingénieur de Recherche aux Laboratoires Jouveinal S.A., qui a été très attentif à mes travaux et qui a accepté d'en être Rapporteur;

- Monsieur le Professeur Daniel CANET, Directeur du Laboratoire de Méthodologie RMN de l'Université de Nancy I, qui a gracieusement accepté d'être Rapporteur pour l'Université;

- Monsieur le Professeur Roger MOHR, du Centre Régional d'Informatique de Nancy, pour son aide précieuse dans le domaine de l'intelligence artificielle et à qui je dois ma formation initiale en informatique;

- Monsieur Alain CARTIER;

- Monsieur le Professeur Jean-Louis RIVAIL que je ne remercierai jamais assez d'avoir dirigé mes recherches.

# Table des Matières

# AVANT-PROPOS

Les travaux, effectués dans le cadre de cette thèse au Laboratoire de Chimie Théorique de l'Université de Nancy I sous la direction de Monsieur le Professeur Jean-Louis RIVAIL, portent sur la mise au point (et surtout la faisabilité) d'une méthode nouvelle de conception de molécules pharmacologiquement actives, alliant intelligence artificielle, techniques statistiques et chimie théorique.

La conception d'un nouveau médicament est une entreprise longue (huit à douze ans) et coûteuse (législation internationale, concurrence, critères économiques, etc.). Cette entreprise comprend de nombreuses étapes:

- sélection d'une molécule *a priori* intéressante,
- synthèse de cette molécule,
- démonstration de son efficacité biologique puis clinique,
- reconnaissance de l'absence d'effets secondaires,
- obtention des autorisations nécessaires à sa commercialisation.

En termes économiques, près de la moitié du coût de lancement d'un nouveau médicament est consacré aux travaux de recherche soit environ 35 millions de dollars en 1985 (il faut en moyenne synthétiser plusieurs milliers de molécules avant d'en commercialiser une) [1].

Depuis quelques années, afin de réduire ce coût et ces délais, de nombreux laboratoires privés et publics orientent leurs efforts de recherche vers de nouvelles techniques informatisées aidant le chimiste à résoudre les problèmes suivants:

- Quelle molécule synthétiser en vue d'une activité biologique donnée ?
- Comment la synthétiser ?
- Comment disposer au mieux des données obtenues au cours de travaux antérieurs sur des molécules analogues ?
- Comment évaluer rapidement et *a priori* l'intérêt d'une molécule nouvelle sans avoir à effectuer une partie du coûteux travail de synthèse chimique et de tests biologiques ?

Ainsi le chimiste ne choisit de synthétiser une molécule qu'après avoir évalué son activité biologique. Cette évaluation est aidée par une documentation sur des molécules analogues, par l'analyse des similitudes structurales entre la molécule considérée et des composés déjà étudiés pour la même activité biologique, par la construction et la comparaison de modèles

moléculaires et/ou l'étude d'une relation éventuelle entre la structure et l'activité.

Dans ce cadre, les travaux entrepris avaient pour objectif de répondre à une partie des questions que se posent le chimiste.

Dans le premier chapitre de ce mémoire sont exposés sommairement les outils informatiques actuels mis à la disposition du chimiste et quelques généralités sur les relations structure-activité. Dans cette partie, nous décrivons surtout la démarche suivie par un chimiste pour résoudre les problèmes cités précédemment car elle est en fait le point de départ de ces travaux.

Les principes du logiciel mis au point et baptisé "SARAH" (pour Structure-Activité: Relations par Apprentissage et Heuristiques) sont décrits en détail dans le deuxième chapitre en insistant particulièrement sur les raisonnements qu'il intègre et sur la détermination de la structure des molécules. Les choix faits seront d'ailleurs vérifiés et commentés dans la partie application de ce mémoire.

L'étude de la famille des benzodiazépines constitue le troisième chapitre où nous noterons l'importance de la géométrie des molécules étudiées pour l'activité biologique considérée à savoir l'activité anti-pentylènetétrazole (composé provoquant des crises d'épilepsie) ou anti-convulsivante. Dans ce chapitre également, figurent des comparaisons entre les estimations de l'activité de molécules tests faites par le logiciel SARAH et celles obtenues par des méthodes couramment employées et purement statistiques.

Pour conclure, nous exposons, dans un quatrième chapitre, les possibilités et les limites du logiciel SARAH qui est comparé aux programmes actuellement utilisés pour la détermination de relations structure-activité.

* *

*

# Chapitre 1

# DES RELATIONS STRUCTURE-ACTIVITE

## 1.1 Définitions

### 1.1.1 Activité d'une molécule - exemple

L'activité d'une molécule par rapport à un processus biologique est en général caractérisée par le score qu'elle a obtenu à un test biologique *in vivo* ou *in vitro* mettant en jeu ce processus.

Une molécule de la famille des *benzodiazépines* est, par exemple, couramment caractérisée biologiquement et *in vivo* par son activité *anti-pentylènetétrazole* qui est définie comme le logarithme décimal de l'inverse de la concentration en benzodiazépine (en mmol/Kg: quantité de produit rapportée à la masse de l'animal test, généralement un rat) nécessaire pour neutraliser les effets (crises d'épilepsie) de l'injection de 125 mg/Kg de pentylènetétrazole chez 50% des animaux traités (test de EVERETT et RICHARDS [2,3]).

Cette activité est donc une donnée quantitative comme nous venons de le voir, mais peut être également une donnée booléenne (molécule active ou inactive). Pour ce qui suit, nous supposerons que l'activité est une donnée chiffrée, son introduction sous forme de donnée booléenne (en ce qui concerne les entrées) étant une extension à venir en ce qui concerne le logiciel SARAH mis au point et qui va être décrit.

### 1.1.2 Structure d'une molécule - exemples

Les méthodes de la chimie quantique [4] (méthodes semi-empiriques [5,6] par exemple) ou de la mécanique moléculaire [7,8] permettent de calculer, généralement à l'aide d'un ordinateur, un grand nombre de caractéristiques moléculaires (géométrie la plus réaliste par optimisation des coordonnées des atomes de la molécule, énergie totale, chaleur de formation, moment dipolaire, polarisabilité totale, volume de van der Waals, anisotropie de la molécule, etc.) ou sub-moléculaires (grandeurs relatives à des atomes ou des groupements d'atomes qui composent la molécule: charge nette, moment dipolaire local, polarisabilité locale, constantes de forces, constantes de couplage spin-spin, spin-orbite, vibration-rotation, etc.).

Tous ces descripteurs ou caractéristiques moléculaires, regroupant des informations électroniques et géométriques, constituent la structure d'une molécule.

Le grand intérêt de ces descripteurs est qu'ils sont calculables et ne nécessitent donc pas la synthèse de la molécule.

### 1.1.3 Relations entre la structure et l'activité

En général, le chimiste dispose d'une collection de molécules analogues étudiées pour une propriété pharmacologique particulière. De ces molécules il connaît à la fois la structure (sous forme d'un ensemble de descripteurs (comme ceux cités précédemment)) et l'activité. A partir de ces données, il essaie d'évaluer l'activité d'une molécule nouvelle non encore synthétisée en recherchant d'éventuelles relations entre ces descripteurs et l'activité biologique étudiée.

La bibliographie, importante dans ce domaine ces dernières années, a permis de mettre en évidence trois types de relations structure-activité:

- Citons en premier lieu les relations quantitatives, sous forme de régressions linéaires multi-variables, qui représentent la majeure partie des recherches sur les relations structure-activité [9,10,11,12,13]. Ces méthodes donnent d'assez bons résultats à condition de manipuler les indices structuraux qui conviennent pour l'activité étudiée.

- Quelques travaux [14] portent sur la détermination de relations qualitatives sous forme d'appartenance ou de non appartenance à un sous-ensemble de molécules, soit actives soit inactives en majorité, correspondant à des valeurs de descripteurs structuraux particuliers (c'est par exemple le résultat d'une analyse en composantes principales, ou d'une analyse d'agrégats). Il est toutefois possible, dans certains cas, de prédire quantitativement l'activité d'une molécule nouvelle en faisant une analyse en composantes principales. Cette approche du problème s'applique bien dans le cas où l'activité est une donnée booléenne.

- Très peu de travaux ont été réalisés dans le domaine des relations qualitatives et quantitatives réunissant les deux types de relations précédents en utilisant une sélection automatique des descripteurs structuraux (les types de relations précédents sont très dépendants de la subjectivité de celui qui les utilise en ce qui concerne le choix des descripteurs). Le principe, qui en plus des méthodes statistiques classiques utilise des techniques d'*intelligence artificielle*, est le suivant: après une première séparation des molécules de référence par rapport à certains descripteurs structuraux trouvés indispensables pour l'activité d'une molécule (à ce stade il est possible de prédire qualitativement l'activité d'une molécule nouvelle), pour chaque sous-ensemble de molécules obtenu est recherchée la meilleure corrélation entre l'activité et une partie des descripteurs structuraux les plus influents sur l'activité mais non rédhibitoires, afin

d'obtenir des valeurs chiffrées de l'activité. Le logiciel SARAH mis au point utilise ce genre de relations qui sera donc expliqué en détail dans le chapitre suivant.

## 1.2 Les outils informatiques à la disposition du chimiste

Ce n'est que depuis peu que sont apparus des systèmes capables de satisfaire la plupart des exigences du chimiste et en particulier grâce au développement des systèmes experts et de l'intelligence artificielle. Trois types de logiciels sont à distinguer [1]:

### 1.2.1 Logiciels de gestion de banques de données

#### La documentation chimique

Des logiciels comme DARC (J. E. DUBOIS et col. - Paris), COUSIN (*Upjohn*) ou CHEMPIX (*Roussel-Uclaf*) permettent, parmi une collection importante de molécules (de quelques milliers à plusieurs millions), soit d'accèder aux informations concernant une molécule ou une famille de molécules ayant une structure particulière, soit de rechercher la famille de molécules possèdant une sous-structure particulière. Ils sont basés sur la représentation de la topologie des molécules par codage de la matrice de connectivité (algorithme de MORGAN [1]).

#### La synthèse assitée par ordinateur (S.A.O.)

Les logiciels de S.A.O. utilisent le principe général de l'analyse rétrosynthétique: la molécule à synthétiser est analysée afin de reconnaître des éléments structuraux (cycles, groupements fonctionnels, associations de groupements fonctionnels) représentatifs des sites réactifs de la molécule, puis une recherche dans un catalogue de transformations chimiques préétabli permet d'obtenir les réactions susceptibles de fabriquer les éléments précédemment reconnus; ces transformations donnent les précurseurs de la molécule à synthétiser.

Ces logiciels, tels que REACCS (*Rhône Poulenc*) ou PASCOP [15] contenant un catalogue de plusieurs dizaines de milliers de réactions, sont plutôt apparentés aux systèmes experts.

### 1.2.2 Les logiciels de calcul des structures moléculaires

Ils permettent, par des calculs de mécanique moléculaire comme MM2 [16], MODEL (*Rhône Poulenc Recherche*) ou par des calculs de chimie quantique (méthodes semi-empiriques ou *ab initio*) comme GEOMO [17], GEOMOS [18], CHIMISTE [19], d'obtenir la géométrie la plus réaliste et les descripteurs structuraux décrits précédemment (voir paragraphe 1.1.2). Notons que les méthodes *ab initio* permettent l'optimisation de géométrie mais avec des temps de calcul très longs.

### 1.2.3 Les logiciels de détermination de relations structure-activité

Ce sont des logiciels qui se développent depuis peu et qui, dans la plupart des cas, utilisent des méthodes statistiques: analyse en composantes principales, analyse factorielle, régression linéaire multi-variables, analyse discriminante [21].

Ces logiciels permettent soit d'obtenir des relations du type:

$$A = K + \sum_{i=1}^{n} C_i P_i \qquad (1.1)$$

où $A$ est l'activité biologique étudiée, $P_i$ est un descripteur structural (grandeur chiffrée caractéristique de la molécule ou d'une partie de la molécule comme celles décrites dans les paragraphes **2.2.3** et **2.2.5**), $C_i$ est le coefficient correspondant et $K$ est une constante - relations dont le coefficient de corrélation est plus ou moins bon -, soit la prédiction qualitative de l'activité d'une molécule nouvelle.

Le plus difficile dans ces méthodes est la sélection des descripteurs structuraux à prendre en compte, nous y reviendrons dans le prochain chapitre.

Très peu de logiciels utilisent une approche *intelligence artificielle* pour traiter ces problèmes de relations structure-activité. Citons cependant le programme CASE [22,23,24] qui utilise comme descripteurs les sous-structures des molécules (-COOH, -$NH_2$, -CN, chaînes aliphatiques hydrocarbonées, etc.). Ce programme "découpe" les molécules d'un lot d'apprentissage en groupements linéaires de trois à douze atomes liés (hydrogènes non compris), et à chaque groupement sont associées sa fréquence d'apparition dans l'ensemble des molécules actives et sa fréquence d'apparition dans l'ensemble des molécules inactives.

Après un tri statistique (distribution binomiale), le programme affecte à chaque groupement un caractère (activant, désactivant, sans influence sur l'activité) et permet la prédiction de la tendance d'une molécule nouvelle (active ou inactive).

Le programme CASE détermine également une corrélation entre l'activité et les fréquences d'apparition des fragments par régression multi-variables [24].

L'avantage de ce programme est que la sélection des descripteurs structuraux (fragments de molécule) est automatique et non limitée; c'est également une des caractéristiques que nous avons retenues pour concevoir le logiciel SARAH.

## 1.3 Les raisonnements du chimiste et l'informatique

### 1.3.1 Les desseins du chimiste

Le travail du chimiste consiste donc à répondre aux questions citées dans l'introduction de ce mémoire :

- Quelle molécule synthétiser en vue d'une activité biologique donnée ? (1)
- Comment la synthétiser ? (2)

- Comment disposer au mieux des données obtenues au cours de travaux antérieurs sur des molécules analogues ? (3)

- Comment évaluer rapidement et *a priori* l'intérêt d'une molécule nouvelle sans avoir à effectuer une partie du coûteux travail de synthèse chimique et de tests biologiques ? (4)

Les questions 2 et 3 sont du domaine de la documentation chimique et de la synthèse assistée par ordinateur. Dans le cadre de ce travail, nous nous efforcerons de résoudre les questions 1 et 4. Pour cela, il nous faut décrire en détail la démarche qu'effectue habituellement le chimiste.

### 1.3.2 L'analyse traditionnelle du chimiste

La pharmacologie moléculaire a mis en évidence le rôle fondamental des récepteurs biologiques et déterminé, dans certains cas [25,26,27,28,29,30], leurs structures (souvent de façon approchée). Elle utilise souvent l'image d'une *clé* et d'une *serrure* pour décrire les interactions entre une *molécule active* et un *récepteur* [31]. Dans de telles études, comme le soulignent F. CHOPLIN [1] et D. E. WALTERS et col. [32], il est donc nécessaire de faire intervenir la géométrie (ou les diverses géométries possibles pondérées par leurs énergies conformationnelles) ainsi que les propriétés électroniques qui gouvernent l'interaction et/ou la réactivité de la molécule étudiée dans le mécanisme biologique considéré.

Il est important de noter qu'un troisième facteur intervient dans de tels mécanismes biologiques, à savoir le transport de la molécule à travers les membranes cellulaires entre une phase aqueuse et une phase lipidique, vers le site d'activité. Cependant, ce mécanisme de transport biologique se ramène à une autre étude de relations structure-activité où l'activité considérée peut-être le coefficient de partage entre la phase contenant le récepteur et la phase par laquelle est assimilée la molécule, puisque le transport à travers les membranes cellulaires est réalisé par des transmetteurs ou des canaux moléculaires qui sont les récepteurs vis-à-vis de cette activité.

Lorsque l'activité est le résultat d'un test *in vivo* (et c'est notre cas), cette donnée biologique prend en compte ce phénomène de transport. Les conclusions de l'étude décrite dans la suite de ce mémoire portent alors sur l'ensemble de ces deux processus biologiques étroitement liés. En l'abscence de données biologiques concernant ces transports, nous restreindrons notre étude aux paramètres structuraux géométriques et électroniques. D'ailleurs, B. P. ROQUES [31], dans ces travaux sur les récepteurs de la morphine et d'enképhalines, souligne l'importance des facteurs stériques et électroniques de la molécule et, plus particulièrement, des substituants d'une sous-structure commune aux molécules (c'est en fait l'analyse faite par le chimiste ou le biologiste), mais parle très peu des processus de transport qui vraisemblablement ralentissent, dans certains cas, les effets biologiques d'une molécule sans les inhiber complètement.

Disposant d'un lot de molécules de *structures* connues et caractérisées par le résultat d'un *test d'activité biologique* (notons ici l'importance de la documentation chimique et des calculs de structures moléculaires) sous forme d'une variable continue (ce sera l'hypothèse faite pour la suite de ces travaux) ou booléenne (comme nous l'avons remarqué précédemment), le chimiste fait l'analyse suivante:

- La *clé* (la molécule active) doit avoir une certaine géométrie pour s'ajuster dans la *serrure* que constitue le site liant du récepteur (qui est généralement une protéine [27]). De ce fait, les molécules du lot doivent avoir en commun une même sous-structure possèdant de surcroît un ou plusieurs hétéro-atomes qui permettent la liaison molécule active-récepteur et que nous appellerons points d'ancrage [12].

- La présence de cette sous-structure n'entraîne pas nécessairement l'activité. Les propriétés électroniques de cette sous-structure, qui peuvent être modulées par la présence de divers substituants, jouent un rôle déterminant qu'il est important de faire apparaître. Par ailleurs, l'encombrement global de la molécule, et en particulier la taille de ces substituants, peut également modifier considérablement la réactivité (gêne stérique à l'accessibilité du site liant du récepteur).

- Le rôle de ces différents facteurs apparaît généralement lorsque nous procèdons à la comparaison systématique des molécules ayant fait l'objet d'une expérimentation (nous pouvons appeler cette étape "superposition" des molécules). Cette comparaison est généralement réalisée par rapport à une molécule de référence. Ainsi I. NAKATSUKA et col. [33] prennent le *diazépam* comme molécule de référence dans leur étude des relations structure-activité dans la famille des *benzodiazépines*.

A ce stade, le chimiste peut déjà bénéficier d'importantes améliorations dues à l'informatique:

- Détermination de géométries moléculaires réalistes (voir paragraphes **1.2.2** et **2.2.2**) et détermination des conformations de plus basses énergies comme le fait remarquer B. P. ROQUES [31].

- Manipulation aisée des structures moléculaires grâce à une bonne représentation dans la mémoire de l'ordinateur (voir paragraphe **2.2.1**).

- Calcul de grandeurs caractéristiques de la structure électronique (voir paragraphes **1.2.2**, **2.2.2** à **2.2.5**).

### 1.3.3 Les espoirs du chimiste et le logiciel SARAH

En confiant à l'ordinateur le soin de procéder aux comparaisons précédentes et à l'élaboration des critères d'activité et/ou d'inactivité, nous pouvons espérer libérer le chimiste d'une tâche fastidieuse qui, de surcroît, est très dépendante de la subjectivité de celui qui l'effectue.

Pour tenter de répondre à ces espoirs, il a été mis au point un logiciel de reconnaissance de formes capable d'effectuer de façon systématique la démarche du chimiste:

- Détermination de la sous-structure moléculaire de base.

- Comparaison des différentes molécules sous les angles propriétés électroniques et géométrie de l'entourage de cette sous-structure.

- Edition de critères d'activité et/ou d'inactivité déduits de cette comparaison.

C'est ce logiciel "SARAH" (pour Structure-Activité: Relations par Apprentissage et Heuristiques) qui est présenté maintenant.

* *

*

# Chapitre 2

# LES PRINCIPES DU LOGICIEL SARAH

## 2.1 Description générale - introduction

Le but de ces travaux était donc l'écriture d'un programme informatique réalisant de façon intelligente la démarche du chimiste dans la conception de molécules pharmacologiquement actives et éditant un ensemble de relations structure-activité (au sens large et non uniquement linéaires), à partir d'un lot de molécules analogues de structures et d'activités biologiques connues, afin de pouvoir prédire l'activité d'une nouvelle molécule non synthétisée.

Autrement dit, ce programme devait effectuer un apprentissage à partir d'exemples répartis en deux classes (les molécules actives et les molécules inactives) avec recherche des règles de discrimination de ces deux classes et construction d'un arbre de décision, la connaissance préalable (celle du chimiste) étant intégrée au programme (ce programme n'a été conçu que pour traiter des problèmes biologiques de détermination de relations structure-activité).

Lorsque nous parlons de relations structure-activité, le premier problème rencontré est de savoir s'il est effectivement possible de prévoir l'activité biologique d'une molécule à partir de ses caractéristiques structurales, ou de définir les éléments structuraux d'une molécule qui induiront une forte activité, par l'étude préalable d'une famille de molécules analogues de structures et d'activités connues.

De nombreux travaux ces dernières années - citons à cet égard l'excellent livre édité par J. K. SEYDEL [34] qui contient 76 publications dans ce domaine - ont montré que cet objectif ambitieux pouvait être atteint notamment au moyen de méthodes statistiques.

Le deuxième problème rencontré - le plus délicat - réside, comme le fait remarquer R. FRANKE [35], dans la sélection des descripteurs structuraux des molécules. En effet, un ensemble "idéal" de paramètres structuraux doit être complet, c'est à dire couvrir tous les types d'interactions médicament-biosystème, général et non restreint à des séries homologues, ces paramètres doivent être bien définis et avoir une signification physico-chimique, et, finalement, cet ensemble doit permettre l'application de techniques de recherche appropriées ou de formalismes mathématiques; en d'autres termes, permettre la détermination des paramètres

importants et la manière dont ils sont reliés à l'activité. Enfin, ces paramètres doivent être disponibles ou accessibles et de préférence calculables pour éviter des synthèses et des mesures coûteuses, ce qui n'est pas si évident.

Le troisième problème rencontré, et qui vient d'être soulevé, est de déterminer comment ces descripteurs structuraux sont liés à l'activité biologique.

La quatrième difficulté est de savoir s'il faut ou non prendre en compte les caractéristiques structurales du récepteur biologique dans ce genre d'étude. Comme il est souvent difficile, voire impossible, de déterminer la structure, même approchée, de la protéine contenant le site liant, il ne sera fait aucune autre hypothèse sur le récepteur que celle du modèle *clé-serrure*.

Le programme à concevoir devait donc déterminer un ensemble de paramètres structuraux ayant ces propriétés, rechercher les plus importants par rapport à l'activité biologique et décrire la manière dont ces descripteurs sont liés à l'activité, sous forme de critères d'activité et/ou d'inactivité afin de prédire les propriétés pharmacologiques d'une molécule nouvelle.

Pour ce faire, la première étape a été la détermination d'une bonne représentation en machine d'une molécule, c'est à dire la plus facile à manipuler dans les différents algorithmes mis au point. En d'autres termes, il s'agissait de trouver une représentation des connaissances bien adaptée à la nature du problème à traiter.

La deuxième étape du travail, ou phase d'apprentissage, a consisté en la détermination automatique de descripteurs structuraux les plus représentatifs des interactions molécule active-récepteur (là intervient la connaissance du chimiste) et en la recherche d'une méthode de comparaison des molécules proche de celle qu'utilise couramment le chimiste (voir paragraphe **1.3**).

La troisième étape enfin, a été la construction d'un arbre de décision, suivie de l'édition de critères d'activité et/ou d'inactivité et leur utilisation pour faire des prédictions sur des molécules nouvelles avant leur synthèse.

Pour résumer les données du problème, nous disposons d'un ensemble d'exemples (ou lot d'apprentissage) bien décrits, répartis en deux classes non vides (les molécules, soit actives soit inactives, de structures et d'activités chiffrées connues) et caractérisant un concept donné (l'activité). Nous disposons également d'une certaine connaissance du domaine, à savoir la connaissance du chimiste pour décrire les interactions molécule active-récepteur. Le but est de trouver des règles de discrimination de ces deux classes, ou, ce qui est équivalent, des descripteurs discriminants, par apprentissage, c'est à dire une recherche à travers un espace de descripteurs, ramenés à une forme booléenne, partiellement ordonnés suivant la relation spécialisation-généralisation. Les descripteurs discriminants sont évalués à l'aide d'une mesure statistique (le $\chi^2$) que nous appellerons *fonction score*.

Remarquons que le fait que l'activité soit chiffrée (c'est le cas dans 50% des tests biologiques) n'est utilisé que pour pouvoir prédire quantitativement l'activité d'une molécule nouvelle. D'ailleurs, lors du premier traitement informatique du lot d'apprentissage, cette donnée sera binarisée (*active/inactive*) grâce à une information biologique supplémentaire: la valeur limite de l'activité ($A_{ai}$) de part et d'autre de laquelle les molécules sont soit actives soit inactives.

C'est ce travail qui est décrit dans ce qui suit.

## 2.2 Sélection des descripteurs structuraux - La connaissance préalable

### 2.2.1 Représentation en machine d'une molécule

Afin de faciliter à la fois la manipulation de la masse importante de données dont nous disposons pour décrire une molécule et la recherche de sous-structures par un algorithme récursif de recherche d'analogies entre deux molécules, une représentation sous forme de graphe a été adoptée pour caractériser les atomes (nœuds du graphe) et les liaisons chimiques entre eux (les liens entre nœuds du graphe).

Chaque atome est caractérisé intrinsèquement par son numéro atomique, ses orbitales atomiques de Slater, sa charge de valence, son électronégativité, son rayon de van der Waals, son degré de coordination, sa position dans le tableau périodique, et, dans la molécule, par la charge nette qu'il porte, ses coordonnées cartésiennes, le nombre et la nature des liaisons chimiques qu'il possède avec ses voisins.

