Reconnaissance de formes moléculaires dans les relations structure-activité

Hervé Mathis

UNIVERSITE de NANCY I

UFR Sciences et Techniques des Matériaux et des Procédés

GFD Chimie et Physico-chimie moléculaires et théoriques

Thèse de Doctorat de l'Université de Nancy I

spécialité Chimie Informatique et Théorique

présentée par Mr Hervé MATHIS

vendredi 13 novembre 1992

Reconnaissance de formes moléculaires dans les relations structure - activité

Membres du jury:

Président :	Mr J.L. RIVAIL	Professeur à l'Université de Nancy I
Rapporteurs :	Mr J.P. DOUCET	Professeur à l'Université de Paris VII
	Mr A. CALVET	Ingénieur de la Société JOUVEINAL
Examinateurs :	Mme M.C. HATON	Professeur à l'Université de Nancy I
	Mr A. CARTIER	Attaché de recherches au CNRS

Cette thèse a été menée au Laboratoire de Chimie Théorique de l'Université de Nancy I. Je remercie ici :

- Monsieur le professeur Jean-Louis Rivail, son directeur, pour l'intérêt qu'il a porté à un travail qui sortait quelque peu des grands axes de recherche du laboratoire et surtout pour sa patience à l'égard d'un logiciel qui fut long à se concrétiser.

- Monsieur le professeur Jean-Pierre Doucet et Monsieur Alain Calvet, qui ont accepté d'être les rapporteurs de ce travail, en dépit de leurs nombreuses tâches.

- Madame le professeur Marie-Christine Haton, pour sa participation à un jury de chimistes auquel elle apporte son expérience en informatique et en intelligence artificielle.

- Monsieur Daniel Rinaldi - dont le regard devient sceptique dès que vous lui annoncez que votre programme fonctionne correctement - pour cette incitation à plus de prudence, et pour sa constante disponibilité.

- Monsieur Alain Cartier, pour ses conseils de tout ordre dans les relations structure - activité comme pour ses calembours les plus savoureux qui soient...

Mes remerciements vont aussi aux autres membres du Laboratoire pour la chaleureuse ambiance de travail qu'ils y font régner.

La relecture de ce manuscrit a été faite par Mr Rivail, Mme Grillot, Mlle Valérie Dillet et Mr Frédéric Bohr. Qu'ils soient remerciés pour ce travail parfois ingrat et pour leurs observations. Je remercie enfin mon propre père à qui sont dues les photographies insérées dans ce rapport.

Avant Propos

Ce rapport présente des travaux consacrés aux relations entre *caractéristiques structurales* et *activité pharmacologique* de différentes molécules. Ils ont été menés dans le cadre d'un Doctorat au Laboratoire de Chimie Théorique de l'Université de Nancy I.

Ce travail a ceci d'original qu'il est mené en un lieu orienté vers la mise au point et l'emploi des méthodes de calcul de la *chimie quantique*, et non dans le cadre plus habituel des laboratoires de l'industrie pharmaceutique. Il tire son origine des constats suivants :

o A l'heure actuelle, les recherches dans le domaine des relations quantitatives structure - activité (que nous abrégerons souvent par l'abréviation anglaise QSAR) sont le plus souvent marquées par l'empirisme. L'expérience et l'intuition des pharmacologues sont aussi déterminants (et efficaces...) que la connaissance scientifique pour mettre au point de nouveaux médicaments. Nous nous efforcerons ici de contribuer à une meilleure compréhension du rapport qui existe entre l'effet thérapeutique d'une molécule active et ses caractéristiques électroniques aussi bien que géométriques. Nous tirerons largement parti de l'expérience acquise au laboratoire en *modélisation moléculaire*.

o La conception des médicaments est une entreprise longue et coûteuse. L'enchaînement des différentes étapes (synthèse d'un composé jugé intéressant, test pharmacologique, recherches puis études de composés analogues, essais cliniques, reconnaissance des effets secondaires, brevet, commercialisation) peut exiger une dizaine d'années. C'est dire que des progrès, mêmes limités, peuvent avoir une certaine incidence économique.

o La puissance des ordinateurs a été considérablement accrue ces dernières années. Ceci justifie l'idée de développer de nouvelles applications fondées essentiellement sur les méthodes de la chimie quantique, par nature très gourmandes en calculs. Ce domaine reste largement à explorer, notamment parce que la collaboration entre personnes concernées (théoriciens et pharmaciens) est peu fréquente du fait de leurs formations très différentes.

— Plan général —

Le domaine des QSAR est un sujet très vaste que nous aborderons à l'échelle de l'angström. Nous nous focaliserons, non pas sur des mécanismes réactionnels mais sur *l'étude des grandeurs géométriques et théoriques susceptibles de gouverner la réaction chimique entre un médicament et un récepteur de l'organisme humain.* Sur le plan topologique, notre propos rappellera souvent le traditionnel modèle *clé-serrure* bien connu des chimistes. Sur le plan électronique, la prise en compte du potentiel, des charges ou de toute autre grandeur nous amènera à des méthodes de reconnaissance de formes plus élaborées.

o Nous rappellerons d'abord les difficultés inhérentes aux recherches en QSAR. Ce sera l'occasion de constater que nous ne savons que très peu de choses sur le sort d'un médicament après son administration. Pour imaginer un modèle d'interaction avec un récepteur, nous serons donc conduit à des hypothèses très simplificatrices. Un regard sur les travaux publiés par ailleurs nous guidera dans le choix des grandeurs moléculaires à étudier et tout autant sur la manière dont il faudra tirer parti de la masse d'informations qu'elles fournissent.

o Un second chapitre sera entièrement consacré aux apports de la chimie quantique à notre travail. Il débutera par un rappel des notions de base avant de présenter les grandeurs électroniques (potentiel, champ, densité, charges, polarisabilités, etc.) dont nous serons amené à observer l'évolution parallèlement à l'activité pharmacologique. Nous ne serons pas de simples utilisateurs de ces grandeurs dont nous pourrions confier le calcul à des logiciels spécialisés. Pour bien les maîtriser et pour les adapter aux besoins propres aux QSAR, nous en programmerons nous-même le calcul dans un nouveau logiciel.

o Dans un travail dont les mots-clés sont : observation, comparaison, topologie, surface, encombrement stérique, il est indispensable d'offrir au chimiste un outil visuel et interactif. Nous avons donc décidé d'inclure dans notre logiciel une interface graphique très complète et dont la réalisation a exigé à elle seule un temps important. Les photographies d'écran insérées dans ce rapport en donneront une idée concrète. Le chapitre 3 expliquera les grandes lignes de sa programmation et sera largement rédigé en pensant au profit que pourraient en tirer d'autres programmeurs. Il n'y sera pas question de pharmacologie.

o L'essentiel de ce rapport sera le compte rendu d'un long travail de programmation mené pour inclure dans notre logiciel des aspects variés de la modélisation moléculaire, mais toujours dans l'optique des relations structure - activité. Nous évoquerons des méthodes pour reconnaître et comparer des sites caractéristiques dans des molécules, pour tenir compte des conformations ou bien encore appliquer quelques notions issues de la théorie des graphes. Nous nous focaliserons ensuite sur les problèmes de surfaces puis sur l'encombrement stérique, selon plusieurs points de vue. La prise en compte des grandeurs électriques, notamment le potentiel, terminera cette partie.

o Le chapitre suivant sera consacré à l'analyse des données calculées dans notre logiciel. Nous verrons d'abord comment reconnaître les variables dont l'évolution semble dépendre de l'activité. Il y sera moins question de statistiques que de classifications. Puis nous en profiterons pour présenter des algorithmes inspirés de l'*intelligence artificielle* qui en reprennent les principaux traits.

o Quelques applications concrètes de notre logiciel seront exposées dans un dernier chapitre. Nous reprendrons une famille de molécules qui a fait l'objet de publications et montrerons en quoi notre outil peut apporter des informations variées et apporter un éclairage probablement assez différent des logiciels de modélisation déjà commercialisés. Notre objectif ne résidera pas en une étude systématique visant à prédire l'activité d'un nouveau composé de cette famille. Nous insisterons plutôt sur l'aspect méthodologique et sur l'idée majeure de cette thèse : concevoir un outil de recherche de molécules actives, qui soit facilement adaptable aux besoins des pharmacologues. Nous avons nommé notre logiciel : OSCAR (*Observation of Structural Characteristics for Activity Relationships*).

* * *

1 Des Relations Structure - Activité

Nous commençons ce chapitre par un aperçu de notre méconnaissance concernant le fonctionnement d'un médicament... Nous expliquerons donc de quelle manière nous aborderons les relations structure - activité et insisterons sur l'étude des analogies structurales entre molécules analogues. Un modèle d'interaction réactif - récepteur sera précisé, de même que ses hypothèses simplificatrices. Un petit compte rendu des travaux publiés par ailleurs permettra de mieux situer notre démarche.

1.1 Principaux aspects

Dire qu'une molécule est active ou non vis-à-vis d'un processus biologique relève surtout d'une observation. Si un prototype de médicament est administré à un rat ou à tout autre animal de laboratoire dans le but de combattre un traumatisme, son *activité* sera estimée de manière subjective. Par exemple, les pharmacologues s'efforceront de mettre en évidence une baisse de la pression artérielle, un dérèglement de la fréquence cardiaque ou encore les effets vasodilatateurs du composé.

Il est délicat de chiffrer numériquement une telle observable. Les tests employés en pharmacie sont d'ailleurs assez révélateurs. Par exemple, certains exprimeront une propriété d'anti-hypertenseur comme la baisse maximale de la pression systolique (en millimètres de mercure) rapportée au nombre de molécules administrées par voie orale à un rat hypertendu d'un certain poids... C'est pourquoi il n'est pas rare de trouver dans la littérature des valeurs un peu contradictoires, fournies sans barres d'erreur ou au contraire avec des incertitudes considérables. Dans d'autres cas, les activités sont simplement données sous une forme purement qualitative : très active, peu active, non mesurable, etc. Nous aurons l'occasion de revenir sur ces classifications plus sommaires, mais qui seront peut-être à préférer à des valeurs numériques continues si ces dernières sont d'une fiabilité douteuse.

Il faut garder à l'esprit la différence majeure entre les effets visibles d'un médicament et les réactions chimiques en cause. Si nous supposons que l'activité relève d'au moins un processus-clé qui survient en un endroit précis de l'organisme humain (par exemple une membrane), il est probable que le résultat de ce processus (passage ou non de la molécule à travers la membrane) est dénaturé avant d'être concrètement observable. Autrement dit, les données dont nous disposons sont non seulement incertaines mais vraisemblablement fausses.

Enfin, s'il paraît évident que l'activité est le résultat d'une succession de réactions chimiques, nous ne savons pas où ces dernières surviennent, dans quelles conditions se succèdent-elles et avec quels *récepteurs* de l'organisme. La pharmacologie moléculaire a beaucoup progressé ces dernières années en indiquant les sites en cause dans certains cas, mais pas au point de livrer des indications topologiques précises à l'échelle de l'angström sur ces récepteurs.

Du fait d'un tel manque d'informations, il est difficile de dépasser le stade de l'observation pour celui de l'explication de l'activité. En fait, une des rares solutions est de raisonner par *analogies* entre des molécules dont les caractéristiques structurales aussi bien que les propriétés thérapeutiques sont comparables. Nous parlerons donc, non pas d'une molécule, mais toujours d'un lot de composés analogues.

1.2 Hypothèses de travail

Nous pouvons diviser le processus biologique en plusieurs étapes :

1. Dès lors qu'un composé est injecté dans l'organisme, une certaine proportion de ses molécules sont dégradées ou plus généralement écartées du processus qui conduit à l'activité. Nous supposerons que cela tient à des phénomènes de transport ou à tout autre effet macroscopique qui intervient à une échelle supérieure à une dizaine d'angströms.

2. En revanche, nous ferons l'hypothèse que tous les composés du lot se comportent de la même manière dans ce trajet initial qui précède la rencontre avec un premier récepteur. Leurs seules différences sont dues en effet à des caractéristiques structurales plus fines.

3. Nous désignerons par récepteur, un site moléculaire de l'organisme sans autres indications sur sa nature. Il s'agira aussi bien d'une molécule partie intégrante d'un tissu que d'une espèce présente dans un liquide biologique. L'essentiel est que nos composés se ressemblent suffisamment pour interagir avec un même site.

4. Nous ne tiendrons pas compte des réactions ou effets secondaires. Nous ne parlerons que d'un seul type de récepteur, du point de vue de la topologie. Plus exactement, il sera le premier lieu où une interaction éventuelle avec nos composés pourra distinguer ces derniers les uns par rapport aux autres.

5. A cette première étape, certains composés réagiront d'une manière jugée positivement, parce que indispensable pour obtenir ultérieurement l'effet thérapeutique. Ces composés se retrouveront au début d'une seconde étape - décrite identiquement - de l'ensemble du processus.

6. D'autres réagiront différemment ou pas du tout et nous ferons alors l'hypothèse qu'ils ne pourront plus, d'une manière ou d'une autre, intervenir dans le processus

conduisant à l'activité. Les composés inactifs peuvent donc être éliminés pour différents motifs.

Nous sommes bien conscients que ces hypothèses sont très restrictives. Cependant, elles représentent un préalable indispensable pour dresser un cadre d'étude plus restreint et imaginer un modèle pour l'interaction de nos composés avec un site.

1. L'aptitude de nos composés à se rapprocher d'un récepteur dépend en premier lieu de leur forme géométrique. A priori, elle varie peu d'un composé à l'autre puisqu'ils sont assez ressemblants. Mais les différences ont un poids croissant au fur et à mesure que se réduit la distance avec le site d'"accueil".

2. La topologie de ce dernier est évidemment déterminante. Les considérations de formes seront moins utiles s'il est d'accès relativement facile que s'il est situé dans une paroi concave. La phase pendant laquelle le réactif s'orientera ou changera de conformation pour tenter de s'adapter au site sera un processus de reconnaissance de formes. Il est possible que d'emblée, certains composés aient une géométrie rédhibitoire dans l'optique d'une interaction positive.

3. Au début de ce processus, il est probable que les interactions électrostatiques soient plus importantes avec le solvant - dont nous ne savons rien dans le cas général - qu'avec le récepteur. Mais à plus courte distance entre les deux partenaires, les phénomènes d'attraction et de répulsion entre ces assemblages de charges que sont les molécules, deviennent déterminants. D'autres composés peuvent être éliminés à cette occasion.

4. Enfin, il est possible que si l'adéquation est toujours bonne entre un composé et le récepteur, des contraintes locales deviennent discriminantes vis-à-vis de l'activité. Ce peut être une distance entre deux substituants qui doit être comprise entre deux bornes, ou bien une distribution de charges à l'origine du mécanisme réactionnel.

1.3 Grandeurs moléculaires. Approche statistique

Il nous faut maintenant examiner les grandeurs moléculaires ou géométriques les plus adaptées à notre modèle. Avant de faire un choix, il est utile de savoir quelles sont les plus utilisées dans les travaux consacrés aux relations structure - activité.

La majeure partie des publications rendent compte de relations *statistiques* [1,2,3] entre effet thérapeutique et grandeurs moléculaires. Dans ce cadre, l'activité est exprimée sous la forme d'une relation linéaire du type :

$$A = \sum_i c_i x_i + K \qquad (1)$$

dans laquelle les x_i sont des descripteurs, les c_i leurs coefficients de pondération et K une constante plus ou moins ajustée. Ces relations offrent l'avantage de la simplicité

et la capacité à donner au moins une idée de l'activité d'une nouvelle molécule, si elles sont préalablement établies pour des composés analogues.

Leur gros inconvénient réside évidemment dans le choix des descripteurs x_i. La première difficulté est de définir les bonnes variables, c'est-à-dire les grandeurs dont on est sûr qu'elles gouvernent l'activité. Or même des grandeurs qui paraissent de prime abord importantes, comme la taille d'un composé, sa lipophilie ou ses grandeurs thermodynamiques n'offrent pas de garanties certaines de ce point de vue. La seconde difficulté tient à l'indépendance de ces variables. S'il semble que la taille et la lipophilie n'ont pas de rapport immédiat entre elles, ce n'est jamais le cas des grandeurs électroniques qui dépendent toujours indirectement des caractéristiques structurales. Il est fréquent de faire abstraction de cette relation de dépendance "cachée" mais sous l'angle statistique, cela est peu rigoureux.

Dans l'hypothèse où nous utilisons des variables "sûres" et "indépendantes", il faut encore s'assurer qu'elles sont toutes prises en compte, ce qui est pratiquement impossible. Dans le meilleur des cas, nous obtiendrons donc une relation statistique partielle, à interpréter avec prudence. Le propre des statistiques est de ne pas pouvoir démontrer formellement une relation de cause à effet ("l'activité dépend de ces variables"). Elle peuvent, au plus, démontrer le contraire, c'est-à-dire que nos x_i ne sont pas adéquats. D'ailleurs peut-être serait-il préférable d'obtenir une succession de résultats négatifs plutôt qu'une bonne régression linéaire qui nous inciterait à des conclusions hasardeuses sur l'importance de certaines variables.

Des améliorations peuvent être apportées à ces études. Une première voie consiste à affiner les traitements statistiques, par exemple en procédant à une *analyse en composantes principales*. Schématiquement, le but est de substituer aux descripteurs x_i, d'autres qui sont mathématiquement indépendants les uns des autres [4]. Une seconde voie consiste en une étude approfondie du rôle de chaque grandeur, afin de mettre en évidence une sorte de hiérarchie "chronologique". L'objectif est d'observer, par exemple, si un critère géométrique n'introduit pas à lui seul des conditions rédhibitoires à l'activité, rendant ainsi inutile l'observation d'autres variables [5]. Le but ultime de ce type de démarche est ainsi de sélectionner intelligemment les grandeurs à considérer avant de les soumettre à une statistique.

Une approche assez différente consiste à corréler l'activité, non avec des grandeurs numériques habituelles, mais avec des propriétés purement qualitatives exprimées le plus souvent sous une forme binaire. Par exemple, des grandeurs géométriques comme le volume ou une surface de contact sont remplacées par des tests sur l'existence de certains fragments de molécule ou certains substituants. Cette méthode de recherche et de comparaison de structures moléculaires est le propre de la plupart des systèmes experts [6,7].

1.4 Des descripteurs plus appropriés

Ce bref aperçu conduit à une seconde observation. Dans toutes ces statistiques, ce sont le plus souvent des grandeurs *macroscopiques* qui sont corrélées avec l'activité [1]. Nous avons déjà cité la lipophilie ou les enthalpies de formation. Il s'y ajoutent aussi des constantes d'équilibres chimiques ou des données cinétiques comme les constantes de vitesse. Les caractéristiques structurales sont souvent réduites à des paramètres empiriques, dont certains sont tabulés (équation de Taft [8,9]). De même les grandeurs électroniques sont simplifiées sous forme de constantes comme dans l'équation de Hammett [10].

Ces grandeurs sont simples et parfois, conduisent à des corrélations satisfaisantes avec l'activité. Mais pour des composés analogues, elles sont évidemment incapables d'expliquer des différences de propriétés dues alors à de petites nuances dans les caractéristiques structurales. Nous avons besoin ici de grandeurs moléculaires intervenant à l'échelle de l'angström et qui, de ce fait, concernent plutôt le champ électrique, les charges ou les polarisabilités atomiques. Ce sont là des variables plus théoriques, moins connues dans le domaine des relations structure - activité et dont il est a priori difficile de trouver un rapport avec un effet thérapeutique. En revanche, elles sont plus rigoureuses puisque intimement liées à la structure moléculaire, et calculables avec une précision suffisante.

Nous devons aussi nous focaliser sur l'encombrement stérique, dont nous avons vu qu'il était susceptible de jouer un rôle déterminant. Dans les publications, cet aspect est abordé souvent d'un point de vue purement qualitatif, notamment pour expliquer a posteriori l'inactivité d'un composé. La raison est relativement simple à comprendre : il n'existe pas de variable unique (la lipophilie, le moment dipolaire) facile à calculer et utilisable dans des séries très différentes de composés. Un travail plus fastidieux parce que plus systématique s'impose. En observant des molécules d'une même famille, il faut déceler toute particularité géométrique remarquable susceptible d'expliquer une propriété biologique (ou son absence). De tels paramètres, nous en ferons alors des *descripteurs* dont nous observerons s'ils évoluent d'une manière particulière avec l'activité. Si cela semble être le cas, il sera possible de les faire intervenir dans des relations statistiques.

1.5 Quel outil de recherche ?

Toutes ces réflexions précisent un peu quelle sera notre démarche. Dans un premier temps, nous nous focaliserons sur ces grandeurs électriques calculables par les méthodes de la chimie quantique. Nous verrons comment les obtenir ou plus exactement comment programmer leur calcul dans un logiciel de QSAR. Nous préciserons aussi le type d'informations que peut apporter chacune.

L'aspect géométrique de l'interaction réactif - récepteur exige un travail d'une nature bien différente. Notre effort doit porter non plus sur de gros calculs en arrière-plan mais sur la réalisation d'une interface graphique adaptée à un travail d'observation. Ce n'est qu'une fois ces deux aspects traités - dans les deux chapitres qui suivent - que nous développerons longuement nos descripteurs.

* * *

2 Activité et Grandeurs Electroniques

2.1 Rappels de chimie quantique

Schématiquement, il y a trois types de méthodes pour étudier les propriétés moléculaires. Nous nous intéresserons uniquement aux méthodes *semi-empiriques* qui se situent entre les calculs *ab-initio* et la *mécanique moléculaire*. Les premières ont l'avantage de fournir des résultats précis pour les grandeurs thermodynamiques et les géométries et permettent même d'accéder aux états de transitions. Elles sont de plus en plus utilisées du fait de la puissance croissante des ordinateurs, mais conduisent encore à des temps de calculs excessifs pour des molécules qui ont quelques dizaines d'atomes, comme les nôtres. Inversement, la mécanique moléculaire convient pour l'étude de plus gros systèmes, comme les enzymes, compte tenu de la courte durée des calculs qu'elle implique. Nous l'emploierons surtout pour distinguer rapidement des conformères. Assez empirique, elle est peu adaptée à l'étude des réactions chimiques et n'est pas assez précise pour calculer des grandeurs moléculaires fines, que nous voulons mettre en relation avec des propriétés biologiques.

2.1.1 Les méthodes du champ auto-cohérent

Les méthodes ab-initio et semi-empiriques reposent sur le même formalisme de la mécanique quantique. Le comportement d'une molécule dans un niveau énergétique E est décrit par une fonction d'onde Ψ, solution de l'équation de Schrödinger :

$$H_t \, \Psi = E \, \Psi \tag{2}$$

H_t est l'opérateur hamiltonien qui inclut les principales contributions à l'énergie totale de la molécule :

$$H_t = T_N + V_{NN} + T_e + V_{eN} + V_{ee} \tag{3}$$

T_N et T_e sont respectivement les énergies cinétiques des noyaux atomiques et des électrons, tandis que V_{NN}, V_{eN} et V_{ee} désignent les potentiels de répulsion entre deux noyaux, d'attraction entre un noyau et un électron et de répulsion entre deux électrons. Compte tenu de la différence de masse entre électrons et noyaux, il est raisonnable de supposer que ces derniers occupent dans l'espace des positions fixes par rapport aux premiers et donc d'éliminer le terme T_N [11]. En outre, V_{NN} ne dépendant que des coordonnées nucléaires, qui ne sont plus des variables de l'hamiltonien, la fonction d'onde recherchée est aussi fonction propre de l'opérateur H,

$$H \, \Psi = E \, \Psi \tag{4}$$

qui ne regroupe que des contributions électroniques à l'énergie de la molécule. En unités atomiques, nous avons plus précisément,

$$H = T_e + V_{eN} + V_{ee} = \sum_i (-\frac{1}{2}\Delta_i + V_{Ni}) + \sum_i \sum_{j<i} \frac{1}{r_{ij}} \quad (5)$$

expression dans laquelle les indices i et j désignent deux électrons. Δ_i est l'opérateur laplacien [1], V_{Ni} la somme des potentiels d'attraction exercés par chaque noyau sur un électron et r_{ij} la distance entre deux d'entre eux. H se décompose donc en un hamiltonien de cœur, somme de termes monoélectroniques, et un potentiel de répulsion, somme de termes biélectroniques.

Ces derniers compliquent considérablement la résolution de l'équation de Schrödinger car ils font intervenir deux variables non séparables. L'*approximation orbitale* consiste alors à simplifier Ψ sous la forme d'un produit de fonctions mathématiques dépendant chacune des coordonnées d'un seul électron [12]. Mais le principe de Pauli imposant à ce produit de changer de signe lors de la permutation de deux d'entre eux, Ψ doit être en fait une combinaison linéaire de telles fonctions, appelée *déterminant de Slater :*

$$\Psi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_a(1) & \chi_b(1) & \dots & \chi_N(1) \\ & \dots & & \\ \chi_a(n) & \chi_b(n) & \dots & \chi_N(n) \end{vmatrix} \quad (6)$$

Dans cette expression, les N fonctions χ sont des *spinorbitales* qui portent sur un électron et n est le nombre total de ces derniers. Chaque χ est le produit de deux fonctions, la première, ψ, dépendant des coordonnées spatiales de l'électron et la seconde, η, de son état de spin (l'une ou l'autre de deux valeurs quantiques qui caractérisent le mouvement de toupie de l'électron sur lui-même).

Dans ces conditions, calculer l'énergie d'une molécule revient à résoudre :

$$E = \langle \Psi \mid H \mid \Psi \rangle = \left\langle \mid \chi_1\chi_2\cdots\chi_N \mid \mid \sum_i H_i^{coeur} + \sum_i \sum_{j<i} \frac{1}{r_{ij}} \mid \mid \chi_1\chi_2\cdots\chi_N \mid \right\rangle \quad (7)$$

qui, après développements [13], conduit à :

$$E = \sum_{k=1}^{N} H_k^c + \sum_{k=1}^{N} \sum_{l<k}^{N} (J_{kl} - K_{kl}) \quad (8)$$

où les indices k et l portent sur les N spinorbitales du déterminant de Slater. E peut être calculée à partir des trois types d'intégrales : [2]

$H_k^c =$ $\langle \chi_k(1) \mid H^c(1) \mid \chi_k(1) \rangle$ (intégrale de cœur)

$J_{kl} =$ $\langle \chi_k(1)\chi_l(2) \mid \frac{1}{r_{12}} \mid \chi_k(1)\chi_l(2) \rangle$ (intégrale coulombienne)

$K_{kl} =$ $\langle \chi_k(1)\chi_l(2) \mid \frac{1}{r_{12}} \mid \chi_l(1)\chi_k(2) \rangle$ (intégrale d'échange)

[1] $\Delta_i = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}$ en coordonnées cartésiennes

[2] 1 et 2 désignent deux électrons quelconques

où ne comptent d'ailleurs que les variables d'espace, l'intégration sur le spin donnant toujours un résultat immédiat, 1 ou 0, selon le cas.

Dans l'état fondamental, Ψ correspond à un extremum de E. Donc, nous pouvons supposer qu'une petite modification d'une de ses spinorbitales - sans remettre en cause ses propriétés mathématiques - ne doit pas faire varier l'énergie de la molécule. En appelant Ψ_{lt} la fonction d'onde dans laquelle χ_l est "altérée" par χ_t, cette condition d'extremum s'écrit :

$$F_{lt} = \langle \Psi \mid H \mid \Psi_{lt} \rangle = 0 \tag{9}$$

où F_{lt} s'écrit avec les relations précédentes :

$$\langle \chi_l(1)|H^c(1)|\chi_t(1)\rangle + \sum_i \left(\langle \chi_l(1)\chi_i(2)|\frac{1}{r_{12}}|\chi_t(1)\chi_i(2)\rangle - \langle \chi_l(1)\chi_i(2)|\frac{1}{r_{12}}|\chi_t(2)\chi_i(1)\rangle \right)$$

ou plus simplement :

$$F_{lt} = \langle \chi_l \mid H^c \mid \chi_t \rangle + \sum_i \left(\langle \chi_l \mid J_i \mid \chi_t \rangle - \langle \chi_l \mid K_i \chi_t \rangle \right) \tag{10}$$

en faisant apparaître deux nouveaux opérateurs J et K. Si nous considérons plus généralement F_{lt} comme un élément d'une matrice carrée de spinorbitales, l'équation (9) devient alors un système de N équations dont les inconnues sont ces N spinorbitales χ_k et dont les valeurs propres ont la dimension d'une énergie :

$$F \chi_k = \epsilon_k \chi_k \tag{11}$$

Ces équations, dues à Hartree et Fock, ne peuvent être résolues que de façon itérative. En effet, l'opérateur F qui permet d'obtenir les solutions χ_k est exprimé en fonction de deux opérateurs, J et K, qui sont eux-mêmes définis à partir des inconnues χ_k. Dans la pratique, la résolution est donc menée par des approximations successives, les fonctions propres calculées à la $n^{ième}$ itération permettant de calculer l'opérateur F de l'étape suivante. C'est la méthode du champ auto-cohérent (*Self Consistent Field* ou SCF).

— Orbitales moléculaires —

Le plus souvent, les orbitales moléculaires ψ du déterminant de Slater sont elles-mêmes construites à partir de *fonctions de base*. Dans l'approximation LCAO (*Linear Combination of Atomic Orbitals*), ce sont les orbitales atomiques - déduites des solutions de l'équation de Schrödinger pour l'atome hydrogénoïde - qui sont retenues :

$$\psi_i = \sum_\mu c_{\mu i} \varphi_\mu \tag{12}$$

Elles sont le produit d'une partie radiale et d'une partie angulaire exprimées dans un système de coordonnées sphériques (r,θ,ϕ) dont chaque atome est l'origine :

$$\varphi_{nlm}(r,\theta,\phi) = N_{nl}(r)\, Y_{lm}(\theta,\phi) \tag{13}$$

n, l et m étant les nombres quantiques à l'origine de la nomenclature de ces orbitales - 1s pour le triplet (1,0,0), 2s pour (2,0,0), 2p pour (2,1,-1...1), etc. - et Y désignant les harmoniques sphériques. Dans la pratique, une forme paramétrée, due à Slater, est très employée :

$$\varphi_{nlm}(r,\theta,\phi)(\zeta) = \frac{(2\zeta)^{n+1/2}}{\sqrt{(2n)!}}\, r^{n-1}\, e^{-\zeta r}\, Y_{lm}(\theta,\phi) \qquad (14)$$

Cette expression décrit pratiquement aussi bien le comportement des électrons à la périphérie de l'atome, donc dans les liaisons chimiques, tout en étant plus facile à manier dans les calculs.

Dans le cas d'une molécule non chargée et dont chaque orbitale est occupée par deux électrons, nous pouvons simplifier les spinorbitales χ en orbitales ψ. La réécriture du système d'équations Hartree - Fock (11) en tenant compte de l'approximation LCAO (12), conduit alors aux *équations de Roothaan* selon un traitement tout-à-fait analogue au précédent :

$$F\, c_k = \epsilon_k\, S\, c_k \qquad (15)$$

Dans cette expression, les inconnues ne sont plus les spinorbitales χ_k mais les coefficients $c_{\mu k}$ des μ orbitales atomiques utilisées. De plus, ces fonctions de base sont quelconques, contrairement aux spinorbitales qui sont mathématiquement orthogonales entre elles. C'est pourquoi il apparaît une matrice S dont chaque élément ij est l'intégrale de recouvrement entre les orbitales atomiques φ_i et φ_j. La résolution du système d'équations nécessite donc une étape supplémentaire où de nouvelles fonctions, orthogonales entre elles, sont déduites des précédentes par une méthode d'orthogonalisation. Les calculs sont menés comme avant, de manière itérative.

— Ab-initio et semi-empiriques —

Les méthodes ab-initio et semi-empiriques ne diffèrent que par les modalités du calcul qui conduit à la résolution des équations Hartree - Fock. Dans les premières, toutes les intégrales et tous les électrons sont pris en compte. Les orbitales de Slater, difficiles à calculer, sont remplacées par des "contractions" de *gaussiennes*, sommes de fonctions mathématiques dont les propriétés facilitent le calcul. Mais il faut surtout retenir que la taille des fichiers sur disque et le temps de calcul augmentent approximativement comme n^4, où n est le nombre de fonctions de base utilisées. En revanche, la qualité des résultats croît beaucoup moins vite et tend vers une certaine limite, compte tenu de l'approximation orbitale.

Dans les méthodes semi-empiriques, les électrons internes ne sont pas pris en compte dans les calculs. Un noyau de soufre aura, par exemple, une charge de 6 unités atomiques. Les intégrales de recouvrement et un grand nombre d'intégrales biélectroniques sont négligées, les différentes méthodes se distinguant par l'étendue des simplifications. De plus, certains éléments de la matrice de Fock ne sont pas calculés mais déterminés empiriquement ou ajustés sur d'autres valeurs expérimentales. Enfin, le temps de calcul

et la place nécessaire sur disque augmentent approximativement comme n^2, où n est le nombre d'orbitales de valence.

Dans la pratique, le choix de la méthode est donc dicté par la taille de la molécule, les informations voulues (charges, géométries optimisées, potentiels d'ionisation, chemins de réaction, etc.), la précision souhaitée pour ces grandeurs et la durée du calcul. Dans notre cas, le premier critère rend obligatoire à lui seul le recours aux méthodes semi-empiriques.

Parmi ces dernières, celle qui procède au maximum de simplifications et qui est donc la plus rapide s'appelle CNDO. Elle fournit de bonnes valeurs pour les distances et longueurs de liaison, de même que pour les charges nettes sur les atomes, mais conduit à des valeurs de l'énergie parfois éloignées des données expérimentales. Ce défaut peut être gênant dans la recherche des conformères d'une molécule. En revanche, les méthodes de type NDDO et leurs variantes plus récentes MNDO [14] sont nettement plus lourdes du fait d'approximations moins poussées. Elles approchent correctement la plupart des grandeurs - énergies de formation, moments dipolaires ou constantes de force - fournissent des géométries réalistes mais peuvent être mises en défaut par des liaisons hydrogènes. Nous avons finalement retenu la méthode PM3 [15,16], une des variantes MNDO les mieux paramétrées et les plus utilisées [1].

2.1.2 Utilisation des méthodes semi-empiriques

GEOMOS [17] et CHIMISTE [18] sont deux des logiciels qui permettent de calculer la fonction d'onde comme bon nombre de grandeurs thermodynamiques ou spectroscopiques. Ils offrent le choix entre la plupart des méthodes semi-empiriques et proposent, en outre, des techniques d'optimisation de géométrie. Nous avons donc décidé de les utiliser et d'obtenir ainsi, pour chacune de nos molécules, les solutions des équations de Roothaan :

- un jeu de vecteurs propres, les coefficients des orbitales atomiques de Slater dans l'expression de chaque orbitale moléculaire,
- les valeurs propres correspondantes, c'est-à-dire les énergies de ces dernières.

Ce sont les résultats les plus importants, d'une part parce qu'ils indiquent la répartition des électrons dans une molécule, "géographiquement" [2] et par niveau énergétique, d'autre part parce qu'ils permettent d'accéder à leur tour à des grandeurs au sens un peu plus physique, exposées dans le paragraphe 2.2.

Pour étudier les relations structure - activité, le recours à GEOMOS était donc un préalable indispensable à l'utilisation de notre propre programme. C'est là un bon

[1] Il est surtout important de conserver la même méthode pour toutes les molécules à étudier

[2] plus exactement leur densité de probabilité de présence dans une région de l'espace

compromis entre le calcul de la fonction d'onde qui est confié à un logiciel conçu pour cela, et son utilisation qui est de notre ressort. Du point de vue de la chimie quantique, OSCAR se limite (!) aux calculs ultérieurs de charges, moment dipolaire, champ et potentiels électriques, etc. L'intégration directe dans notre logiciel des optimisations de géométrie et de l'essentiel de la mécanique SCF est un travail de grande ampleur, qui sort du cadre des QSAR et qui est en cours au moment de la rédaction de ce rapport.

2.2 Propriétés à considérer

Une des grandeurs moléculaires les plus intéressantes pour nous est le *potentiel électrostatique*, car il joue probablement un rôle déterminant dans la rencontre site actif - site récepteur. Si nous reprenons l'approche d'un réactif imaginée au paragraphe 1.2, le potentiel doit gouverner l'interaction probablement après les critères géométriques globaux, qui éliminent les molécules de formes inadéquates, mais avant les critères locaux, qui sélectionnent les dernières candidates. Encore une fois, ce schéma est assez réducteur - il est peu vraisemblable qu'une et une seule propriété décide de l'interaction à un éloignement donné du récepteur - mais bien utile. Le potentiel a évidemment un sens en tout lieu. Mais à des distances "lointaines", il détermine certainement plus l'interaction du réactif avec son environnement - les molécules du liquide biologique - qu'avec un récepteur éventuel. Et à "faible" distance, sa valeur, même forte en valeur absolue, est probablement sans conséquence s'il existe par ailleurs une contrainte stérique rédhibitoire.

Nous reviendrons amplement sur les différentes manières proposées par OSCAR pour observer l'évolution du potentiel et du champ électriques dans une série de molécules. Il s'agira là encore de regarder s'ils permettent de définir des critères discriminants vis-à-vis de l'activité dans les deux cas.

2.2.1 Potentiel électrique

Dans l'immédiat, rappelons comment calculer le potentiel en tout point M d'une molécule (ou du moins pas trop près d'un atome) :

$$V(M) = \sum_{noyaux\ k} \frac{Z_k}{r_{kM}} - \langle \Psi \mid \frac{1}{r_M} \mid \Psi \rangle \qquad (16)$$

Dans cette expression, Z_k est la charge de l'atome k et Ψ, la fonction d'onde multiélectronique. Dans le formalisme de Roothaan, le potentiel s'écrit aussi :

$$V(M) = \sum_{noyaux\ k} \frac{Z_k}{r_{kM}} - \sum_{OM\ i} n_i \langle \psi_i \mid \frac{1}{r_M} \mid \psi_i \rangle \qquad (17)$$

où n_i désigne le nombre d'électrons décrits par l'orbitale moléculaire i. Si nous reprenons ensuite l'approximation LCAO (équation 12), nous avons :

$$V(M) = \sum_{noyaux\ k} \frac{Z_k}{r_{kM}} - \sum_i n_i \sum_\mu c^*_{\mu i} \sum_\lambda c_{\lambda i} \langle \varphi_\mu \mid \frac{1}{r_M} \mid \varphi_\lambda \rangle \qquad (18)$$

où les indices μ et λ portent cette fois sur les orbitales atomiques de Slater.

Le calcul des seules contributions électroniques au potentiel fait donc intervenir deux types d'intégrales biélectroniques, selon que les orbitales φ_μ et φ_λ portent sur deux atomes différents, ou bien sur le même [19]. Dans la pratique, nous négligerons les premières, d'une part pour rester cohérent avec les approximations propres aux méthodes semi-empiriques, d'autre part pour réduire la durée des calculs qui seront menés en mode interactif avec OSCAR. Les résultats seront forcément entachés d'une certaine erreur, mais comme elles interviendront vraisemblablement dans le même sens pour toutes nos molécules, les comparaisons resteront valables.

2.2.2 Densité électronique

Kier [20] a présenté un point de vue différent dans lequel le récepteur reconnaît des régions, similaires du point de vue de la densité électronique, et que peuvent présenter des molécules de formes comparables. Depuis, les chimistes utilisent le terme de *pharmacophore* pour désigner une telle région, et par extension, l'assemblage des atomes qui est requis pour obtenir un effet biologique particulier. La densité électronique livre pourtant une information différente du potentiel électrique et dont il n'est pas certain qu'elle soit plus importante. C'est d'abord une grandeur strictement positive, et non algébrique, et qui, comme son nom l'indique, rend compte de la concentration des électrons au voisinage des atomes. Elle se mesure donc à des distances inférieures à celles du potentiel, ce qui en fait une grandeur plus locale.

Numériquement, la densité électronique en un point M est :

$$\rho(M) = \sum_i n_i \psi^*_i(M) \psi_i(M) \qquad (19)$$

qui, dans le cadre LCAO (équation 12), s'écrit aussi :

$$\rho(M) = \sum_i n_i \sum_\mu c^*_{\mu i} \varphi^*_\mu(M) \sum_\lambda c_{\lambda i} \varphi_\lambda(M) \qquad (20)$$

En utilisant la *matrice densité* dont les éléments sont issus des vecteurs propres et définis par :

$$P_{\mu\lambda} = \sum_i n_i c^*_{\mu i} c_{\lambda i} \qquad (21)$$

nous obtenons finalement deux expressions plus simples, aussi bien pour le potentiel électrique que pour la densité électronique :

$$\rho(M) = \sum_\mu \sum_\lambda P_{\mu\lambda} \varphi^*_\mu(M) \varphi_\lambda(M) \qquad (22)$$

$$V(M) = \sum_{noyaux\ k} \frac{Z_k}{r_{kM}} - \sum_{\mu} \sum_{\lambda} P_{\mu\lambda} \langle \varphi_\mu \mid \frac{1}{r_M} \mid \varphi_\lambda \rangle \qquad (23)$$

Elles montrent [1] d'ailleurs que le calcul de la densité électronique demande nettement moins de temps que celui du potentiel et qu'à des distances conséquentes des atomes, il est probable que la densité soit faible et assez comparable d'une molécule à l'autre. Elle ne pourra donc être un bon critère de comparaison que dans des cas particuliers, par exemple dans des zones concaves, voire dans des "poches".

2.2.3 Polarisabilité moléculaire

La densité électronique nous conduit naturellement à rechercher une interprétation plus fine du comportement d'une molécule à partir de sa fonction d'onde. Nous abordons là un sujet plus délicat, puisqu'il s'agit d'exploiter des données quantiques, moins physiques que des grandeurs électriques et donc moins faciles à appréhender. De plus, elles sont beaucoup plus sensibles aux erreurs commises durant les calculs SCF. En bref, nous abordons là des descripteurs dont il est rarement question en QSAR et dont l'emploi comme variables à corréler avec l'activité requiert plus de prudence.