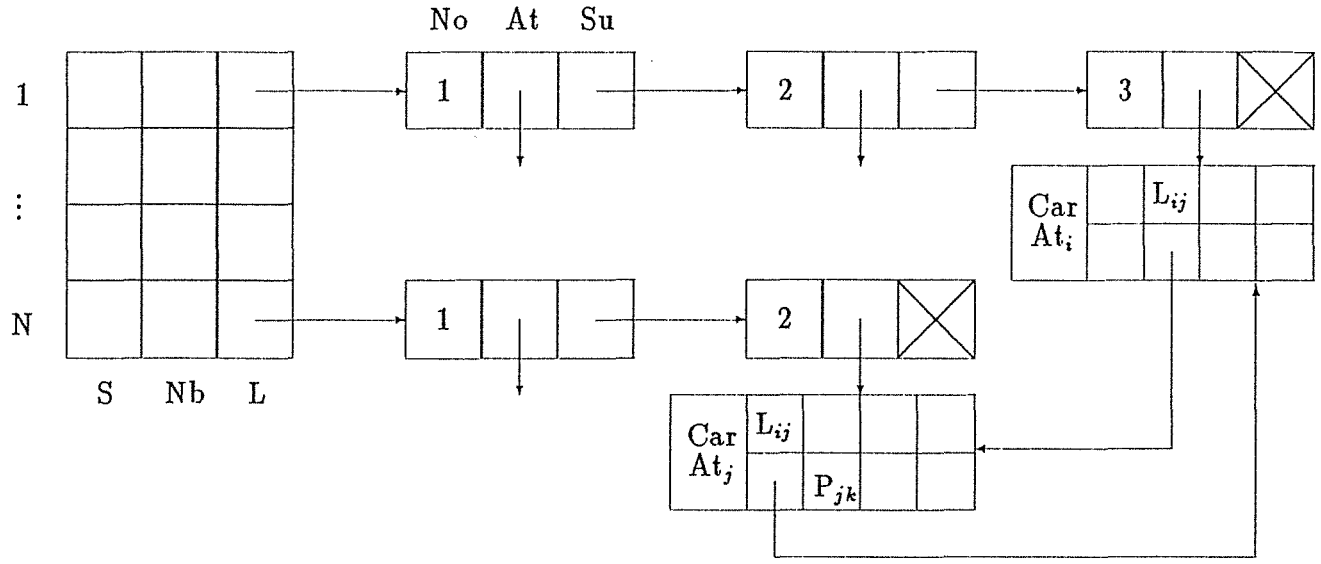
Cette représentation sous forme de graphe permet d'une part un accès rapide à un élément chimique particulier de la molécule grâce à la table des éléments et aux listes chaînées des atomes correspondantes, et d'autre part un parcours facile de toute la molécule, en ayant accès aux caractéristiques de chaque atome, par l'intermédiaire des structures liées entre elles (pointeurs) représentant les atomes et les liaisons chimiques.

Ce graphe fait partie d'une structure plus importante (voir Figure 2.1, page suivante) contenant également l'activité (qui, rappelons le, est chiffrée), l'ensemble des orbitales moléculaires calculées sur la base des orbitales atomiques -nous allons y revenir dans le paragraphe **2.2.2**-, un ensemble de grandeurs moléculaires calculées et que nous allons énumérer dans le paragraphe **2.2.3**, un ensemble de valeurs locales calculées sur l'enveloppe de van der Waals, de la manière que décrivent les paragraphes **2.2.4** et **2.2.5**, et concernant la géométrie et les propriétés électroniques de régions de la molécule, et l'ensemble des atomes faisant partie de la sous-structure commune aux autres molécules de la famille.

Représentation d'une molécule en mémoire:

- Activité
- Grandeurs moléculaires caractéristiques
- Grandeurs locales calculées sur l'enveloppe de van der Waals :
  - * géométriques
  - * électroniques
- Atomes du motif commun à la famille de molécules
- Nombre d'éléments chimiques différents (N)
- Graphe représentant les atomes et les liaisons chimiques

Graphe:




- S : Symbole chimique d'un élément de la molécule
- Nb : Nombre d'atomes de symbole S dans la molécule (numérotés de 1 à Nb)
- L : Pointeur vers la liste des atomes de symbole S
- No : Numéro de l'atome dans la liste des atomes de même symbole S
- At : Pointeur vers la structure représentant l'atome $S_{No}$
- Su : Pointeur vers l'atome de symbole S suivant
- Car $At_k$ : Caractéristiques, décrites à la page précédente, de l'atome k
- $L_{ij}$ : Type de la liaison entre l'atome i et l'atome j (simple, aromatique, double, triple)
- $P_{jk}$ : Pointeur vers la structure représentant l'atome k, voisin de l'atome j

Figure 2.1

### 2.2.2 Détermination de la géométrie la plus réaliste et calcul des propriétés électroniques

La base adoptée, pour obtenir un ensemble de descripteurs à analyser, est la détermination de la géométrie des molécules et des orbitales moléculaires qui décrivent leurs électrons. Pour déterminer la géométrie la plus réaliste et calculer les propriétés électroniques d'une molécule, plusieurs méthodes peuvent être utilisées (voir paragraphe **1.2.2**). En ce qui nous concerne, ce sont les méthodes de la chimique quantique (méthodes semi-empiriques) qui ont retenu notre attention pour trois raisons:

- Elles étaient disponibles au laboratoire [17].
- Elles donnent de meilleurs résultats que la mécanique moléculaire.
- Elles sont moins coûteuses en temps de calcul que les méthodes *ab initio*.

Le programme GEOMO [17] utilise ces méthodes dont le principe général, dans le cadre du formalisme *Hartree-Fock non restreint (UHF)*, est le suivant [5,6,36,37]:

- Les spin-orbitales moléculaires sont développées sur la base des orbitales atomiques des atomes qui composent la molécule (approximation LCAO (Linear Combination of Atomic Orbitals)):

$$\Phi_i^\alpha = \sum_\nu C_{\nu i}^\alpha \chi_\nu \tag{2.1}$$

$$\Phi_i^\beta = \sum_\nu C_{\nu i}^\beta \chi_\nu \tag{2.2}$$

- La résolution de l'équation de Schrödinger conduit à rechercher les valeurs et vecteurs propres de l'hamiltonien de Hartree-Fock dont les éléments dans la base des orbitales atomiques $\{\chi_\nu\}$ sont:

$$F_{\mu\nu}^\gamma = H_{\mu\nu} + \sum_{\lambda,\sigma} [P_{\lambda\sigma}\langle\mu\nu \mid \lambda\sigma\rangle - P_{\lambda\sigma}^\gamma\langle\mu\sigma \mid \lambda\nu\rangle] \ , \tag{2.3}$$

où

$$P_{\mu\nu}^\gamma = \sum_{OM\ occ.\ i} C_{\mu i}^\gamma \, C_{\nu i}^\gamma \quad \text{U.A.}, \tag{2.4}$$

($\gamma$ représente la fonction de spin $\alpha$ ou $\beta$) et

$$P_{\mu\nu} = P_{\mu\nu}^\alpha + P_{\mu\nu}^\beta \tag{2.5}$$

sont les éléments de la matrice densité et

$$H_{\mu\nu} = \int_\tau \chi_\mu^* \, \hat{H}_{coeur} \, \chi_\nu \, d\tau \tag{2.6}$$

$$\hat{H}_{coeur} = \sum_{e^-\ j} \left( -\frac{1}{2}\Delta_j - \sum_{noyaux\ k} \frac{Z_k}{r_{jk}} \right) \text{ U.A.} \tag{2.7}$$

sont les éléments de la matrice de l'Hamiltonien de coeur. Les intégrales d'interactions électroniques sont représentées par:

$$\langle\mu\nu \mid \lambda\sigma\rangle = \int_{\tau_1}\int_{\tau_2} \chi_\mu(1)\chi_\nu(1) \, \frac{1}{r_{12}} \, \chi_\lambda(2)\chi_\sigma(2) \, d\tau_1 d\tau_2 \quad \text{U.A..} \tag{2.8}$$

- La résolution des équations de Berthier-Pople-Nesbet [38] permet le calcul des valeurs et vecteurs propres de l'opérateur $\hat{F}$:

$$\sum_{\nu}(F^{\alpha}_{\mu\nu} - \epsilon^{\alpha}_{i} S_{\mu\nu})\ C^{\alpha}_{\nu i} \ = \ 0 \tag{2.9}$$

$$\sum_{\nu}(F^{\beta}_{\mu\nu} - \epsilon^{\beta}_{i} S_{\mu\nu})\ C^{\beta}_{\nu i} \ = \ 0 \tag{2.10}$$

où $S_{\mu\nu}$ est l'intégrale de recouvrement des orbitales atomiques $\chi_{\mu}$ et $\chi_{\nu}$, et $\epsilon^{\alpha}_{i}$ et $\epsilon^{\beta}_{i}$ sont les valeurs propres, énergies des spin-orbitales $\Phi^{\alpha}_{i}$ et $\Phi^{\beta}_{i}$ vecteurs propres associés. Dans le cas courant des systèmes à couches fermées, où chaque orbitale est occupée par deux électrons de spins opposés, nous obtenons les équations bien connues de Roothaan:

$$\sum_{\nu}(F_{\mu\nu} - \epsilon_{i} S_{\mu\nu})\ C_{\nu i} \ = \ 0 \tag{2.11}$$

En ce qui nous concerne, pour la partie application, ce sont ces dernières équations qui ont été utilisées. Leur résolution constitue le calcul SCF (champ auto-cohérent) qui est simplifié par l'application d'une des hypothèses INDO, CNDO/2, MINDO/3, MNDO [5].

- La géométrie la plus stable de la molécule est atteinte en minimisant l'énergie totale de la molécule, par rapport aux coordonnées des atomes, par la méthode des gradients conjugués [6]:

$$E_t \ = \ \frac{1}{2}\sum_{\mu,\nu}\left[P^{\alpha}_{\mu\nu}(H_{\mu\nu} + F^{\alpha}_{\mu\nu}) + P^{\beta}_{\mu\nu}(H_{\mu\nu} + F^{\beta}_{\mu\nu})\right] \ + \ \frac{1}{2}\sum_{B}\sum_{A\neq B}\frac{Z_A Z_B}{R_{AB}} \quad \text{U.A.} \tag{2.12}$$

où $R_{AB}$ est la distance entre les noyaux $A$ et $B$. Les résultats de ces calculs sur les molécules d'un lot donné donnent une très bonne approche des structures géométrique et électronique de ces molécules.

Ces résultats sont approchés compte tenu des hypothèses faites pour les obtenir, néanmoins, les erreurs sont toujours faites dans le même sens (surestimation de l'énergie) et si la même hypothèse est utilisée pour toutes les molécules, les comparaisons entre différentes grandeurs ont une signification réelle.

### 2.2.3 Calcul de descripteurs théoriques globaux

Du point de vue du chimiste, lorsque nous étudions des interactions molécule active-récepteur biologique, l'image très souvent vérifiée de la *clé* et de la *serrure* conduit à faire intervenir des caractéristiques géométriques et électroniques de la molécule.

La première idée qui vient à l'esprit est d'utiliser des grandeurs moléculaires comme le *volume de van der Waals* ($VVDW$) et l'*anisotropie* de la molécule (ou sa "forme", $FORME$) pour décrire les caractéristiques géométriques, le *moment dipolaire* ($\vec{\mu}$), le *tenseur de polarisabilité* ($\bar{\bar{\alpha}}$), l'*anisotropie de polarisabilité* ($\beta$), l'*énergie de la plus haute orbitale occupée* ($HOMO$),

l'*énergie de la plus basse orbitale inoccupée* (*LUMO*) et d'autres grandeurs plus théoriques comme la *superdélocalisabilité électrophile totale* (*SDEL*) et l'*auto-polarisabilité atomique totale* (self-atom polarisability, *SAPO*) de la molécule pour décrire les caractéristiques électroniques.

D'ailleurs, ces grandeurs moléculaires sont souvent utilisées pour établir des relations structure-activité quantitatives; citons par exemple les travaux de A. CARTIER [9], de E. J. LIEN [10], de T. BLAIR [11], de L. BUYDENS [12] et de I. LUKOVITS [39].

Les approches statistiques précédentes ont montré que ces descripteurs moléculaires, s'ils sont liés à l'activité, le sont vraisemblablement par une relation linéaire multi-variables (équation 1.1). Dans cette hypothèse, ces descripteurs globaux seront donc pris en compte bien qu'une critique en soit faite dans les paragraphes suivants.

Ces caractéristiques moléculaires sont également calculées par des méthodes de la chimie quantique après détermination de la géométrie et des orbitales moléculaires, méthodes disponibles au laboratoire sous forme de programmes [9,40,41]:

$$VVDW = \frac{1}{3}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} r^3(\theta,\phi)\ \sin\theta\ d\theta\ d\phi \tag{2.13}$$

où $r(\theta,\phi)$ est la distance de l'origine du repère à l'enveloppe de van der Waals dans la direction $(\theta,\phi)$, supposée constante pour l'élément de surface de l'enveloppe défini par l'angle solide $d\omega = \sin\theta\ d\theta\ d\phi$ (si $d\theta$ et $d\phi$ sont suffisamment petits);

$$FORME = \frac{b\,c}{a^2} \tag{2.14}$$

où $a$, $b$ et $c$ sont les moments principaux d'inertie obtenus en remplaçant les masses atomiques par les numéros atomiques (pour accroître l'influence des atomes d'hydrogène): $a$, $b$ et $c$ sont les valeurs propres, classées par ordre décroissant, du tenseur d'inertie $\bar{\bar{I}}$ dont les éléments sont:

$$I_{kl} = I_{lk} = \sum_{atomes\ i} Z_i \left(\delta_{kl} r_i^2 - r_i^k r_i^l\right) , \tag{2.15}$$

les coordonnées des atomes étant prises dans le repère du centre de gravité $G$ ($\overrightarrow{OG} = \frac{\sum_i Z_i \vec{R}_i}{\sum_i Z_i}$)

et $r_i^k$ et $r_i^l$ étant les composantes du rayon vecteur $\vec{r}_i$ de l'atome $i$ suivant les directions $k$ et $l$ ($x$, $y$ ou $z$) du système d'axes de référence $(G,x,y,z)$;

$$\vec{\mu} = \Big(\sum_{noyaux\ i} Z_i\Big)\,\overrightarrow{OG^+} - \Big(\sum_{OM\ occ.\ k} n_k\Big)\,\overrightarrow{OG^-} \quad \text{U.A.} \tag{2.16}$$

où $n_k$ est le nombre d'électrons décrits par la spin-orbitale moléculaire $\Phi_k^\gamma$, $O$ est l'origine du repère, $G^-$ et $G^+$ sont les barycentres des charges négatives et positives, et avec, dans le cas

de couches fermées (deux électrons par orbitale moléculaire), dans un système à $n$ électrons:

$$x_{G-} = \frac{1}{n}\sum_{\mu}\sum_{\nu} P_{\mu\nu}\,\langle\chi_\mu \mid x \mid \chi_\nu\rangle\,, \tag{2.17}$$

$$y_{G-} = \frac{1}{n}\sum_{\mu}\sum_{\nu} P_{\mu\nu}\,\langle\chi_\mu \mid y \mid \chi_\nu\rangle\,, \tag{2.18}$$

$$z_{G-} = \frac{1}{n}\sum_{\mu}\sum_{\nu} P_{\mu\nu}\,\langle\chi_\mu \mid z \mid \chi_\nu\rangle\,, \tag{2.19}$$

et

$$x_{G+} = \frac{\sum_i Z_i\, x_i}{\sum_i Z_i}\,, \tag{2.20}$$

$$y_{G+} = \frac{\sum_i Z_i\, y_i}{\sum_i Z_i}\,, \tag{2.21}$$

$$z_{G+} = \frac{\sum_i Z_i\, z_i}{\sum_i Z_i}\,; \tag{2.22}$$

$$\bar{\bar{\alpha}} = \frac{4}{n_v}\bar{\bar{Q}}^2\,, \tag{2.23}$$

$$\alpha = \frac{1}{3}(\alpha_{xx}+\alpha_{yy}+\alpha_{zz})\,, \tag{2.24}$$

$$Q_{kl} = \sum_{\mu\nu} P_{\mu\nu}\left[\langle\chi_\mu \mid r_k r_l \mid \chi_\nu\rangle - \frac{1}{2}\sum_{\lambda\sigma} P_{\lambda\sigma}\langle\chi_\mu \mid r_k \mid \chi_\lambda\rangle\langle\chi_\nu \mid r_l \mid \chi_\sigma\rangle\right] \text{ U.A.}\,, \tag{2.25}$$

dans le cas de couches fermées, où $r_k$ et $r_l$ représentent les composantes du rayon vecteur $\vec{r}$ suivant les directions $k$ et $l$ ($x$, $y$ ou $z$) du système d'axes de référence (les autres symboles ont la même signification que dans le paragraphe **2.2.2**), $n_v$ est le nombre d'électrons de valence;

$$\beta^2 = \frac{1}{2}\left[(\alpha_{xx}-\alpha_{yy})^2+(\alpha_{yy}-\alpha_{zz})^2+(\alpha_{zz}-\alpha_{xx})^2\right]; \tag{2.26}$$

*HOMO* et *LUMO* sont des orbitales moléculaires calculées de la façon exposée dans le paragraphe **2.2.2**;

$$SDEL = \sum_{l=1}^{n}\sum_{OM\ occ.\ i}\sum_{m=1}^{N_l}\frac{(C^l_{im})^2}{\epsilon_i} \text{ U.A.}\,; \tag{2.27}$$

$$SAPO = 4\sum_{l=1}^{n}\sum_{OM\ occ.\ i}\sum_{OM\ inocc.\ j}\sum_{a=1}^{N_l}\sum_{b=1}^{N_l}\frac{C^l_{ai}\,C^l_{aj}\,C^l_{bi}\,C^l_{bj}}{\epsilon_i-\epsilon_j} \text{ U.A.} \tag{2.28}$$

où $n$ est le nombre d'atomes dans la molécule et $N_l$ le nombre d'orbitales atomiques de valence de l'atome $l$.

Ces grandeurs calculées constituent le champ *descripteurs globaux* de la structure *molécule*. Cependant un problème ce pose: l'analyse de ces descripteurs est-elle suffisante pour rendre compte de l'activité biologique et établir des relations avec l'activité ?

C'est ce que nous allons voir maintenant.

### 2.2.4 Calcul de caractéristiques géométriques

Comme nous l'avons observé précédemment (voir paragraphe **1.3.2**), les molécules, étudiées pour une activité biologique précise, doivent avoir en commun une sous-structure particulière possèdant en général un ou plusieurs hétéro-atomes. Cependant l'influence de certains substituants est primordiale dans le mécanisme biologique d'interactions avec un récepteur. L'analyse doit donc se porter de préférence au niveau *sub-moléculaire* ou *atomique*. C'est pourquoi plusieurs auteurs [9,10,11,12] se sont intéressés à des descripteurs plus locaux caractérisant des atomes ou des groupements d'atomes comme la *charge nette* d'un atome particulier de la structure commune, la *superdélocalisabilité électrophile* d'un groupement, etc. Malheureusement le défaut de ces descripteurs est qu'ils sont très dépendants des molécules étudiées et de leur rigidité, qu'ils sont essentiellement de type électronique et qu'ils sont très dépendants de la subjectivité du chimiste. Or, compte tenu du modèle *clé-serrure* utilisé, il est certain que des contraintes stériques jouent un rôle important sinon essentiel (la gêne stérique à l'accessibilité du site actif est très souvent un facteur rédhibitoire de l'activité même si les facteurs électroniques sont favorables). Le problème a donc été de modéliser de façon générale les effets stériques locaux.

La forme d'une molécule est généralement assimilée à celle de son enveloppe de van der Waals [41] qui est déterminée en "lissant" les surfaces externes des sphères de van der Waals accollées que constituent les atomes. Les rayons de van der Waals, caractéristique intrinsèque des atomes, utilisés sont ceux qui apparaissent dans les tables et dont voici les valeurs pour les atomes les plus couramment rencontrés en chimie organique et en biologie:

| | | | | | |
|---|---|---|---|---|---|
| H<br>1.20 Å | B<br>1.73 Å | C<br>1.70 Å | N<br>1.55 Å | O<br>1.52 Å | F<br>1.47 Å |
| | | Si<br>2.10 Å | P<br>1.80 Å | S<br>1.80 Å | Cl<br>1.75 Å |
| | | | As<br>1.85 Å | Se<br>1.90 Å | Br<br>1.85 Å |
| | | | | | I<br>1.98 Å |

Tableau 2.1

Rayons de van der Waals des principaux atomes rencontrés

Citons, par exemple, les travaux de D. E. WALTERS et A. J. HOPFINGER [32] qui utilisent cette enveloppe de van der Waals dans leur analyse de formes sur la famille des carbamates

inhibiteurs de de l'acéthylcholinestérase (voir Figure 2.2 ci-dessous).

Pour modéliser les facteurs stériques au voisinage de la structure commune (motif commun), il a donc été choisi d'utiliser cette enveloppe de la façon suivante:

Pour chaque atome $A_i$ du motif commun de la molécule de référence (le chimiste utilise très souvent une molécule de référence [33] dans ce type d'étude pour effectuer les comparaisons entre molécules) est défini l'ensemble de ses voisins $\{V_j\}$.
Les deux points d'intersection ($P_{ij}$ et $P_{ji}$) de chaque droite ($A_i$ $V_j$) avec l'enveloppe de van der Waals de la molécule sont ensuite déterminés afin d'obtenir les deux distances $D_{i,ij}$ et $D_{i,ji}$ de cet atome $A_i$ à chacun des ces points (voir Figure 2.3 ci-dessous).

Les distances $D_{i,ij}$ et $D_{i,ji}$ représentent en quelque sorte les dimensions de la molécule dans les directions privilégiées, déterminées dans la molécule de référence, que sont les liaisons "atomes du motif commun-substituants".

Cette méthode permet de limiter le nombre de descripteurs géométriques à l'ensemble le plus représentatif des contraintes stériques.

Représentation de la "forme" d'une molécule:

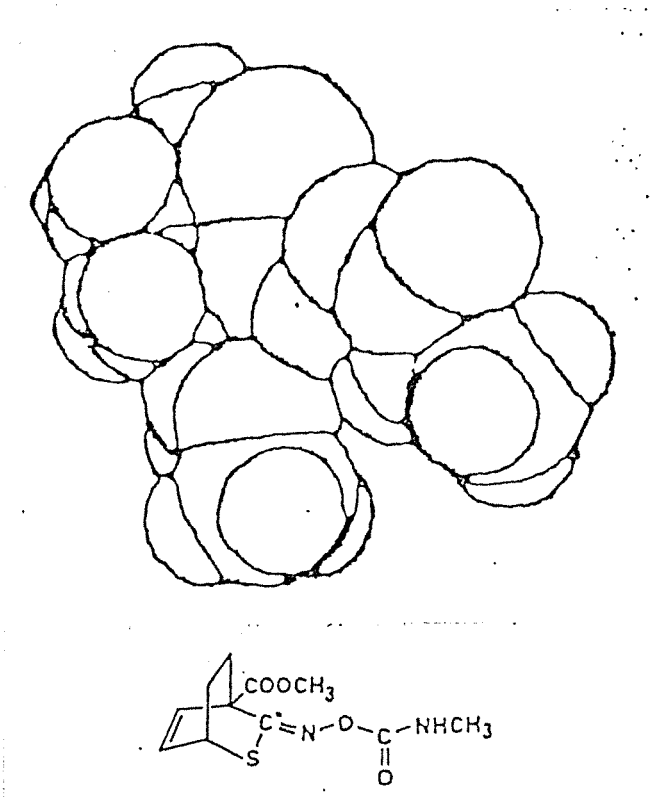
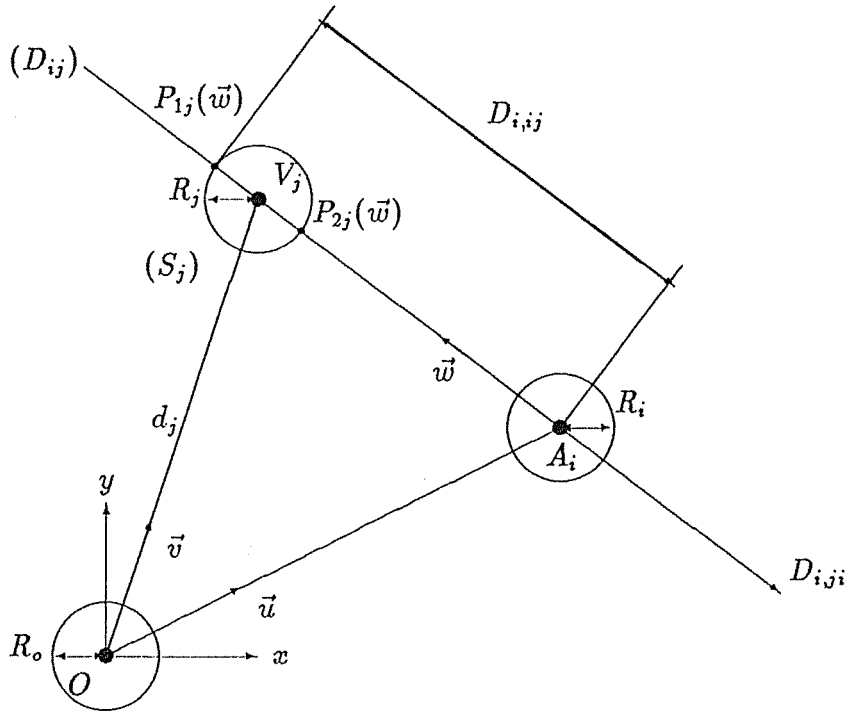



Figure 2.2

Principe de détermination des descripteurs $D_{i,ij}$:




$$(S_j) : \; x^2 + y^2 + z^2 - 2(x.x_{V_j} + y.y_{V_j} + z.z_{V_j}) + d_j^2 - R_j^2 \; = \; 0 \qquad (2.29)$$

$$(D_{ij}) : \begin{cases} x = w_x.t + x_{A_i} \\ y = w_y.t + y_{A_i} \\ z = w_z.t + z_{A_i} \end{cases} \qquad (2.30)$$

$O$ est l'atome du motif commun choisi comme origine,

$A_i$ un atome du motif commun et $V_j$ un des voisins de $A_i$.

Pour chaque atome $k$ de la molécule sont calculés les points $P_{1k}(\vec{w})$ et $P_{2k}(\vec{w})$, intersections de la droite $(D_{ij})$ (une direction privilégiée) et de la sphère de van der Waals $(S_k)$. La valeur de $D_{i,ij}$ ($i$: à partir de $A_i$; $ij$: direction $A_i \rightarrow V_j$) sera la valeur maximale des mesures algébriques $\overline{A_iP_{1k}(\vec{w})}$ et $\overline{A_iP_{2k}(\vec{w})}$; $D_{i,ji}$ correspondant à l'opposé de la valeur minimale de ces mesures algébriques.

Remarque: Pour simplifier la figure, nous n'avons pas mis d'atome derrière l'atome $V_j$ lié à $A_i$ et nous avons pris l'exemple de la molécule de référence.

Figure 2.3

Le modèle utilisé conduit le chimiste à schématiser les variations de l'activité en fonction d'un descripteur géométrique qui lui est corrélé de la façon suivante:

Il ne faut pas qu'un morceau de la *clé* soit trop gros ni trop petit sinon l'ajustement à la *serrure* ne peut se faire. C'est ce qu'indique la figure 2.4, page suivante.

Ce type de descripteurs géométriques ($D_{i,ij}$) est donc caractérisé par les deux bornes $D_{i,ij}^{min}$ et $D_{i,ij}^{max}$ qu'il faut déterminer. Précisons également que l'activité peut dépendre de plusieurs paramètres $D_{i,ij}$ différents; dans ce cas, nous pouvons aboutir à une conclusion du type:

"Une molécule sera active si la valeur du descripteur $D_{k,kl}$ est comprise entre $D_{k,kl}^{min}$ et $D_{k,kl}^{max}$ et si la valeur du descripteur $D_{m,mn}$ est comprise entre $D_{m,mn}^{min}$ et $D_{m,mn}^{max}$."

Variations de l'activité en fonction d'un paramètre géométrique corrélé:

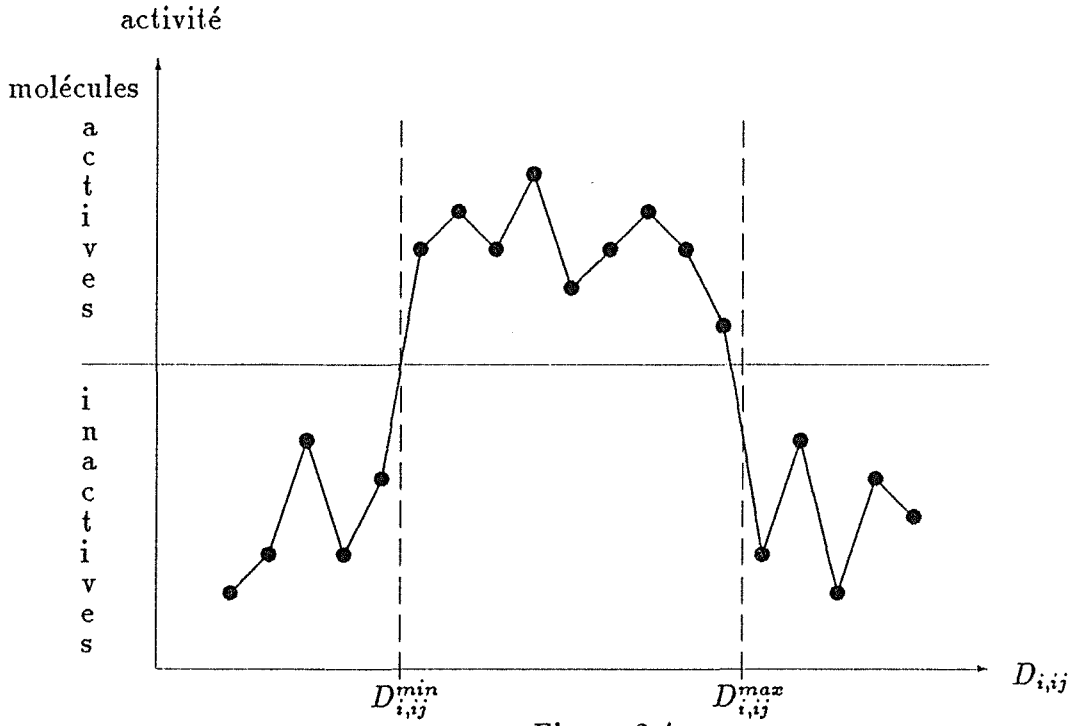



Figure 2.4

### 2.2.5 Calcul de caractéristiques électroniques locales

Pour modéliser les effets électroniques locaux, mis en jeu lors d'interactions molécule active-récepteur biologique, et dus aux substituants de la sous-structure commune et aux atomes qui la composent, il a été décidé, en se référant à un modèle d'interactions électrostatiques inter-moléculaires, de calculer, également en ces points $P_{ij}$ et $P_{ji}$, le champ et le potentiel électriques. Cette tâche est réalisée par des sous-programmes de GEOMO disponibles au laboratoire. Rappelons simplement que le potentiel et le champ électriques s'expriment en fonction des orbitales moléculaires (ici en Unités Atomiques):

$$V_{noyaux}(M) = \sum_{noyaux\ i} \frac{Z_i}{R_{iM}} \tag{2.31}$$

$$\vec{E}_{noyaux}(M) = \sum_{noyaux\ i} \frac{Z_i\ \vec{R}_{iM}}{R_{iM}^3} \tag{2.32}$$

$$V_{élec}(M) = -\sum_{e^- j} \langle \Phi^\gamma(j) \mid \frac{1}{r_{jM}} \mid \Phi^\gamma(j) \rangle \tag{2.33}$$

$$\vec{E}_{élec}(M) = -\vec{\nabla} V_{élec}(M) \tag{2.34}$$

$$V(M) = V_{noyaux}(M) + V_{élec}(M) \tag{2.35}$$

$$\vec{E}(M) = \vec{E}_{noyaux}(M) + \vec{E}_{élec}(M) \tag{2.36}$$

où $\Phi^\gamma(j)$ est la spin-orbitale décrivant l'électron $j$ et $M$ le point où sont calculés le potentiel $V(M)$ et le champ $\vec{E}(M)$ électriques.
Dans le cas de l'approximation LCAO et pour un système en couches fermées (deux électrons ($e^-$) par orbitale moléculaire), le potentiel électronique devient:

$$V_{élec}(M) = -\sum_{\mu} \sum_{\nu} P_{\mu\nu} \langle \chi_\mu \mid \frac{1}{r_M} \mid \chi_\nu \rangle \quad \text{U.A..} \tag{2.37}$$

L'importance du potentiel et du champ électriques au niveau de l'enveloppe de van der Waals de la molécule a, par ailleurs, été mise en évidence dans l'étude des interactions *soluté-solvant* (calcul de chaleur de solvatation, de solubilité, etc.) [42].
De même, G. H. LOEW utilise des surfaces équipotentielles dans ses études théoriques de relations structure-activité [43].
Ceci conforte leur utilisation dans la description d'interactions molécule active-récepteur biologique.

Les études statistiques prenant en compte d'autres descripteurs électroniques locaux (voir début du paragraphe **2.2.4**) font supposer que, si le potentiel et/ou le champ électriques sont liés à l'activité, ils le sont probablement par une relation linéaire multi-variables:

$$A = \sum_k C_k \, Q_{élec}(P_{ij}) + Cste \tag{2.38}$$

où $Q_{élec}(P_{ij})$ est une grandeur électronique locale (potentiel ou champ) calculée en un point $P_{ij}$ décrit dans le paragraphe **2.2.4** précédent.

Signalons également d'autres facteurs électroniques locaux pris en compte mais qui sont plus ou moins liés aux précédents: les *charges nettes* des atomes constituant la sous-structure commune. La charge nette $q_A$ d'un atome $A$ s'exprime, en fraction de la charge d'un électron, de la manière suivante (approximation LCAO):

$$q_A = Z_A - \sum_{OM \ occ. \ i} n_i \sum_{\mu \in A} \left[ C_{i\mu}^2 + \sum_{B \neq A} \sum_{\nu \in B} C_{i\mu} C_{i\nu} S_{\mu\nu} \right] \tag{2.39}$$

où $n_i$ est le nombre d'electrons décrits par l'orbitale moléculaire $\Phi_i$ (1 ou 2).

Remarquons que si une grandeur $Q_{élec}(P_{ij})$ est un facteur déterminant pour l'activité et

non en concurrence avec un paramètre géométrique, la relation linéaire $A = \alpha\, Q_{élec}(P_{ij}) + \beta$ sera un bon critère de discrimination entre molécules actives et molécules inactives; nous allons y revenir.

Cette remarque s'applique également aux descripteurs globaux décrits dans le paragraphe **2.2.3.**

La figure 2.5 ci-dessous montre les variations de l'activité biologique en fonction d'un paramètre électronique local ou global qui lui est corrélé et qui peut discriminer les deux classes de molécules.

Variations de l'activité en fonction d'un paramètre électronique corrélé:

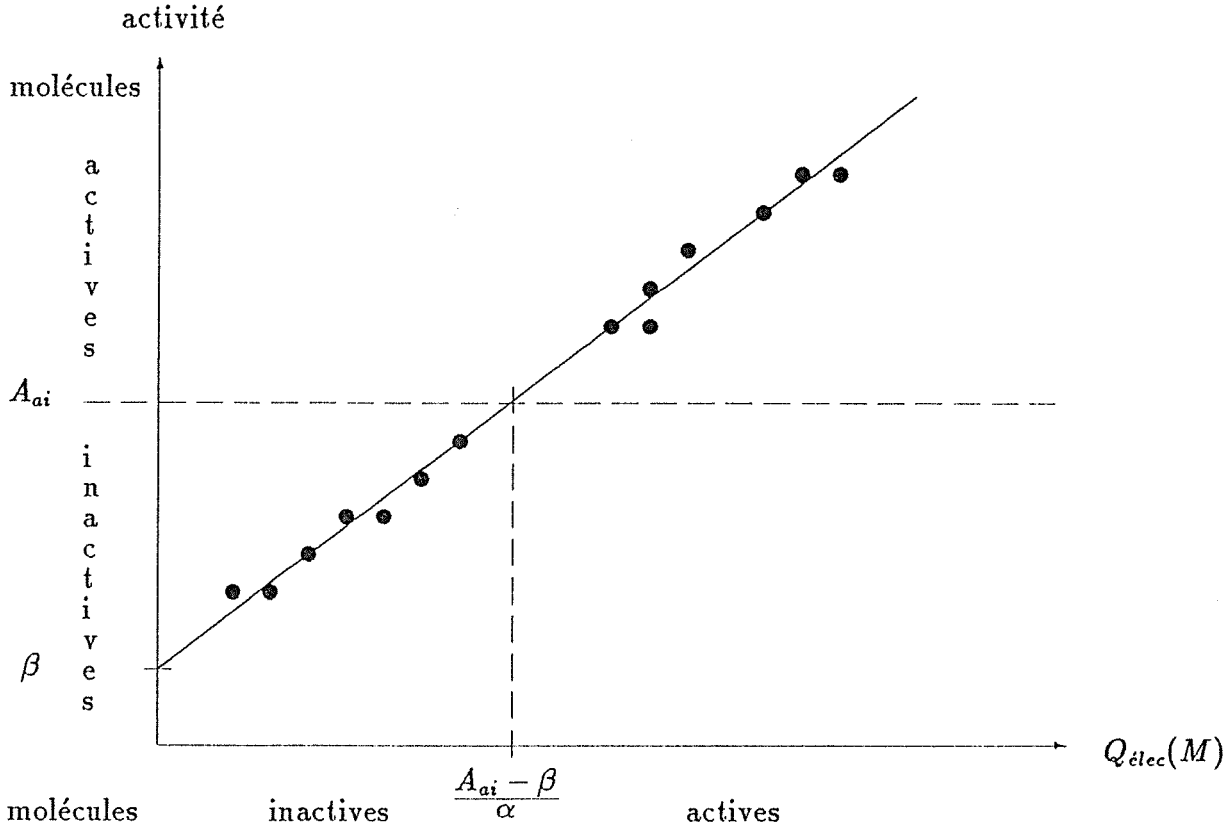



Figure 2.5

## 2.3 La phase d'apprentissage

Dans tout programme d'apprentissage de ce type, la première étape est de décrire correctement et complètement la connaissance préalable et de la classer si elle est hiérarchisée. Cette étape consiste également à mettre au point une stratégie d'utilisation de cette connaissance.

La deuxième étape est de trouver une bonne représentation des exemples en utilisant - et

c'est ce que nous avons fait pour le logiciel SARAH - des descripteurs sous forme booléenne ou en tout cas sous une forme permettant de leur attribuer facilement une note sur leur caractère discriminant.

La troisième étape est de déterminer une bonne *fonction score* permettant de sélectionner les meilleurs descripteurs et d'éditer les règles de discrimination des objets de classes différentes.

Il est donc bon ici de résumer la connaissance du chimiste dans le domaine des interactions molécule active-récepteur biologique:

- le modèle *clé-serrure* est, pour le moment, la meilleure façon de décrire ces interactions et permet d'établir les règles qui suivent,
- présence d'une sous-structure particulière, possèdant un ou plusieurs hétéro-atomes, chez toutes les molécules étudiées pour une activité donnée,
- les caractéristiques géométriques des substituants de cette sous-structure d'abord, leurs propriétés électroniques ensuite et les caractéristiques moléculaires enfin, sont les paramètres parmi lesquels se trouvent les descripteurs discriminant les deux classes de molécules (les actives et les inactives).

En ce qui concerne la stratégie d'analyse, la démarche est la suivante:

- recherche de la sous-structure commune (ou motif commun) aux molécules du lot d'apprentissage avec consultation éventuelle de l'utilisateur chimiste pour valider la sous-structure déterminée, et en s'intéressant d'abord à la mise en correspondance de motifs contenant un ou plusieurs hétéro-atomes (points d'ancrage),
- recherche (molécule ayant l'activité la plus proche de l'activité moyenne) ou demande à l'utilisateur de la molécule de référence qui va définir les directions privilégiées dans lesquelles sont calculées les grandeurs électroniques et géométriques locales (voir paragraphes **2.2.4** et **2.2.5**),
- détermination, dans cette sous-structure, d'un repère orthonormé lié à la molécule et translatable d'une molécule à l'autre afin de pouvoir analyser un ensemble de descripteurs cohérents,
- construction d'un ensemble de descripteurs locaux concernant la géométrie et les propriétés électroniques des substituants du motif commun en calculant, pour chaque atome $A_i$ du motif, les grandeurs $D_{i,ij}$, $D_{i,ji}$, $V(P_{ij})$, $V(P_{ji})$, $\vec{E}(P_{ij})$ et $\vec{E}(P_{ji})$ - les directions associées à ces grandeurs sont les directions privilégiées définies dans la molécule de référence (couples $(A_i, V_j)$, $j$ allant de 1 au nombre de voisins de $A_i$) - , ces descripteurs s'ajoutant aux grandeurs moléculaires calculées par ailleurs,
- la recherche des descripteurs discriminants doit d'abord se faire à travers l'ensemble des paramètres géométriques locaux calculés puis binarisés (voir paragraphe **2.3.3** suivant) car, si certains de ces descripteurs sont liés à l'activité, ils sont beaucoup plus discriminants que les autres, et pour peu que le *score* obtenu par un de ces paramètres soit supérieur à un seuil de

satisfaction ajustable par l'utilisateur, il est inutile d'explorer les descripteurs électroniques locaux et globaux.

Cette connaissance préalable, cette modélisation des descripteurs sous une forme booléenne et cette stratégie ont été intégrées au logiciel SARAH et sont développées dans ce qui suit.

### 2.3.1 Recherche de sous-structures communes

Comme nous l'avons remarqué, compte tenu du modèle utilisé, il est important de pouvoir faire apparaître ce qu'il y a de commun entre les molécules d'un ensemble, dotées de la même propriété biologique. Cette similitude peut se situer à différents niveaux:

- existence d'un même assemblage d'atomes,
- existence d'assemblages d'atomes différents mais possèdant des caractéristiques physico-chimiques comparables (géométrie, propriétés électroniques, etc.) qui doivent être précisées au cours de la comparaison et peuvent être définies de façon plus ou moins restrictive (atomes de même nature ou d'électronégativités voisines, etc.).

La méthode de recherche de la sous-structure commune (ou du motif commun) mise au point réalise la comparaison des graphes, que constituent les molécules, à différents niveaux en ce qui concernent les nœuds des graphes (les atomes). Pour les liens entre ces nœuds, les liaisons chimiques (simples, doubles ou aromatiques et triples) ayant des réactivités différentes, il a été choisi de ne mettre en correspondance entre les graphes que des liaisons du même type, les liaisons doubles et aromatiques composant un même type. Par contre, la comparaison au niveau des atomes est plus souple:

- Un premier niveau de recherche de la sous-structure, est la mise en correspondance entre les graphes, d'atomes identiques (ayant le même numéro atomique).
- Dans le deuxième niveau, ce sont les propriétés électroniques des atomes qui sont comparées, c'est à dire que des atomes, même de natures différentes, peuvent être appariés pour peu qu'ils aient des électronégativités (relatives par rapport au carbone (électronégativité absolue du carbone: 2.5 Pauling): atomes plus ou moins électronégatifs que le carbone) voisines (de même signe plus exactement), c'est par exemple le cas du fluor (1.5 Pauling), de l'oxygène (1 Pauling), de l'azote (0.5 Pauling) et du chlore (0.5 Pauling) qui peuvent être appariés; dans ce niveau, la possibilité de mettre ou non en correspondance des atomes d'électronégativités voisines, mais n'appartenant pas à la même ligne du tableau périodique, est également offerte.
- Le troisième niveau concerne la géométrie ou plus précisément le degré de coordination des atomes, il est possible, dans ce niveau, d'apparier des halogènes différents, des atomes d'oxygène avec des atomes de soufre (groupements C=O et C=S ou -O- et -S-) et des atomes d'azote de groupements amines avec des atomes de phosphore de groupements phosphines; ce niveau permet la mise en évidence d'un motif possèdant une certaine forme, sans pour autant imposer une recherche trop contraignante.

Une des raisons pour lesquelles la représentation des molécules sous forme de graphes a été choisie, est la mise au point de l'algorithme de recherche de la sous-structure commune (recherche de sous-graphes communs à plusieurs graphes).

Lors de la détermination du motif commun, plusieurs possibilités s'offrent à l'utilisateur:
- choix d'un des niveaux de recherche précédemment énoncés,
- prise en compte ou non des atomes d'hydrogène dans le motif,
- spécification d'un type d'atomes particulier (généralement un hétéro-atome) devant se trouver dans la sous-structure,
- sélection d'un motif parmi plusieurs éventuellement trouvés,
- modification du motif obtenu en supprimant un ou plusieurs atomes.

Il faut également préciser que le motif commun doit avoir des caractéristiques minimales qui sont évidentes:
- un nombre minimum d'atomes (fixé par exemple à deux),
- présence d'au moins un hétéro-atome dans le motif (point d'ancrage).

Pour décrire plus précisément l'algorithme mis au point, précisons d'abord qu'il s'agit d'une fonction (SOUS_GRAPHE) prenant comme données deux graphes et retournant le sous-graphe commun.

Pour déterminer la sous-structure commune à toutes les molécules, le programme procède de la façon suivante :
- recherche du motif commun aux deux premières molécules $G_1$ et $G_2$ (avec éventuellement l'intervention de l'utilisateur dans le cas où plusieurs solutions sont possibles) grâce à la fonction SOUS_GRAPHE et donc construction d'un premier sous-graphe $SG_1$,
- recherche du sous-graphe $SG_i$ commun au graphe $SG_{i-1}$ et à la molécule $G_{i+1}$ pour $i$ allant de 2 au nombre de molécules moins un.

La fonction SOUS_GRAPHE, pour rechercher le sous-graphe commun aux deux graphes $G_1$ et $G_2$ passés en paramètres, fait appel à TESTE_ET_CONSTRUIT_GRAPHE, fonction récursive prenant comme données deux structures représentant deux atomes: $A_1$ du graphe $G_1$ et $A_2$ de $G_2$, et retournant un ensemble de sous-graphes possibles à partir de $A_1$ et $A_2$ si ces deux atomes peuvent être mis en correspondance compte tenu des critères de recherche choisis par l'utilisateur.

Au premier appel de TESTE_ET_CONSTRUIT_GRAPHE, dans SOUS_GRAPHE, les deux atomes appariés sont deux hétéro-atomes (points d'ancrage). Au nième appel, la fonction récursive vérifie si, compte tenu des exigences de l'utilisateur, les deux atomes passés en paramètres peuvent être appariés, si c'est le cas, pour chaque atome voisin de $A_1$, est

recherché un homologue dans l'ensemble des voisins de $A_2$ en tenant compte du type de liaison chimique entre l'atome $A_i$ et ses voisins. Si cet homologue existe, la fonction s'appelle récursivement avec pour paramètre ce nouveau couple. Si plusieurs voisins de $A_2$ sont envisageables pour un voisin de $A_1$, ils sont pris en compte, ce qui correspond à autant d'appels internes et donc de sous-graphes. Notons également que, pour éviter les bouclages infinis, chaque atome est marqué après avoir été apparié. Au retour de la fonction récursive, dans la fonction SOUS_GRAPHE, si plusieurs motifs ont été mis en évidence, il est demandé à l'utilisateur de choisir celui qu'il juge le meilleur.

Une fois la sous-structure établie, se pose le problème de la "superposition" des molécules; c'est le point que nous allons aborder maintenant.

### 2.3.2 Mise en correspondance des sous-structures communes

Pour pouvoir analyser un ensemble de descripteurs cohérents, il est nécessaire de les calculer dans un même système d'axes, c'est à dire que la sous-structure commune à toute les molécules soit munie d'un repère orthonormé dans lequel seront calculées les coordonnées des atomes.

Le motif commun trouvé constitue également un graphe analogue à celui d'une molécule. Ce graphe tout entier est recherché dans chaque molécule du lot grâce à l'algorithme, cité dans le paragraphe **2.3.1** précédent, en partant du ou des hétéro-atomes ou points d'ancrage qu'il possède et en utilisant le même niveau de recherche qui a permis de l'établir. Les atomes du motif sont alors numérotés de la même façon dans chaque molécule.

Les comparaisons des molécules se faisant à partir de la molécule de référence choisie par l'utilisateur ou déterminée automatiquement, un repère orthonormé est automatiquement déterminé dans la sous-structure trouvée dans la molécule de référence, à partir de trois atomes non alignés en utilisant de préférence les points d'ancrage.
Ces trois atomes étant déterminés, il est alors possible de faire le changement de repère dans toutes les molécules du lot d'apprentissage.
A partir de la molécule de référence, les descripteurs locaux (décrits dans les paragraphes **2.2.4** et **2.2.5**) peuvent alors être calculés dans chaque molécule.

Après la détermination de ces descripteurs qui, soulignons-le encore, est automatique, c'est à dire non soumise à la subjectivité du chimiste et non fonction de la série de molécules (pour une autre famille de molécules, d'autres paramètres seront calculés en se basant simplement sur le modèle), la phase d'apprentissage proprement dite peut commencer.

### 2.3.3 Détermination des règles de discrimination - Construction d'un arbre de décision

Comme cela a déjà été dit, il est nécessaire de binariser les descripteurs calculés. Rappelons également qu'une donnée biologique supplémentaire est fournie par l'utilisateur au logiciel, à savoir la valeur limite de l'activité ($A_{ai}$) définie comme suit (Figure 2.6 ci-dessous):

La donnée biologique supplémentaire:

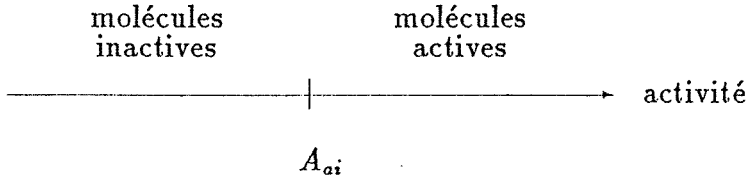



Figure 2.6

Cette donnée permet de mettre sous une forme booléenne l'activité.

Pour binariser les descripteurs des molécules, il est nécessaire de déterminer, pour les paramètres géométriques locaux $D_{i,ij}$, les bornes $D_{i,ij}^{min}$ et $D_{i,ij}^{max}$ (voir Figure 2.4, paragraphe **2.2.4**), et pour les autres paramètres (essentiellement électroniques), la borne $B_{ai} = \frac{A_{ai} - \beta}{\alpha}$ (voir Figure 2.5, paragraphe **2.2.5**). Ainsi, pour tout descripteur, il est possible de définir deux états de *position*:

- pour tout descripteur géométrique local $D_{i,ij}$: *entre* les bornes $D_{i,ij}^{min}$ et $D_{i,ij}^{max}$ ou *à l'extérieur*,
- pour tout descripteur électronique: *avant* la borne $B_{ai}$ ou *après*.

Autrement dit, il existe deux *spécialisations* d'un concept que nous pouvons nommer *position* s'appliquant à tous les descripteurs:

- un premier concept que nous appellerons *position-descripteurs géométriques*,
- un second concept nommé *position-descripteurs électroniques*.

Ces deux concepts sont des *généralisations* des concepts de descripteurs (grandeurs calculées dans des directions particulières) notés *position-type de descripteurs-descripteur* D qui sont eux mêmes des *généralisations* des concepts de positions des valeurs des descripteurs par rapport à leurs bornes caractéristiques: *position-type de descripteurs-descripteur* D-*position* 1 et *position-type de descripteurs-descripteur* D-*position* 2, *position* 1 et *position* 2 étant soit *entre* et *à l'extérieur*, soit *avant* et *après*. Les exemples sont évidemment des *spécialisations* de tout ces concepts (voir Figure 2.7, page suivante).

L'apprentissage consiste à trouver le concept le plus général ou les concepts les plus généraux discriminant au mieux les deux classes d'exemples, c'est à dire discriminant le concept *activité*, *généralisation* des deux concepts *active* et *inactive* que *spécialisent* également les exemples.

Notons que le concept le plus général, c'est à dire vérifié par tous les exemples, est le concept que nous appellerons *position-activité*. Quelque soit l'exemple, il possède une *activité* (la molécule est soit *active* soit *inactive*) et quelque soit un descripteur particulier de l'exemple, il est toujours possible de le caractériser par une ou deux bornes et donc de définir deux positions particulières (soit *entre* et *à l'extérieur*, soit *avant* et *après*) (spécialisations du concept *position*) dont une caractérise l'exemple.