Quelles informations intéressantes peut-on espérer dans l'optique d'une interaction site actif - site récepteur ? A priori, ce sont essentiellement des résultats locaux, susceptibles d'expliquer des mécanismes réactionnels concernant des molécules qui ont déjà satisfait les autres critères pour agir avec le récepteur. La seule donnée suffisamment générale pour jouer un rôle antérieur est la polarisabilité moléculaire. Elle rend compte de l'aptitude de chaque molécule à ajuster sa structure électronique sous l'effet d'un champ électrique dû essentiellement à la présence d'un solvant et d'un récepteur.

Si μ et μ_0 désignent respectivement les moments dipolaires résultant et permanent d'une espèce soumise à un champ E, la réponse de la molécule se déduit de l'expression simplifiée :

$$\mu = \mu_0 + \alpha E + k\beta E^2 + \ldots \qquad (24)$$

dans laquelle α désigne la polarisabilité moléculaire [21]. Des essais préliminaires ont montré que cette grandeur variait surtout avec certains substituants, comme les halogènes.

[1] rappelons que ces expressions ne sont pas valables trop près d'un atome puisqu'elles font intervenir des orbitales de Slater

2.2.4 Energies des orbitales frontières

Les premières orbitales concernées dans une interaction chimique sont les orbitales frontières [22], c'est-à-dire la dernière orbitale moléculaire occupée par des électrons (*HOMO: highest occupied molecular orbital*) et la première qui soit vacante (*LUMO: lowest unoccupied molecular orbital*). Elles sont faciles à mettre en évidence à l'issue d'un calcul SCF-Roothaan, car dans les fichiers de résultats, elles apparaissent séparées par un "saut" énergétique.

Schématiquement, deux cas de figure sont envisageables. Les substitutions nucléophiles sont explicables par une "fusion" progressive de la *HOMO* d'un réactif avec la *LUMO* d'un récepteur. Ainsi dans une SN_2 typique, le réactif OH^- s'approche d'une liaison C - X et s'attaque à une orbitale vacante et antiliante située du côté opposé à X. Inversement, les substitutions électrophiles résultent d'une interaction entre la *LUMO* d'un réactif et la *HOMO* d'un récepteur. Dans une SE_2 l'attaque électrophile de H^+ porte sur une orbitale liante C - C pour conduire à un état de transition à trois centres. Reste que dans notre cas, le fait que nous ne connaissions pas le récepteur nous empêche de nous rapprocher d'un cas d'école. Les perturbations et les mouvements moléculaires invoquent simplement les mêmes éléments de symétrie que leurs orbitales frontières.

La première idée qui s'impose est de regarder si nos molécules se distinguent les unes des autres par un écart énergétique appréciable entre les deux orbitales frontières. En première approximation, une différence importante serait significative d'une molécule plus stable donc moins apte à réagir. Ce résultat serait alors à rapprocher de l'examen de grandeurs thermodynamiques, comme l'énergie totale ou la chaleur de formation.

La comparaison entre molécules ne peut porter que sur des différences d'énergies entre orbitales frontières, non sur leurs valeurs absolues. En effet, ces dernières dépendent indirectement du nombre d'atomes de chaque espèce et ce dernier peut varier de plus d'une dizaine. Il faudra cependant vérifier l'absence d'espèces qui se distingueraient surtout par des niveaux énergétiques sensiblement plus bas (ou plus haut). Dans ce cas, il serait vraisemblable que l'interaction avec le récepteur s'effectue selon d'autres modalités. Il est possible de montrer [23] qu'elle serait d'autant plus favorable que les populations électroniques du réactif et du récepteur seraient différentes.

Regarder comment évolue l'écart énergétique *HOMO - LUMO* apporte une information insuffisante. Les orbitales moléculaires voisines, "au-dessous" de la première ou "au-dessus" de la seconde, sont également à prendre en compte, d'une part parce que leurs énergies sont très proches, d'autre part parce qu'elles peuvent présenter des symétries plus adaptées à un mécanisme réactionnel. Par exemple, des mouvements électroniques peuvent intervenir entre la *HOMO* et la *LUMO+1* de deux partenaires. Ces considérations interdisent donc la définition d'un descripteur unique à corréler avec l'activité. C'est pourquoi, nous avons surtout prévu dans le logiciel OSCAR, de guider l'utilisateur par l'affichage de graphes, où l'activité est portée en fonction de la différence d'énergie entre des orbitales à choisir.

2.2.5 Charges et polarisabilités atomiques

La meilleure information que fournissent les orbitales moléculaires est la connaissance des charges portées par chaque atome. Si P désigne la matrice densité, nous avons :

$$q_A = \sum_{\mu} P_{\mu\mu} \tag{25}$$

où la sommation ne prend en compte que les éléments de la matrice densité relatif aux orbitales de valence de l'atome A. Il faut ensuite tenir compte des électrons internes pour obtenir une charge totale. Dans des molécules où prédomine la résonance, il est ainsi possible de mettre en évidence des effets électroniques bien plus fins que ce que prévoient les tables (tableau 1) ou les paramètres ρ de l'équation empirique de Hammett [10,24] [1].

Cette information peut être affinée par un calcul de polarisabilités pour estimer la facilité avec laquelle se déforme le nuage électronique local sous l'action d'un champ extérieur. Deux solutions sont possibles selon la précision souhaitée pour cette grandeur. La première consiste à les calculer par une méthode variationnelle pendant les itérations SCF [25] , puis à récupérer les résultats dans le logiciel OSCAR comme pour les vecteurs propres. La seconde, plus rapide mais moins précise, propose une estimation de l'*autopolarisabilité* de chaque atome, définie sous une forme simplifiée à partir des vecteurs propres :

$$SP_A = 4 \sum_i \sum_j \frac{\sum_\mu \sum_\lambda c_{\mu i} c_{\lambda i} c_{\mu j} c_{\lambda j}}{(\epsilon_i - \epsilon_j)} \tag{26}$$

Dans cette expression, les indices i et j portent sur les orbitales moléculaires d'énergies ϵ_i et ϵ_j, tandis que les indices μ et λ ne concernent que les orbitales de valence de l'atome [26]. Dans la pratique, ce calcul est parfois même restreint aux seules orbitales frontières. Cette variante est également implantée dans notre programme, avec, comme d'habitude, des possibilités d'affichage plus explicite.

Ces données peuvent être réunies dans des variables plus générales, la somme des valeurs absolues des charges nettes, le moment dipolaire, la polarisabilité globale de la molécule et son anisotropie ou encore la valeur absolue de la somme des autopolarisabilités [26]. Elles sont plus souvent utilisées dans des relations statistiques et là encore, toutes peuvent être obtenues dans notre logiciel.

---

[1] En série benzénique, les constantes de vitesse de réaction des dérivés méta et para substitués sont fonction de deux paramètres, ρ et σ, fonctions respectivement du type de réaction et du substituant :

$$log\left(\frac{K}{K_0}\right) = \rho\,\sigma$$

où K_0 est la constante de vitesse dans le cas de l'acide benzoïque. ρ et σ sont tabulés

Tableau 1: Effets électroniques des principaux substituants

(comparaisons qualitatives [24])

	accepteur mésomère	inductif attracteur	inductif accepteur	donneur mésomère
$-N^{\oplus} \equiv N$	xxxx	xxxx		
$-NO_2$	xxx	xxx		
$-N^{\oplus}Me_3$		xxxx		
$-CF_3$		xxxx		
$\succ C=O$	xx	xx		
$-C\equiv N$	xx	xxx		
$-CO_2Me$	x	xx		
$-SO_2Me$		xxx		
-F		xxx		xx
-Cl		xx		x
-Br		xx		x
-I		x		x
$-C_6H_5$	x	x		x
-Me			x	x
$-CMe_3$			xx	
-SMe	x	x		x
-OMe		x		xx
$-NMe_2$				xxx
$-SiMe_3$			xxx	
$-O^{\ominus}$			xxx	xxxx

2.2.6 Délocalisation électronique

Le défaut du calcul des charges nettes est qu'il ne détaille pas la répartition de cette charge en fonction des orbitales moléculaires. Il est donc souhaitable d'examiner plus précisément la délocalisation électronique. Une première idée est de calculer le "poids" d'un atome dans une orbitale donnée i, c'est-à-dire la proportion suivante :

$$\Gamma = \frac{\sum_{\mu} c_{\mu i}^2}{\sum_j c_{ji}^2} \tag{27}$$

dans laquelle l'indice j porte sur la totalité des orbitales atomiques tandis que μ ne concerne que celles de l'atome. Cette solution est rapide et l'option d'affichage correspondante met bien en évidence les orbitales délocalisées ou au contraire celles qui sont concentrées au niveau d'un atome (le doublet libre d'un azote par exemple).

Même si l'observation s'avère peu concluante, c'est-à-dire si l'activité ne puisse être mise en rapport avec une délocalisation électronique particulière, il est souhaitable d'examiner aussi les orbitales voisines, surtout pour des systèmes résonnants. Dans

cette optique, le logiciel OSCAR offre des graphes pour étudier le poids d'un atome en fonction des niveaux énergétiques mais comme il devient difficile de résumer une information plus abondante, nous avons prévu le calcul des descripteurs suivants :

$$SDE = 2 \sum_i \frac{\sum_\mu c_{\mu i}^2}{\epsilon_i} \tag{28}$$

SDE donne une idée de la force avec laquelle chaque atome retient les électrons de la molécule (l'indice i porte sur les orbitales moléculaires occupées, μ ne concerne que les orbitales de l'atome). Là encore, il est fréquent de limiter le calcul à la seule *HOMO*.

$$SDN = 2 \sum_i \frac{\sum_\mu c_{\mu i}^2}{\epsilon_i} \tag{29}$$

Inversement, SDN représente l'aptitude de chaque atome à céder des électrons (l'indice i porte cette fois sur les orbitales moléculaires inoccupées). Ces deux descripteurs sont appelés *super-délocalisabilités* électrophiles et nucléophiles respectivement. Des essais ont montré qu'ils fournissaient en général une information très comparable aux charges nettes.

Il existe enfin un dernier moyen pour étudier la densité électronique dans une molécule. La démarche est nettement différente puisqu'il s'agit non plus d'exploiter les orbitales issues du calcul SCF-Roothaan mais de les transformer sous la forme d'*orbitales localisées* plus adaptées à l'étude des réactions chimiques. Toutes ces méthodes sont fondées sur des critères plus ou moins physiques comme par exemple, la minimisation de l'énergie d'interaction des électrons [27], ou bien la maximisation de la distance séparant les centres de gravité de chaque orbitale [28]. L'une et l'autre sont disponibles au laboratoire [18].

* * *

3 Graphisme moléculaire

3.1 Préliminaires

Nous avons choisi d'utiliser le langage FORTRAN pour tout le travail de programmation mené dans le cadre de cette thèse, et tout particulièrement pour sa partie graphique. Bien que moins moderne et moins structuré que d'autres, c'est le plus répandu et le mieux connu dans l'industrie. Ceci est important puisque nous ne souhaitions pas faire du logiciel OSCAR un "produit" à livrer clés en main mais un outil facilement modifiable par des pharmacologues avertis aux délices de la programmation. C'est aussi dans cet ordre d'idées que nous avons écrit un code source relativement simple (pour éviter les particularités des compilateurs) et surtout très modulaire. La majorité des algorithmes ont été inclus dans des bibliothèques d'emploi indépendant et beaucoup plus général. Les programmeurs intéressés par les détails techniques en trouveront un résumé à l'annexe A.

Le système d'exploitation utilisé au laboratoire s'appelle AIX. Mais nous nous sommes efforcé de ne pas y faire appel durant l'écriture du programme. Le but était, là encore, d'anticiper les problèmes d'adaptation à d'autres environnements plus répandus dans l'industrie, comme MVS ou d'autres versions d'UNIX. L'adoption de normes dans ce domaine, solution à bien des problèmes sordides de compatibilité, en est encore aux balbutiements comme en témoignent les dissensions actuelles entre constructeurs. C'est pourquoi nous insistons tant sur la portabilité des programmes qui évite de perdre parfois beaucoup de temps.

Des stations de travail d'une puissance d'environ 20 mégaflops [1] et un outil de programmation prévus en particulier pour le graphisme étaient disponibles au laboratoire (tout en étant de plus en plus répandus ailleurs). Ils sont indispensables pour réaliser un outil d'étude des molécules axé sur l'observation, l'intuition, la souplesse et la facilité d'emploi. Une présentation plus détaillée en est nécessaire avant de développer les algorithmes écrits pour nos besoins.

3.2 Graphisme sur ordinateur

En informatique, une des normes élaborées pour les programmes axés sur le graphisme est la norme PHIGS, abréviation de *Programmer's Hierarchical Interactive Graphics System*. C'est en premier lieu, un ensemble de concepts qui décrivent les Objets dans l'Espace : que sont-ils, comment les construire, où les situer, qui les observe.

[1] millions d'opérations traitées par un microprocesseur en une seconde

Il s'agit ensuite de formuler de façon aussi systématique comment ces objets pourront être déformés, déplacés, en bref comment agir dessus. De nombreuses autres actions sont à envisager, comme l'éclairage, la superposition des objets, leurs réactions face à tel ou tel événement... Ainsi, la norme PHIGS est un *formalisme* de toutes ces situations théoriques. Puis elle définit une sorte de cahier de charges (à respecter) pour traduire en programmation toutes ces situations en autant de *procédures.*

Dans notre cas, il sera à peine moins abstrait de considérer le chimiste comme un observateur. Les objets qui l'intéressent, par exemple des molécules, seront matérialisés sur une scène, face à lui. Il pourra alors les choisir, les sélectionner, les désigner, modifier même l'angle d'observation, avant de les soumettre à d'autres traitements.

Dans ce cadre théorique et général, il n'est pas question de telle ou telle marque d'ordinateur mais uniquement de stations de travail virtuelles et en trois dimensions... C'est aux constructeurs de fournir un outil de programmation, fondé sur la norme, qui prendra en compte ces questions bassement matérielles. Pour notre part, nous disposions du progiciel graPHIGS disponible sur des machines IBM RISC 6000 et fonctionnant sous *X Windows.*

Se familiariser avec cet outil de développement est un travail de longue haleine pour les néophytes. Une fois maîtrisés les concepts de la norme PHIGS, il faut mener de très nombreux essais pour les mettre en pratique. Ce n'est donc que petit à petit qu'a pris forme l'organisation d'un programme interactif qui soit un assemblage cohérent de procédures de modélisation moléculaire.

3.3 Dessins de molécules

graPHIGS se présente essentiellement comme une bibliothèque de procédures auxquelles le programmeur fait appel dans son code source. Ces *subroutines* se répartissent en plusieurs ensembles.

Pendant la session [1], des requêtes permettent d'interroger l'ordinateur quant à ses capacités réelles : c'est une conséquence de la norme évoquée ci-dessus. Connaître la palette des couleurs disponibles, les épaisseurs possibles pour les tracés de courbes, le mode de dessin (en pointillé, en tirets, etc.) n'est pas futile. Les réponses déterminent la facilité avec laquelle le programmeur pourra dessiner des liaisons chimiques de force variable ou bien des sphères pour matérialiser les atomes. Ce sont deux des nombreux exemples qui l'obligent à prévoir bien des lignes de code en fonction des stations de travail.

[1] laps de temps durant lequel l'utilisateur exécute le programme

D'autres *inquiries* ont pour but de connaître l'*état* de l'ordinateur à un moment donné. Dans la norme PHIGS, ce terme précise si l'ordinateur est en train de construire un objet, de calculer l'apparence qu'il aura ou bien s'il s'occupe de dessiner sur l'écran. En fait, les stations de travail sont le plus souvent dotées d'au moins deux processeurs qui se répartissent ces différentes tâches. C'est pourquoi, il faut considérer la simultanéïté d'une action en plusieurs lieux. A l'écran, l'utilisateur désigne - par exemple - un plan défini par trois atomes ; en mémoire, une procédure calcule une grandeur moléculaire dans ce plan ; enfin l'écran affiche les points correspondant à une certaine valeur de cette grandeur. De plus, l'asynchronisme des périphériques est souvent à l'origine de déphasages entre données en mémoire et données affichées. Ceci contribue encore à la diversité des situations que le programmeur doit traiter.

Pour dessiner une molécule, ou tout autre objet, des procédures permettent de créer des *structures*. Ce sont des sortes de boites qui contiennent des *éléments*, d'une part des *primitives graphiques*, d'autre part leurs *attributs*. Nos molécules seront autant de boîtes, constituées de lignes pour les liaisons interatomiques, de textes d'annotations à côté de chaque atome, de sphères déduites de leurs rayons de covalence et de Van der Waals. Les attributs préciseront les couleurs, les tailles, les types (tracé des lignes, polices de caractères, motifs de remplissage, etc.) de ces primitives graphiques. Les conventions habituelles des chimistes (halogènes en vert, liaisons -Hydrogène en pointillés par exemple) ont été généralement respectées.

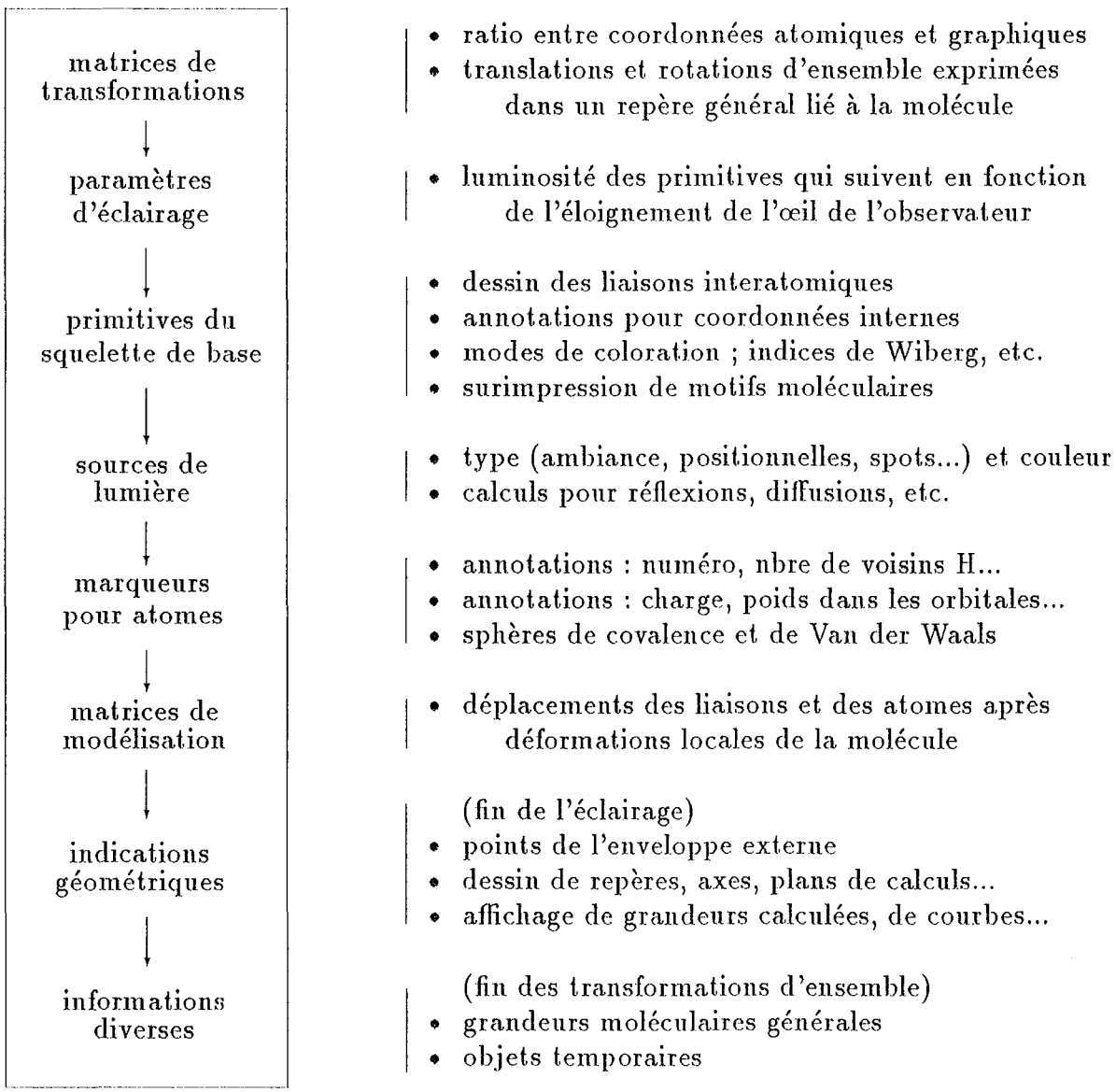
Il faut éviter de faire un dessin trop riche d'indications car l'affichage serait vite surchargé. Et pourtant, il doit inclure a priori le maximum d'informations susceptibles d'intéresser l'utilisateur. Pour résoudre ce dilemme, graPHIGS permet au programmeur d'insérer des éléments particuliers dans la structure. Ce sont des mentions d'appartenance ou de non appartenance de certaines primitives à des *classes de visibilité*. Ainsi les différentes annotations qui concernent un atome (son numéro, une de ses coordonnées ou sa charge par exemple) appartiendront chacune à une classe. Un système de filtres, géré séparément par le programmeur, sélectionnera ensuite les classes à rendre visibles ou invisibles. L'utilisateur pourra donc prendre connaissance de chacune des informations précédentes, individuellement et non toutes simultanément.