L'arbre des concepts sur les molécules:

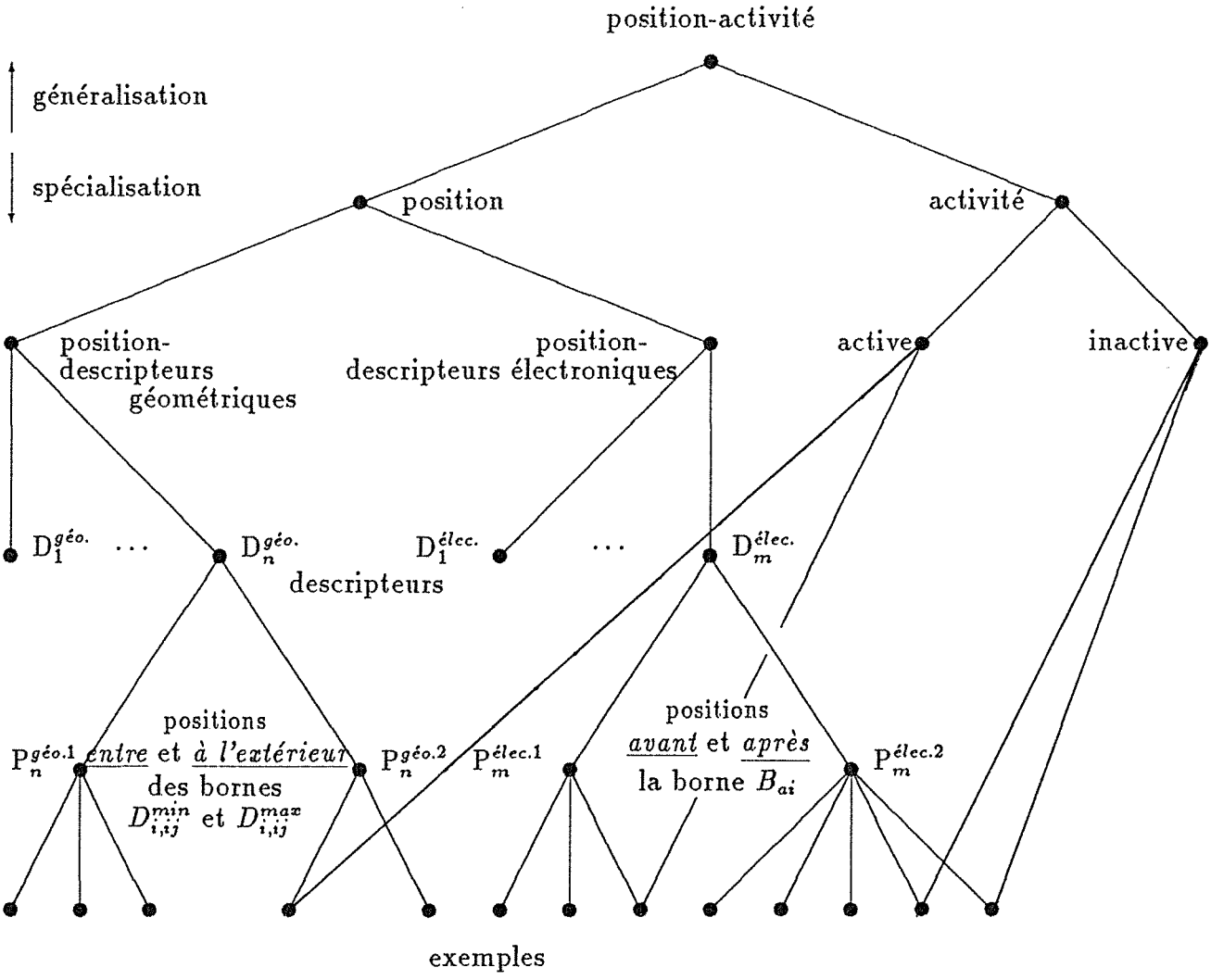



Figure 2.7

Cette technique d'apprentissage, qui se schématise par un arbre des concepts sur les exemples comme celui utilisé dans notre étude et représenté par la figure 2.7 sur la page précédente, est souvent utilisée en intelligence artificielle [59,60]. Citons par exemple les travaux d'O. GASCUEL (programme PLAGE) [44,45,61].

Pour rechercher les concepts les plus généraux discriminant au mieux les deux classes de molécules, il est nécessaire d'attribuer une *note* à chaque concept. La meilleure *fonction score* que l'on puisse obtenir en spécialisant les descripteurs est, comme le fait remarquer O. GASCUEL [45], la mesure du $\chi^2$:

Pour un concept booléen $C$ donné, sont dénombrés les exemples de la première classe vérifiant $C$, ne vérifiant pas $C$, les exemples de la seconde classe vérifiant $C$, ne vérifiant pas $C$; ce qui est résumé par le tableau de contingence $T$ suivant:

| Concept $C$ | actif | inactif |
|---|---|---|
| vrai | $a$ | $b$ |
| faux | $c$ | $d$ |

Un concept "idéal" est un concept possèdant un tableau $T$ dont une des deux diagonales contient deux zéros. La *note* (ou le *score*) attribuée à $C$ et qui correspond à son aptitude à discriminer les deux classes est:

$$\begin{aligned} N &= a+b+c+d \\ F_{vrai} &= a+b \\ F_{faux} &= c+d \\ F_{actif} &= a+c \\ F_{inactif} &= b+d \\ score &= \frac{1}{N}\sum_{i=vrai}^{faux}\sum_{j=actif}^{inactif}\frac{\left(T_{ij}-\frac{F_iF_j}{N}\right)^2}{\frac{F_iF_j}{N}} \end{aligned} \tag{2.40}$$

A chaque descripteur binarisé ($D_{i,ij}$, $Q_{élec}(M)$, descripteur global) est donc attribuée une *note*.

En fait, à chaque étape de la construction de l'arbre de décision (arbre des concepts généraux les plus discriminants), tous les *scores* ne sont pas évalués. En effet, l'heuristique suivante:

"si un concept de type *géométrique* obtient un *score* supérieur à un *seuil de satisfaction* ajustable alors il est retenu et l'exploration de l'ensemble des autres types de concepts n'est plus envisagée pour cette étape",

permet d'accélérer la construction de l'arbre. Cette heuristique est basée sur une des règles de la connaissance préalable qui spécifie que lorsqu'un critère géométrique est fortement lié à l'activité, il est rédhibitoire. Cette heuristique est également appliquée aux autres types de descripteurs, après avoir envisagé tous les descripteurs géométriques locaux, car ils sont explorés par ordre décroissant d'importance (voir introduction de ce paragraphe **2.3**).

La construction de l'arbre de décision se fait de la manière suivante:

- Initialement, disposant de tous les exemples du lot d'apprentissage, une première recherche dans tout l'espace des descripteurs permet de déterminer le plus discriminant (celui ayant obtenu le meilleur *score*).
- L'ensemble des exemples est alors divisé en deux sous-ensembles: molécules ayant des descripteurs vérifiant le concept trouvé et molécules n'ayant pas de descripteur vérifiant ce concept.
- Cette démarche est alors appliquée de nouveau au deux sous-ensembles obtenus en éliminant de l'espace des concepts envisageables, celui qui vient d'être sélectionné.

Ce processus récursif s'arrête lorsque nous avons, soit obtenu satisfaction, c'est à dire une règle $R$ de séparation des exemples des deux classes (taux de majorité $T_R$ de molécules d'une même classe supérieure à un taux seuil $S_1$, donné par l'utilisateur, dans un sous-ensemble), soit un nombre insuffisant de molécules dans un sous-ensemble, où la satisfaction n'est pas atteinte, pour continuer la séparation (nombre inférieur à un certain pourcentage $S_2$, également fourni par l'utilisateur, du nombre d'exemples initial) - les molécules d'un tel sous-ensemble peuvent être considérées comme non classées par manque d'exemples -.
Les feuilles de l'arbre de décision correspondent donc aux sous-ensembles qui viennent d'être décrits; il est alors possible, à ce stade, de prédire qualitativement l'activité d'une molécule nouvelle en parcourant l'arbre de décision pour la placer dans un des sous-ensembles.

Les règles obtenues (une règle est la suite de branches de l'arbre partant de la racine (l'ensemble du lot d'apprentissage) vers une feuille particulière (un des sous-ensembles décrits ci-dessus)) sont donc des *conjonctions* de critères discriminants. A chaque règle $R$, il est possible d'affecter un pourcentage de validité $PV_R$ défini comme suit:
$PV_R$ est le nombre $N_{CR}$ d'exemples de la classe majoritaire $C$ dans le sous-ensemble caractéristique de $R$ (de cardinal $N_R$) divisé par le nombre total $N_C$ d'exemples de $C$ et multiplié par le taux $T_R = \frac{N_{CR}}{N_R}$:

$$PV_R = 100\frac{N_{CR}^2}{N_C N_R}\ \%. \tag{2.41}$$

Nous pouvons alors dire qu'une molécule nouvelle vérifiant la règle $R$ a $PV_R\%$ de chance d'appartenir à la classe $C$.

L'étape suivante consiste donc à établir, cette fois, des relations quantitatives entre l'activité

chiffrée et les descripteurs également quantitatifs.

### 2.3.4 Détermination de corrélations entre facteurs électroniques et activité

Puisque les molécules, appartenant à des sous-ensembles de l'arbre de décision différents, vérifient des règles différentes, il était logique d'établir autant de relations quantitatives que de sous-ensembles.

Comme dans la grande majorité des travaux, les relations envisagées sont des régressions linéaires multi-variables qui ne sont, parfois, pas trop mauvaises (en particulier si elles sont obtenues sur un ensemble restreint d'exemples voisins), nous avons décidé de garder ce type de corrélations.

Compte tenu de la connaissance préalable d'une part et du modèle de variation de l'activité en fonction d'un paramètre géométrique local qui lui est lié d'autre part, ce type de descripteurs ne sera pas envisagé dans les régressions qui ne porteront donc que sur les descripteurs de type électronique.

Le principe de la méthode utilisée est celui décrit dans le livre de E. DIDAY [21]:
- recherche du descripteur le mieux corrélé à l'activité,
- recherche des autres descripteurs qui améliorent cette corrélation et qui répondent aux caractéristiques suivantes: descripteurs non corrélés aux précédents déjà trouvés et qui possèdent des intervalles, dits de confiance, dont les largeurs sont inférieures à une certaine fraction ajustable de leurs valeurs. Pour chaque sous-ensemble de l'arbre de décision est alors calculée une corrélation (équation 2.38, paragraphe **2.2.5**). Il est alors possible de prédire, cette fois quantitativement, l'activité d'une molécule nouvelle.

## 2.4 Analyse d'une molécule nouvelle - Mise à jour des critères de discrimination

Cette partie porte sur l'utilisation du logiciel SARAH en tant qu'outil de prévision de propriétés biologiques et d'amélioration des règles de discrimination et/ou des régressions associées.
Le programme est, en effet, bien sûr capable de prédictions mais aussi de s'adapter, s'il y a lieu, à la nouvelle structure d'une molécule, soit non testée biologiquement, soit constituant un nouvel exemple, par une mise à jour de l'arbre de décision, mise à jour qui doit être confirmée par l'utilisateur.

### 2.4.1 Estimation de l'activité d'une molécule non encore testée biologiquement

L'analyse d'une molécule nouvelle est simple:

- La première étape (après avoir réuni toutes les données concernant sa structure et calculées par des programmes annexes comme GEOMO) consiste à évaluer, par rapport à la molécule de référence et à la sous-structure origine de l'activité qui a été repérée dans cette nouvelle molécule (voir paragraphe **2.3.2**), les descripteurs à considérer.
Dans cette première étape, notons que si le motif commun, défini lors de l'apprentissage, ne se retrouve pas exactement dans la molécule nouvelle, une autre sous-structure commune peut être mise en évidence après approbation de l'utilisateur puis un nouvel arbre de décision peut alors être construit.

- Une fois la molécule à tester modélisée (descripteurs calculés et binarisés), elle est analysée en regardant à quelle(s) branche(s) de l'arbre de décision correspond la valeur booléenne d'un descripteur de cette molécule *spécialisation* d'un concept discriminant. Certaines branches de l'arbre sont alors parcourues jusqu'à placer la molécule dans un des sous-ensembles feuilles de l'arbre, c'est à dire, déterminer la règle $R$ qu'elle vérifie. Ceci permet de donner sa tendance (active ou inactive) avec un certain pourcentage de validité $PV_R\%$.

- Enfin, la régression linéaire multi-variables, déterminée dans le sous-ensemble en question, permet de chiffrer son activité et le degré de corrélation (coefficient de corrélation) entre l'activité et les descripteurs structuraux.

### 2.4.2 Mise à jour des règles de discrimination et/ou des corrélations associées

#### Remarques préalables

Toutes les données (activités et structures des molécules) et tous les résultats (arbre de décision et prédictions sur des molécules non testées biologiquement) sont mémorisés sur fichiers gérés par le logiciel SARAH.

#### Confirmation ou infirmation d'une prédiction

Après avoir prédit l'activité d'une molécule, le chimiste peut disposer du résultat du test biologique. Il peut alors le soumettre au programme comme une confirmation ou une infirmation de la prédiction.

Dans le cas de la confirmation d'une règle, le logiciel affine la corrélation associée en prenant en compte les valeurs des descripteurs de cette molécule.

Dans le cas d'une infirmation, le lot d'apprentissage, après adjonction de la molécule nouvelle, est analysé complètement en reconstruisant l'arbre de décision puis en calculant les nouvelles régressions linéaires multi-variables associées aux feuilles.
Deux résultats peuvent alors être obtenus:

- soit de nouvelles règles ont été trouvées,

- soit l'arbre de décision n'a pas changé, ce qui signifie que la molécule est considérée comme une exception puisque le pourcentage $T_R$ de la règle $R$ qu'elle suit est toujours supérieur au seuil $S_1$ fixé par l'utilisateur. Dans ce cas, seule la corrélation associée à $R$ est modifiée.

### Adjonction d'une nouvelle molécule d'activité connue

Le logiciel procède de la manière suivante:

- La molécule est d'abord analysée comme si son activité était inconnue (voir paragraphe **2.4.1**).

- Si l'activité biologique est en accord avec les prédictions, seule la corrélation associée à la règle vérifiée est affinée (le pourcentage de validité de la règle est également modifié).

- Si la prédiction est mauvaise, nous nous retrouvons dans le cas décrit plus haut: soit la molécule est considérée comme une exception, soit de nouvelles règles sont recherchées.

### Explication de l'exception

Les exemples constituent une connaissance incomplète, c'est à dire une description non exhaustive du concept à apprendre. Cependant, les règles déduites de cette connaissance sont tout à fait cohérentes par rapport au lot d'apprentissage. Cette constatation est d'ailleurs le propre de tous les programmes d'apprentissage à partir d'exemples. Il est donc possible de rencontrer des échecs lors de prévisions. Ces échecs peuvent, par la suite (données biologiques nouvelles), soit provoquer une remise en cause de la connaissance apprise, soit être considérés comme des exceptions jusqu'à ce qu'il y ait possibilité d'établir de nouvelles règles compte tenu des seuils de satisfaction exigés.

## 2.5 Le coté technique: langages utilisés et système hôte

La plus grande partie du logiciel SARAH a été écrite en PASCAL.

Les sous-programmes de calcul du champ et du potentiel électriques, de détermination de la base d'orbitales atomiques à utiliser et des coefficients de Slater associés, fonctions de l'hypothèse de calcul demandée, ont été tirés du programme GEOMO [17] et adaptés au logiciel SARAH. Ils ont été écrits en FORTRAN 77.

Des interfaces avec le système hôte, à savoir le système *AIX*, ont également été écrites, cette fois en C, ce qui permet d'utiliser d'autres programmes comme CHIMISTE (programme contenant une importante interface graphique conviviale) [19], GEOMO, GOMHELP [20] à partir du logiciel SARAH.

Le tout forme un poste de travail complet pour chimiste, implantable sur IBM RT PC.

*  *

*

# Chapitre 3

# APPLICATION A LA FAMILLE DES BENZODIAZEPINES

L'intérêt grandissant des *benzodiazépines* (**bdz** en abrégé), que reflette une importante bibliographie, concernant autant les relations structure-activité [11,12,29,47] que des études biologiques ou médicales du comportement [3,25,26,27,28,33], nous a amené à étudier cette famille de molécules.

Les résultats, obtenus grâce au logiciel SARAH, qui constitue, rappelons-le, une approche nouvelle du problème des relations structure-activité, pourront ainsi facilement être analysés et comparés aux nombreux résultats publiés.

## 3.1 Mécanisme et activité biologiques étudiés

### 3.1.1 Propriétés pharmacologiques des benzodiazépines

Les *benzodiazépines* constituent une importante famille de molécules découverte récemment dont les "chefs de file" sont le *diazépam* (Valium ®, première molécule à être commercialisée en 1963), le *clonazépam* (Rivotril ®), le *chlordiazépoxide* (Librium ®), le *nitrazépam* (Mogadon ®), le *médazépam* (Nobrium ®), le *flunitrazépam* (Rohypnol ®) et l'*oxazépam* (Serax ®).

Ces molécules sont maintenant connues pour leurs propriétés pharmacologiques. En effet, dans cette famille, nous trouvons des calmants, des neuroleptiques, des anxiolytiques, des tranquilisants, des anti-épileptiques, etc.

### 3.1.2 Structure des benzodiazépines

Une propriété supplémentaire de ces molécules, est qu'elles sont constituées d'une structure de base composée de trois cycles (voir Figure 3.1, page suivante), ce qui leur confère une certaine rigidité du point de vue géométrique. Cette caractéristique permet une étude plus facile et plus détaillée.

Structure générale d'une 1,4-benzodiazépine:

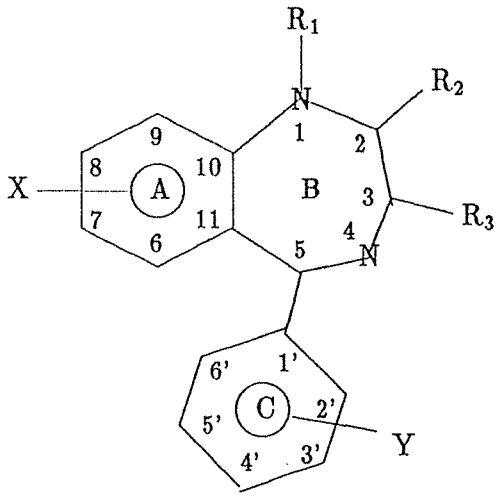



Figure 3.1

Les premières études sur ces molécules ont consisté à rechercher les susbstituants $R_1$, $R_2$, $R_3$, X, Y et leur position, qui leur conféraient une activité importante; citons, par exemple, les travaux de W. SIEGHART [25] et de W. MILKOWSKI [50] sur lesquels nous reviendrons.

Tous les travaux réalisés sur ces molécules soulignent la présence d'une sous-structure commune et l'importance des substituants, surtout de leur taille. Ceci renforce l'image de la *clé* et de la *serrure* utilisée pour concevoir le programme.

### 3.1.3 Le test d'activité

Le test d'activité le plus couramment utilisé pour mesurer les effets biologiques des benzodiazépines, est le test de EVERETT et RICHARDS [2]. Il s'agit d'un test *in vivo*, appliqué généralement à des rats. Il consiste à mesurer la quantité $C$, ramenée à la masse de l'animal (mmol/Kg) , de molécule active à injecter, nécessaire pour neutraliser, chez 50% des animaux traités, les effets d'une première injection d'une quantité précise (125 mg/Kg) d'un composé provoquant des crises d'épilepsie (le *pentylènetétrazole*).
L'activité d'une benzodiazépine est en fait le logarithme décimal de l'inverse de cette concentration $C$ ($-log_{10}C$).
Toutes les molécules que nous avons étudiées étaient caractérisées par cette activité biologique.

### 3.1.4 Un mot sur les récepteurs biologiques

Suite à leur découverte en 1977, un grand nombre d'études, au niveau moléculaire, ont porté sur la localisation dans le cerveau, la caractérisation et la détermination de la structure des récepteurs des benzodiazépines [25,26,27,28,29,33].

Ces études ont surtout montré qu'il était très difficile de déterminer la structure de tels récepteurs biologiques.
Néanmoins, un certain nombre de points ont été éclaircis:

- Les récepteurs des benzodiazépines ont été localisés dans les synapses des neurones du système nerveux central. Ils sont très abondants dans le cerveau antérieur (système cortical, système limbique, corps calleux) ainsi que dans le cervelet, moins abondants dans le cerveau postérieur [26].

- Leur structure et leur nature probables ont été modélisées grâce à un marquage au tritium des benzodiazépines: il s'agit vraisemblablement d'une protéine [27] dont W. HAEFELY et ses collaborateurs [29] décrivent la structure complexe dont elle fait partie. Cette structure est constituée de récepteurs de l'acide $\gamma$-aminobutyrique (GABA), qui constitue le transmetteur, et de récepteurs de benzodiazépines proprement dits, interconnectés par des canaux à ions chlorure (voir Figure 3.2 ci-dessous). Généralement, c'est ce complexe lui même qui est considéré lorsque le terme récepteur de benzodiazépines est employé.

- Ces excellents travaux [29] ont conduit également à la description théorique du fonctionnement des récepteurs de benzodiazépines (voir Figures 3.3 et 3.4, pages suivantes).

Structure et fonctionnement théoriques des récepteurs de benzodiazépines [29,48]:

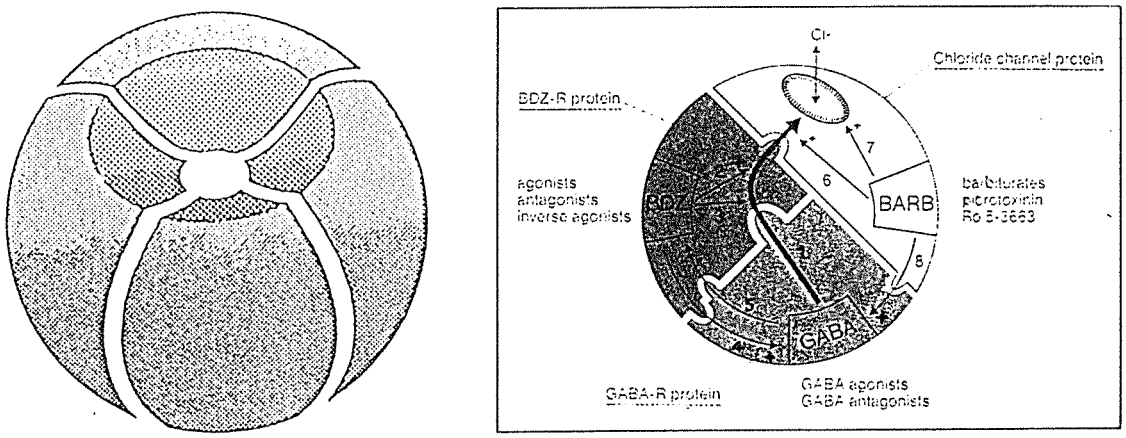
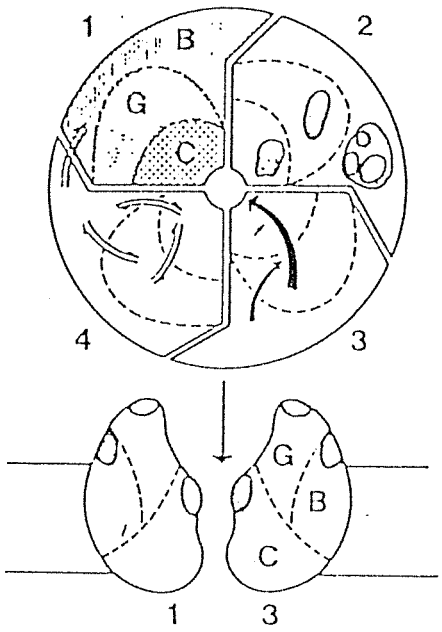



Figure 3.2

Fonction des canaux à ions $Cl^-$:




Partie 1:

B: Domaine des sites liants pour les benzodiazépines.
C: Domaine des canaux à $Cl^-$.
G: Domaine des sites liants pour le GABA.

Partie 2:

Pour chaque domaine est indiqué le site actif liant:
le site liant du récepteur de benzodiazépines (site liant pour les agonistes, les antagonistes et les agonistes inverses) est probablement composé de sous-sites distincts mais voisins.

Partie 3:

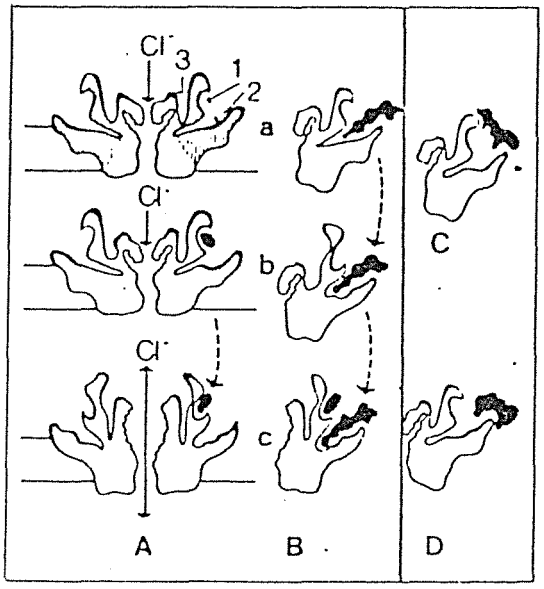
Fonction principale du complexe: ouverture des canaux à $Cl^-$ induite par le GABA (grande flèche), régulation de ce processus par les benzodiazépines agonistes et agonistes inverses (flèches moyenne) et régulation des propriétés des canaux à $Cl^-$ par l'intermédiaire des sites liants du domaine C (liants pour la picrotoxine).

Partie 4:

Les quatre interactions bidirectionnelles entre les domaines.

Figure 3.3

Mécanisme théorique de l'interaction entre une molécule active (benzodiazépine) et le récepteur de benzodiazépines:




Aa: Coupe du complexe récepteur inactif (canaux fermés)

- 1: sites liants pour le GABA
- 2: sites liants pour les benzodiazépines
- 3: sites liants pour la picrotoxine (domaine des canaux à $Cl^-$).

Ab: Le GABA est reconnu par son récepteur ce qui provoque une modification de la conformation (isomérisation par exemple) de ce dernier et l'ouverture des canaux à $Cl^-$ (Ac) laissant entrer davantage de $Cl^-$ à l'intérieur de la cellule; c'est le stimulus pharmacologique: il y a hyperpolarisation de la membrane et sans doute diminution de l'excitabilité neuronale.

Ba: Un agoniste est reconnu par le récepteur de benzodiazépines, la conformation de ce dernier est alors modifiée (Bb). Cette perturbation seule ne peut provoquer l'ouverture des canaux mais renforce le couplage entre le récepteur GABA-ergique et le ionophore chlorique; l'excitabilité neuronale diminue.

C: Interaction avec un agoniste inverse, molécule bloquant l'activité biologique en donnant une conformation défavorable (diminution du couplage récepteur GABA-ergique/canal); l'excitabilité neuronale augmente.

D: Interaction avec un antagoniste (haute affinité avec le site benzodiazépinique): inhibition des interactions avec un agoniste ou un agoniste inverse mais pas de l'action du GABA (du couplage).

Figure 3.4

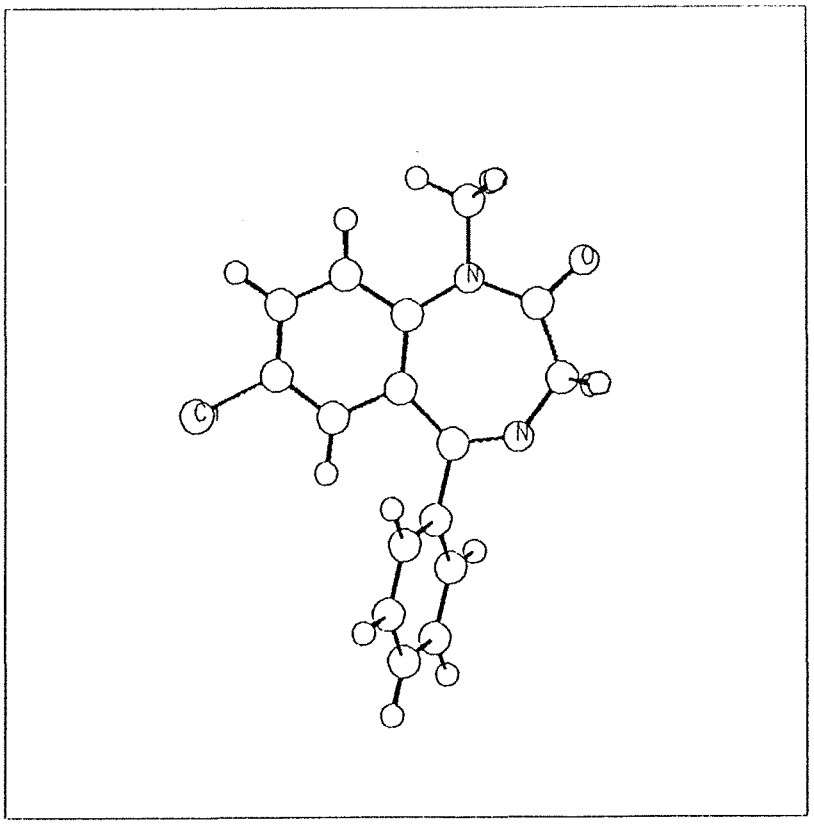
Ces travaux mettent bien en évidence le modèle *clé-serrure* utilisé et dont nous allons voir l'application maintenant.

## 3.2 Détermination des critères d'activité

### 3.2.1 Remarques sur le calcul de la structure des molécules

La géométrie et la structure électronique de chaque molécule étudiée ont été calculées par le programme GEOMO [17] dans l'hypothèse MNDO (Modified Neglect of Differential Overlap), hypothèse la plus satisfaisante pour ces calculs. Notons qu'en ce qui concerne la géométrie, le cycle C fait un angle d'environ 45 degrés avec le plan de l'autre cycle benzénique A (voir Figure 3.5 ci-dessous).

Géométrie optimisée du diazépam:




- Optimisation de géométrie de la molécule par GEOMO.
- Visualisation grâce à l'interface graphique de CHIMISTE.

Figure 3.5

A partir de ces résultats (coordonnées optimisées des atomes, valeurs et vecteurs propres de l'hamiltonien de Hartree-Fock, énergie totale, chaleur de formation, moment dipolaire, population électronique de chaque atome), ont été calculés les descripteurs globaux, enumérés dans le paragraphe **2.2.3**, grâce aux programmes POL180 (calcul de la polarisabilité $\bar{\bar{\alpha}}$ et de $\beta$), VOLUME (calcul du volume de van der Waals), INDICE2 (calcul de l'énergie de la $HOMO$ et de la $LUMO$, de la superdélocalisabilité électrophile totale $SDEL$ et de l'auto-polarisabilité atomique totale $SAPO$) et INDICE3 (calcul de la forme de la molécule $FORME$) [49].

### 3.2.2 Le lot d'apprentissage

Une caractéristique essentielle d'un bon lot d'apprentissage est qu'il soit composé d'un nombre important d'exemples (au moins quarante à cinquante). Nous nous sommes donc efforcés, dans un premier temps, de réunir le plus grand nombre possible de données biologiques sur les benzodiazépines [3,11,14,46,50], soit plus d'une centaine de molécules testées biologiquement (test de EVERETT et RICHARDS).

Les exemples d'un bon lot d'apprentissage ne doivent pas être soumis à une sélection préalable de l'utilisateur, cependant les molécules répertoriées sont loin de constituer un ensemble "objectif" puisqu'elles ont été sélectionnées et testées par des chimistes et des biologistes. Cela explique certainement le fait que nous disposons d'un nombre beaucoup plus important de molécules qualifiées d'actives par rapport au nombre de benzodiazépines inactives.

Etant limités par l'espace de stockage nécessaire au grand nombre de données concernant ces molécules et par les temps de calcul (il faut en moyenne quinze heures, sur *IBM RT PC 6150*, pour calculer complètement une molécule (grâce aux programmes cités dans le paragraphe **3.2.1** précédent) sans parler de la sélection de la configuration de plus basse énergie (la plus probable)), nous n'avons pas pris en compte toutes les molécules à notre disposition. Néanmoins, cette sélection a été aléatoire ou presque, puisque toutes les molécules inactives, vu leur faible nombre, ont été prises.
Il fallait effectivement, comme dans tout programme d'apprentissage de ce genre, des exemples répartis en deux classes, c'est à dire des molécules actives et des molécules inactives, sans qu'il y ait trop de déséquilibre entre ces deux classes.

Pour tester le programme, nous avons aléatoirement écarté de l'ensemble des molécules prises en compte, des exemples des deux classes.

Les tableaux suivants indiquent les molécules constituant le lot d'apprentissage (Tableau 3.1) et les molécules constituant le lot test (Tableau 3.2).

| molécule | activité | X | $R_1$ | $R_2$ | $R_3$ | Y |
|---|---|---|---|---|---|---|
| chlordiazépoxide [1] Librium ® ($N_4$->O,$N_1$=$C_2$) | 1.5737 | 7-Cl | | $NHCH_3$ | $H_2$ | H |
| clonazépam Rivotril ® | 3.2948 | 7-$NO_2$ | H | =O | $H_2$ | 2'-Cl |
| démoxépam [2] ($N_4$->O) | 1.6728 | 7-Cl | H | =O | $H_2$ | H |
| diazépam Valium ® | 2.3100 | 7-Cl | $CH_3$ | =O | $H_2$ | H |
| flunitrazépam Rohypnol ® | 2.4164 | 7-$NO_2$ | $CH_3$ | =O | $H_2$ | 2'-F |

| | | | | | | |
|---|---|---|---|---|---|---|
| lorazépam | 3.2054 | 7-Cl | H | =O | H,OH | 2'-Cl |
| [ médazépam Nobrium ® | 1.5870 | 7-Cl | $CH_3$ | $H_2$ | $H_2$ | H |
| [ nitrazépam Mogadon ® | 2.7497 | 7-$NO_2$ | H | =O | $H_2$ | H |
| nordazépam | 1.6539 | 7-Cl | H | =O | $H_2$ | H |
| [ oxazépam Serax ® | 2.5000 | 7-Cl | H | =O | H,OH | H |
| témazépam | 2.6300 | 7-Cl | $CH_3$ | =O | H,OH | H |
| [ triazolam [3] Halcion ® | 3.5896 | 7-Cl | -C($CH_3$)= | =N-N= | $H_2$ | 2'-Cl |
| 4'-fluorodiazépam | -0.4220 | 7-Cl | $CH_3$ | =O | $H_2$ | 4'-F |
| 7-deschlorodiazépam | -0.5051 | H | $CH_3$ | =O | $H_2$ | H |
| bdz15 | -0.5300 | H | H | =O | $H_2$ | H |
| bdz16 | 2.8804 | 7-Cl | $CH_3$ | =O | $H_2$ | 2'-Cl |
| bdz17 | -0.0917 | 8-Cl | H | =O | $H_2$ | H |
| bdz18 | 2.6917 | 7-$NO_2$ | $CH_3$ | =O | $H_2$ | H |
| bdz19 | -0.4600 | 9-$NO_2$ | H | =O | $H_2$ | H |
| bdz20 | 2.4829 | 7-$CF_3$ | H | =O | $H_2$ | H |
| bdz21 | 0.0000 | 7-Cl | $CH_3$ | =O | $H_2$ | 2',4'-Cl |
| bdz22 | 2.1971 | 7-Br | H | =O | $H_2$ | H |
| bdz23 | 3.4601 | 7-Cl | H | =O | $H_2$ | 2'-F |
| bdz24 | 1.6027 | 7-Cl | H | =O | $H_2$ | 2'-$OCH_3$ |
| bdz25 | 2.4934 | 7-$NO_2$ | H | =O | $H_2$ | 2'-F |
| bdz26 | 2.9691 | 7-$NO_2$ | H | =O | $H_2$ | 2'-$NO_2$ |
| bdz27 | -0.5000 | 7-F | H | =O | $H_2$ | H |
| bdz28 | -0.4800 | 7,9-$CH_3$ | H | =O | $H_2$ | H |
| bdz29 | -0.4400 | 7-Cl | H | =O | $H_2$ | 4'-F |
| bdz30 | 0.1549 | 7-$CH_3$ | H | =O | $H_2$ | H |
| bdz31 | 2.7200 | 7-$NO_2$ | $CH_3$ | =O | $H_2$ | 2'-$CF_3$ |
| bdz32 | -0.2017 | 7-$NH_2$ | $CH_3$ | =O | $H_2$ | H |
| bdz33 | 1.8958 | 7-$CF_3$ | H | $H_2$ | $H_2$ | H |
| bdz34 | 1.5900 | 7-Cl | H | =S | $H_2$ | 2'-Cl |
| bdz35 | 2.6800 | 7-Cl | $CH_3$ | =S | $H_2$ | 2'-Cl |
| bdz36 | -0.4100 | 7-phényle | H | =O | $H_2$ | H |
| bdz37 | 0.1311 | 6-Cl | H | =O | $H_2$ | H |
| bdz38 | -0.4700 | 9-Cl | H | =O | $H_2$ | H |
| bdz39 | -0.4253 | 7-Cl | H | =O | $H_2$ | 4'-$OCH_3$ |
| bdz40 | -0.3724 | 7-Cl | H | =O | $H_2$ | 2',4'-Cl |
| bdz41 | 1.1500 | 7-$SCH_3$ | H | =O | $H_2$ | H |
| bdz42 | 1.5613 | 7-$NO_2$,9-$CH_3$ | H | =O | $H_2$ | H |
| bdz43 | 1.5651 | 7,9-Cl | H | =O | $H_2$ | H |
| bdz44 | 2.8821 | 7-Cl | H | =O | $H_2$ | 2'-Cl |
| bdz45 | 1.5675 | 7-Cl | H | =O | $H_2$ | 2'-$CH_3$ |
| bdz46 | 0.8142 | 7-$NO_2$ | H | =O | $H_2$ | 3'-$NO_2$ |
| bdz47 | 3.9200 | 7-$NO_2$ | $CH_3$ | =O | $H_2$ | 2'-Cl |
| bdz48 | 2.4800 | 7-CN | $CH_3$ | =O | $H_2$ | H |
| bdz49 | 1.1500 | 7-Cl | H | =O | H,$CH_3$ | H |

Tableau 3.1
Le lot d'apprentissage

| molécule | activité | X | $R_1$ | $R_2$ | $R_3$ | Y |
|---|---|---|---|---|---|---|
| flurazépam [4] / Dalamane ® | 2.4000 | 7-Cl | $-(CH_2)_2-N(Et)_2$ | =O | $H_2$ | 2'-F |
| bdz51 | 1.0100 | $7-SOCH_3$ | H | =O | $H_2$ | H |
| bdz52 | -0.2800 | $7-SO_2CH_3$ | H | =O | $H_2$ | H |
| bdz53 [4] | -0.4300 | 7-Cl | H | =O | H,Et | H |
| bdz54 | -0.3600 | 7-Cl | H | =O | H,phényle | H |
| bdz55 | 2.3000 | 7-CN | H | =O | $H_2$ | H |
| bdz56 | 2.000 | $7-NO_2$ | H | =O | H,OH | H |
| bdz57 | 2.6977 | $7-NO_2$ | H | =O | $H_2$ | $2'-CF_3$ |
| bdz58 | 0.6000 | $7-SCH_3$ | $CH_3$ | =O | $H_2$ | H |
| bdz59 | 2.8800 | 7-Cl | $CH_3$ | =O | $H_2$ | 2'-F |
| bdz60 | -0.2700 | 7-F | $CH_3$ | =O | $H_2$ | 2'-F |
| bdz61 | 2.8200 | $7-N(CH_3)_2$ | $CH_3$ | =O | $H_2$ | 2'-Cl |
| bdz62 | 0.8512 | 7-Cl | H | =O | $H_2$ | $3'-CH_3$ |

Tableau 3.2
Le lot test

1: liaison double entre l'azote numéro 1 et le carbone 2
l'azote numéro 4 porte un oxygène (oxydation)
2: l'azote numéro 4 porte un oxygène
3: les substituants $R_1$ et $R_2$ forment un cycle
4: "Et" représente un groupement éthyle $C_2H_5$

### 3.2.3 Le motif commun

Une fois toutes les informations, concernant les benzodiazépines du lot d'apprentissage, saisies, le logiciel SARAH procède, en faisant intervenir l'utilisateur, à la détermination du motif commun à toutes les molécules. Cette recherche est assez rapide compte tenu de la stratégie utilisée (voir paragraphe **2.3.1**).

Dans un premier temps nous avons choisi le premier niveau de recherche à notre disposition, c'est à dire une recherche stricte, ne mettant en correspondance que des atomes de même nature. Nous avons également décidé de ne pas prendre en compte les atomes d'hydrogène, qui pouvaient être présents dans la sous-structure commune, puisque les positions des substituants X, Y, $R_1$, $R_2$ et $R_3$ sont très variables d'une molécule à l'autre.

Comme nous pouvions nous-y attendre au vu de la structure générale des 1,4-benzodiazépines (Figure 3.1, paragraphe **3.1.2**), le motif édité par le programme a été le suivant (Figure 3.6 ci-dessous):

Sous-structure commune aux 1,4-benzodiazépines:

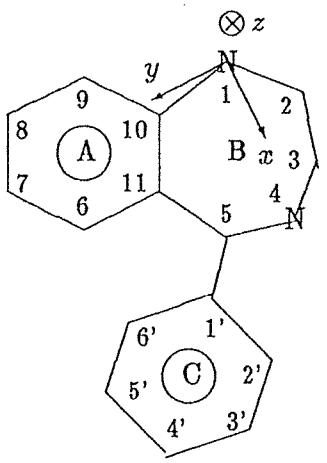



Figure 3.6

Ce motif commun, compte tenu, là encore, de la structure générale des benzodiazépines, est très satisfaisant. De plus les autres niveaux de recherche de la sous-structure conduisent au même motif.

Dans l'étape suivante, le programme repère, dans chaque molécule, cette sous-structure et y accroche le système d'axes orthonormés (voir Figure 3.6, page précédente), déterminé automatiquement et défini par les trois atomes $N_1$, $N_4$ et $C_7$ (axe x: $N_1$->$N_4$, axe y: orthogonal à x et le plus proche de la direction $N_1$->$C_{11}$, axe z: orthogonal à x et à y). La phase de recherche des règles de discrimination des molécules actives et des molécules inactives peut alors commencer.

### 3.2.4 Les règles de discrimination obtenues

Comme cela a déjà été dit, l'utilisateur dispose d'une donnée biologique supplémentaire qu'il doit fournir au programme: la valeur de la borne $A_{ai}$ (voir Figure 2.6, paragraphe **2.3.3**). En ce qui concerne l'activité anti-pentylènetétrazole ou anti-convulsivante des benzodiazépines, tous les auteurs citent une valeur de cette borne comprise entre 0.1 et 0.2. Nous avons choisi de "couper la poire en deux" en prenant $A_{ai}$ égale à 0.15.

Avant d'énoncer les résultats, rappelons que le programme ne connaît, à cet instant, que le motif commun repéré dans chaque molécule, la géométrie des molécules et leurs orbitales moléculaires, et l'ensemble des descripteurs globaux.

A ce stade le programme détermine automatiquement l'ensemble des descripteurs locaux (voir paragraphes **2.2.4** et **2.2.5**), qui ont un poids plus important que les descripteurs globaux, et dans lequel seront donc d'abord recherchés les plus discriminants. Ces descripteurs sont calculés par rapport à la molécule de référence bdz24 sélectionnée automatiquement.

Lors de l'exécution, les seuils de satisfaction donnés au programme étaient les suivants:

- La séparation est supposée terminée lorsqu'il ne reste plus à séparer que 5% du nombre de molécules initial (49), c'est à dire 2 molécules.
- La séparation est également finie lorsqu'il y a moins de 5% de molécules minoritaires dans un sous-ensemble (si par exemple le programme aboutit à un sous-ensemble contenant 1 molécule inactive et 20 actives, il y a arrêt de la recherche, la satisfaction étant atteinte).
- Un descripteur géométrique (ou d'un autre type si aucun descripteur géométrique n'a été trouvé suffisamment discriminant) est considéré comme essentiel (l'exploration de l'espace des autres descripteurs est terminée) s'il a obtenu une *note* (le $\chi^2$) supérieure à 0.85 (descripteur discriminant à plus de 85%).

Les deux premiers seuils ont permis de fixer les tests d'arrêt de l'algorithme en respectant le souhait de tous les chimistes, à savoir prédire la propriété biologique d'une molécule avec une probabilité de plus de 80%. Le troisième seuil fixé correspond, en quelque sorte à ce désir.

Les critères de séparation molécules actives/molécules inactives lors de cette phase d'apprentissage furent les suivants (voir aussi Figure 3.7, page suivante):

- L'enveloppe de van der Waals d'une molécule active a une dimension comprise entre 3.20 et 5.28 Angströms dans la direction du substituant du carbone en position 7 (distance entre $C_7$ et l'enveloppe de van der Waals dans cette direction) et une dimension inférieure à 6.85 Angströms dans la direction du substituant du carbone 4' (distance entre $C_5$ et l'enveloppe de van der Waals dans la direction $C_5$->$C_{1'}$, ce qui correspond effectivement à la taille du substituant du carbone 4').

Toutes les molécules actives du lot d'apprentissage vérifient cette règle $R_1$ qui ne s'applique à aucune molécule inactive, c'est pourquoi, par rapport à l'ensemble des exemples, cette règle possède un pourcentage de validité $PV_{R_1}$ de 100%.

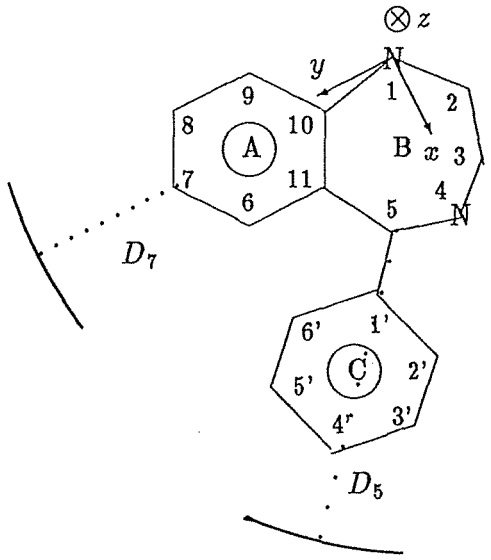
- Les molécules inactives vérifient une des deux règles suivantes:

* La dimension de l'enveloppe de van der Waals dans la direction du substituant du carbone en position 7 est inférieure à 3.20 Angströms ou supérieure à 5.28 Angströms (distance entre $C_7$ et l'enveloppe de van der Waals dans cette direction). Pour cette règle $R_2$, $PV_{R_2}$ est égal à 66,67% (les deux tiers des molécules inactives la vérifient).

* La dimension de l'enveloppe de van der Waals dans la direction du substituant du carbone en position 7 est comprise entre 3.20 et 5.28 Angströms et la dimension dans la

direction du substituant du carbone 4' supérieure à 6.85 Angströms (distance entre $C_5$ et l'enveloppe de van der Waals dans la direction $C_5$->$C_{1'}$, taille du substituant du carbone 4'). Pour cette règle $R_3$, $PV_{R_3}$ est égal à 33.33% (une molécule inactive sur trois la vérifie).

Critères discriminants:




$R_1$ : (3.20 Å < $D_7$ < 5.28 Å) et ($D_5$ < 6.85 Å) $\Longrightarrow$ molécule active; $PV_{R_1}$ = 100%
$R_2$ : (3.20 Å > $D_7$) ou ($D_7$ > 5.28 Å) $\Longrightarrow$ molécule inactive; $PV_{R_2}$ = 66.67%
$R_3$ : (3.20 Å < $D_7$ < 5.28 Å) et ($D_5$ > 6.85 Å) $\Longrightarrow$ molécule inactive; $PV_{R_3}$ = 33.33%

Figure 3.7

L'arbre de décision obtenu:

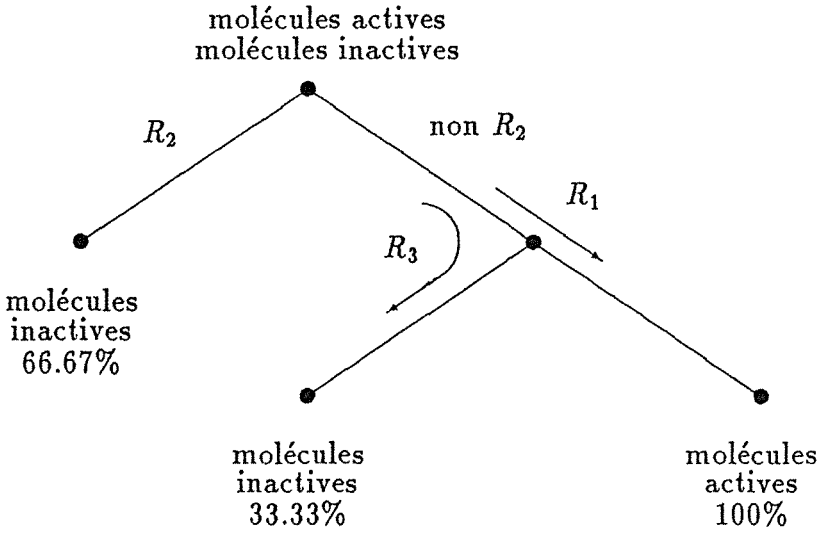



Figure 3.8

### 3.2.5 Les régressions associées aux règles

Après la construction de l'arbre de décision (représenté sur la figure 3.8 ci-dessus) et donc séparation du lot d'apprentissage en sous-ensembles homogènes aux seuils de satisfaction près, le programme procède, sur chaque sous-ensemble ou feuille de l'arbre, à un calcul de régression linéaire multi-variables en ne considérant que les descripteurs de type électronique (voir paragraphe **2.3.4**).

Pour notre application, nous avons obtenu les trois corrélations suivantes:

$$\begin{aligned} R_1 : A \;=\; & -0.789\ [\pm 0.137] \times V_{élec}(C_{3'}, C_{4'} \rightarrow C_{3'}) \\ & +3.074\ [\pm 0.704] \times E_{élec}(C_{11}, C_5 \rightarrow C_{11}) \\ & -1.677\ [\pm 0.940] \times V_{élec}(N_4, C_3 \rightarrow N_4) \\ & -0.427\ [\pm 0.195] \times V_{élec}(C_5, C_{1'} \rightarrow C_5) \\ & -1.107\ [\pm 0.324] \end{aligned} \tag{3.1}$$

Coefficient de corrélation associé: $r_1 = 0.83$.

L'ensemble des molécules actives sur lequel a été calculée cette régression était le suivant (uniquement toutes les molécules actives):
chlordiazépoxide, clonazépam, démoxépam, diazépam, flunitrazépam, lorazépam, médazépam, nitrazépam, nordazépam, oxazépam, témazépam, triazolam, bdz16, bdz18, bdz20, bdz21, bdz23, bdz24, bdz25, bdz26, bdz30, bdz31, bdz33, bdz34, bdz35, bdz41, bdz42, bdz43, bdz44, bdz45, bdz46, bdz47, bdz48, bdz49.

$$\begin{aligned} R_2 : A \;=\; & 0.556\ [\pm 0.000] \times E_{élec}(C_6, C_{11} \rightarrow C_6) \\ & -0.241\ [\pm 0.000] \times V_{élec}(C_8, C_8 \rightarrow \text{substituant}) \\ & +104.756\ [\pm 0.000] \times q_{C_{3'}} \\ & +0.486\ [\pm 0.000] \times E_{élec}(C_{2'}, C_{2'} \rightarrow C_{3'}) \\ & -4.381\ [\pm 0.000] \times q_{C_{2'}} \\ & -0.015\ [\pm 0.000] \times E_{élec}(C_{5'}, C_{5'} \rightarrow C_{6'}) \\ & -0.004\ [\pm 0.000] \times E_{élec}(N_4, N_4 \rightarrow C_5) \\ & -0.002\ [\pm 0.000] \times E_{élec}(C_6, \text{substituant} \rightarrow C_6) \\ & +5.350\ [\pm 0.000] \end{aligned} \tag{3.2}$$

Coefficient de corrélation associé: $r_2 = 1.00$.