La figure 1 montre les types d'éléments qui composent la structure graPHIGS d'une molécule. Il peut y en avoir plus de 2000.

Dans ce schéma général, l'ordre des primitives graphiques va de celles qui sont le moins susceptibles de changer (le squelette de la molécule) jusqu'aux dessins créés temporairement. Un clic sur deux atomes consécutifs sera suivi du tracé, en fin de structure, d'une ligne entre eux avec indication de la distance interatomique.

Par ailleurs, cette organisation est affinée en fonction de la station de travail utilisée. Les sphères atomiques sont ainsi remplacées sur des IBM RT PC par des disques 2D dont il faut ensuite modifier l'orientation à chaque mouvement 3D de la molécule.

Figure 1: Structure graPHIGS d'une molécule



Notons que l'ordre d'insertion des éléments est important car lors de la mise à jour de l'affichage sur écran, la structure est *parcourue* séquentiellement. Les derniers objets du dessin apparaîtront donc au premier plan, même s'ils sont plus éloignés de l'œil de l'observateur que les premiers. C'est là un sérieux défaut car l'utilisateur pourrait se tromper d'atome en désignant de la souris celui qu'il croit au premier plan. Seules des stations de travail dotées de coprocesseurs 3D corrigent correctement ces inconvénients grâce à un *z-buffer*. L'image est modifiée dans cette zone mémoire locale pour tenir compte de la profondeur ou dessiner les objets avec des codes de couleur tels qu'ils donnent une impression de relief (technique de *depth cueing*). Pour les autres stations de travail, le programmeur doit ruser avec l'ordre des éléments...

Il existe finalement trois moyens de créer puis de modifier une molécule à l'écran :

o utiliser la technique des classes de visibilité, rapide à l'affichage et qui a l'avantage de ne pas modifier les données graphiques elles-mêmes. Nous l'avons surtout adoptée pour les modes de *design (balls and sticks, dreiding, CPK*, etc.) dans lesquels l'utilisateur visualise certaines annotations des atomes et d'autres concernant les liaisons. Y recourir s'impose également pour alléger l'affichage en cas d'opérations lourdes : pour faire tourner une molécule autour d'un axe, il vaut mieux cacher momentanément son enveloppe de Van der Waals.

o pratiquer de l'*édition* de structure. Certains éléments, appelés *labels*, repèrent des portions de dessin que le programmeur peut remplir de nouvelles données graphiques. L'opération est plus radicale, aussi efficace et un peu moins rapide à l'affichage. C'est en général de cette façon que seront insérées dans le dessin, les données acquises sur une molécule en cours d'étude. L'affichage d'indices de Wiberg entre liaisons lorsqu'on connaîtra la matrice densité, la superposition d'un motif commun après comparaison de structures moléculaires, le dessin de la surface de contact en sont autant d'exemples.

o employer des *éléments-matrices de modélisation* qui peuvent altérer localement les coordonnées de certaines primitives graphiques. Il est possible ainsi de dilater une liaison (par homothétie), d'éloigner (par translation) ou d'orienter (par rotation) un substituant. Plus généralement, il sera facile d'utiliser des systèmes locaux de coordonnées permettant ensuite de ne traiter que des parties d'une molécule.

3.4 Visualisation

Nous savons maintenant construire graphiquement des molécules et les modifier. Pour les visualiser, il faut associer à chaque structure une *vue* qui précise :

- le système de coordonnées utilisé par les primitives et la partie de l'écran dans laquelle l'utilisateur observera la molécule,
- le mode de visualisation (projection parallèle ou perspective à partir d'un point de fuite) et la position de l'œil de l'observateur,
- une matrice de transformation 3D permettant de modifier l'image dans son ensemble,
- *l'activité* qui permet d'ignorer momentanément des molécules en les rendant invisibles,
- la transparence ou l'opacité, rendant possibles les superpositions et donc les comparaisons entre molécules,

- le mode d'affichage des faces cachées et d'éclairage des surfaces par les sources de lumière (qui demandent obligatoirement un coprocesseur graphique 3D sur la station de travail).

Ces caractéristiques peuvent être choisies une fois pour toutes par le programmeur. Mais l'expérience montrant que l'utilisateur a toujours besoin de changer l'affichage en fonction des circonstances (observation, comparaison, réglages pour copies d'écran sur une imprimante spécialisée, etc.), nous avons inclus dans le logiciel OSCAR des procédures qui traitent ce problème. Fonctionnant indépendamment du contenu de chaque vue (des molécules ou n'importe quel autre objet en trois dimensions), elles constituent d'ailleurs un utilitaire d'emploi très général (présenté simultanément avec OSCAR en annexe B).

A l'usage, il est apparu nécessaire de prévoir pour chaque molécule, non pas une mais deux vues. Une structure, dont les primitives sont purement 2D, est associée à la première. Elle inclut simplement un titre, une zone de commentaires et une case de clic qui donne accès à divers utilitaires. Une seconde structure, contenant la molécule proprement dite, est associée à la seconde vue. Ce sont ces dernières, et elles seules, qui feront l'objet des transformations 3D.

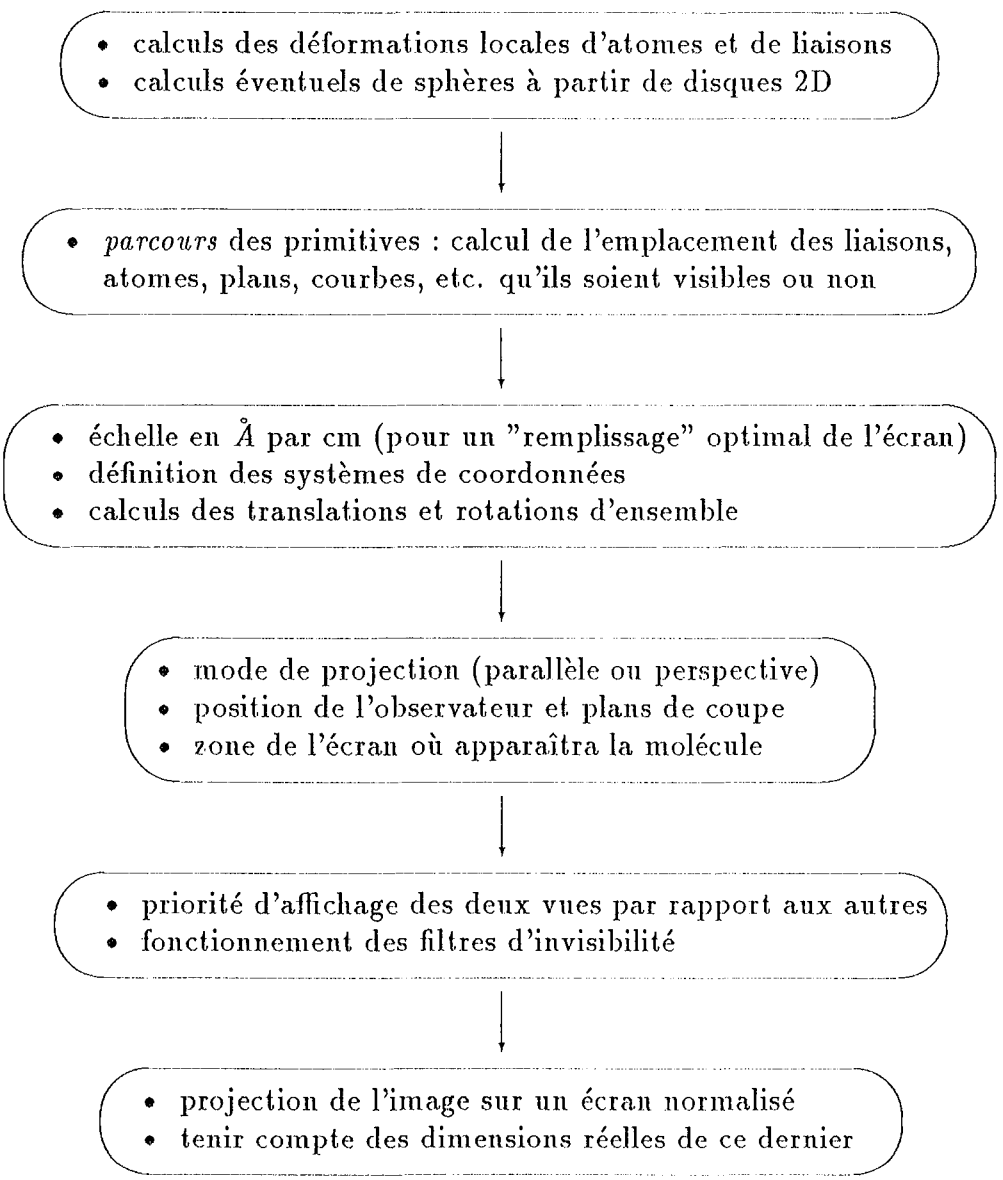
3.4.1 Les tables traçantes

Celles-ci sont également considérées comme des stations de travail qui reçoivent a priori les mêmes images qu'un écran. Leurs caractéristiques physiques et leur mode de fonctionnement étant naturellement très différents, les vues devront être le plus souvent modifiées. Il faudra également s'assurer de la compatibilité des primitives graphiques et de leurs attributs (absence de 3D évidemment, mode de coloration, finesse des dessins, etc.). Quant à la mise à jour de l'affichage, elle consiste, non à dessiner sur une feuille de papier mais à créer sur disque un *Graphics Data File* que l'utilisateur imprimera séparément avec une commande du système d'exploitation.

3.4.2 L'affichage

La *mise à jour* de l'affichage est une opération complexe mais identique dans son principe pour toutes les stations de travail. Le programmeur doit la demander explicitement car les fonctions de graPHIGS n'agissent que sur des données en mémoire. Elle consiste en une succession de calculs matriciels, d'autant plus longs que le dessin de la molécule est riche d'informations. La figure 2 en détaille le déroulement pour chaque couple de vues associées à une molécule.

Figure 2: Mécanisme de l'affichage sur une station de travail



3.5 Dialogue utilisateur - ordinateur

Examinons brièvement les conditions dans lesquelles l'utilisateur communique avec une application graphique. La question a deux aspects, matériel et logiciel. Dans le premier cas, il s'agit simplement de rappeler les périphériques dont dispose l'utilisateur : un clavier, une souris, éventuellement un pavé de touches lumineuses et des rotateurs, commodes pour faire tourner des objets en trois dimensions.

Du point de vue logiciel, les actions de l'utilisateur sont formalisées dans le cadre de la norme PHIGS : localiser (dans un plan 2D), relever, orienter, choisir, désigner, répondre (au clavier) ou encore se croiser les bras. Le plus souvent, une application interactive fonctionne comme une boucle d'*événements*. Elle attend, pendant un laps de temps déterminé, un geste de l'utilisateur et l'interprète - lorsqu'il survient - comme une combinaison simultanée des classes précédentes. Par exemple, un clic de l'utilisateur sur un atome est interprété comme un choix (quel bouton de la souris ?), une désignation (quel atome ?) et une localisation (où est-il ?). Tous ces événements doivent être prévus et traités par le programmeur. C'est un peu lourd lorsque les événements sont multiples, comme ci-dessus, ou simultanés comme dans le cas de manipulations consécutives des rotateurs. Enfin, il ne reste qu'à aiguiller l'action vers la procédure qui y répondra.

De cette façon, l'application garde une structure rationnelle mais pas trop rigide. Son exécution ne se déroule pas linéairement, mais plutôt sous la forme de petites tâches lancées depuis la boucle et qui y reviennent. De plus, elle se prête facilement à des adaptations ultérieures, soit pour y introduire de nouvelles procédures de traitement des molécules, soit pour introduire le parallélisme dans l'exécution du programme. A priori, rien n'empêchera de lancer simultanément deux procédures bien distinctes.

3.5.1 Convivialité...

Le deuxième aspect de l'interactivité concerne les messages transmis par le programme à l'utilisateur ou inversement les données que doit fournir ce dernier. Malheureusement, autant graPHIGS est riche de procédures pour manipuler des objets en trois dimensions, autant il en a peu pour afficher simplement du texte ou plus généralement organiser des dialogues entre machine et utilisateur. La comparaison entre des environnements bien différents est instructive. Concevoir une telle interface se fait en quelques jours sur un simple Macintosh (qui, en revanche, offre beaucoup moins de possibilités en 3D), un peu moins vite sous XWindows, en quelques semaines avec graPHIGS...

Dans notre cas, il a fallu consacrer un temps important à la programmation d'alertes, de boites de dialogues et de menus déroulants, travail un peu en dehors de notre sujet mais qui n'en est pas moins indispensable. Les alertes sont utilisées dès qu'il faut

avertir l'utilisateur quant au bon déroulement des tâches qu'il a lancées ; les dialogues lui permettent d'introduire des noms de fichiers, des valeurs numériques ou de cocher des options de calcul ; les menus permettent de "se diriger" parmi les principales fonctions d'une application. De plus amples précisions n'ont pas leur place ici, mais les programmeurs pourront se reporter au mode d'emploi et aux nombreux commentaires qui ornent ma prose FORTRAN. Dans la pratique, l'utilisateur se familiarisera très vite avec tous ces outils.

Chacun se fera une meilleure idée de toutes ces notions de graphisme et d'interactivité en se reportant aux photographies insérées dans ce rapport : elles sont le résultat de nombreux mois de programmation. A l'intention des programmeurs désireux d'en faire le lien avec nos procédures, nous indiquons les plus importantes de ces dernières, qui leur sont accessibles :

Tableau 2: Procédures graphiques de base

gbahsc	taille des textes d'annotations dans les dessins
gbawev	attente et tri des événements par classe. Premier traitement
gbawev2	*idem* + réponses aux seuls événements rotateurs
gbawst	boucle d'événements provenant d'entrées *strings*
gbawvl	boucle d'événements provenant de rotateurs
gbclws	fermetures de stations de travail (écran, plotter)
gbdmgf	affichage d'un message, choix d'une fonction
gbdplt	transformations propres aux tables traçantes (pour utiliser des formats autres que A4 pour les dessins)
gbegst	saisies de chaînes de caractères au clavier
gbegvl	saisies de valeurs numériques à l'aide de rotateurs
gbexsc	exécution directe d'une commande du système d'exploitation
gbfct	définition d'une table de couleurs (RGB, HSV...)
gbfdct	fonctionnement du *depth cueing* : luminosité = fct (éloignement)
gbflst	définition de sources de lumière (ambiante, directionnelle, spots, etc.)
gbgcp	calculs de points de contrôle pour affichage de *B spline* surfaces. Sert pour les enveloppes moléculaires
gbgcp2	variante avec calcul préalable des axes principaux d'inertie
gbgiev	gestion (propre) des interruptions en cours de calcul
gbgkv	calcul de vecteurs *knot* pour interpolations par fonction *B spline* en une ou deux dimensions, courbes de Bézier, etc.
gbinch	initialisation des pavés de fonctions, souris, claviers, etc.
gbinlc	préparation de la souris pour localiser tout objet
gbinpl	initialisation d'une table traçante (variables communes)
gbinsk	préparation de la souris pour relever des points, images, etc.
gbinst	initialisation d'une chaîne de caractères (prompt, buffer, saisie)
gbinv	définition améliorée d'une vue. Utilisation du *z-buffer*

Tableau 3: Procédures graphiques de base (suite)

gbinws	initialisation de la station principale (variables communes)
gbivlr	initialisation d'un rotateur pour fonctionnement en rotation
gbivlt	*idem* en translation
gbivlz	*idem* en zoom (homothéties)
gbmqse	édition dans une structure : récupération des données stockées
gbpcf	filtres de désignation, de visibilité ou de *highlighting*
gbper	gestion (sans plantage...) des erreurs graPHIGS
gbplv	impression de vues graphiques sur table traçante
gbpm3	affichage de points avec options diverses (paquets, double précision...)
gbpper	gestion personnelle des erreurs graPHIGS
gbqspc	conversion des couleurs en fonction des stations de travail
gbqtw	taille d'un texte en fonction des caractéristiques d'une vue
gbqvo	connaître le plan auquel est affiché une vue
gbrotv	rotation interactive d'une vue. Effet d'animation
gbrqst	saisie simplifiée d'une chaîne de caractères
gbrse	recherche d'un élément (primitive ou attribut) dans une structure
gbrv	recherche de vues possédant certaines caractéristiques
gbscv	homothétie interactive d'une vue. Effet d'animation
gbsvo	choix de l'ordre d'affichage (par plan) des vues
gbsb	dessin d'un bouton de désignation
gbsl	dessin d'un cadre avec légende
gbtv	translation d'une vue
gbtx2	affichage de textes pour commentaires
gbsd	gestion des classes d'invisibilités. Traitement plus complet
gbsows	sélection et ouverture d'une station de travail
gbupws	mise à jour des stations de travail (modes multiples)
gbvmp2	système de coordonnées pour vues. Types de projection, angle d'observation, *z-buffer* et transparence partielle
gbzvcs	zoom de tout l'affichage de l'écran. Utile pour photos.
gp..	routines fictives pour *link* avec versions anciennes de graPHIGS

Tableau 4: Procédures graphiques de l'interface utilisateur

giadu	alerte définie par le programmeur (message momentané)
giaefn	choix d'un fichier sur disque (avec vérifications)
giafd	alerte avant destruction d'un fichier
giaqf	dialogue pour informations sur des fichiers
giaqt	alerte avant de quitter un module (ou le programme)
giavd	alerte avant de détruire des éléments graphiques
giawev	attente d'un événement ; *routine* principale du logiciel OSCAR qui indirectement traite 40 % des requêtes
gibd	construction d'un dialogue façon Macintosh
gidcni	dialogue pour renumérotation d'atomes
gidd	destruction d'un dialogue
giddlc	dialogue pour choix des paramètres d'éclairage
gidm	destruction d'un menu déroulant
gidplv	dialogue pour impression d'un graphe, de courbes, etc.
gipd	affichage et utilisation d'un dialogue OSCAR - utilisateur
gipm	gestion d'un menu (déroulements, etc.)
gipra	utilisation de la partie droite de l'écran ; routine qui indirectement s'occupe de toute la visualisation
giqd...	caractéristiques des contrôles (case cochée, etc.)
giqm	caractéristiques d'un menu déroulant
gisd	insertion dans un dialogue de tous ses "contrôles" (chaînes, barres de défilement, cases à cocher, etc.)
gism...	construction de menus déroulants
gisra	dessin de la partie droite de l'écran (boutons de désignation, légendes, etc.)
. . .	

Tableau 5: Procédures graphiques pour graphes de fonctions

gfbfs...	construction (options standard) d'un graphe y = fct (x) *routine* de base pour observation de toute variable numérique
gfddfs...	*routines* de construction des menus *pop-ups* propres à ce module. Indépendant du logiciel OSCAR
gfdofs	gestion de l'affichage (invisibilité, etc.) propre au module
gfpfs	traitement automatique des événements concernant cet utilitaire : design, interpolations, statistiques, etc. OSCAR lui-même est donc déchargé de ces opérations...
gfdfs	destruction d'un graphe ; options de sauvegarde des données
gficfs	insertions de commentaires pour les calculs menés dans le module
gfiifs...	*routines* de gestion des graphes
. . .	

4 Activité et Formes moléculaires

Nous engageons ici un long chapitre qui détaille les rouages du logiciel OSCAR. Nous commencerons par regarder comment trouver des sites caractéristiques puis les comparer d'une molécule à l'autre. Cette question sera toujours examinée dans deux cas de figure. Le premier est celui d'un chimiste qui utilise le logiciel en mode *interactif* et qui donc peut l'aider en se montrant précis dans ses requêtes. Dans le second cas, le programme doit mener toutes les opérations géométriques (recherches, comparaisons, superpositions, problèmes de conformères) en *arrière-plan* (en *batch*) c'est-à-dire automatiquement et sans intervention de l'utilisateur. Ce travail est indispensable dans l'optique future d'un logiciel connecté à des bases de données, donc susceptible de traiter une masse d'informations autrement plus considérable. Nous verrons que là résident les problèmes de programmation les plus délicats.

Nous détaillerons au fur et à mesure les paramètres géométriques les plus significatifs de la topologie des molécules : surface de contact, distances entre substituants, encombrement stérique, volume occupé dans la "poche" d'un site récepteur, etc. Nous ajouterons ensuite à ces considérations topologiques la prise en compte des grandeurs électroniques, notamment du potentiel.

4.1 Recherche de sites caractéristiques

Il est important de bien choisir les types de données à manipuler dans la mémoire centrale d'un ordinateur. Elles doivent faciliter l'écriture de procédures rapides, fiables et susceptibles de modifications fréquentes pendant la mise au point. Pour notre travail, nous avons à identifier des atomes, des liaisons chimiques, des cycles et finalement des *motifs*, agencements particuliers et remarquables d'atomes et de liaisons.