L'ensemble des molécules inactives sur lequel a été déterminée cette corrélation était le suivant:

7-deschlorodiazépam, bdz15, bdz17, bdz19, bdz27, bdz28, bdz32, bdz36, bdz37, bdz38.

$$\begin{aligned} R_3 : A = & \ 1.129\ [\pm 0.000] \times E_{élec}(C_6, C_6 \rightarrow C_{11}) \\ & +36.898\ [\pm 0.000] \times q_{C_7} \\ & +0.056\ [\pm 0.000] \times V_{élec}(C_{5'}, C_{5'} \rightarrow \text{substituant}) \\ & +0.003\ [\pm 0.000] \times E_{élec}(C_{11}, C_5 \rightarrow C_{11}) \\ & +1.354\ [\pm 0.000] \end{aligned} \tag{3.3}$$

Coefficient de corrélation associé: $r_3 = 1.00$.

L'ensemble des molécules inactives sur lequel a été obtenue cette équation était le suivant:
4'-fluorodiazépam, bdz21, bdz29, bdz39, bdz40.

Note: $q_X$ représente la charge nette de l'atome X, $Q_{élec}(X, X \rightarrow Y)$ (ou $Q_{élec}(X, Y \rightarrow X)$) correspond au champ ou au potentiel électrique au point de l'enveloppe de van der Waals, intersection de cette enveloppe avec la demi-droite définie par l'atome X et la direction X→Y (ou Y→X).

### 3.2.6 Analyse des résultats et vérification des hypothèses formulées pour le choix des descripteurs

**La règle $R_1$**

En observant la figure 3.1 du paragraphe **3.1.2**, et en se rappelant que le cycle C fait un angle d'environ 45 degrés avec le plan de l'autre cycle benzénique, nous constatons, en ce qui concerne la règle $R_1$, qu'une molécule active possède certainement un substituant sur le carbone numéro 7, mais que ce substituant ne doit être ni "trop petit", ni "trop gros". Une molécule active possède également un "petit" substituant sur le carbone numéro 4'.

Lorsque nous examinons le lot d'apprentissage, nous constatons effectivement que chez toutes les molécules actives, le carbone 4' porte un atome d'hydrogène et que, s'il y a un autre substituant sur cet atome, la molécule est inactive. Nous remarquons également que toutes les molécules actives possèdent un groupement de taille "moyenne" en position 7. Les molécules 7-deschlorodiazépam, bdz15, bdz17, bdz19, bdz27, bdz28, bdz30, bdz32, bdz37, bdz38 ont un "petit" substituant en position 7 (H,F,$NH_2$,$CH_3$) et la molécule bdz36 possède un "gros" substituant en 7 (un cycle benzénique). Notons cependant que le groupement $CH_3$ doit avoir la taille limite puisque la molécule bdz28 est inactive (cela est certainement dû au méthyle en position 9 qui modifie un peu la géométrie) alors que la molécule bdz30 est active (activité faible néanmoins).

Du point de vue de la structure électronique des molécules actives, lorsque nous analysons la corrélation associée à la règle $R_1$, nous remarquons l'importance du potentiel électrique créé sur la partie de l'enveloppe de van der Waals entourant le substituant du carbone $C_{2'}$ et proche de l'atome d'azote $N_4$ ($V_{élec}(C_{3'}, C_{4'} \rightarrow C_{3'})$). Plus ce potentiel est faible (négatif), plus, apparemment, la molécule est active. La présence d'un groupement, possèdant des électrons libres et assez délocalisés, sur le carbone 2' conduit probablement à une molécule fortement active, pour autant que la régle $R_1$ soit vérifiée. C'est le cas des groupements F,Cl,$CF_3$,$NO_2$, et, donnant une activité plus faible, CH3 et OCH3.

L'intensité du champ électrique au voisinage du substituant sur le carbone numéro 8 ($E_{élec}(C_{11}, C_5 \rightarrow C_{11})$) semble également avoir une importance. Une molécule, vérifiant $R_1$ est d'autant plus active que ce champ, créé vraisemblablement par les substituants (hydrogène ou autre), du cycle benzénique A, en position 7, 8 et 9, est intense.

Les valeurs du potentiel électrique au niveau de l'enveloppe de van der Waals du substituant en position 4' et des substituants $R_1$ et $R_2$ ($V_{élec}(N_4, C_3 \rightarrow N_4)$ et $V_{élec}(C_5, C_{1'} \rightarrow C_5)$) modulent également l'activité biologique d'une benzodiazépine. Plus ces valeurs sont faibles (négatives), plus l'activité biologique de la molécule est importante. La présence de groupements possèdant des électrons libres, et assez délocalisés, en $R_1$ et $R_2$, induira probablement une forte activité. Nous constatons effectivement qu'une molécule, comme le diazépam, ayant un carbonyle ($R_2$) est plus active qu'une molécule très voisine, comme le médazépam, n'ayant de différent que deux atomes d'hydrogène pour groupement $R_2$. Un groupement $R_1$ ayant les mêmes propriétés électroniques est également souhaitable pour concevoir une molécule active (voir la forte activité du triazolam).

### La règle $R_2$

Cette règle indique que si une molécule possède un groupement soit "trop volumineux", soit "trop petit" en position 7, elle a toutes les chances d'être inactive. L'activité d'une molécule vérifiant cette règle dépend probablement et essentiellement du champ électrique au niveau du substituant en 7, du potentiel électrique sur l'enveloppe de van der Waals du substituant en 8, de la charge nette portée par le carbone 3', du champ électrique au niveau du substituant en 4' et de la charge nette portée par le carbone 2' (les autres paramètres retenus pour cette corrélation ont de faibles coefficients et donc interviennent peu).

### La règle $R_3$

Bien qu'une molécule vérifie la première partie de la règle $R_1$ (substituant "moyen" en position 7), si elle possède un substituant différent d'un hydrogène en position 4', elle sera certainement inactive (la molécule bdz29 possède un atome de fluor en 4' et est inactive par exemple). Pour cet ensemble de molécules, l'activité est corrélée au champ électrique au niveau de l'enveloppe de van der Waals des substituants $R_2$ et $R_3$, à la charge nette portée

par l'atome de carbone numéro 7, au potentiel électrique créé au niveau du substituant en position 5' (vraisemblablement fonction également des groupements en 4'et 6') et du champ électrique créé aux environs du substituant en position 8.

### Comparaison avec les résultats publiés

Beaucoup d'auteurs ont étudié l'influence des divers substituants $R_1$, $R_2$, $R_3$, X et Y. Parmi eux, W. SIEGHART [25] fait remarquer que des substituants comme Cl ou $NO_2$ en position 7 induisent une bonne activité. Il montre aussi qu'un groupement $CH_3$ en position 1 augmente l'affinité avec le récepteur. Ces conclusions rejoignent effectivement les résultats que nous avons obtenus.

W. HAEFELY et ses collaborateurs [29] classent les différents groupements possibles en position 7 suivant l'activité décroissante des molécules qui les portent:
$NO_2 > CF_3$ Br > CN > Cl > $N(CH_3)_2$ > $SOCH_3$ > SButyle > $SCH_3 \gg CH_3$ > H > $SO_2CH_3$ > phényle > F.
Cette classification correspond à la première partie de la règle $R_1$ qui montre qu'un "petit" substituant (H, F, $CH_3$) ou un "gros" groupement ($SO_2CH_3$, phényle) en position 7 conduit à une molécule inactive.

Les travaux de W. HAEFELY ont également montré l'influence de la nature du substituant en 2'. Il classe ces groupements de la façon suivante:
Cl > F > Br > $NO_2$ > $CF_3$ > H > $OCH_3$ > $CH_3$.
Ceci est à comparer avec la corrélation 3.1 associée à la règle $R_1$ de laquelle nous avons déduit qu'un groupement, portant une charge nette négative et délocalisée, en position 2' favorise l'activité.

Dans les mêmes travaux, W. HAEFELY remarque qu'un substituant en position 4' est fortement désactivant, ce qu'il n'a pu expliquer en terme d'effets électroniques. Cela laisse supposer qu'il s'agit plutôt, comme le montrent les deuxièmes parties des règles $R_1$ et $R_3$, d'effets stériques.

W. MILKOWSKI [50] observe également qu'un substituant en position autre que 2', sur le cycle benzénique C, diminue fortement l'activité de la molécule, ce que nous pouvons associer aux secondes parties des règles $R_1$ et $R_3$.

### Conclusion sur le modèle choisi

Les résultats obtenus font tout d'abord ressortir l'importance de la géométrie dans des processus tels que les interactions entre molécule active et récepteur biologique. Nous observons également que le modèle des variations de l'activité biologique en fonction d'un descripteur

géométrique qui lui est lié, utilisé dans le programme (voir Figure 2.4, paragraphe 2.2.4), est satisfaisant.

En ce qui concerne les propriétés électroniques des molécules, et en particulier des substituants de la sous-structure commune, nous constatons qu'elles ne constituent pas des conditions suffisantes de l'activité, mais qu'elles peuvent moduler plus ou moins fortement l'affinité avec le récepteur biologique.

Notons également qu'aucun descripteur global n'apparaît dans ces résultats, ce qui confirme le fait que les études de tels processus biologiques doivent se porter au niveau submoléculaire.

Le modèle *clé-serrure*, utilisé pour concevoir le logiciel SARAH, est donc assez proche de la réalité, et la démarche qui consiste à comparer les caractéristiques géométriques d'abord, puis électroniques des substituants d'un motif commun aux molécules dotées d'une propriété pharmacologique particulière, n'est pas une approche utopique du problème des interactions molécule active-récepteur biologique, mais au contraire possède des fondements certains.

### 3.2.7 Conception d'une nouvelle benzodiazépine active

Compte tenu de la règle $R_1$ et de la corrélation associée, nous pouvons énoncer, avec une certaine assurance, qu'une nouvelle benzodiazépine active devra posséder un substituant de taille "moyenne" (comme $NO_2$, Cl, CN) en position 7, portant une charge nette relativement élevée afin de créer un champ intense dans cette région de la molécule.

La molécule active devra également posséder un "petit" substituant en position 4', qui sera certainement un atome d'hydrogène.

La présence d'un substituant ayant des électrons libres délocalisés en position 2', sera nécessaire pour une activité importante, car il créera un potentiel électrique négatif au voisinage de l'atome d'azote 4.

Les substituants $R_1$ et $R_2$, pour lesquels aucune dimension limite n'a été trouvée, devront certainement être chargés négativement afin d'avoir un potentiel électrique négatif au niveau de l'enveloppe de van der Waals de ces groupements.

Ces conclusions ne prennent pas en compte la toxicité éventuelle des molécules. Cependant, cette toxicité est une activité biologique, de surcroît chiffrée (dose létale), mais négative, des molécules. Le problème peut donc également être traité par le programme SARAH, par l'étude de cet autre processus biologique.

Les processus d'interactions benzodiazépine-récepteurs ne sont probablement pas les seuls mis en jeu. Comme nous l'avons déjà dit, dans de tels mécanismes biologiques, il y a également des phénomènes de transport de la molécule médicamenteuse vers le site liant du récepteur. Les conclusions de cette étude recouvrent donc ces deux types de processus biologiques.

### 3.2.8 Le récepteur et les mécanismes dans tout cela ?

Pour concevoir le logiciel SARAH, aucune hypothèse, autre que celle du modèle utilisé, n'a été faite. Cependant, en considérant ce modèle, le meilleur actuellement, et les conclusions du programme, il est peut-être possible d'approcher plus en détail la structure du récepteur biologique, impliqué dans l'activité étudiée, et les mécanismes biologiques associés.

Si une benzodiazépine vérifie la règle $R_1$, nous avons observé que plus le potentiel au voisinage de l'atome d'azote numéro 4 était négatif, plus son activité était importante. Ceci peut signifier que la première étape du phénomène est l'attraction de cette zone de la molécule par une partie du site liant de la protéine réceptrice.
Cette liaison est également favorisée par un potentiel électrique négatif autour de l'atome d'azote numéro 1 et des substituants $R_1$ et $R_2$, montrant probablement une deuxième liaison entre cette autre zone de la molécule et une seconde partie du site liant, voisine de la première.

A partir de ces remarques, nous pouvons éventuellement fournir l'hypothèse suivante:
Une benzodiazépine s'oriente, par rapport à son récepteur, probablement en positionnant le cycle B en avant, le cycle A étant à l'arrière de façon à réaliser une liaison entre les *points d'ancrage* ($N_1$ et $N_4$) et le site liant.
Cette configuration du complexe montrerait alors, en prenant l'image d'une *serrure* pour le récepteur, qu'un substituant en position, autre que 2' sur le cycle C, et en particulier en 4', gêne considérablement l'introduction de la *clé*; c'est effectivement une des conclusions de notre étude.

Quant au groupement "moyen" en position 7 sur le cycle A, l'explication pourrait être la suivante:
L'approche décrite ci-dessus ne peut se faire correctement que si un substituant en position 7 stabiblise la molécule. Un groupement "trop gros" reste bloqué à l'entrée de la cavité réceptrice et le processus biologique ne peut se produire, alors qu'un groupement "trop petit" ne stabilise pas assez la molécule.

Bien sûr, ce ne sont que des hypothèses qui doivent attendre une confirmation expérimentale éventuelle pour être validées. De plus, les mécanismes de transport sont présents avant les interactions entre la molécule et le récepteur et doivent imposer des caractéristiques structurales à la molécule active, caractéristiques prises en compte dans notre étude mais qu'il est difficile de distinguer de celles nécessaires à la formation du complexe molécule active-

récepteur conduisant au stimulus biologique. "La balle est dans le camp des chimistes et des biologistes."

## 3.3 Estimation de l'activité de molécules nouvelles

Après l'obtention de ces critères d'activité et d'inactivité, il nous fallait les valider en estimant l'activité des molécules que nous avions écartées (voir Tableau 3.2, paragraphe **3.2.2**). C'est ce que nous avons fait, en voici les résultats (Tableau 3.3 ci-dessous):

| molécule | tendance en % (active ou inactive) | activité estimée | activité expérimentale | règle vérifiée | $A_{calc}$-$A_{exp}$ |
|---|---|---|---|---|---|
| flurazépam | active à 100% | 2.4673 | 2.4000 | $R_1$ | 0.0673 |
| bdz51 | active à 100% | 1.5281 | 1.0100 | $R_1$ | 0.5181 |
| bdz52 | inactive à 66.66% | 0.1105 | -0.2800 | $R_2$ | 0.3905 |
| bdz53 | active à 100% | 1.8266 | -0.4300 | $R_1$ | 2.2566 |
| bdz54 | active à 100% | 1.6891 | -0.3600 | $R_1$ | 2.0491 |
| bdz55 | active à 100% | 1.8888 | 2.3000 | $R_1$ | -0.4112 |
| bdz56 | active à 100% | 2.0645 | 2.0000 | $R_1$ | 0.0645 |
| bdz57 | active à 100% | 2.5816 | 2.6977 | $R_1$ | -0.1161 |
| bdz58 | active à 100% | 1.6222 | 0.6000 | $R_1$ | 1.0222 |
| bdz59 | active à 100% | 3.0073 | 2.8800 | $R_1$ | 0.1273 |
| bdz60 | inactive à 66.66% | -2.8826 | -0.2700 | $R_2$ | -2.6126 |
| bdz61 | active à 100% | 2.5073 | 2.8200 | $R_1$ | -0.3127 |
| bdz62 | active à 100% | 1.8677 | 0.8512 | $R_1$ | 1.0165 |

Tableau 3.3
Les estimations des molécules du lot test

Dans l'ensemble, les estimations sont tout à fait satisfaisantes excepté les prédictions sur les molécules bdz53 et bdz54. Néanmoins, ces deux échecs peuvent s'expliquer:

Comme l'indique W. SIEGHART [25], les benzodiazépines portant un substituant "volumineux" en position 3 ont une très faible affinité avec le récepteur biologique. Les molécules bdz53 et bdz54 sont les seules à posséder un "gros" groupement en position 3 ($R_3$), c'est à dire que dans le lot d'apprentissage, il n'y avait pas d'exemple ayant cette caractéristique. Le programme n'a donc pas pu, lors de la phase d'apprentissage, mettre en évidence ce critère d'inactivité. Autrement dit, le logiciel ne peut connaître ce qu'il n'a pas appris.
Cependant, en ajoutant ces molécules au lot d'apprentissage, en donnant des seuils de satisfaction assez contraignants, ce critère apparaîtrait sûrement.

La prévision de la tendance, active ou inactive, des molécules est exacte à près de 85% (2 échecs pour 13 tests). Du point de vue quantitatif, les résultats sont un peu moins bons, mais dans ce domaine de pointe, tout le monde semble d'accord pour dire que des prévisions avec plus de 80% de réussite sont excellentes.

En outre, plus nous introduirons de molécules dans le lot d'apprentissage, meilleure sera la prévision chiffrée (amélioration des régressions).

## 3.4 Comparaison des résultats du logiciel SARAH et des résultats obtenus par des méthodes statistiques

Il est intéressant maintenant de comparer les résultats du logiciel SARAH avec ceux obtenus par des méthodes statistiques.

### 3.4.1 Régression linéaire multi-variables

La méthode statistique la plus employée est la régression linéaire multi-variables [9,10,11,34]. Après analyse de ces travaux, nous avons décidé de prendre, outre les descripteurs globaux déjà cités (voir paragraphe **2.2.3**), les données locales suivantes: auto-polarisabilité atomique, superdélocalisabilité électrophile, charge nette et polarisabilité des atomes numéros 1, 4, 5, 6, 7, 10, 11, 2', 6'. Ces descripteurs ont été choisis arbitrairement (pour ces études, le chimiste est obligé de faire des choix) - certains ont d'ailleurs été utilisés par les auteurs référencés ci-dessus - en raison de leur nature et de leur position (hétéro-atomes, atomes portant des substituants variables d'une molécule à l'autre, atomes gardant le même environnement (au sens des atomes voisins) chez toutes les benzodiazépines). La méthode utilisée était celle décrite dans le paragraphe **2.3.4**.

Les descripteurs retenus par cette méthode étaient la superdélocalisabilité de l'atome 6 ($SD_6$), la polarisabilité de l'atome 7 ($POL_7$) et l'anisotropie de la molécule ($FORM$):

$$\begin{aligned} A \;=\; & 2.473\ [\pm 0.565] \times SD_6 - 1.163\ [\pm 0.381] \times POL_7 \\ & +7.623\ [\pm 2.564] \times FORM + 15.828\ [\pm 3.842]\ . \end{aligned} \qquad (3.4)$$

Le coefficient de corrélation multiple $r$ était de 0.681 ($A$ représente l'activité).

A partir de cette relation faiblement corrélée, les activités des molécules du lot test ont été estimées (Tableau 3.4 ci-dessous):

| molécule | activité estimée | activité expérimentale | $A_{calc}$-$A_{exp}$ |
|---|---|---|---|
| flurazépam | 2.7790 | 2.4000 | 0.3790 |
| bdz51 | 2.2120 | 1.0100 | 1.2020 |
| bdz52 | 2.9940 | -0.2800 | 3.2740 |
| bdz53 | 2.0290 | -0.4300 | 2.4590 |
| bdz54 | 1.8950 | -0.3600 | 2.2550 |
| bdz55 | 0.6280 | 2.3000 | -1.6720 |
| bdz56 | 2.5550 | 2.0000 | 0.5550 |

| | | | |
|---|---|---|---|
| bdz57 | 3.5620 | 2.6977 | 0.8643 |
| bdz58 | 0.7780 | 0.6000 | 0.1780 |
| bdz59 | 1.8440 | 2.8800 | -1.0360 |
| bdz60 | 0.5480 | -0.2700 | 0.8180 |
| bdz61 | 1.0550 | 2.8200 | -1.7650 |
| bdz62 | 2.0410 | 0.8512 | 1.1898 |

Tableau 3.4

Les estimations des molécules du lot test par régression multi-variables

Nous constatons rapidement que les calculs statistiques donnent des résultats moins bons (taux de réussite: 69%) que ceux du logiciel SARAH qui a estimé qualitativement, avec près de 85% de réussite, les activités des molécules du lot test et qui, si nous écartons les deux molécules responsables des échecs, conduit à des prévisions quantitatives satisfaisantes.

T. BLAIR et ses collaborateurs [11], ont obtenu, en ne considérant que les charges nettes des atomes $N_1$, $C_2$, $C_3$, $N_4$ et O (ils n'ont utilisé que des molécules possèdant un carbonyle (groupement $R_2$), ce qui était une contrainte supplémentaire), ainsi que le moment dipolaire moléculaire $\mu$, la corrélation suivante:

$$A = 42.2\ [\pm 35.9] \times q_O - 0.389\ [\pm 0.103] \times \mu + 18.3\ [\pm 11.9]\ , \qquad (3.5)$$

avec un coefficient de corrélation multiple $r$ de 0.7715.

I. LUKOVITS et L. ÖTVÖS [39] ont corrélé l'énergie de la HOMO à l'activité de benzodiazépines ayant les caractéristiques structurales suivantes: substituant X en position 7, substituant Y en position 2' et $H_2$ comme substituant $R_3$.
Ils en ont conclu que l'activité biologique de cette famille restreinte de benzodiazépines augmentait avec l'énergie de la HOMO.

P. A. BOREA [47] s'est intéressé à un ensemble encore plus restreint de benzodiazépines possèdant un carbonyle ($R_2$). Il a utilisé les descripteurs suivants: $\pi$, somme des constantes hydrophobiques de HANSCH [51] des substituants, $\sigma_7$, constante de HAMMETT [52] du substituant en position 7, $F_{2'}$, combinaison linéaire des constantes de champ de TAFT [53] du substituant en 2' et $E_{S2'}$, constante stérique de TAFT [54] de ce substituant.
La corrélation qu'il obtient est bonne, mais il est impossible de l'étendre à d'autres benzodiazépines compte tenu du choix des molécules et des descripteurs:

$$\begin{aligned} A = {} & -0.388[\pm 0.269] + 0.643[\pm 0.122] \times \pi + 2.787[\pm 0.339] \times \sigma_7 \\ & +2.962[\pm 0.427] \times F_{2'} + 0.638[\pm 0.129] \times E_{S2'}\ , \end{aligned} \qquad (3.6)$$

avec un coefficient de corrélation multiple $r$ de 0.918.

La qualité de cette corrélation est certainement due au fait que toutes les molécules

étudiées étaient très voisines. Remarquons également que la prise en compte de la lipophilie (constantes de HANSCH) a probablement amélioré la régression (prise en compte des phénomènes de transport cités dans le paragraphe **1.3.2**).

Pour parler un peu plus en détail de ces constantes de HANSCH [55], notons qu'il s'agit de coefficients de partage d'une substance chimique entre une phase aqueuse et une phase lipidique séparées, dans le cas de cellules vivantes, par une membrane. Ces coefficients d'origine expérimentale ont été tabulés pour différentes substances. Lorsque les mesures, pour des molécules particulières (médicaments), ne sont pas possibles ou n'ont pas été effectuées, ils peuvent être calculés approximativement, en sommant les coefficients de HANSCH caractéristiques des groupements constitutifs de ces molécules.

Nous n'avons pas réussi à reproduire les conclusions de ces auteurs, soit parce que certains descripteurs n'ont pas été retenus par l'analyse statistique, soit parce qu'ils étaient trop dépendants d'une sous famille de benzodiazépines ou trop approximativement calculés pour les molécules étudiées.

### 3.4.2 Analyse d'agrégats

La deuxième méthode d'analyse de données classique couramment utilisée est l'analyse discriminante [21]. Nous avons donc décidé de comparer les résultats de cette méthode avec ceux obtenus par SARAH. En utilisant les mêmes variables et le même lot d'apprentissage que pour l'analyse statistique décrite au début du paragraphe **3.4.1** précédent, grâce au programme BMDP7M [56], nous avons déterminé des sous-ensembles ou agrégats de molécules correspondant à des valeurs particulières de certains descripteurs puis estimé qualitativement (il est rare de pouvoir chiffrer les estimations avec cette méthode) les activités des molécules du lot test (Tableau 3.5):

| molécule | activité estimée | activité expérimentale |
|---|---|---|
| flurazépam | inactive | 2.4000 |
| bdz51 | active | 1.0100 |
| bdz52 | active | -0.2800 |
| bdz53 | active | -0.4300 |
| bdz54 | active | -0.3600 |
| bdz55 | active | 2.3000 |
| bdz56 | active | 2.0000 |
| bdz57 | active | 2.6977 |
| bdz58 | active | 0.6000 |
| bdz59 | active | 2.8800 |
| bdz60 | inactive | -0.2700 |
| bdz61 | inactive | 2.8200 |
| bdz62 | active | 0.8512 |

Tableau 3.5

Les estimations des molécules du lot test par analyse d'agrégats

Le taux de réussite obtenu par cette méthode est de 61% et est donc moins bon que celui obtenu par le logiciel SARAH, à savoir 85%.

### 3.4.3 Conclusion sur les méthodes statistiques

Ceci montre la faiblesse des méthodes statistiques classiques:
- sélection arbitraire des descripteurs nécessaire,
- application à un ensemble de molécules très semblables (les descripteurs ne sont pas automatiquement adaptés à des molécules, ayant les mêmes propriétés biologiques que les molécules utilisées pour établir la corrélation, mais possèdant des géométries un peu différentes).

De plus ces méthodes permettent difficilement une meilleure compréhension des mécanismes étudiés.

## 3.5 Extension de la famille de molécules à d'autres 1,4-diazépines

Pour montrer ce que nous pouvons appeler "la souplesse" ou "l'adaptabilité" du logiciel SARAH face à des situations un peu différentes de celle dans laquelle il se trouvait lors de l'apprentissage, nous avons soumis au programme, après la phase d'apprentissage, des molécules, de structures différentes de celles des molécules de l'ensemble des exemples, afin qu'il estime leur activité *anti-pentylènetétrazole.*

Comme nous l'avons décrit dans le paragraphe **2.4**, le programme, lorsqu'il est confronté à une telle situation, recommence une analyse en déterminant un nouveau motif commun "inclus" dans la sous-structure précédemment mise en évidence et, cette fois, adapté à la nouvelle molécule.