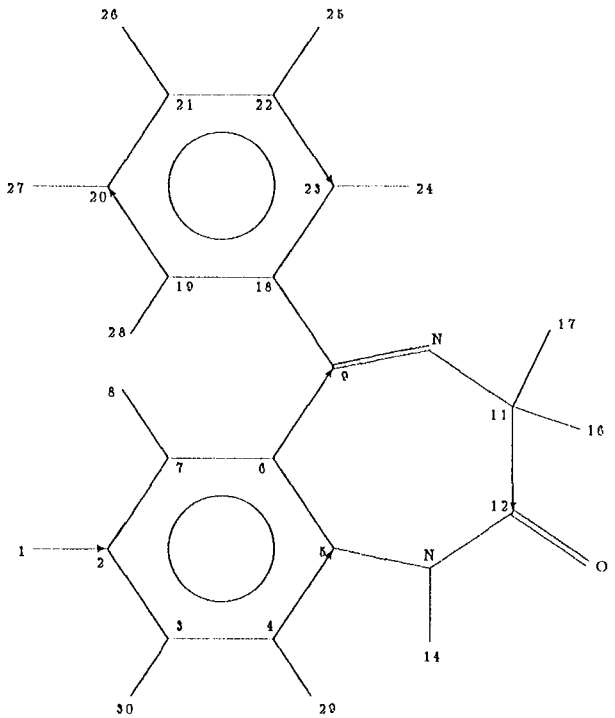
Les programmeurs font souvent usage de *graphes* - dont les atomes sont les nœuds et les liaisons chimiques les connections - pour décrire une molécule. Mais cette technique utilise en général l'allocation dynamique de mémoire au cours de l'exécution d'un programme et des variables spéciales appelées pointeurs. Elle n'est pas acceptée par tous les compilateurs FORTRAN et entraine souvent les programmeurs à définir des données redondantes ou écrire un source [1] compliqué, peu lisible ou difficilement réutilisable. J'ai donc opté pour des définitions plus simples et des algorithmes récursifs, fondés simplement sur l'idée que "les voisins d'un atome sont les voisins de ses voisins immédiats" [2].

[1] Le lecteur ne s'étonnera pas de trouver ce mot au masculin dans un contexte informatique

[2] sont voisins deux atomes dont l'éloignement est inférieur à la somme de leurs rayons de covalence plus une tolérance de 0.40 angström

Cette méthode est à la base du *parcours* d'une molécule, dont une première application est le jeu de coordonnées internes (longueurs de liaisons, angles et angles dièdres) établi par le logiciel OSCAR dans une benzodiazépine (figure 3). L'ordre dans la numérotation des atomes y est caractéristique d'un algorithme récursif.

Figure 3: Structure de base des 1,4 benzodiazépines



La technique peut être ensuite affinée en introduisant des critères qui précisent pour chaque atome, une hiérarchie de ses voisins. Par exemple, le parcours doit-il privilégier, ou non, un atome terminal par rapport au membre d'un cycle ? Il s'agira ainsi de définir *le chemin descriptif* de la molécule le plus approprié pour mener une tâche donnée.

Ainsi souhaitons-nous :

- décrire surtout le contour de la molécule ou rechercher au contraire le ou les cycles qui constituent sa structure de base,
- identifier les substituants, s'enquérir de leurs dimensions et des distances qui les séparent,
- repérer ceux qui sont en porte-à-faux ou dans toute autre position facilement accessible à un réactif,
- inversement découvrir les zones concaves de la molécule, là où l'encombrement stérique d'un réactif sera déterminant.

Un jeu optimal de coordonnées internes étant adopté dans chacun de ces cas, il s'agit ensuite d'examiner l'enchaînement des éléments, des longueurs de liaisons et surtout de comparer les angles de valence et angles dièdres consécutifs. L'algorithmique saura ainsi identifier les cycles en lisant une succession d'angles de sens et de valeurs similaires, même s'ils sont tendus (valeurs plus faibles que celles prévues par les degrés de coordination). Les substituants ramifiés seront mis en évidence par de fréquents changements de direction dans le parcours. Inversement, la faible variation d'une coordonnée cartésienne indiquera un substituant plus linéaire. La reconnaissance des sites très accessibles, ou au contraire des zones plus fermées de la molécule, sera moins facile mais conduite selon les mêmes raisonnements.

4.1.1 Coordonnées internes et conformères

En écrivant toutes ces procédures dans le logiciel OSCAR, nous avons vite été confronté au problème des conformères. Il n'est pas possible de considérer uniquement chaque molécule dans une géométrie idéale et optimisée. Il suffit parfois d'une seule rotation autour d'une liaison clé pour changer une grandeur localement ou modifier l'accessibilité d'un site. Le chemin descriptif doit donc être défini plus rigoureusement en précisant des bornes de variation pour certaines coordonnées internes.

De ce point de vue, le cas de la benzodiazépine dessinée figure 3 est assez instructif. Comme le deuxième cycle aromatique est symétrique et semble bien séparé du reste de la molécule, il serait raisonnable de penser que sa rotation doit se faire facilement. Autrement dit, la valeur de l'angle dièdre 6-9-18-19 serait sans importance dans notre algorithme de parcours. Cependant, des calculs de vérification ont mis en évidence une interaction appréciable entre les orbitales de valence des atomes 18, 19, 24 et 28 d'une part, 7, 6, 9 et 10 d'autre part. Ce surcroît de stabilité, dû à la délocalisation électronique du cycle benzodiazépinique, fait apparaître un saut énergétique suffisamment important pour figer la population des molécules dans les deux conformères d'angle dièdre 6-9-18-19 de + / - 90 degrés.

La substitution d'un chlore en position 28 augmente encore ce saut mais surtout, fait disparaître la symétrie du cycle benzénique. Contrairement au cas précédent, le logiciel devra donc distinguer les deux conformères, c'est-à-dire supposer que l'enchaînement des coordonnées internes 2-7-6-9-18-19/28 concerne deux motifs distincts, comme s'ils étaient dans deux parties de la molécule. Ceci est d'autant plus indispensable que l'activité pharmacologique n'est peut-être due qu'à une seule des deux espèces, ou pire encore que l'une la favorise tandis que l'autre s'y oppose (garder à l'esprit le fait que les pharmacologues ne fournissent qu'une seule valeur expérimentale.). Par exemple, elle pourrait être le fait du rôle conjugué de l'atome de chlore et de la zone comprise entre les atomes d'azote qui - la figure ne le montre guère - est déformée hors du plan de la feuille.

Tableau 6: Equilibre conformationnel dans une rotation autour d'une liaison simple

d'après la relation de Boltzmann à 25°C

Δ G (kcal / mol)	Pourcentage du conformère le moins favorisé
0	50 %
1	~ 16 %
2	~ 5 %
3	~ 0.5 %
4	~ 0.1 %

Dans la pratique, les "barrières" de rotation pourront être calculées, soit par les méthodes semi-empiriques (paragraphe 2.1), soit par la mécanique moléculaire. Cette dernière méthode (dont le principe est sommairement rappelé en annexe C) présente l'avantage de fournir plus rapidement une estimation acceptable des différences d'énergie entre conformères.

4.1.2 Autres méthodes

Il est utile de rappeler, pour comparaison, les algorithmes évoqués dans les publications pour représenter une molécule en mémoire. Il est fait usage le plus souvent de tables de connectivité ou d'autres modes de *codage* [29,30,31] qui sont moins évocateurs de la géométrie dans l'espace. D'ailleurs, ils contiennent rarement les valeurs numériques des coordonnées internes. En revanche, ils se prêtent mieux par nature à des manipulations fondées sur *la théorie des graphes* [32,33,34].

Ces graphes moléculaires offrent les avantages mais aussi les inconvénients de la simplicité. Il arrive souvent qu'un graphe corresponde à plusieurs molécules ou que les informations structurales soient si bien résumées qu'il devient difficile de comparer deux espèces. C'est pourquoi des auteurs s'efforcent d'affiner ces modes de description, le plus souvent dans une optique précise [30,35]. Une des applications les plus connues, due à Morgan [36], attribue à chaque molécule une représentation unique appelée forme canonique. Elle est à l'origine du *Chemical Abstracts System Chemical Registry Service* [37].

A ce stade, nous disposons donc de procédures qui peuvent, en mémoire centrale, traiter des molécules et y déceler toute particularité géométrique remarquable. Ce problème de *reconnaissance* étant résolu, intéressons-nous maintenant à celui de la *comparaison* des motifs entre molécules.

4.2 Comparaison de motifs moléculaires

Il est assez facile pour le chimiste d'apprécier visuellement les différences entre deux formes dans l'espace. S'il lui faut rechercher un motif ressemblant à un cycle aromatique, il lui viendra à l'esprit de s'intéresser - par exemple - à un cycle pyridinique. En revanche, écrire un programme qui mène la même démarche d'intuition (qu'est ce qui pourrait ressembler ?) et de comparaison (est ce que ça y ressemble effectivement ?) est un problème beaucoup plus ardu.

Le chimiste compare fréquemment des composés dont la structure de base est identique. Dans ce cas simple, il suffit de noter la liste des substituants possibles et d'indiquer au logiciel OSCAR les atomes de base ou atomes *correspondants*. Nous appelons ainsi ceux qui, d'une molécule à l'autre, occupent les mêmes positions dans l'espace, indépendamment de leur nature ou de leur numérotation. Si nous imposons aux molécules le même et unique système de coordonnées fondé sur trois atomes correspondants, elles adoptent alors la même orientation dans l'espace, rendant immédiate la comparaison de leurs sites. Cette démarche est détaillée dans le paragraphe 4.3.

Cependant, il est apparu intéressant de s'affranchir de cette restriction qui fait qu'en QSAR, les études portent le plus souvent sur des familles de composés ressemblants. Ne devrait-on pas comparer directement des sites locaux, arrangements originaux et spécifiques de quelques atomes présents dans des molécules par ailleurs très différentes ? Un pharmacologue, étudiant quelques molécules, serait certainement intéressé par une démarche qui, en arrière plan et sans la subjectivité du chimiste, rechercherait dans des bases de données des composés incluant le même motif caractéristique.

Cette question est l'objet des travaux de *screening* de molécules [38,39,40]. Le chimiste en choisit une, de référence, dont il connaît la table de connectivité, puis une série d'espèces à lui comparer. L'idée est de parcourir leur graphe en notant 1 pour chaque nœud ou connection similaire à celle du motif de référence ou 0 dans le cas contraire. L'avantage est d'obtenir des *bitmaps*, succession de données logiques qui sont très rapides à traiter dans des comparaisons [41,42,43]. Mais comme les graphes sont eux-mêmes des résumés succincts de l'information structurale des molécules, la méthode n'est pas toujours d'une très grande fiabilité. Les problèmes les plus fréquents portent sur la confusion des liaisons (doubles, aromatiques), les particularités structurales (doublets libres) et surtout les conformères ou stéréoisomères.

En fait, il semblerait que le meilleur moyen de s'affranchir des limitations propres aux graphes moléculaires consiste à inclure des règles d'apprentissage (ou d'*heuristique*) qui utilisent une base de connaissances la plus étendue possible. Ce sont là les caractéristiques des *systèmes experts* dont SECS + ALCHEM (synthèse de molécules organiques complexes) [44], DENDRAL (recherche de structures à partir de formules brutes) ou GENSAL [45] sont quelques applications bien connues des chimistes. Très sophistiquées, elles ne peuvent guère être réutilisées ou même servir de sources d'inspiration dans un travail plus modeste comme le nôtre. Protégées, écrites dans des langages

dédiés à l'intelligence artificielle, elles nécessitent aussi des investissements autrement plus considérables en temps comme en personnes.

4.2.1 Programmation

Ces considérations expliquent les essais menés séparément avec OSCAR. Pour comparer des molécules de géométrie quelconque, nous avons repris la technique récursive exposée dans le paragraphe 4.1 mais en complétant les procédures par de multiples tests.

Les analogies portent d'abord sur les éléments, en particulier les hétéroatomes. A priori, ils seront différenciés en fonction de leur position dans la classification périodique. Les critères essentiels sont :

- l'effet inductif et l'effet mésomère,
- le rayon de covalence, donc la taille,
- le degré de coordination et la présence d'électrons non appariés.

Tableau 7: Electronégativités des principaux éléments

d'après Pauling

H	B	C	N	O	F
2.1	2.0	2.5	3.0	3.5	4.0
		Si	P	S	Cl
		1.8	2.1	2.5	3.0
			As	Se	Br
			2.0	2.4	2.8

Même s'ils sont simples à utiliser, ces critères sont souvent contradictoires, à commencer par les deux premiers. Ainsi les substituants =O et -Cl - qu'on essaie parfois d'interchanger dans les synthèses qui conduiront à de nouvelles molécules - ont pratiquement la même électronégativité. Mais il est bien évident que cette donnée n'est pas utilisable dans un même test, selon que les éléments sont terminaux ou en alpha de liaisons insaturées.

Les analogies entre motifs portent aussi sur les distances interatomiques. Mais pour comparer deux liaisons - dans deux molécules - il ne suffit pas de calculer une différence de longueur [46]. Les valeurs tabulées dans la littérature démontrent qu'une même valeur peut se rapporter à deux liaisons C = C bien différentes d'un point de vue électronique et qu'inversement des écarts importants peuvent être tolérés dans des liaisons simples de cycles [47,48].

Les angles de valence et les angles dièdres présentent des difficultés similaires. Trop de cas particuliers sont à envisager pour espérer mettre au point un algorithme d'usage général. L'exemple des cycles aromatique et pyridinique est assez révélateur. Le nombre d'atomes de chaque cycle est différent, un hétéroatome intervient dans le second et les angles de valence y sont modifiés par une certaine tension du cycle. Du reste, la similitude des propriétés chimiques n'est-elle pas explicable surtout par une disposition géométrique d'ensemble qui confère à un modeste petit doublet le même rôle qu'une double liaison ?

La solution retenue en définitive est donc un compromis dans lequel OSCAR propose des voies de recherche pour trouver des structures moléculaires comparables. Un premier tri, sommaire, permet de retenir des molécules qui ont le même nombre de cycles et le même nombre de substituants que celles qu'étudie le chimiste. La suite de la méthode est surtout interactive, l'utilisateur imposant successivement des critères de similitude parmi ceux énoncés plus haut. Il lui est même possible de préciser un degré de *sévérité* (cet atome doit il être exactement identique ou accepte-t-on qu'un soufre remplace cet oxygène ?) pour rendre la démarche moins rigide. Il y a donc multiplication des combinaisons possibles, à tel point que le logiciel peut retrouver finalement un même motif dans des composés très différents mais au prix d'une certaine perte de temps. [1]

4.2.2 Les indices de connectivité

Avant de détailler les problèmes d'encombrement stérique dans un site, il me semble utile de lever une possible ambiguïté. Le traitement matriciel utilisé dans le code source de OSCAR pour manipuler et comparer des jeux de coordonnées internes, n'a rien de commun avec les matrices de connectivité, objets de nombreuses publications sur les relations structure-activité. Si dans les deux cas, l'idée est de définir de bons *descripteurs* structuraux de chaque molécule, leur finalité est complètement différente.

Dans le logiciel OSCAR, ces descripteurs sont seulement utilisés comme intermédiaires de calculs en mémoire. En revanche, les matrices introduites par Randic [49] et Rouvray [50] et dont les éléments sont appelés *indices topologiques*, sont destinées à intervenir dans des relations statistiques "Structure - Activité" en tant que telles. Développés par Kier et Hall [51], ces indices doivent bien résumer l'information structurale d'une molécule - tâche d'autant plus délicate qu'on en attend un usage plus général - avant d'être corrélés à des grandeurs macroscopiques. De nombreuses publications [51,52,53] rendent compte de ce genre d'études, parfois concluantes mais souvent limitées à des exemples particuliers.

[1] dont il faut bien reconnaître qu'elle peut être rédhibitoire si l'utilisateur ne limite pas le domaine de recherche en précisant des conditions d'analogies assez restrictives. En bref, nous avons bien trouvé une solution mais pas la panacée.

Nous n'avons pas repris cette démarche de relations statistiques. Elle reste empirique et néglige le plus souvent les effets géométriques et électroniques locaux. En revanche, nous nous sommes inspiré des indices topologiques mais avec l'idée de les définir de façon plus parlante pour le chimiste et non avec un formalisme théorique qui ne leur donne pas vraiment de signification physique. La présentation systématique de ces descripteurs structuraux est l'objet du paragraphe 4.4.

4.3 Superposition de molécules

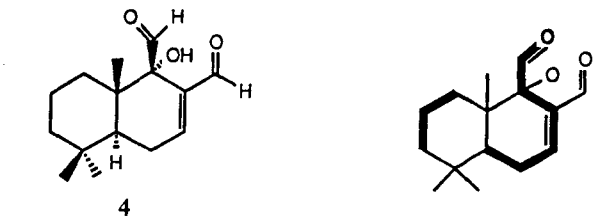
Après recherche et mise en évidence d'un site caractéristique commun à plusieurs molécules, il est temps d'examiner l'interaction possible avec un site récepteur. Rappelons avant toute chose que nous travaillons dans le cadre des hypothèses très restrictives du paragraphe 1.2 : ne sachant pratiquement rien sur la localisation et les caractéristiques géométriques ou électroniques du ou des récepteurs, la seule démarche possible est de procéder par *analogies* entre des molécules dont on connaît par ailleurs la valeur de l'activité pharmacologique.

Nous devons préalablement amener nos molécules en coïncidence, c'est-à-dire les *superposer* de manière à observer ce qui les distingue les unes des autres d'un point de vue géométrique. Comme précédemment, il faut distinguer d'une part le cas où le chimiste travaille en mode interactif, d'autre part le cas où cette tâche doit être menée en arrière-plan. Dans le premier, le chimiste étudie des composés, analogues ou non, qui présentent le même site caractéristique. Il lui suffit d'établir la liste des atomes *correspondants*, c'est-à-dire de noter les numéros de ceux qui occupent, d'une molécule à l'autre, les mêmes positions dans l'espace. Tous ne peuvent pas satisfaire cette condition à l'image des deux oxygènes d'un groupement nitro qui n'ont pas d'équivalent dans une molécule où ce substituant est remplacé par une fonction cyano. La liste ne peut donc inclure qu'une partie des atomes, ce qui s'avère peu gênant dans la pratique. Ceci fait, il est possible d'imposer n'importe quel système de coordonnées fondé sur trois de ces atomes, la liste des correspondances permettant de l'appliquer, de la même manière, aux autres molécules. Elles adoptent alors la même orientation dans l'espace, leur superposition définitive ne demandant plus qu'une translation entre leurs origines. Cette solution a le mérite d'être très souple, l'utilisateur d'OSCAR étant libre de choisir la disposition dans l'espace la plus appropriée à toute comparaison.

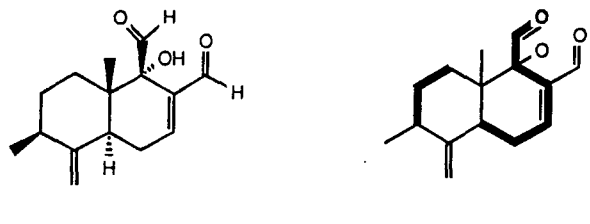
S'il devient nécessaire de superposer des molécules dans des procédures fonctionnant en arrière-plan, la question devient plus délicate. L'utilisateur n'est plus là pour définir une orientation commune ou ajuster finement les positions de chacune. Au problème de la reconnaissance d'un même motif s'ajoute donc celui de la mise en coïncidence des atomes et des liaisons, avec tous les degrés de liberté que cela suppose. Une première solution consisterait à calculer la forme 3D de chaque molécule, c'est-à-dire l'ensemble des sphères de Van der Waals ou de covalence, puis à estimer numériquement des différences de volumes ou de surfaces locales. En les minimisant, la méthode fourni-

rait une orientation commune à nos molécules. Mais il apparaît immédiatement que ces calculs risquent d'être bien longs et d'aboutir en définitive à des systèmes de coordonnées arbitraires, a priori indépendant de l'orientation des liaisons ou des éléments de symétrie, et pas forcément bien adaptés localement.

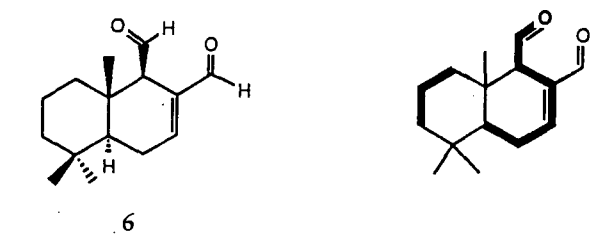
Figure 4: Recherche puis superposition d'un motif commun de taille déterminée



4



5



6

Une autre solution est d'étendre la technique proposée par Varkony [54,55] pour des molécules planes. Elle inclut le calcul puis une classification de toutes les coordonnées internes dans chaque molécule. Une nouvelle sous-structure est construite en ajoutant successivement des paires d'atomes dont les coordonnées internes sont retrouvées dans chaque classe. L'opération est répétée pour toutes les combinaisons possibles, le but étant d'obtenir un *motif-modèle* constitué du plus grand nombre d'atomes, de structure commune à toutes les molécules et doté d'un système de coordonnées adéquat. La dernière étape consiste à superposer chaque molécule sur ce modèle, plutôt que de les superposer entre elles. Les atomes du modèle servent alors de *points d'ancrage*, leurs coordonnées cartésiennes restant fixes. Dans chaque molécule sont identifiés les atomes correspondants grâce à la classification précédente, avant de leur imposer les coordonnées des points d'ancrage. Une matrice de passage entre les deux systèmes de coordonnées est finalement calculée pour déplacer correctement les autres atomes. L'expérience a

montré que cette méthode est d'autant plus fiable qu'il y a peu de liaisons de longueur très similaire mais que la durée des calculs (notamment pour la classification) croît très vite avec le nombre d'atomes.

Notons que les publications consacrées à ce problème de superposition proposent des techniques différentes car fondées sur une représentation des molécules en mémoire centrale à l'aide de graphes. Danzinger et Dean [56] cherchent à minimiser la somme des distances entre atomes correspondants par un traitement matriciel fondé sur les indices topologiques évoqués plus haut. Sheridan [57] a amélioré la méthode pour tenir compte des conformations possibles. Mais dans chaque cas leurs applications ont porté sur des familles de composés analogues dont l'orientation commune est obtenue globalement et non pas en fonction d'un site local. Nous retiendrons donc la difficulté de mise au point d'algorithmes simultanément fiables et généraux.

4.4 Décrire la topologie d'un site actif

Dès lors que nos molécules sont orientées de la même façon dans l'espace, par exemple face à un site récepteur, nous pouvons comparer leur forme externe. L'idée est de regarder si les molécules les plus actives, ou au contraire les plus inactives, se singularisent par une topologie localement différente. Il est possible que pour une série donnée de molécules, la propriété biologique à étudier soit due à une forme plus ou moins concave du site actif, à la taille d'un certain substituant, ou bien encore à une déformation par rapport à un axe. En bref, nous voulons savoir si un petit changement dans la géométrie du site "optimise" sa disposition face à un site récepteur, à l'image d'une clé dont la forme serait plus adaptée à une serrure.

Quel paramètre utiliser pour décrire un tel changement ? Il vient à l'esprit la distance entre des atomes ou des substituants du site, l'éloignement de la surface de contact de la molécule mesuré dans certaines directions de l'espace, ou bien encore la valeur d'un angle ou d'un angle dièdre significatif. Mais quel qu'il soit, le critère géométrique idéal sera celui qui s'avère *discriminant* vis-à-vis de l'activité, c'est-à-dire tel que sa connaissance pour une nouvelle molécule nous permette de prédire son activité avec un risque d'erreur minimum.

Dans le paragraphe 5.1, nous reviendrons largement sur cette condition, en particulier sur les tests qui montreront dans quels cas les descripteurs que nous allons présenter maintenant se révèleront satisfaisants.

4.4.1 Eloignement des contours

Un bon critère pour évaluer l'encombrement stérique est la mesure de la distance entre la surface du site actif et un point A_1 situé au cœur de la molécule. La surface est constituée ici à partir des sphères de Van der Waals centrées sur chaque atome. Même si leur définition est empirique, les rayons de Van der Waals donne une assez bonne idée de l'encombrement de chaque élément (tableau 8). En revanche, le strict "assemblage" des sphères met en évidence des zones anguleuses, à l'intersection de deux ou trois d'entre elles. Nous reviendrons sur ce défaut un peu plus loin.

Tableau 8: Rayons de Van der Waals des principaux éléments (en angströms)

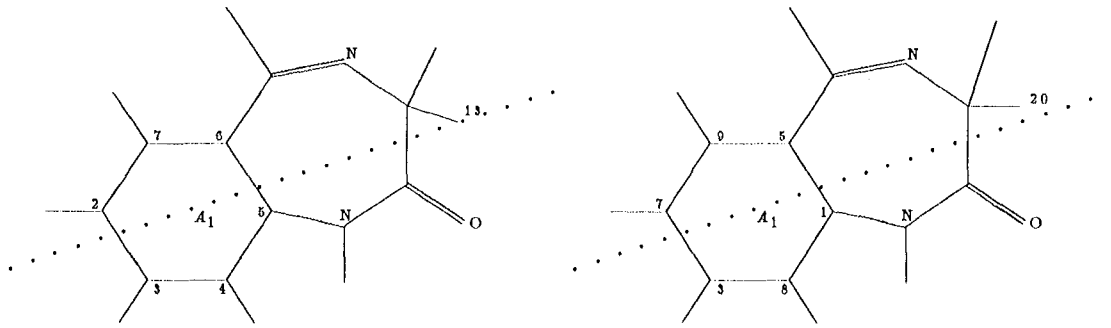
H	B	C	N	O	F
1.20	1.50	1.70	1.55	1.52	1.47
		Si	P	S	Cl
		2.00	2.10	1.80	1.75
			As	Se	Br
			2.00	2.00	1.85

Le point A_1 peut être choisi très librement dans une première molécule, dite de *référence*. Une première condition est qu'il soit simplement inclus dans les limites de l'enveloppe pour qu'il y ait toujours deux points d'intersection de celle-ci avec toute droite passant par A_1. L'autre condition est de procéder de sorte qu'il conserve *la même position géométrique* lors du passage d'une molécule à une autre. Elle sera satisfaite en superposant d'abord les deux molécules puis en exprimant les coordonnées cartésiennes de A_1 comme une combinaison linéaire des coordonnées de quelques atomes *correspondants*. La figure 5 illustre ce report du point A_1 d'une benzodiazépine à une autre, les correspondances entre atomes de base étant rappelées dans le tableau qui la suit.

Dans la première molécule, A_1 est défini à partir des atomes de carbone du premier cycle benzénique. Il peut donc figurer en septième position du tableau. Nous avons ensuite imposé le système de coordonnées fondé sur les septième, quatrième et cinquième points correspondants A_1, C_5 et C_6. A_1 est l'origine, le vecteur $\vec{i}$ est orienté vers C_5, le vecteur $\vec{j}$ lui est perpendiculaire et dans le même sens que $\vec{A_1C_6}$ puis le vecteur $\vec{k}$ est déduit pour obtenir un trièdre orthonormé direct. Nous avons enfin tracé une droite passant par A_1 et un deuxième atome, H_{13}, pour calculer l'éloignement de l'enveloppe de la molécule dans cette direction de l'espace.

Pour calculer la même valeur dans une seconde molécule, nous devons la superposer à la première, c'est à dire lui appliquer le système de coordonnées fondé sur les septième, quatrième et cinquième points, A_1, C_1 et C_5 puis tenir compte du décalage entre les deux origines. Il ne reste plus alors qu'à reporter la direction de calcul avant de déterminer le

Figure 5: Eloignement relatif de la surface de contact dans deux molécules



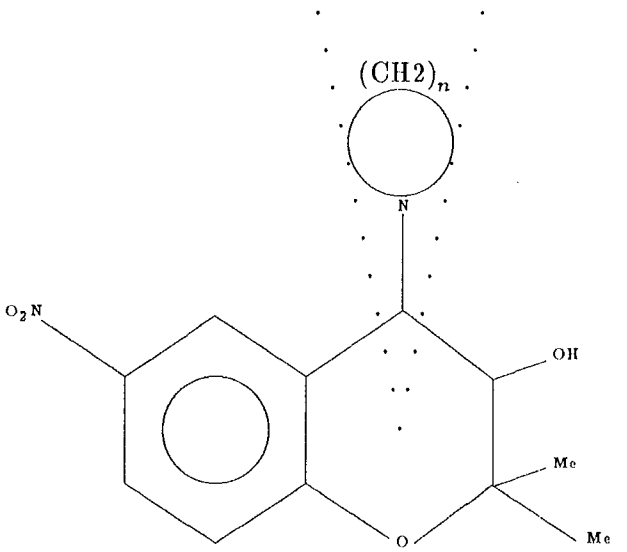
	correspondances entre atomes de base					
molécule 1	2	3	4	5	6	7
molécule 2	7	3	8	1	5	9

nouveau point d'intersection avec l'enveloppe. Ainsi nous obtenons un premier descripteur calculable sur plusieurs molécules, une distance point de référence - site actif. Dans la pratique, le logiciel OSCAR peut ensuite réïtèrer la procédure sur plusieurs directions passant par A_1, la succession des points d'intersections sur la surface donnant alors une bonne comparaison des contours de chaque site moléculaire.

Cette méthode est d'emploi très général. Elle ne demande qu'une table de correspondances à établir au début de l'étude d'un lot de molécules et un point de référence que l'utilisateur peut placer là où il veut. En outre, nous avons prévu une variante dans laquelle le descripteur n'est pas une distance A_1P_i mais plutôt une moyenne de distances calculées dans plusieurs directions divergentes à partir de A_1. Ce critère, moins fin, est en revanche plus fiable pour estimer l'encombrement stérique d'un substituant asymétrique, plus ou moins gros d'une molécule à l'autre et libre de tourner autour d'une liaison. Nous l'avons testé sur des benzopyranes (figure 6) qui ne diffèrent les uns des autres que par la taille du substituant de l'atome d'azote. La moyenne de plusieurs distances "barycentre du cycle pyrane - enveloppe de Van der Waals" mesurées dans un angle solide orienté vers les atomes de carbone du substituant, s'est avérée un descripteur correct. En effet, elle permet de retrouver des résultats publiés sur ces molécules [58,59], à savoir qu'un éloignement de la surface trop faible ou au contraire trop important coïncide avec une moindre activité des molécules.

L'éloignement du site actif par rapport à un point de référence est donc un critère géométrique simple à calculer et digne d'intérêt puisque susceptible d'influer sur l'activité. Cependant, il faut garder à l'esprit les limites propres à ce genre de méthode. Les molécules que nous avons considérées, les benzodiazépines comme les benzopyranes, ont une structure de base rigide et leurs conformères sont dus à la rotation d'un seul groupe substituant : autant dire que nous n'avons pas choisi les exemples les plus compliqués. Dans le cas des benzopyranes, nous avons pris soin d'inclure dans le test précédent

Figure 6: Descripteurs dans les Benzopyranes



des espèces aux substituants compacts et non pas linéaires. Il faut bien reconnaître qu'une molécule comme Bzp - $N(CH_2)_3Me$ a des rotamères tellement distincts - quant aux régions de l'espace qu'ils occupent - que la mesure de l'éloignement de la surface risque de donner n'importe quelle valeur.

L'autre inconvénient de ces descripteurs est qu'ils se prêtent peu à des calculs en arrière-plan. En effet, OSCAR peut difficilement trouver lui-même des points de référence qui occupent une position géométrique remarquable - notamment en matière de symétries - et que le chimiste trouvera, lui, intéressants. C'est donc ce dernier qui doit bien préciser les critères géométriques à étudier, le logiciel se chargeant ensuite des calculs nécessaires. Ce point est important à souligner car si OSCAR permet de calculer pratiquement n'importe quelle grandeur, c'est à l'utilisateur de s'assurer qu'elles sont assez significatives et d'en interpréter les résultats avec prudence.

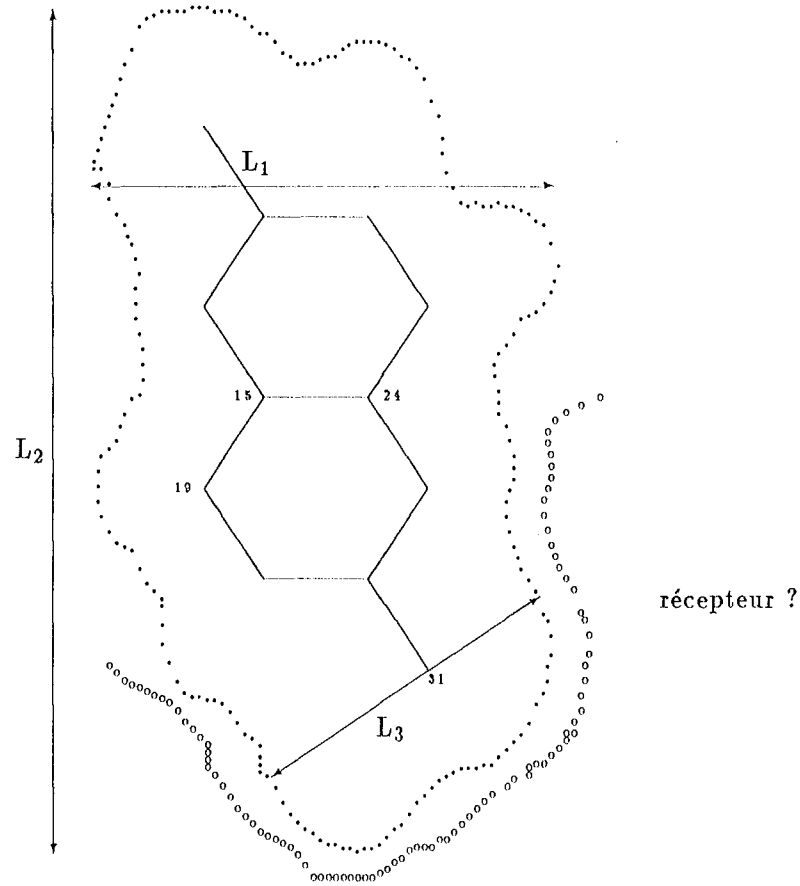
4.4.2 Développements

Des améliorations peuvent être apportées à ces descripteurs, à commencer par une définition plus fine de la surface moléculaire. Ainsi Richards [60] a-t-il proposé de prendre en compte le solvant sous la forme d'une molécule supposée parfaitement sphérique et qui, en "roulant" sur les sphères de Van der Waals, dessinerait une surface de contact aux formes moins anguleuses [1]. Cette définition est plus réaliste puisqu'elle fait intervenir l'environnement et a l'avantage de *lisser* la surface en éliminant les creux,

[1] ce modèle, qui peut paraître simpliste, nous convient en l'absence d'informations précises sur le solvant. De plus le rayon de ce dernier est un paramètre modifiable dans notre programme

notamment aux intersections entre deux ou trois sphères [61]. Ce point étant important dans nos calculs de descripteurs, nous avons jugé utile d'adapter dans le logiciel deux algorithmes fondés sur cette définition [62,63,64].

Figure 7: Surface moléculaire et site récepteur



Dans ces conditions, il est possible d'étendre la technique précédente en mesurant la distance entre deux points d'intersection d'une direction préférentielle avec la surface moléculaire. Il ne s'agit plus de mesurer l'éloignement du site actif par rapport à un point de la molécule mais de regarder si celle-ci a des dimensions telles qu'elle puisse s'inscrire dans un site récepteur [65,66]. L'observation peut porter soit sur la structure entière, soit sur une partie concave ou convexe, comme l'illustrent respectivement les descripteurs L_1 et L_3 de la figure 7. Pour définir le premier, il suffit de prendre l'atome 15 comme point A_1 et une ligne directrice parallèle à celle qui passe par l'atome 24. Pour le second, l'utilisateur pourra choisir les atomes 31 et 15 tout en indiquant au logiciel de "basculer" la direction de 90 degrés. Dans les deux cas, cette plus grande liberté dans le choix des directions de l'espace nécessite la programmation dans OSCAR de calculs supplémentaires de matrices locales de translations et de rotations.

4.4.3 Surface de contact

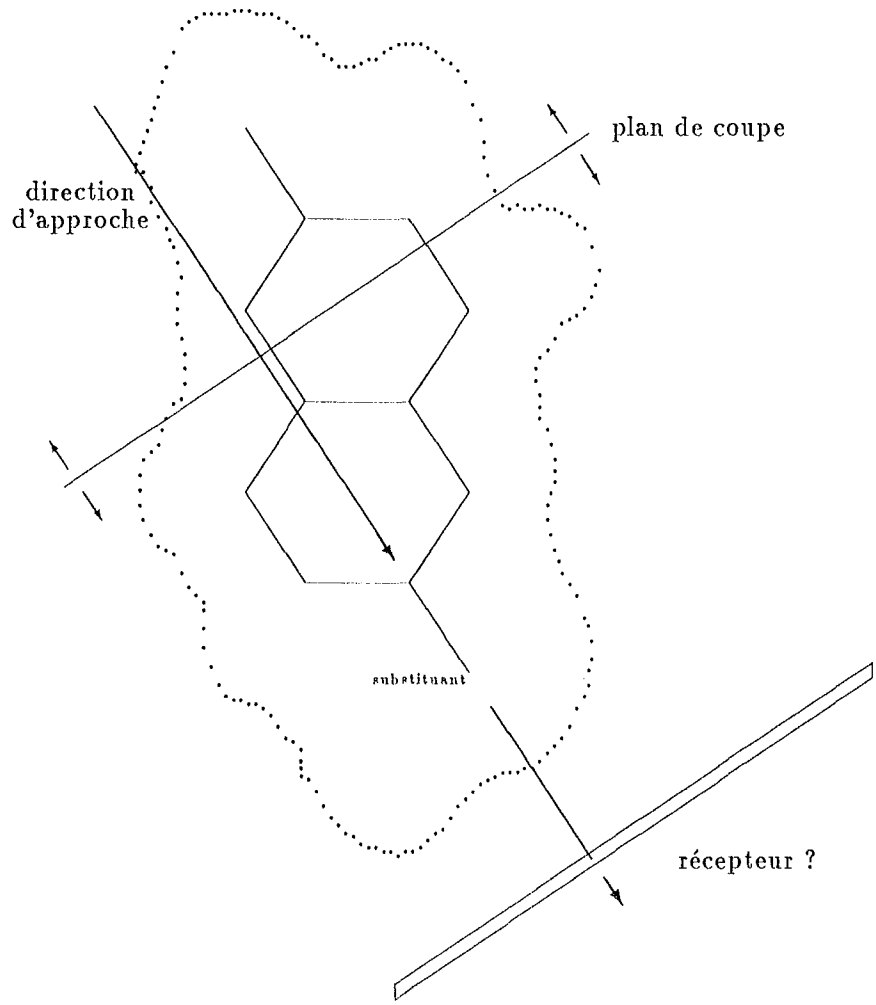
L'avantage des descripteurs précédents est qu'ils sont fiables, calculables rapidement et d'une manière très automatique. En réalité, le problème réside surtout dans le fait que le chimiste ignore s'ils sont vraiment intéressants... Ainsi le paramètre L_3 de la figure 7 semble-t-il être le critère idéal, mais seulement si le site récepteur se trouve réellement là où nous l'avons dessiné. Dans le cas contraire, le graphe de l'activité en fonction de sa valeur n'aura guère de significations, ressemblant probablement à un fouillis de points comme celui de la figure 14 (paragraphe 5.1). Dans la pratique, l'utilisateur du logiciel OSCAR devra donc étudier bien des directions géométriques, avant de trouver les critères qui semblent gouverner l'interaction avec le site récepteur.

Pour écourter ces tatônnements, des auteurs ont développé des descripteurs de l'encombrement stérique moins ciblés que les nôtres tout en restant fiables [52]. Les paramètres MSD (pour *minimal steric difference*) sont fondés sur l'hypothèse que l'affinité d'une molécule pour un récepteur est une fonction linéairement décroissante de la somme des volumes non superposables entre leurs contours respectifs. En d'autres termes, il s'agit de calculer numériquement le volume de l'interstice entre sites actif et récepteur puis d'y ajouter le volume de chaque bosse ou de chaque creux qui nuisent à une bonne imbrication des deux formes. Le principal inconvénient est qu'en l'absence d'informations topologiques sur le récepteur, c'est concrètement la molécule la plus active qui fournit la forme de référence. A cela s'ajoute la difficulté de calculer précisément ces petits volumes successifs, sans oublier les problèmes préalables de conformations. Un algorithme dont le code source est publié (!) est exposé à ce sujet [52]. Des corrélations satisfaisantes entre ces descripteurs et l'activité ont finalement pu être obtenues sur des séries de petites molécules.

Cette idée nous a paru intéressante à approfondir de la manière suivante. Pour comparer les formes de plusieurs molécules face à un récepteur inconnu, une solution est de raisonner en termes de *surface projetée*. Comme précédemment, il s'agit de définir une direction dont nous supposons qu'elle est orientée vers le récepteur. Et nous voulons savoir comment évolue l'encombrement stérique de chaque molécule lorsqu'elle s'approche de ce dernier le long de cette direction (pour simplifier, nous supposons que cette approche s'effectue de face et non de biais). Ainsi, pourrons-nous regarder si une distinction entre molécules, actives et inactives par exemple, coïncide avec un encombrement moindre à partir d'une certaine distance.