### 3.5.1 Estimation de l'activité du thiènol[3,2-e][1,4]diazépine

Le *thiènol[3,2-e][1,4]diazépine* fait partie, comme son nom l'indique, de la famille des 1,4-diazépines recouvrant également les 1,4-benzodiazépines. Il possède les mêmes propriétés pharmacologiques que les molécules déjà citées. W. HAEFELY et ses collaborateurs [29] ont estimé qualitativement son activité *anti-pentylènetétrazole*: la molécule possède une activité moyenne (entre 1 et 1.5).

#### Géométrie de la molécule

L'optimisation de géométrie par GEOMO conduit à la structure du *thiènol[3,2-e][1,4]diazépine* décrite par la figure 3.9 suivante (visualisation par l'interface graphique de CHIMISTE).

Géométrie du thiènol[3,2-e][1,4]diazépine:

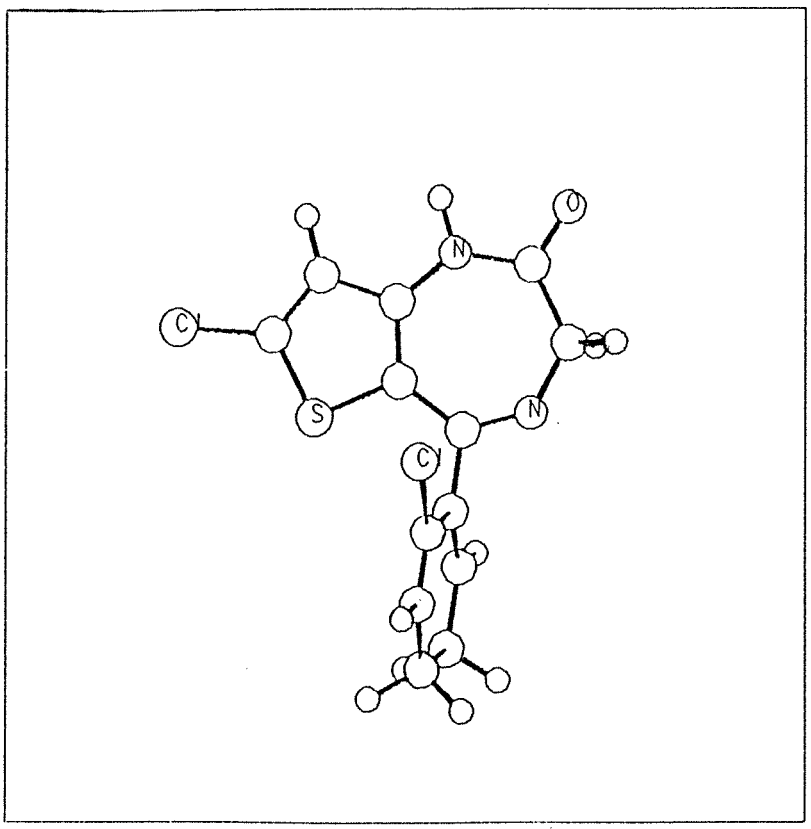



Figure 3.9

### Le nouveau motif commun

Comme cela était prévisible au vu de la structure du thiènol[3,2-e][1,4]diazépine, le logiciel SARAH, ne pouvant retrouver dans la molécule le motif commun précédemment déterminé (voir Figure 3.6, paragraphe **3.2.3**), a recherché une nouvelle sous-structure commune aux molécules du lot d'apprentissage et à la nouvelle molécule. Cette sous-structure est représentée par la figure 3.10 suivante.

Sous-structure commune aux 1,4-diazépines déjà étudiées:

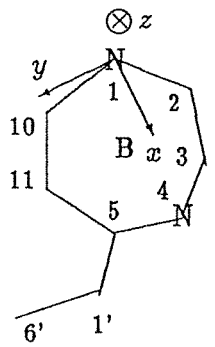
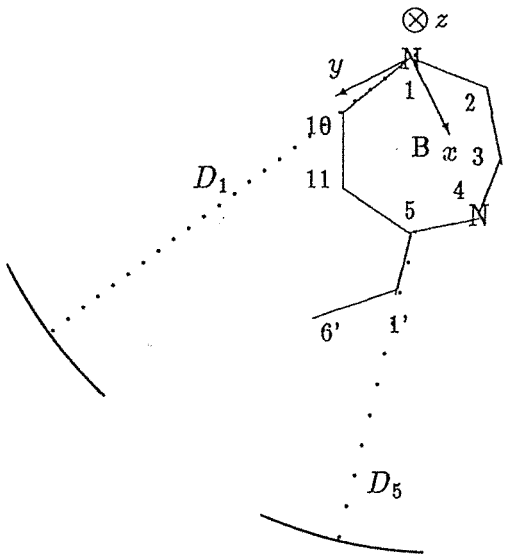



Figure 3.10

**Les nouvelles règles de discrimination**

Les nouvelles règles apprises par SARAH, à partir du même lot d'apprentissage et des mêmes seuils de satisfaction que lors de la première exécution, furent les suivantes (notons que le repère déterminé par le programme et lié au motif est resté le même puisque, comme pour l'analyse précédente, les atomes utilisés pour le définir étaient les deux atomes d'azote et l'atome de cabone 11):

règles de discrimination:




$R_1$ : (7.02 Å < $D_1$ < 9.33 Å) et (5.43 Å < $D_5$ < 6.85 Å) $\Longrightarrow$ molécule active; $PV_{R_1} = 97.14\%$

$R_2$ : (7.02 Å > $D_1$) ou ($D_1$ > 9.33 Å) $\Longrightarrow$ molécule inactive; $PV_{R_2} = 53.33\%$

$R_3$ : (7.02 Å < $D_1$ < 9.33 Å) et (($D_5$ > 6.85 Å) ou ($D_5$ < 5.43 Å)) $\Longrightarrow$ molécule inactive; $PV_{R_3} = 40\%$

Figure 3.11

L'arbre de décision obtenu:

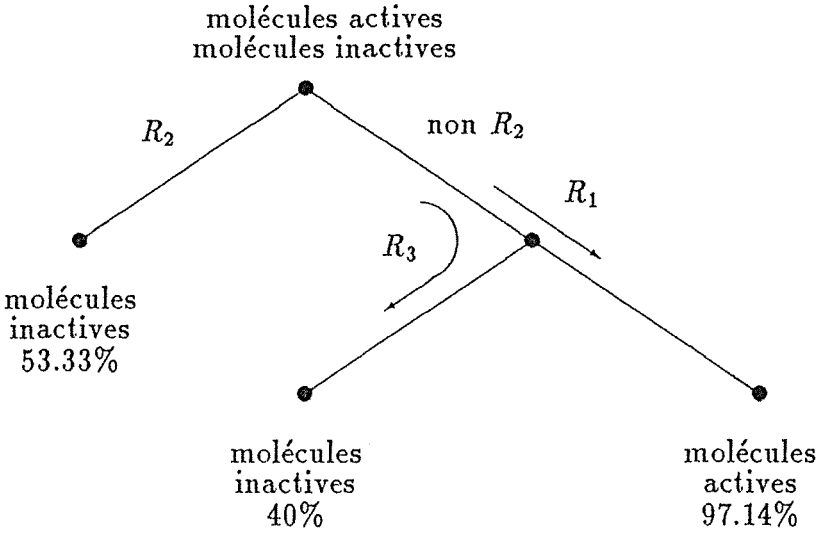



Figure 3.12

Si nous comparons ces résultats aux précédents (voir paragraphe **3.2.4**), nous constatons, qu'en ce qui concerne la géométrie des molécules, nous avons obtenu des règles très semblables et qui correspondent toujours à la dimension du groupement en position 7 des molécules du lot d'apprentissage et à la taille du substituant en 4'. Pour qu'une molécule soit active, il faut une dimension "moyenne" dans la direction azote 1 - carbone 10 (entre 7.02 et 9.33 Å) et dans la direction carbone 5 - carbone 1' (entre 5.43 et 6.85 Å) soit un petit substituant en position 4' si nous observons les molécules du lot d'apprentissage (voir Tableau 3.1, paragraphe **3.2.2**).

Dans le lot d'apprentissage, les molécules vérifiant la règle $R_1$ étaient: chlordiazépoxide, clonazépam, démoxépam, diazépam, flunitrazépam, lorazépam, médazépam, nitrazépam, nordazépam, oxazépam, témazépam, triazolam, bdz16, bdz18, bdz20, bdz21, bdz23, bdz24, bdz25, bdz26, bdz28, bdz30, bdz31, bdz33, bdz34, bdz35, bdz41, bdz42, bdz43, bdz44, bdz45, bdz46, bdz47, bdz48, bdz49; le pourcentage de validité de cette règle était de 97,14% et non plus de 100% comme cela était le cas lors de la première analyse. Ceci s'explique par le fait que toutes les molécules actives plus une molécule inactive, à savoir bdz28, vérifient cette règle. Cette molécule bdz28 a été considérée cette fois comme une exception à la règle $R_1$; en effet, nous avions décidé de tolérer 5% de molécules minoritaires dans les sous-ensembles feuilles de l'arbre de décision. Cette exception provient du fait que, comme nous l'avons déjà remarqué (voir paragraphe **3.2.6**), cette molécule, comme bdz30 qui est très peu active, porte un groupement $CH_3$ en position 7 d'où une ambiguïté pour la placer dans un des deux sous-ensembles: molécules vérifiant $R_1$ et molécules vérifiant $R_2$ - lors de la première analyse, bdz28 avait été placée correctement mais le critère discriminant (distance $D_7$, voir Figure 3.7) était légèrement différent de celui obtenu lors du second traitement du lot d'apprentissage (distance $D_1$, voir Figure 3.11) ce qui explique le changement de position de bdz28 -.

Pour cette règle, la corrélation calculée était:

$$\begin{aligned} R_1 : A \;=\; & 2.283\ [\pm 0.500] \times V_{élec}(C_{11}, C_5 \rightarrow C_{11}) \\ & -0.869\ [\pm 0.234] \times V_{élec}(C_{1'}, \text{substituant} \rightarrow C_{1'}) \\ & -1.221\ [\pm 0.368] \times E_{élec}(C_{1'}, \text{substituant} \rightarrow C_{1'}) \\ & +1.899\ [\pm 0.950] \end{aligned} \tag{3.7}$$

Coefficient de corrélation associé: $r_1 = 0.77$.
note: Le substituant de $C_{1'}$ mentionné correspond, pour les molécules du lot d'apprentissage, au carbone $C_{2'}$.

Les caractéristiques électroniques mises en évidence par cette corrélation, tout comme lors de la première analyse, sont celles des groupements en position 7, 8, 9 du cycle A (en particulier en position 7) et celles du voisinage de l'azote $N_4$, c'est à dire du substituant en position 2'.

La règle $R_2$, avec un pourcentage de validité de 53.33%, était vérifiée par les molécules inactives: 7-deschlorodiazépam, bdz15, bdz17, bdz19, bdz27, bdz36, bdz37, bdz38 et la corrélation calculée était:

$$\begin{aligned} R_2 : A \;=\; & 0.620\ [\pm 0.001] \times E_{élec}(C_{11}, C_{11} \rightarrow \text{substituant}) \\ & -0.207\ [\pm 0.001] \times V_{élec}(C_{10}, C_{10} \rightarrow \text{substituant}) \\ & +0.224\ [\pm 0.002] \times V_{élec}(C_3, C_3 \rightarrow \text{substituant}) \\ & -1.285\ [\pm 1.084] \end{aligned} \tag{3.8}$$

Coefficient de corrélation associé: $r_2 = 0.98$.
note: Le substituant de $C_{11}$ correspond au carbone $C_6$, le substituant de $C_{10}$ correspond au carbone $C_9$ et le substituant de $C_3$ est le groupement se situant du même coté du presque plan cycle benzodiazépinique que le substituant en 2'.

Dans le sous-ensemble de molécules cité ci-dessus, outre la disparition de bdz28, par rapport à la première analyse, nous constatons également la disparition de la molécule bdz32 qui se retrouve dans le sous-ensemble associé à la règle $R_3$ - c'est pour cela, d'ailleurs, que le descripteur discriminant $D_5$ possède une borne inférieure -. Ceci est dû, là encore, à la légère différence entre les descripteurs $D_7$ et $D_1$ (la valeur de la distance $D_1$ pour bdz32, molécule inactive, étant en accord avec la première partie de la règle $R_1$, bdz32 a été séparée des molécules actives à cause de la valeur de la distance $D_5$ qui se trouvait être plus petite que la valeur de ce descripteur chez chaque molécule active).

Les propriétés électroniques mises en évidence chez les molécules vérifiant $R_2$ étaient celles des groupements en position 7, 8 et 9 (surtout en position 7) et celles de l'entourage de $N_4$ dues au substituant en 2'.

La corrélation associée à la règle $R_3$ ($PV_{R_3} = 40\%$) et calculée sur le sous-ensemble de

molécules inactives: 4'-fluorodiazépam, bdz21, bdz29, bdz32, bdz39, bdz40, était:

$$\begin{aligned} R_3 : A = & -0.481\ [\pm 0.000] \times V_{élec}(C_{11}, \text{substituant} \rightarrow C_{11}) \\ & +0.204\ [\pm 0.000] \times E_{élec}(N_4, C_3 \rightarrow N_4) \\ & -0.107\ [\pm 0.000] \times V_{élec}(N_4, C_3 \rightarrow N_4) \\ & +0.022\ [\pm 0.000] \times E_{élec}(N_1, C_2 \rightarrow N_1) \\ & -0.407\ [\pm 0.000] \end{aligned} \tag{3.9}$$

Coefficient de corrélation associé: $r_3 = 1.00$.

Cette corrélation met en évidense l'influence des propriétés électroniques des substituants des atomes de carbone $C_2$ et $C_3$, du cycle C et du cycle A.

**Estimation de l'activité du thiènol[3,2-e][1,4]diazépine et des molécules du lot test**

A la suite de cette mise à jour des règles de discrimination des molécules actives et des molécules inactives, SARAH a estimé l'activité du thiènol[3,2-e][1,4]diazépine, ce qui était en fait notre requète initiale: règle vérifiée $R_2$, probabilité de 53.33% d'être inactif, valeur de l'activité: -0.2455. Cette estimation erronée est, comme cela a déjà été signalé dans le paragraphe **3.3**, la conséquence de l'absence, dans le lot d'apprentissage, d'exemples possèdant des caractéristiques voisines de celles de cette molécule. Néanmoins pour s'assurer de la validité des nouvelles règles obtenues, nous avons décidé d'estimer l'activité des molécules du lot test (Tableau 3.2, paragraphe **3.2.2**); en voici les résultats (Tableau 3.6):

| molécule | tendance en % (active ou inactive) | activité estimée | activité expérimentale | règle vérifiée | $A_{calc}$-$A_{exp}$ |
|---|---|---|---|---|---|
| flurazépam | active à 97.14% | 1.5937 | 2.4000 | $R_1$ | -0.8063 |
| bdz51 | active à 97.14% | 1.1058 | 1.0100 | $R_1$ | 0.0958 |
| bdz52 | inactive à 53.33% | -0.7280 | -0.2800 | $R_2$ | -0.4480 |
| bdz53 | active à 97.14% | 1.3934 | -0.4300 | $R_1$ | 1.8234 |
| bdz54 | active à 97.14% | 1.2266 | -0.3600 | $R_1$ | 1.5866 |
| bdz55 | active à 97.14% | 1.6866 | 2.3000 | $R_1$ | -0.6134 |
| bdz56 | active à 97.14% | 1.8273 | 2.0000 | $R_1$ | -0.1727 |
| bdz57 | active à 97.14% | 2.2822 | 2.6977 | $R_1$ | -0.4155 |
| bdz58 | active à 97.14% | 1.5391 | 0.6000 | $R_1$ | 0.9391 |
| bdz59 | active à 97.14% | 2.4805 | 2.8800 | $R_1$ | -0.3995 |
| bdz60 | inactive à 53.33% | -0.4161 | -0.2700 | $R_2$ | -0.1461 |
| bdz61 | active à 97.14% | 2.2533 | 2.8200 | $R_1$ | -0.5667 |
| bdz62 | active à 97.14% | 1.6480 | 0.8512 | $R_1$ | 0.7968 |

Tableau 3.6

Les nouvelles estimations des molécules du lot test

Nous retrouvons les mêmes erreurs d'estimation (bdz53 et bdz54) que lors de la première analyse, erreurs qui s'expliquent d'ailleurs de la même façon (voir paragraphe **3.3**). Par

ailleurs, l'écart type entre les valeurs expérimentales et les valeurs calculées vallant 0.846 est meilleur que celui calculé lors de la première analyse et qui était de 1.245.

### 3.5.2 Estimation de l'activité du pyrrolo[3,4-e][1,4]diazépine

W. HAEFELY et ses collaborateurs [29] ont étudié le *pyrrolo[3,4-e][1,4]diazépine* ou *prémazepam*, ils ont affecté à cette molécule une activité moyenne.

#### Géométrie de la molécule

L'optimisation de géométrie par GEOMO conduit à la structure du *prémazépam* décrite par la figure 3.13 suivante (visualisation par l'interface graphique de CHIMISTE).

Géométrie du prémazépam:

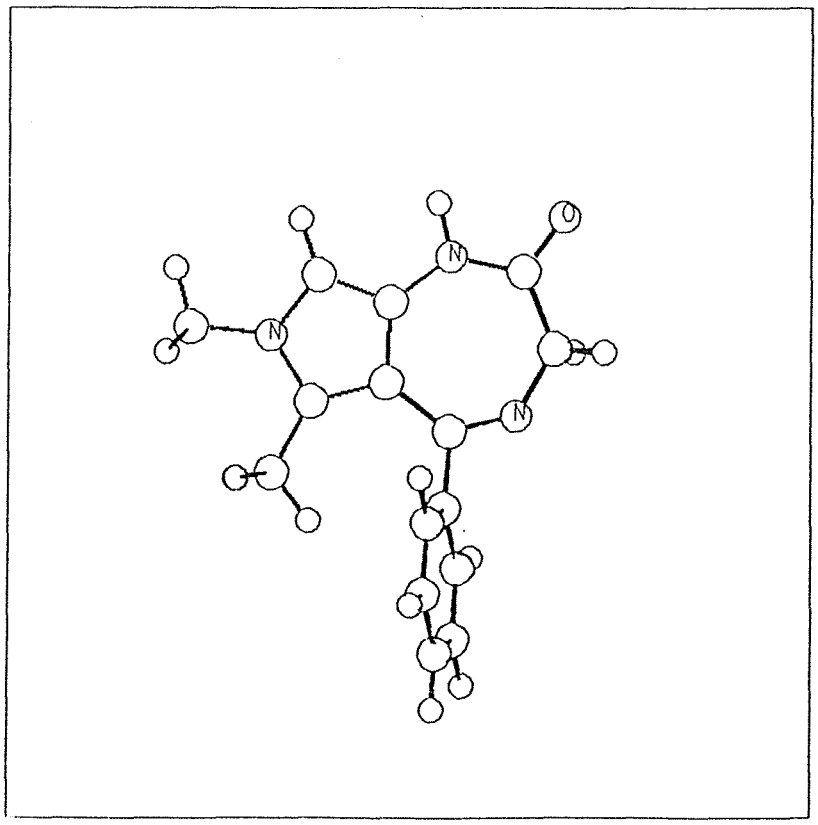



Figure 3.13

#### Estimation de l'activité

Le logiciel SARAH a retrouvé dans le prémazépam le nouveau motif commun obtenu après l'analyse du thiènol[3,2-e][1,4]diazépine, le programme n'a donc pas eu besoin de "réapprendre" de nouvelles règles. SARAH a estimé l'activité de cette molécule à -0.8330 (règle vérifiée $R_2$, probabilité de 53.33% d'être inactive). Ce nouvel échec s'explique comme celui de l'estimation

de l'activité du thiènol[3,2-e][1,4]diazépine: ces deux molécules possèdent un cycle à cinq atomes portant des petits substituants, accollé au cycle B d'où une distance $D_1$ inférieure à la borne inférieure caractéristique de ce descripteur et calculée lors de la phase d'apprentissage.

Nous constatons un point faible du logiciel SARAH à savoir que lors des échecs, les écarts avec la réalité peuvent être importants. Ceci ne remet pas en cause les principes du programme qui donne des règles tout à fait cohérentes avec la connaissance représentée par les exemples du lot d'apprentissage. Répètons le, le logiciel peut se tromper s'il n'a pas appris. Malgré tout, une mise à jour des règles et corrélations obtenues sera possible grâce à l'apport des données biologiques chiffrées relatives au prémazépam et au thiènol[3,2-e][1,4]diazépine, ce qui conduira à une amélioration des estimations de l'activité d'autres 1,4-diazépines.

### 3.5.3 Comparaison avec les méthodes statistiques

Nous avons fait une rapide comparaison avec les estimations obtenues à partir d'un programme de régression multi-variables en n'utilisant que les descripteurs globaux, les descripteurs locaux de la première analyse statistique (voir début du paragraphe **3.4.1**) ne pouvant plus être utilisés puisque le motif commun avait changé. Les résultats obtenus étaient plus mauvais encore que ceux obtenus par la première régression multi-variables: 54% de réussite contre 69%.

Quant aux molécules thiènol[3,2-e][1,4]diazépine et prémazépam, les estimations par cette méthode étaient: 1.8630 pour l'activité de la première molécule et -0.5590 pour la seconde. La première estimation est correcte mais, en ce qui concerne les taux de réussite, le logiciel SARAH, avec 4 échecs sur 15 essais (73%), donne de meilleurs résultats que ce programme de regression multi-variables (7 échecs sur 15 essais soit 53%).

Nous constatons, là encore, les faiblesses des programmes d'analyse de données classiques (régression linéaire multi-variables, analyse discriminante), moins bien adaptés à ces problèmes de relations structure-activité.

* *

*

# Chapitre 4

# POSSIBILITES ET LIMITES DU LOGICIEL SARAH

## 4.1 Les possibilités actuelles du logiciel

Dans son état actuel, le logiciel SARAH, aidé, en ce qui concerne les calculs de géométries et de structures électroniques, par des programmes comme GEOMO [17] ou CHIMISTE [19], est capable d'analyser un lot de molécules dotées d'une propriété biologique particulière - molécules réparties en deux classes: les actives et les inactives - d'en déduire une sous-structure commune et les descripteurs géométriques et électroniques locaux pouvant avoir une influence sur l'activité.

C'est une des caractéristiques les plus importantes du programme. En effet, dans la quasi totalité des études de relations structure-activité, les descripteurs utilisés sont sélectionnés arbitrairement par les chimistes et sont donc très subjectifs.

Cette analyse est calquée sur celle du chimiste qui utilise, dans la majorité des cas, l'image d'une *clé* et d'une *serrure* pour décrire les interactions molécule active-récepteur biologique. C'est ce qui fait l'originalité de ce programme.

Dans cet ensemble de descripteurs, le programme recherche les plus discriminants, c'est à dire séparant au mieux les molécules actives des molécules inactives. Cette recherche est réalisée en s'appuyant sur le modèle sus-dit, montrant que la géométrie a une plus grande importance que les autres paramètres, lors d'interactions molécule active-récepteur.

Cette analyse permet d'éditer des règles vérifiées soit par des molécules actives, soit par des molécules inactives, et appelées règles de discrimination des deux classes d'exemples.

A partir de ces règles, le programme SARAH est déjà capable de déterminer qualitativement l'activité biologique d'une nouvelle molécule avec une certaine probabilité (molécule active ou molécule inactive).

Lors de cette étape, il y a création d'un arbre de décision et le lot d'apprentissage est décomposé en sous-ensembles de molécules ayant la même tendance (active ou inactive). Pour chaque sous-ensemble, le logiciel procède à une analyse statistique (calcul d'une régression linéaire multi-variables) en ne prenant en compte que les descripteurs de la structure électro-

nique des molécules (compte tenu du modèle utilisé, les descripteurs géométriques ne peuvent être corrélés linéairement). Ces corrélations permettent d'estimer quantitativement l'activité d'une molécule nouvelle non testée biologiquement.

Une fois la phase d'apprentissage terminée, le programme peut d'abord déterminer la tendance (active ou inactive) avec une certaine probabilité, puis estimer quantitativement l'activité d'une nouvelle molécule non testée biologiquement et qui lui est présentée.

Lors de l'adjonction d'autres molécules à la base d'exemples ou lors de l'analyse d'une molécule nouvelle, le logiciel SARAH est capable de mettre à jour la totalité ou une partie des règles et des corrélations obtenues après l'analyse du lot d'apprentissage, en recherchant une nouvelle sous-structure commune et des critères plus discriminants et/ou en affinant les corrélations précédemment obtenues.

Le cas de la modification des règles de discrimination est rencontré lorsqu'une molécule, dont la structure est un peu différente de celle des molécules précédemment introduites (le motif commun n'est plus tout à fait le même), est soumise au programme.

## 4.2 Les développements futurs

Un des développements futurs du logiciel est l'analyse automatique des résultats puis l'édition d'une liste d'atomes et/ou de substituants, avec leur position sur le motif commun, donnant des molécules actives.

Il est également prévu d'étendre le programme à des problèmes pour lesquels les données biologiques sont uniquement qualitatives.

Il est possible aussi que la génération, par le logiciel, de descripteurs théoriques supplémentaires nouveaux, à définir compte tenu des connaissances des chimistes en évolution permanente, (modélisant, par exemple, le transport de la molécule) conduise à des résultats améliorés.

Pour l'instant, nous n'envisageons pas la modification de la stratégie d'apprentissage tirée du modèle *clé-serrure*. Cependant, il semblerait, d'après C. LIBERSA et J. CARON [48], que les récepteurs n'aient pas une structure figée et devraient plutôt être considérés comme des protéines réceptrices à conformation spatiale variable. Ce concept de structure "à géométrie variable" selon les conditions du milieu pourrait expliquer la diversité des réactions pharmacologiques observées sans pour autant amener à l'implication de nouveaux sous-types de récepteurs (en ce qui concerne les récepteurs de benzodiazépines, il est souvent question de récepteurs de type I ou de type II ou alors de type périphérique (système ventriculaire) ou de type central (système cortical et système limbique)).