Dans une session du logiciel OSCAR, le chimiste pourra choisir une molécule singulière, comme la plus active, puis un de ses axes de symétrie comme direction de référence. Elle sera alors reportée dans les autres molécules selon la méthode exposée dans le paragraphe 4.4.1. La direction pourra être aussi définie directement à partir d'atomes de base ou comme d'habitude de combinaisons linéaires d'atomes de base, auquel cas les molécules ne seront pas orientées strictement de la même manière. La première variante est plus appropriée pour comparer l'approche d'ensemble des molécules, la seconde pour s'intéresser plus précisément au déplacement d'un substi-

Figure 8: Encombrement stérique à l'approche d'un récepteur



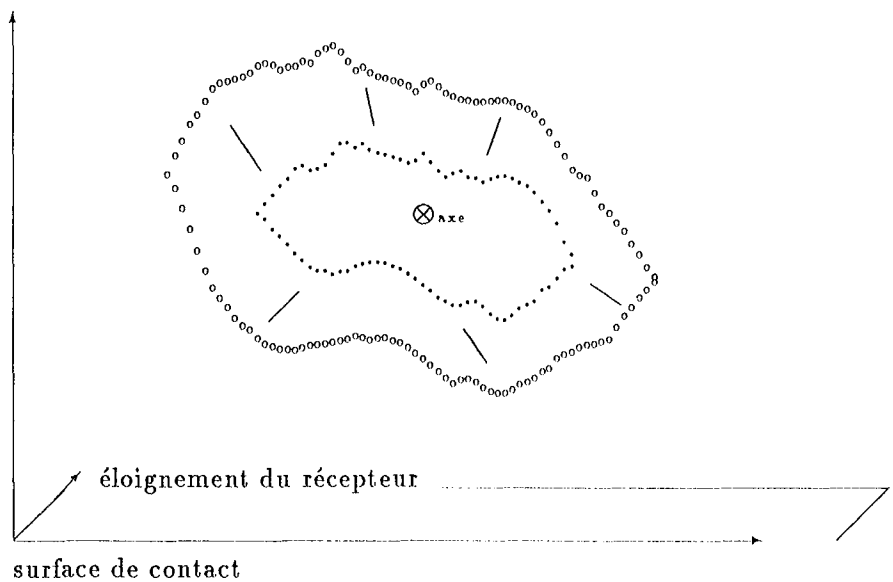
tuant (figure 8).

Nous avons besoin de deux plans parallèles entre eux et perpendiculaires à la direction d'approche. Le récepteur sera supposé proche du premier, mais sans autres hypothèses quant à sa forme. Les molécules "passent" à travers du second ce qui permet de mesurer la surface qu'elles occupent au fur et à mesure de leur rapprochement. En fait, du point de vue de la programmation, nous envisagerons la situation inverse, dans laquelle le deuxième plan est déplacé le long de l'enveloppe de chaque molécule fixe. Et seuls les points situés à l'avant de ce plan de coupe serviront à déterminer la surface cumulée et croissante avec la distance entre les deux plans.

En projetant ces points sur le premier plan, OSCAR fournit une image comme celle de la figure 9. Même si l'effet de perspective n'est pas très bien restitué, elle donne une idée assez réaliste de la place que prend chaque molécule. Reprenons alors ce que nous avons imaginé au paragraphe 1.2, toujours dans l'hypothèse où le récepteur est

réellement situé dans la direction choisie. Une espèce s'approche à quelques angströms de lui et, là, débute un processus de reconnaissance de formes. La molécule expose une partie de sa surface au récepteur, peut tourner sur elle-même (autour de la direction de référence) pour mieux s'adapter à la topologie de ce dernier, mais finalement "décider" de poursuivre son chemin si elle ne lui ressemble pas assez. OSCAR rend compte de cette situation avec une image obtenue pour la plus faible distance entre les deux plans d'observation. Si réactif et récepteur s'accordent bien, nous pouvons imaginer que le premier se rapproche du second et que le logiciel indique alors si cela se fait sans encombre - cas d'une molécule allongée - ou avec le risque de "coincer" plus ou moins vite - cas de molécules plus rondes. Si en définitive, les deux formes restent compatibles à distance plus réduite, ce seront vraisemblablement des critères géométriques plus locaux ou des critères électroniques qui décideront d'une interaction ou non.

Figure 9: Rapprochement d'une molécule d'un récepteur



Comparer ces images avec celles que donne chacune des autres molécules est délicat - s'agissant de formes dans l'espace - et fastidieux - compte tenu de leur nombre. Avec le logiciel OSCAR, une première solution est de "diviser" l'affichage, c'est-à-dire d'observer jusqu'à quatre molécules simultanément et de lancer le calcul en parallèle sur chacune. Dans ces conditions, la surface de contact est dessinée par tracés successifs en se rapprochant progressivement du récepteur, ce qui facilite les comparaisons. De plus, à titre d'aide, des valeurs numériques peuvent être calculées, l'aire en $\AA^2$ ou bien un des rayons allant de la direction vers le pourtour de la surface. L'essentiel est de regarder si les molécules inactives ne sont pas celles qui précisément ne peuvent pas se rapprocher du récepteur à moins d'une "distance - seuil" que le logiciel pourrait fournir. Dans le paragraphe 5.2, nous reprendrons ce type d'étude en intégrant le potentiel électrostatique dans cette approche du récepteur.

4.4.4 Critères géométriques locaux

Il se peut qu'une étude systématique d'un lot de molécules avec les descripteurs qui précèdent ne soit pas concluante. En d'autres termes, les tests exposés dans le paragraphe 5.1 montrent qu'aucun des critères géométriques à longue distance ne semble être déterminant dans l'interaction molécule - récepteur.

Cet échec relatif peut avoir trois causes :

1. L'activité pharmacologique des molécules n'est pas la conséquence plus ou moins directe du succès ou de l'échec d'un seul processus bien déterminé de reconnaissance de formes. Par exemple, un récepteur peut sélectionner les réactifs qui lui conviennent, les autres trouvant néammoins un autre site, différent sur le plan topologique mais équivalent sur le plan pharmacologique. Nous ne ferons pas d'autres hypothèses sur des processus complexes et si peu connus, mais il est bien évident que les descripteurs que nous avons présentés deviennent caduques, incapables de rendre compte de ce qui se passe réellement...

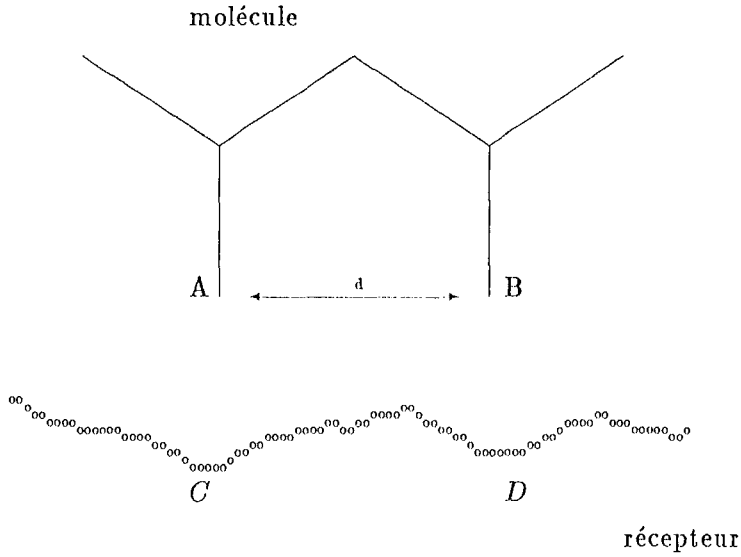
2. Il est raisonnable de penser que nos molécules, ayant la même structure de base, vont tout de même intervenir aux mêmes endroits de l'organisme humain. Cependant, ce ne sont pas des critères à longue distance qui sélectionnent les molécules mais des facteurs plus locaux. Ce qui signifie qu'elles ont toutes une forme d'ensemble satisfaisante mais qu'en se rapprochant du récepteur, celui-ci impose d'autres conditions, plus précises, portant par exemple sur des distances entre substituants voisins.

3. Il est vraisemblable que finalement, ce soient des critères de nature électronique et non pas géométrique qui décident du sort de la réaction. L'inconvénient est que la frontière est bien difficile à établir entre les deux : qui peut dire par exemple si à courte distance du récepteur, une bonne imbrication des sites l'emporte sur la répulsion des électrons ?

Revenons au second cas. Les critères géométriques locaux susceptibles de jouer un rôle sont des distances-clés, entre deux substituants par exemple (figure 10) ou bien entre un atome et un plan de symétrie. Les tests préliminaires du premier descripteur ont d'ailleurs été menés dans un cadre différent des relations structure-activité puisqu'il s'agissait de comparer les aptitudes respectives de quelques dérivés de l'acide benzoïque à se physisorber sur une surface de fluoro-apatite [67].

Pour ces essais, nous connaissions la surface réceptrice, c'est-à-dire les sites possibles de fixation. Dans nos calculs, nous avons considéré l'accrochage, non pas d'une seule molécule de réactif, mais de plusieurs molécules à peu près parallèles entre elles. Leurs différentes conformations ont été envisagées en tenant compte des barrières de rotation et de l'encombrement stérique de la région. Nous avons finalement constaté que la variation de distance entre deux substituants de ces réactifs était très discriminante vis-à-vis de la mesure du volume que pouvait fixer la surface. Même s'il est impossible

Figure 10: Site récepteur et distance entre substituants



de prouver formellement une relation de cause à effet entre ce descripteur et ce résultat expérimental, cette corrélation semble particulièrement significative.

Techniquement, ce descripteur est immédiat à calculer. Il suffit de prendre la distance entre deux atomes correspondant d'une molécule à l'autre et de tenir compte de leur rayon de covalence. Comme pour les directions préférentielles, ce paramètre est d'emploi général et facile à utiliser pour rechercher des différences locales de topologie. De plus, il peut être étendu aux autres coordonnées internes, comme la mesure d'un angle clé, celle d'un angle dièdre ou encore la distance entre un atome et un plan de symétrie. Ces critères sont souvent utiles et seront largement utilisés dans le dernier chapitre.

4.5 Formes électriques

Après cet examen détaillé des caractéristiques topologiques du site actif, il est temps de prendre en compte les propriétés électroniques. Nous nous intéresserons surtout au potentiel électrostatique, dont nous avons vu au paragraphe 2.2 qu'il était important, et que nous savons calculer à partir de la base d'orbitales atomiques et de la matrice densité des orbitales moléculaires. Regardons d'abord comment manipuler cette propriété sur plusieurs molécules simultanément, selon plusieurs points de vue. Nous évoquerons ensuite une démarche plus originale, qui nous permettra de parler de formes topologiques *et* électriques.

4.5.1 Courbes équipotentielles

La première idée qui vient à l'esprit est de tracer des courbes isopotentielles dans chaque molécule puis de regarder en quoi elles diffèrent. Il s'agit d'abord de choisir trois atomes de base non alignés (ou trois points définis à partir d'atomes de ce type) pour avoir un plan de référence. D'autres plans, parallèles et équidistants, répartis "devant" et "derrière", peuvent être ajoutés à cette occasion. L'utilisateur doit ensuite fournir une grille de points dans chaque plan, c'est-à-dire une origine et un pas d'incrémentation selon ses deux axes. OSCAR calculera alors le potentiel en chaque point ainsi défini.

Pour alléger l'affichage, il est souhaitable de choisir quelques valeurs particulières du potentiel et de demander à ne voir que les points où il est proche d'une de ces valeurs. Il est possible aussi de demander un affichage par lignes, plutôt que par points, grâce à une méthode d'interpolation. L'utilisateur dispose par ailleurs des classes de visibilité (page 26) pour sélectionner ce qu'il veut observer, plusieurs courbes isopotentielles dans un même plan ou bien une courbe isopotentielle dans des plans parallèles. N'oublions pas que moins l'affichage est sollicité, plus le calcul est rapide.

Ces courbes ont le mérite de fournir une information assez évocatrice. Les zones de potentiel positif et négatif sont clairement distinguées, avec les conventions habituelles (rouge / bleu). L'utilisateur verra facilement les "nœuds" qui peuvent apparaître au cœur même des molécules, contrairement aux courbes d'isodensité. Pour faciliter ensuite la comparaison avec d'autres molécules, il sera encore possible de "diviser" l'affichage et donc d'observer des tracés menés en parallèle sur plusieurs espèces. Des illustrations en sont données pages 67 et 68.

Ce travail de programmation un peu fastidieux est nécessaire puisqu'il fournit un des rares outils de comparaison entre molécules qui soit assez "parlant". De plus, il est offert par la plupart des logiciels de modélisation moléculaire. Cependant la qualité des dessins obtenus ne doit pas faire oublier que ces comparaisons sont purement visuelles et donc qualitatives. Autrement dit, elles ne peuvent pas être menées en arrière-plan et si le chimiste étudie toute une série de composés, il sera vite confronté à une très grande masse d'informations. Pour un lot d'une cinquantaine de molécules, il est hors de question de définir un critère systématique sur le potentiel qui distinguent des espèces éventuellement assez différentes.

4.5.2 Points remarquables

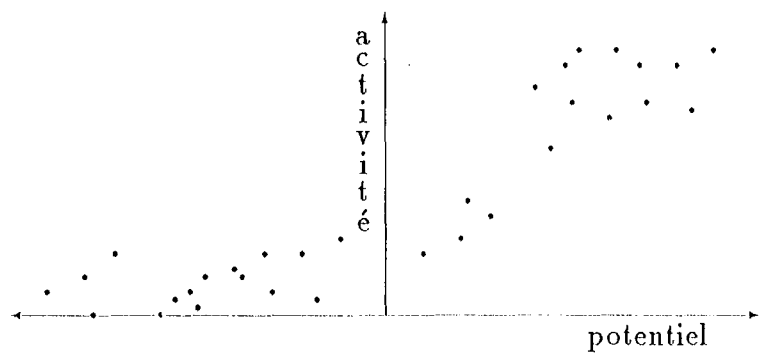
Compte tenu de la surabondance des résultats fournis par les courbes équipotentielles, il faut réduire le nombre de points où le potentiel est calculé sans trop perdre d'informations. Une bonne idée est de n'utiliser que des points de la surface moléculaire définie au paragraphe 4.4. Nous pouvons même les distinguer en trois catégories selon

qu'ils appartiennent à des zones convexes de cette surface, à des zones "cols" ou bien à des zones concaves. Du point de vue d'un récepteur éventuel, ils n'ont d'ailleurs pas la même importance puisque les premiers lui sont plus accessibles que les derniers.

Dans la pratique, il est malheureusement difficile d'exploiter pleinement cette classification. Si avec le logiciel OSCAR, nous savons calculer ces points et indiquer la catégorie à laquelle chacun appartient, nous ne pouvons pas établir de correspondances entre les différentes molécules. Autrement dit, nous n'avons pas l'assurance que le calcul du potentiel mené en un point le soit dans une position strictement comparable dans chacune des autres espèces, comme pour les atomes de base. Après de multiples essais de programmation, nous n'avons pu contourner cette difficulté que par une démarche complètement différente exposée à partir du paragraphe 5.2.

Dans l'immédiat, nous proposons de reprendre la technique des directions préférentielles du paragraphe 4.4.1. Pour chaque axe choisi par l'utilisateur dans une molécule de référence, nous obtenons un point sur une partie concave ou non de la surface moléculaire. Le potentiel y est alors calculé. Le "report" de ce point dans les autres molécules donne après une idée de l'évolution du potentiel, relativement exacte mais très ciblée localement. Dans une session du logiciel OSCAR, l'utilisateur obtiendra un graphe analogue à ceux du paragraphe 5.1, à ceci près que la nature algébrique du potentiel en augmente l'intérêt.

Figure 11: Activité et Potentiel en un point de l'espace



L'utilisateur pourra faire simultanément les deux comparaisons, portant d'une part sur l'éloignement de la surface moléculaire, d'autre part sur la valeur locale du potentiel. Comme les deux descripteurs sont d'une nature différente, géométrique et électronique, ils ne sont pas redondants. Le chimiste pourra donc énoncer des conditions comme par exemple : "pour qu'une molécule soit active, il faut que le potentiel mesuré à 1.8 angströms du substituant X dans la direction du premier axe principal d'inertie soit positif" ou bien encore "le potentiel mesuré dans la direction du substituant NO_2 n'a pas d'importance pour estimer l'activité".

Cette manière de comparer le potentiel en des points remarquables présente un risque. Il est possible que dans une molécule, il y ait une variation sensible de la grandeur dans la région située autour du point. Dans ce cas, une première variante consiste à considérer plutôt une moyenne du potentiel dans cette région mais pose alors le problème de la délimitation, forcément arbitraire, de cette zone. De plus, faire une moyenne de valeurs et rechercher ensuite une variation de cette moyenne dans plusieurs molécules est plutôt antinomique.

Une meilleure idée est de calculer le gradient du potentiel, donc les composantes locales du champ électrique. Le programme procédera de sorte que la première soit orientée comme la direction de référence et vers l'extérieur de la molécule, les deux autres lui étant perpendiculaires. Si ces composantes s'avèrent sensiblement différentes, entre elles ou d'une molécule à l'autre, elles démontreront l'intérêt des lignes de champ comme descripteurs électroniques. C'est pourquoi, nous avons inclus cette seconde variante dans notre logiciel.

Numériquement, nous avons :

$$\vec{E} = -\vec{\nabla} V \tag{30}$$

Les contributions électroniques font donc intervenir les trois composantes u :

$$E_u(M) = \sum_{\mu} \sum_{\lambda} P_{\mu\lambda} \frac{\partial}{\partial u} \left(\langle \varphi_\mu \mid \frac{1}{r_{uM}} \mid \varphi_\lambda \rangle \right) \tag{31}$$

expressions qui laissent entrevoir un calcul plus complexe et sensiblement plus long [19]. Dans un premier temps, l'utilisateur peut d'ailleurs se contenter de la première composante. Globalement, elles fournissent une information un peu différente du potentiel, et peut-être même plus fine. En effet, il est possible que des molécules similaires soient sensibles, non pas à une grandeur électrique scalaire, mais plutôt à son gradient, plus ou moins élevé et orienté.

Concrètement, ces composantes peuvent être affichées sur un graphe en fonction de l'activité, comme pour le potentiel. Graphiquement, une option permet le dessin des lignes du champ électrique.

4.5.3 Domaines équipotentiels

Mezey [68,69,70,71] a proposé ces dernières années, une méthode assez différente. L'idée n'est pas d'observer la variation du potentiel en des points correspondants mais au contraire de comparer des ensembles de points équipotentiels. La difficulté est donc reportée sur un autre aspect qui est celui de la *similarité* de formes moléculaires.

Contrairement aux enveloppes de Van der Waals ou de Connolly - qui sont déterminées "au mieux" - une surface équipotentielle est différentiable. En effet, nous avons

là une fonction continue dont nous pouvons calculer les dérivées premières et secondes. Si P et T désignent respectivement un point quelconque de la surface et le plan qui y est tangent, nous pouvons utiliser la formule de Taylor pour exprimer le potentiel en un point voisin de P et situé à une "altitude" h de T :

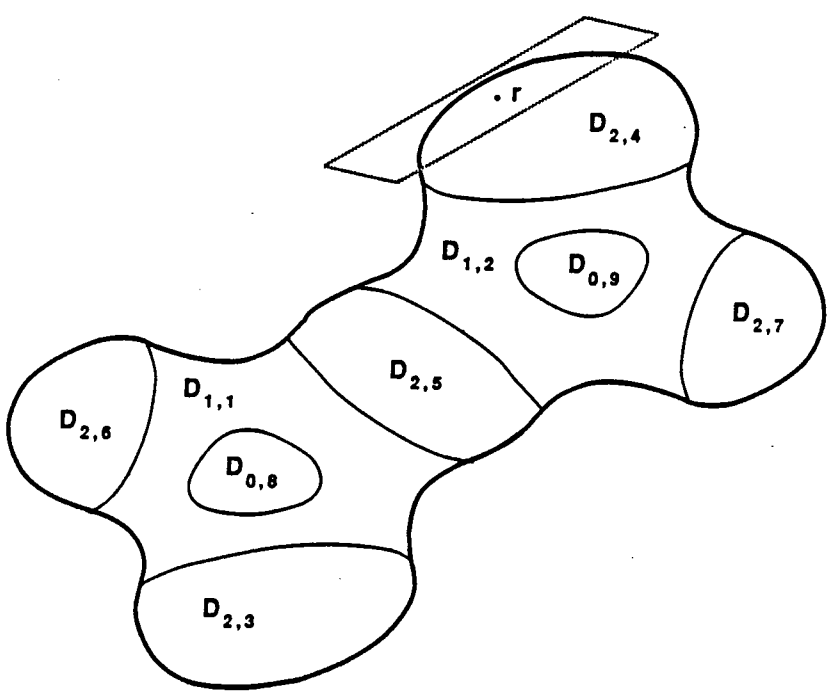
$$f(x+h) = f(x) + h f'(x) + \frac{1}{2} h^2 f''(x) + \ldots \tag{32}$$

relation qui s'écrit aussi :

$$f(x+h) \approx c - b\,x + \frac{1}{2}\, x\, A\, x \tag{33}$$

où c, b et A désignent respectivement la grandeur scalaire, le gradient de la fonction et la matrice de ses dérivées secondes (*Hessian matrix*). Une fois cette dernière diagonalisée, elle fournit comme valeurs propres deux rayons de courbure de la surface au point P. Si les deux sont négatifs, le point est concave ; si les deux sont positifs, il est convexe ; s'ils sont de signes opposés, il s'agit d'un col. Le nombre de valeurs propres négatives est donc un critère pour classer un point et diviser ainsi la surface en *domaines* regroupant des points de même type. Dans la figure 12, ces domaines sont doublement indexés, par concavité puis par taille décroissante.