B. P. ROQUES [31] pense également que le récepteur s'adapte à la molécule et propose un

modèle *fermeture-éclair*, à la place du modèle *clé-serrure*, pour décrire les interactions entre une molécule active et un récepteur biologique. Néanmoins, les changements conformationnels conduisant à la structure biologique active liée au site actif du récepteur affecteraient peu le squelette peptidique et concerneraient essentiellement les chaînes latérales du récepteur.

Nous attendons des confirmations expérimentales de l'un ou l'autre des deux modèles avant d'apporter des modifications importantes au logiciel SARAH.

## 4.3 Les limites du logiciel

Lorsque les lots de molécules sont trop petits, le programme SARAH peut conduire assez souvent à des échecs dans les prévisions, surtout lorsque toute la gamme des exemples n'est pas représentée dans le lot d'apprentissage. Cependant, ceci est le propre de tout programme d'apprentissage par l'exemple puisque la connaissance représentée par les exemples n'est pas exhaustive. Néanmoins les règles éditées sont cohérentes avec le lot d'apprentissage.

Dans le cas d'échecs, les prévisions peuvent être très éloignées de la réalité.

La notion de *motif commun* peut s'avérer ambiguë lorsque les molécules à traiter ont une structure de base non rigide. En effet il est alors difficile de faire coïncider les molécules en fonction de ce motif et de leur attacher un repère translatable. Une solution envisageable dans ce cas est la réduction de la sous-structure à un groupement d'atomes concervant une géométrie voisine d'une molécule à l'autre. C'est d'ailleurs une possibilité offerte par SARAH.

Quoiqu'il en soit, les résultats de l'apprentissage peuvent toujours être remis en cause soit par l'apport de nouvelles molécules testées expérimentalement, soit par la modification de la sous-structure commune.

## 4.4 Comparaison du logiciel SARAH avec les programmes connus actuellement

Très peu de programmes spécifiques à l'analyse des relations entre la structure des molécules et leurs propriétés pharmacologiques existent.

En effet, pour la majorité des travaux dans ce domaine, ce sont des programmes d'analyse de données qui sont utilisés. Rappelons tout de même le programme CASE [22,23,24,46] qui semble utiliser une approche *intelligence artificielle* pour traiter ce problème.

Il apparaît, au vu de ces travaux, que le logiciel SARAH est un des rares programmes alliant *intelligence artificielle* et *analyse statistique*.

Répètons le, un des nombreux avantages du logiciel SARAH par rapport aux autres pro-

grammes, réside dans l'utilisation d'un modèle maintenant adopté par tous les chimistes et biologistes, le modèle *clé-serrure*.

Cette connaissance du domaine conduit à la recherche d'une sous-structure commune aux molécules étudiées, puis à la détermination de directions privilégiées (les liaisons avec les substituants de ce motif) afin de ne déterminer puis sélectionner automatiquement, parmi l'ensemble très important des descripteurs des molécules, que ceux susceptibles d'être reliés à l'activité.

L'explosion combinatoire est ainsi évitée, d'autant plus que des heuristiques, déduites du modèle utilisé, permettent encore d'accélérer l'apprentissage.

L'ensemble des descripteurs n'est donc pas limité ni subjectif comme l'est celui des paramètres choisis par le chimiste pour une étude statistique, et est en partie constitué de caractéristiques géométriques qui apparaissent rarement dans les approches statistiques classiques du problème et dont nous avons vu l'importance (voir paragraphe **3.2.6**).

Notons cependant, que les programmes de calculs statistiques essaient maintenant de contourner la subjectivité de l'utilisateur en prenant en entrée un très grand nombre de données et en recherchant les meilleures corrélations à $N$ variables, $N$ étant fixé par l'utilisateur.

Une autre caractéristique du logiciel SARAH, absente des autres programmes, est la possibilité d'intervention de l'utilisateur lors des phases clés de l'apprentissage (recherche plus ou moins souple du motif commun, choix de seuils plus ou moins contraignants, remise en cause de prévisions).

Le programme mis au point se distingue aussi par le fait qu'il peut prendre en compte l'exception (le nombre d'exceptions est fonction des seuils de satisfaction donnés) jusqu'à ce qu'un nombre suffisant (défini par les seuils sus-dits) de ces dites exceptions, puisse conduire à une nouvelle règle de discrimination.

## 4.5 Conclusion

Afin de satisfaire les désirs des chimistes et surtout de réduire les coûts de lancement de nouveaux médicaments, il a été élaboré un logiciel (SARAH), d'un type nouveau, d'aide à la conception de molécules pharmacologiquement actives, alliant *intelligence artificielle* (en particulier des techniques d'apprentissage [57,58]) et *analyse statistique*.

Ce programme est basé sur la connaissance qu'ont les chimistes dans le domaine des relations structure-activité, décrivant les interactions entre une molécule pharmacologiquement active et un récepteur biologique.

Les résultats obtenus sur une famille de molécules, les *benzodiazépines* découvertes il y a environ vingt cinq ans mais dont les mécanismes biologiques qu'elles induisent viennent récemment d'être approchés (moins de dix ans), et les améliorations envisagées, semblent très prometteurs de l'avenir de ce logiciel. En effet, les prévisions réalisées dépassent celles

des programmes aujourd'hui à la disposition des chimistes.

En déterminant objectivement les structures géométrique et électronique fines d'une molécule très active, le programme SARAH peut également permettre de mieux comprendre les phénomènes biologiques induits par les interactions entre cette molécule et le récepteur et d'approcher la structure même de ce récepteur.

* *

*

# Bibliographie


[1] *L'ordinateur en chimie*
François CHOPLIN
*Pour la Science*, Novembre 1985, pp. 50-60

[2] G. M. EVERETT and R. K. RICHARDS
*J. Pharmacol.*, No. 81, 1944, pp. 402

[3] *Anticonvulsivant Properties of Chlordiazepoxide, Diazepam and Certain Other 1,4-Benzodiazepines*
R. F. BANZIGER
Department of Pharmacology, Roche Reseach Division, Hoffmann-La Roche inc., Nutley 10, New Jersey, U.S.A.
*Arch. int. Pharmaccdyn.*, Vol. 154, No. 1, 1965, pp. 131-137

[4] *Mécanique Quantique*
L. LANDAU et E.LIFCHITZ
Editions MIR, Moscou 1966

[5] *Application des méthodes semi-empiriques à l'étude des propriétés électroniques moléculaires et des interactions entre molécules*
Daniel RINALDI
Thèse présentée à l'Université de Nancy I en 1975

[6] *Versatile Techniques for Semi-Empirical SCF-LCAO Calculations Including Minimization of Energy*
Daniel RINALDI
Laboratoire de Chimie Théorique, Université de Nancy I
54506 Vandœuvre lès Nancy cedex
*Computers & Chemistry*, Vol. 1, 1976, pp. 109-114

[7] *Molecular Mechanics, the Method and Its Underlying Philosophy*
Donald B. BOYD and Kenny B. LIPKOWITZ
Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285
Indiana-Purdue University, Indianapolis, IN 46205
*Journal of Chemical Education*, Vol. 59, No. 4, 1982, pp. 269-274

[8] *Molecular Mechanics, Illustration of its Application*
Philip J. COX
School of Pharmacy, Robert Gordon's Institute of Technology, Aberdeen, AB9 IFR Scotland
*Journal of Chemical Education*, Vol. 59, No. 4, 1982, pp. 269-274

[9] *Electronic Descriptors in Quantitative Structure-Activity Relationships*
Alain CARTIER and Jean-Louis RIVAIL

Laboratoire de Chimie Théorique, U.A. au C.N.R.S. No. 510 "Interactions Moléculaires", Université de Nancy I, B.P. 239, 54506 Vandœuvre-Lès-Nancy cedex (France)
*Chemometrics and Intelligent Laboratory Systems*, No. 1, 1987, pp. 335-347

[10] *Structure-Activity Correlations for Anticonvulsivant Drugs*
Eric J. LIEN
School of Pharmacy, University of Southern California, University Park, Los Angeles, California 90007
*Journal of Medicinal Chemistry*, Vol. 13, No. 6, 1970, pp. 1189-1191

[11] *Electronic Factors in the Structure-Activity Relationship of Some 1,4-Benzodiazepin-2-ones*
T. BLAIR and G. A. WEBB
Department of Chemical Physics, University of Surrey, Guildford, Surrey, England
*Journal of Medicinal Chemistry*, Vol. 20, No. 9, 1977, pp. 1206-1210

[12] *Pharmacological Activity of Neuroleptic Drugs and Physicochemical, Topological and Quantum Chemically Calculated Parameters: a QSAR Study*
Lutgarde BUYDENS, D. Luc MASSART and Paul GEERLINGS
Vrije Universiteit Brussel, Farmaceutisch Instituut, Laarbeeklaan 103, B-1090 Brussel
Vrije Universiteit Brussel, Algemene Chemie, Fakulteit Wetenschappen, Pleinlaan 2, B-1050 Brussel
*Eur. J. Med. Chem. - Chim. Ther.*, Vol.21, No.1, 1986, pp. 35-43

[13] *Mutagenicity of Substituted (o-Phenyllenediamine)platinum Dichloride in the Ames Test. A Quantitative Structure-Activity Analysis*
Corwin HANSCH, Benjamin H. VENGER and Augustine PANTHANANICKAL
Department of Chemistry, Pomona College, Claremont, California 91711
*J. Med. Chem.*, No. 23, 1980, pp. 459-461

[14] *Decomposition of Pharmacological Activity Indives into Mutually Independent Components Using Principal Component Analysis*
István LUKOVITS and Antal LOPATA
Central Research Institute of Chemistry, Hungarian Academy of Sciences, 1025 Budapest, Pusztaszeri u. 59, Hungary
*J. Med. Chem.*, No. 23, 1980, pp. 449-459

[15] Communication préliminaire: C. LAURENÇO et G. KAUFMANN
*Tetrahedron Lett.*, 2243, (1980)
*Synthèse Assistée par Ordinateur de la Phosphacarnegine: Etablissement du Plan de Synthèse avec l'Aide de PASCOP*
C. LAURENÇO, L. VILLIEN et G. KAUFMANN
Laboratoire de Modèles Informatiques Appliqués à la Synthèse, E.R.A. 671 du C.N.R.S. - Université Louis Pasteur - 4 Rue Blaise Pascal - 67008 STRASBOURG Cedex (FRANCE)

[16] N.L. ALLINGER and Y.H. YUH
Department of Chemistry, University of Georgia
Athens, Georgia 30602
QCPE 010

[17] Daniel RINALDI
Laboratoire de Chimie Théorique, Université de Nancy I
54506 Vandœuvre lès Nancy cedex
QCPE 290

[18] Daniel RINALDI, Alain CARTIER et Philippe E. J. HOGGAN
Laboratoire de Chimie Théorique, Université de Nancy I
54506 Vandœuvre lès Nancy cedex
à paraître au QCPE

[19] Marilia T. C. MARTINS COSTA
Laboratoire de Chimie Théorique, Université de Nancy I
54506 Vandœuvre lès Nancy cedex
à paraître au QCPE, thèse en préparation

[20] Alain CARTIER
Laboratoire de Chimie Théorique, Université de Nancy I
54506 Vandœuvre lès Nancy cedex
à paraître au QCPE

[21] *Eléments d'analyse de données*
E. DIDAY, J. LEMAIRE, J. POUGET et F. TESTU
Editeur: DUNOD

[22] *Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules*
Gilles KLOPMAN
Contribution from the Chemistry Department, Case Western Reserve University
Cleveland, Ohio 44106, U.S.A.
*J. Am. Chem. Soc.*, Vol. 106, No 24, 1984, pp. 7315-7321

[23] *Causality in Structure-Activity Studies*
Gilles KLOPMAN and Alexander N.KALOS
Contribution from the Chemistry Department, Case Western Reserve University
Cleveland, Ohio 44106, U.S.A.
*Journal of Computational Chemistry*, Vol. 6, No. 5, 1984, pp. 492-506

[24] *Use of the Computer Automated Structure Evaluation Program in Determining Quantitative Structure-Activity Relationships Within Hallucinogenic Phenylalkylamines*
Gilles KLOPMAN and Orest T. MACINA
Contribution from the Chemistry Department, Case Western Reserve University
Cleveland, Ohio 44106, U.S.A.
*J. Theor. Biol.*, No. 113, 1985, pp. 637-648

[25] *Affinity of various ligands for benzodiazepine receptors in rat cerebellum and hippocampus*
Werner SIEGHART and Annemarie SCHUSTER
Department of Biochemical Psychiatry, Psychiatrische Universitätsklinik, Vienna, Austria
*Biochemical Pharmacology*, Vol. 33, No.24, 1984, pp. 4033-4038

[26] *Benzodiazepine Receptors*
Claus BRAESTRUP and Mogens NIELSEN
A/S Ferrosan, Soeborg, Denmark and Sct. Hans Mental Hospital, Roskilde, Denmark
*Clinical Neuropharmacology*, Vol. 8, Suppl. 1, 1985, pp. S2-S7

[27] *Anxiety and the benzodiazepine receptor*
Trevor R. NORMAN and Graham D. BURROWS
Department of Psychiatry, University of Melbourne, Austin Hospital, Heidelberg, Victoria 3084, Austria

*Progress in Brain Research*, Vol. 65, 1986, pp. 73-90
J.M. Van Ree and S. Matthysse (Eds.)

[28] *Comparison of Typical and Atypical Benzodiazepines on the Central and Peripheral Benzodiazepine Receptors*
Yasuo WATANABE, Takeshi SHIBUYA, Salahadine KHATAMI and Bernard SALAFSKY
Department of Pharmacology, Tokyo Medical College, Tokyo 160, Japan
Department of Biochemical Sciences, University of Illinois College of Medicine at Rockford, Rockford IL 61107-1897, U.S.A.
*Japan. J. Pharmacol.*, No. 42, 1986, pp. 189-197

[29] *Recent Advances in the Molecular Pharmacology of Benzodiazepine Receptors and in the Structure-Activity Relationships of Their Agonists and Antagonists*
Willy HAEFELY, Emilio KYBURZ, Max GERECKE and Hanns MÖHLER
Department of Pharmaceutical Research, F. Hoffmann-La Roche & Co. Ltd., Basel, Switzerland
*Advances in Drug Research*, Vol. 14, 1985, pp. 165-322

[30] *Anxiolytics, Anticonvulsivants and Sedative-Hypnotics*
Michael WILLIAMS and Naokata YOKOYAMA
Research Department, Pharmaceuticals Division, CIBA-GEIGY Corporation, Summit, NJ 07901
*Annual Repports in Medicinal Chemistry*, No. 21, 1986, pp. 11-20

[31] *Reconnaissance Moléculaire et Conception Rationnelle de Molécules Pharmacologiquement Actives*
Bernard P. ROQUES
Département de Chimie Organique, U. No. 266 INSERM, U.A. No. 498 CNRS, UER des Sciences Pharmaceutiques et biologiques, 4 Avenue de l'Observatoire, 75270 Paris Cedex 06
*La Vie des Sciences*, Comptes Rendus de l'Académie des Sciences, Série Générale, Tome 4, No. 3, Mai-Juin 1987, pp. 193-209

[32] *Case Studies of the Application of Molecular Shape Analysis to Elucidate Drug Action*
D. Eric WALTERS and A. J. HOPFINGER
Department of Medicinal Chemistry, Searle Reseach and Development, 4901 Searle Parkway, Skokie, IL 60077 U.S.A.
*Journal of Molecular Structure (Theochem)*, No. 134, 1986, pp. 317-323

[33] *Benzodiazepines and their Metabolites: Relationship between Binding Affinity to the Benzodiazepine Receptor and Pharmacological Activity*
Iwao NAKATSUKA, Hiroshi SHIMIZU, Yukio ASAMI, Terufumi KATOH, Akira HIROSE and Akira YOSHITAKE
Takarazuka Research Center, Sumitomo Chemical Co. Ltd., 2-1, 4-Chome, Takatsukasa, Takarazuka, Hyogo, 665, Japan
*Life Sciences*, Vol. 36, 1985, pp. 113-119

[34] *QSAR and Strategies in the Design of Bioactive Compounds*
Proceedings of the Fifth European Symposium on Quantitative Structure-Activity Relationships Bad Segeberg 1984
edited by J. K. SEYDEL

[35] *QSAR Parameters*
Rainer FRANKE
Academy of Sciences of the GDR, Research Centre for Molecular Biology and Medicine, Institute

of Drug Research, Berlin, GDR
in [34], pp. 59-78

[36] D. R. HARTREE, *Proc. Cambridge Phil. Soc.*, **24**, 89 (1928)
V. FOCK, *Z. Physik*, **61**, 126 (1930)

[37] *Approximate Molecular Orbital Theory*
John A. POPLE and David L. BEVERIDGE
*McGRAW-HILL BOOK COMPANY*, 1970

[38] G. BERTHIER, *C. R. Acad. Sci. Paris*, **238**, 91 (1954); *J. Chim. Phys.*, **51**, 363 (1954)
J. A. POPLE and R. K. NESBET, *J. Chem. Phys.*, **22**, 571 (1954)

[39] *Correlation Between the Protein Binding, Biological Activity and Energy of the Highest Occupied Molecular Orbital of Benzodiazepine Derivatives*
István LUKOVITS and L. ÖTVÖS
Central Research Institute for Chemistry of the Hungarian Academy of Sciences, Pusztaszeri út 59-67, 1025 Budapest, Hungary
*Studia Biophysica*, Berlin, Band 69, Heft 3, 1978, S. 187-191

[40] *Calcul Théorique des Polarisabilités Electroniques Moléculaires. Comparaison des Différentes Méthodes*
Daniel RINALDI et Jean-Louis RIVAIL
Laboratoire de Chimie Théorique, E.A.R. au C.N.R.S. No. 22 "Interactions Moléculaires", Université de Nancy I, B.P. 239, 54506 Vandœuvre-Lès-Nancy cedex (France)
*Theoret. Chim. Acta (Berl.)*, No. 32, 1974, pp. 243-251

[41] *Calcul des Grandeurs et Coefficients Liés à la Forme d'une Molécule Définie à partir des Rayons de Van Der Waals des Atomes. Cas Particulier du Volume Moléculaire*
Bernard TERRYN et Jean BARRIOL
Laboratoire de Chimie Théorique, E.A.R. au C.N.R.S. No. 22 "Interactions Moléculaires", Université de Nancy I, B.P. 239, 54506 Vandœuvre-Lès-Nancy cedex (France)
*Journal de Chimie Physique*, Vol. 78, No. 3, 1981, pp. 207-212

[42] Bernard TERRYN
Laboratoire de Chimie Théorique, Université de Nancy I, B.P. 239, 54506 Vandœuvre-Lès-Nancy cedex
Thèse en préparation

[43] *Theoritical Structure-Activity Studies of β-Carboline Analogs. Requirements for Benzodiazepine Receptor Affinity and Antagonist Activity*
Gilda H. LOEW, John NIENOW, JOHN A. LAWSON, Lawrence TOLL and Edward T. UYENO
Life Sciences Division, SRI International, Menlo Park, California 94025
*Molecular Pharmacology*, Vol. 28, 1985, pp. 17-31

[44] *PLAGE: A Way to Give and Use Knowledge in Learning*
Olivier GASCUEL
Orsay
in *EWSL 1*, 1986

[45] *Critères pour Elaguer la Recherche lorsque la Complétude et la Cohérence ne sont pas Requises*
Olivier GASCUEL
CRIM, 860 rue de St. Priest, 34100 Montpellier
Article tiré des *Actes des Troisièmes Journées Françaises de l'Apprentissage*, 1988

[46] *Use of Artificial Intelligence in Structure-Activity Correlations of Anticonvulsant Drugs*
Gilles KLOPMAN and Renato CONTRERAS
Department of Chemitry, Case Western Reserve University, Cleveland, Ohio 44106
*Molecular Pharmacology*, Vol. 27, 1984, pp. 86-93

[47] *Structure-Activity Relationships in 1,4-Benzodiazepines*
P. A. BOREA
Istituto di Farmacologia dell'Università di Ferrara
*Boll. Soc. It. Sper.*, Vol. 57, 1981, pp. 103-107

[48] *BENZODIAZEPINES: de la Recherche à la Clinique*
Résumés d'exposés.
Laboratoires ROCHE, 52 Boulevard du Parc 92521 NEUILLY-SUR-SEINE Cedex

[49] Ensemble de programmes écrits en FORTRAN 77
par Alain CARTIER
Laboratoire de Chimie Théorique, Université de Nancy I, B.P. 239, 54506 Vandœuvre-Lès-Nancy cedex

[50] *1,4-Benzodiazepines and 1,5-Benzodiazocines. VII. Synthesis and Biological Activity*
W. MILKOWSKI, H. LIEPMANN and H. ZEUGNER
Kali-Chemie AG, Pharmaceutical Division, Chemical Dpt., D-3000 Hannover, W.-Germany
M. RUHLAND
Kali-Chemie AG, Pharmaceutical Division, Pharmacological Dpt., D-3000 Hannover, W.-Germany
M. TULP
Duphar B.V., Pharmacological Dpt., 1381 CP WEESP, The Netherlands
*Eur. J. Med. Chem. - Chim. Ther.*, Vol. 20, No 4, 1985, pp. 345-358

[51] C. HANSCH, *Acc. Chem. Res.*, 1969, 2, 232

[52] H. H. JAFFE', *Chem. Rev.*, 1953, 53, 191

[53] C. HANSCH, A. LEO, S. H. UNGER, K. H. KIM, D. NIKAITANI and E. J. LIEN, *J. Med. Chem.*, 1973, 16, 1207

[54] R. W. TAFT, in *Steric Effect in Organic Chemistry*, M. S. NEWMAN Ed., John WILEY, New York, 1956

[55] *Partition Coefficients and their Uses*
Albert LEO, Corwin HANSCH and David ELKINS
Department of Chemistry, Pomona College, Claremont, California 91711
*Chemical Reviews*, Vol. 71, No. 6, 1971, pp. 525-615

[56] *BMDP Statistical Software* R. JENNRICH and P. SAMPSON
in W. J. DIXON (Editor), University of California Press, Berkeley, 1983, p 519

[57] *Actes des deuxièmes Journées Françaises de l'Apprentissage*
Chamrousse, 12-13 mars 1987
Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle
Institut IMAG, B.P. 68, 38402 Saint Martin d'Hères cedex

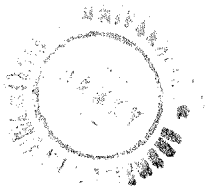
[58] *Actes des troisièmes Journées Françaises de l'Apprentissage*
Cassis sur Mer, 5-6 mai 1988
Groupe Représentation et Traitement des Connaissances, CNRS

GRECO/PRC-IA (pôle 3)
31 Chemin Joseph Aiguier, 13402 MARSEILLE cedex

[59] *Inductive Learning of Relational Productions*
Steven A. VERE
University of Illinois at Chicago Circle
in *Pattern-Directed Inference Systems*, 1978

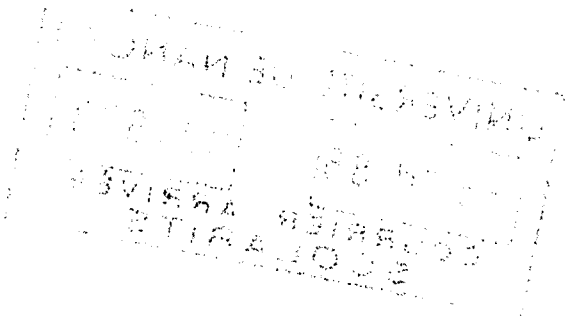
[60] *A Theory and Methodology of Inductive Learning*
Ryszard S. MICHALSKI
Department of Computer Science, University of Illinois, Urbana, IL 61801, USA
*Artificial Intelligence*, No. 20, 1983, pp. 111-161

[61] *Discriminer sur des Descriptions Structurelles dans un Environnement Incertain*
Olivier GASCUEL
Grenoble
dans *AFCET - RFIA 5*, 1985, pp. 1263-1271



NOM DE L'ETUDIANT : ROZOT Roger

NATURE DE LA THESE : Doctorat de l'Université de NANCY I en Chimie Informatique et Théorique



VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 11 JUIL. 1988 n° 1233

LE PRESIDENT DE L'UNIVERSITE DE NANCY I


R. MAINARD

## RESUME

Un nouveau logiciel, SARAH, alliant intelligence artificielle et méthodes d'analyse statistique, a été mis au point. Il a pour objectif d'aider le chimiste à concevoir des molécules pharmacologiquement actives.

Le principe de ce programme est le suivant: en se basant sur l'analyse traditionnelle du chimiste observant des processus biologiques, il réalise un apprentissage du concept d'activité à partir d'exemples répartis en deux classes: les molécules actives et les molécules inactives.

Les molécules du lot d'apprentissage ont été calculées par les méthodes de la chimie quantique et testées biologiquement. A partir de ces données, le logiciel SARAH, après avoir construit un arbre de décision et donc séparé les molécules en sous-ensembles homogènes par rapport à l'activité, édite des règles de discrimination des deux classes de molécules en prenant surtout en compte les caractéristiques stériques des molécules, liées à l'accessibilité du site actif du récepteur biologique, mais aussi leurs propriétés électroniques locales qui gouvernent leur réactivité. Pour chaque sous-ensemble, le programme corrèle ensuite l'activité aux propriétés électroniques locales des molécules.

Ces règles et ces corrélations permettent au logiciel d'estimer quantitativement l'activité biologique de nouvelles molécules qui lui sont soumises, avant de les synthétiser.

La connaissance apprise n'est pas figée, mais au contraire peut être remise en cause ou affinée par l'apport de nouvelles données biologiques.

Les résultats obtenus lors de l'application à la famille des 1,4-benzodiazépines sont en accord avec les données expérimentales et ont montré la faiblesse des méthodes classiques d'analyse de données.



## MOTS-CLES

Intelligence artificielle - Apprentissage à partir d'exemples répartis en deux classes - Arbre de décision - Règle de discrimination - Analyse de données - Chimie quantique - Structure moléculaire - Calcul SCF - Enveloppe de van der Waals - Potentiel et champ électriques - Récepteur biologique - 1,4-benzodiazépines - Relations structure-activité