Figure 12: Domaines d'une surface équipotentielle



Chaque domaine D_i est délimité par des points frontières, qui constituent un sous-domaine D_i^c. Ils sont utiles pour définir une *relation de voisinage* de la manière suivante : D_i et D_j sont voisins si, et seulement si, la réunion des ensembles $D_i \cap D_j^c$ et $D_j \cap D_i^c$ n'est pas vide. Toute l'information sur les domaines peut finalement être condensée sous la forme d'une *matrice de forme* symétrique, dont chaque élément diagonal s_{ii} caractérise

la concavité de D_i, tandis que les éléments s_{ij} valent 1 ou 0 selon que D_i et D_j sont voisins ou non. Dans le cas de la figure 12, cette matrice est celle qui suit.

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Comparer les surfaces isopotentielles de deux molécules revient donc à comparer deux matrices de formes. Si elles apparaissent assez différentes mais qu'elles peuvent être "rapprochées" en permutant simultanément des lignes ou des colonnes dans l'une d'elles, cela signifie que les arrangements de parties concaves, convexes et cols sont similaires, et que la dissemblance est due pour l'essentiel à la taille de ces domaines. Le nombre de permutations nécessaires peut même servir pour quantifier approximativement le degré de ressemblance de deux molécules.

Cette technique a fait l'objet récemment de quelques applications [72,73,74,75]. Elle suscite beaucoup d'intérêt car elle résout de façon élégante la difficulté de définir des critères numériques fiables pour distinguer des formes "électriques" dans l'espace. Cependant, elle présente des inconvénients de deux types.

Elle est très gourmande en calculs. Nous devons d'abord "enfermer" chaque molécule dans une grille de points où le potentiel sera calculé. Une valeur tabulée de cette grandeur sera choisie pour ne conserver que la surface équipotentielle correspondante (ce qui revient à dire qu'il faudra probablement traiter plusieurs valeurs du potentiel successivement). Puis nous aurons à trier les points, d'une part en fonction de leur proximité, d'autre part en fonction de leur concavité. Ceci suppose le calcul des dérivées premières, puis des dérivées secondes, enfin la diagonalisation de la matrice "hessienne". En bref, établir une matrice de forme doit demander plusieurs heures de calculs sur nos ordinateurs [1], ce qui est tout de même beaucoup pour une seule molécule.

L'utilisation de telles matrices présente des difficultés conceptuelles. Elles sont définies de manière intuitive donc très subjective, à l'image des raisonnements propres à l'intelligence artificielle. Par exemple, la relation de voisinage pourrait être définie différemment, notamment pour tenir compte des distances entre domaines ou encore de leurs orientations les uns par rapport aux autres. De plus, même s'il n'est pas trop difficile d'imaginer un algorithme de comparaison entre matrices [76] (*The Journal of Graph Theory* offre bien des idées de programmation), il est préférable que leurs éléments dia-

[1] Une architecture parallèle serait d'ailleurs assez appropriée

gonaux et non diagonaux aient la même signification. Ici, les deux propriétés, concavité et voisinage, semblent peu comparables.

— Procédures —

Comme pour la partie consacrée au graphisme moléculaire, nous indiquons ici les *routines* les plus significatives du logiciel OSCAR, du moins pour celles dont le source est accessible aux programmeurs :

Tableau 9: Procédures de calcul pour traiter des molécules

mccac	calcul des charges et du moment dipolaire
mccawm	indices de Wiberg (force des liaisons)
mccbci	calcul d'intégrales bicentriques monoélectroniques
mccdm	calcul de la matrice densité
mcced	densité électronique en un point
mccep	champ et potentiel électriques en un point
mcchd	super-délocalisabilités électrophiles / nucléophiles
mccmo	calcul d'orbitales moléculaires en un point
mccngf	*n-gamma function*
mccrm	calculs de matrices de rotations
mccrmh	rotation des harmoniques sphériques
mccrsf	partie radiale d'une orbitale de Slater
mccsao	orbitales atomiques de Slater en un point
mccsap	auto-polarisabilités atomiques sur HOMO et LUMO
mccsh	calculs d'harmoniques sphériques
mcqam1	paramètres de Slater pour méthode AM1
mcqavs	données sur la couche de valence de chaque élément
mcqcnd	paramètres de Slater pour méthode CNDO
mcqmi0	*idem* pour méthode MINDO
mcqmi2	*idem* pour méthode MINDO corrigé
mcqmi5	*idem* pour méthode MINDO 3
mcqmn0	*idem* pour méthode MNDO 0
mcqmn1	*idem* pour méthode MNDO 1
mcqmvs	données sur la couche de valence d'une molécule
mcqpm3	paramètres de Slater pour méthode PM3
mcqqn	données sur les orbitales atomiques de valence ; nomenclature
mcqzp	paramètres de Slater d'orbitales atomiques
mcrfap	lecture de polarisabilités atomiques sur fichier
mcwcao	valeurs d'orbitales atomiques de valence
mcwced	calcul de courbes / plans d'isodensité électronique
mcwcep	calcul de courbes / plans d'isopotentiel
mcwcmo	valeurs d'orbitales moléculaires de valence
	densité de probabilité de présence dans l'espace

Tableau 10: Procédures de base pour traiter des molécules

mbcabm	comparaison d'atomes appartenant à des molécules différentes
mbccac	conversion coordonnées internes → coordonnées cartésiennes
mbciac	établissement de jeux de coordonnées internes (récursif)
mbcmv	calcul données géométriques (volume, anisotropie, etc.)
mbgabl	longueur moyenne d'une liaison entre deux éléments (en fonction de l'ordre et de l'état physique du composé; tables)
mbgbo	ordre d'une liaison entre deux éléments
mbgcm	recherche de structures cycliques dans une molécule
mbqcmp	comparaison de séquences d'atomes entre molécules
mbgea	angles tabulés entre atomes
mbgmdm	calcul des dimensions d'une molécule
mbgec	données de la classification périodique des éléments
mbgec2	*idem*
mbgna	liste des voisins immédiats d'un atome
mbgsna	liste exhaustive des voisins d'un atome ; identification d'un substituant
mbmm2...	*routines* de mécanique moléculaire algorithmes inspirés des travaux de Allinger
mbqana	recherche de voisins d'atomes (récursif)
mbqcm	recherche de motifs communs à plusieurs molécules
mbqimc...	algorithmes de parcours (récursif)
mbqps	structure de base et substituants d'une molécule
mbqcs	calcul de l'enveloppe d'une molécule en présence d'un solvant algorithme fondé sur la définition de Connolly
mbqisw	intersections de l'enveloppe moléculaire avec des directions préférentielles de l'espace
mbqms	calcul de l'enveloppe d'une molécule en présence d'un solvant algorithme fondé sur la définition de Pascual et Silla
mbrcd	lecture du fichier d'une molécule (format CHIMISTE)
mbrcs	lecture de l'enveloppe moléculaire sur fichier
mbrd	lecture de fichiers de données en fonction de leur suffixe
mbrdfs	lecture de données sur tout un lot de molécules
mbrgd	lecture d'une molécule. Fichier au format GEOMOS (coord. internes)
mbrgg	*idem.* Fichier au format GAUSSIAN
mbrgk	*idem.* Fichier au format KGNGRAF
mbrgm	*idem.* Fichier au format MAD
mbrgr	*idem.* Fichier de résultats de GEOMOS (orbitales moléculaires)
mbrms	lecture d'une enveloppe moléculaire sur fichier (2)
mbviad	vérification des distances interatomiques
mbwcdf	Sauvegarde du fichier d'une molécule au format CHIMISTE
mbwdf	Sauvegarde de fichiers en fonction de leur suffixe
mbwfs	écriture de données sur tout un lot de molécules
mbwgdf	Sauvegarde du fichier d'une molécule en coordonnées internes

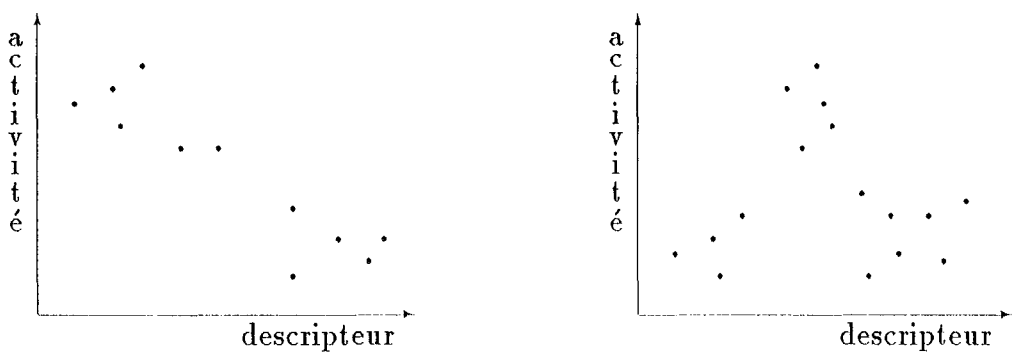
* * *

5 Analyse des données

5.1 Bons et mauvais descripteurs géométriques

Après cette énumération (non exhaustive) des paramètres géométriques qui décrivent plus ou moins bien la topologie du site, il est temps d'examiner les tests qui rendent compte de leur *qualité*. Pour chaque descripteur, cela consiste à observer comment évolue l'activité donc à tracer un simple graphe. Ensuite, il s'agit de calculer numériquement la probabilité de commettre une certaine erreur en estimant l'activité d'un nouveau composé pour lequel est calculée une nouvelle valeur du descripteur. Dans une session du logiciel OSCAR, l'utilisateur obtiendra parfois des graphes ressemblant à ceux de la figure 13, chaque point désignant une molécule.

Figure 13: Descripteurs géométriques discriminants vis-à-vis de l'activité



Les deux tracés semblent indiquer que l'activité ne puisse être obtenue que si la valeur du paramètre porté en abscisse est située dans des intervalles bien déterminés. Dans le premier cas, elle doit être supérieure à un certain seuil ; dans le second, elle doit rester comprise entre deux bornes. Il est donc raisonnable de penser que si un graphe similaire est obtenu pour une série de N molécules, la mesure de son descripteur dans une N + *unième* permettra d'estimer l'activité de cette dernière avec une faible probabilité de se tromper.

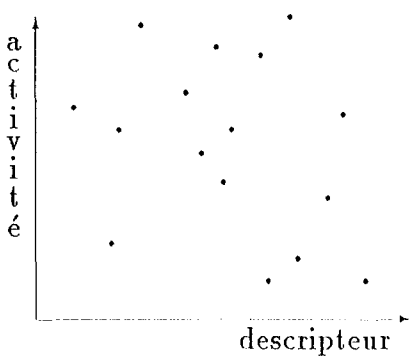
C'est exactement le genre de situations que nous recherchons. Pour une famille donnée de composés, OSCAR passera donc en revue les descripteurs géométriques possibles - soit automatiquement en arrière-plan, soit sur demande explicite de l'utilisateur - pour regarder ceux qui fournissent des graphes aussi intéressants. Même si la figure 13 illustre des cas d'école (bonne séparation des composés actifs et inactifs), ces situations ne sont pas purement théoriques. La première peut être constatée après mesure de

l'éloignement d'un certain substituant : l'activité exigerait qu'il ne soit pas trop encombrant de manière à dégager l'accès au récepteur. La seconde situation serait le propre d'une activité liée à la distance entre deux atomes séparés par une zone concave. Une "pince" trop étroite ou trop ouverte pourrait gêner tout "accrochage" sur le récepteur qui exigerait un écartement bien déterminé (on sait qu'il existe dans l'organisme humain des sites récepteurs dont les dimensions sont exactement adaptées au passage d'une certaine molécule et non d'une autre).

5.1.1 Corrélations numériques

Dans la pratique, l'utilisateur du logiciel OSCAR peut choisir de visu le critère géométrique qui l'intéresse et demander le tracé du graphe correspondant. Si le résultat est proche du fouillis de points qu'illustre la figure 14, il n'aura besoin d'aucune valeur numérique pour lui indiquer à quel point ce descripteur est peu satisfaisant (mais après tout, c'est aussi un résultat de constater que cette variable n'a aucun rapport avec l'activité). En revanche, ce calcul est nécessaire dans les situations pas très franches où l'activité semblerait évoluer d'une façon ou d'une autre... Et il est indispensable pour chaque essai de paramètre mené en arrière-plan.

Figure 14: Descripteur géométrique non discriminant vis-à-vis de l'activité



Calculer la séparation entre molécules actives et inactives revient à étudier en statistiques si les deux distributions de points (c'est-à-dire les valeurs du descripteur utilisé) sont similaires ou non. Utiliser les moyennes ou les variances pour les comparer est cependant d'autant moins précis que l'intervalle de variation du paramètre est étendu. C'est une première raison pour laquelle nous en avons limité l'usage. La seconde tient aux valeurs numériques données pour les activités. Les pharmacologues tirent ces valeurs de tests pharmacologiques plutôt empiriques et fournissent parfois des barres d'erreurs très grandes. Du fait de ce sérieux problème, il nous faut alors des tests plus complets que les *Student's t* et autres statistiques.

La question est la suivante : les deux ensembles de valeurs proviennent-ils ou non d'une même fonction ? Plus rigoureusement, peut-on prouver, et si oui jusqu'à quel point, que les deux distributions de valeurs (du descripteur et de l'activité) n'ont rien à voir l'une avec l'autre ? [1] Pour cela, deux cas sont à envisager selon que l'activité est donnée sous forme de valeurs continues ou *discrètes.*

l'activité est donnée par des valeurs continues :

Les critères possibles pour distinguer les deux distributions sont la mesure de l'aire qui les sépare, le décompte des intersections des deux courbes, la mesure de l'écart maximal entre les deux, etc. [77] Chacun a ses avantages et ses inconvénients, le dernier, dû à Kolmogorov [78], étant un des plus simples. Outre la mesure de la "distance" maximale entre les deux distributions, il fournit une valeur numérique (*significance*) qui indique à quel point le test est valide. Cette valeur est comprise entre 0 et 1, correspondant aux cas extrêmes des figures 14 et 13 (premier tracé). Dans la pratique, l'utilisateur du logiciel OSCAR peut appliquer ce test pour tout graphe affiché à l'écran.

l'activité est donnée par des valeurs discrètes :

Il peut s'avérer indispensable de trier les molécules en deux classes - molécules actives et inactives - ou plus finement encore. C'est le cas lorsque les valeurs numériques de l'activité diffèrent de plus d'un ordre de grandeur d'une molécule à l'autre ou lorsque les barres d'erreurs sont trop élevées. Mais cela peut être aussi intéressant lorsque le graphe indique des fluctuations significatives de l'activité pour de petits intervalles de variation du descripteur. Paradoxalement, la *discrétisation* des valeurs peut rendre la comparaison des deux distributions plus fiable même si cette opération constitue en elle-même une perte d'informations [79].

Après avoir classé les molécules selon le descripteur et selon l'activité - l'utilisateur étant libre de mener lui-même ce classement - la comparaison des distributions se fait grâce à un *tableau de contingences.* OSCAR y inscrit dans chaque case (i,j) le nombre d'espèces appartenant simultanément aux classes i et j. Dans l'hypothèse où les deux variables n'ont aucun rapport, les molécules devraient se répartir uniformément dans les cases du tableau. Inversement, si les deux variables sont totalement dépendantes, ce dernier devrait contenir beaucoup de 0, probablement selon une des diagonales. Le χ^2 donne une estimation du degré de discrimination mais nous lui préférons le coefficient de Cramer, compris entre 0 et 1 et qui dépend moins du nombre total de molécules.

Le test du χ^2 présente l'inconvénient d'être surtout qualitatif. En particulier, il ne permet pas de comparer directement la discrimination de deux variables x par rapport à une même grandeur y. Pour cela, il faut un test quantitatif plus fin, fondé, par exemple,

[1] Il est peut-être bon de rappeler que ces statistiques ne peuvent jamais démontrer la réalité d'une relation de dépendance y = fct (x). En revanche, elles peuvent éventuellement prouver le contraire, c'est à dire que deux variables n'ont rien de commun. Dans le cas d'une corrélation linéaire, la mesure du coefficient de régression de Pearson ne suffit pas et doit être complété par un test de validité. Et même après, rien ne prouve qu'une corrélation, même *bonne* ET *significative*, soit *réelle* pour autant.

Tableau 11: Tableau de contingences Activité - Descripteur géométrique

	$x_m<x_1$ ...	$x_i<x_m<x_{i+1}$	... $x_m>x_N$
$a_m<a_1$			
. . .			
		décompte des	
$a_j<a_m<a_{j+1}$		molécules	
. . .		(ou nombre d'*observations*)	
$a_m>a_L$			

sur la notion d'*entropie* [78,80,81].

5.1.2 L'entropie informationnelle

Imaginons la situation suivante : pour reconnaître une molécule parmi d'autres, plusieurs questions sont posées successivement à une personne qui y répond simplement par oui ou par non. Chacune de ses réponses a un "poids" différent selon qu'elle cerne plus ou moins bien la molécule recherchée, donc qu'elle élimine une fraction plus ou moins grande des molécules possibles. Numériquement, ce poids peut être défini par :

$$w_i = - \ln (p_i)$$

où p_i désigne la proportion des solutions restantes (entre 0 et 1). Le logarithme, qui accorde donc à la réponse un poids compris entre 0 et $+\infty$, permet en plus de donner la même valeur à une réponse qui élimine 83 % des solutions qu'à deux réponses qui en éliminent 50 % puis 67 %.

Inversement, l'intérêt d'une question dépend des réponses qu'elle suscite. S'il n'y en a qu'une, elle est caduque et son poids nul. S'il y en a deux, le poids est fonction de la probabilité d'avoir une réponse ou l'autre, puis de leur intérêt respectif. Supposons par exemple que la question soit : "Cette molécule est-elle active ?", que deux réponses soient possibles, "oui" et "non" et que dans un lot de molécules, 15 % d'entre elles soient actives. Si la question est posée à la personne précédente, la réponse "non" sera donc la plus probable (0.85), la molécule recherchée se trouvant ensuite parmi les 85 % qui restent. Inversement, la réponse "oui" sera peu probable (0.15) mais réduira bien davantage le nombre de solutions possibles (15 %). Plus généralement, l'intérêt d'une question (a), ou son "entropie", s'écrira donc sous la forme :

$$h(a) = \sum_{réponses\ a} p_a\ w_a = - \sum_{réponses\ a} p_a\ \ln (p_a) \qquad (34)$$

où il apparaît que $h(a)$ varie entre 0 et $\ln(n_a)$, maximum atteint avec un nombre n_a de réponses possibles, toutes équiprobables de $1/n_a$.

Nous posons maintenant une seconde question - "Cette molécule a-t-elle une forte valeur pour un critère géométrique donné ?", qui admet également deux réponses "oui" et "non". La probabilité d'avoir simultanément les réponses x et a, étant p_{xa}, la relation précédente donne aussi l'intérêt de deux questions posées simultanément :

$$h(x,a) = -\sum_{x,a} p_{xa} \ln (p_{xa})$$

Si elles sont bien distinctes, cet intérêt est maximal et vaut alors $h(x) + h(a)$ car les probabilités d'obtenir chaque réponse sont indépendantes : $p_{xa} = p_x\, p_a$. Si inversement, les deux questions sont complètement redondantes, elles offrent le même intérêt et les poser simultanément n'apporte rien de plus : $h(x) = h(a) = h(x,a)$.

De même, le poids de deux questions posées successivement s'écrit :

$$h(a \text{ "sachant" } x) = h(a|x) = \sum_x p_x \sum_a \frac{p_{xa}}{p_x} \ln \frac{p_{xa}}{p_x} = \sum_{x,a} p_{xa} \ln \frac{p_{xa}}{p_x}$$

qui permet de déduire la relation :

$$h(x,a) = h(x) + h(a|x) = h(a) + h(x|a)$$

En fait, poser (x) d'abord ne peut que réduire l'intérêt de poser (a) ensuite, proportionnellement au *degré d'association* des deux questions. Cette proportion de (a) qui devient redondante lorsque (x) est connue, peut se formuler :

$$u(a|x) = \frac{h(a) - h(a|x)}{h(a)}$$

Elle doit varier entre 0, si les deux n'ont aucun rapport, et 1, si (x) prédit complètement (a).

En considérant symétriquement les deux questions, les relations précédentes permettent finalement de définir le degré de dépendance de deux variables x et a :

$$u(x,a) = 2\, \frac{h(x) + h(a) - h(x,a)}{h(x) + h(a)}$$

ou aussi

$$u(x,a) = \frac{h(x)\, u(x|a) + h(a)\, u(a|x)}{h(x) + h(a)} \qquad (35)$$

où $u(x,a)$ apparaît comme une somme pondérée de leurs degrés de redondance.

Nous aurons donc une mesure quantitative du degré de dépendance entre un descripteur géométrique et l'activité de nos molécules en nous posant d'une part n_a questions "Cette molécule a-t-elle une activité comprise entre deux bornes a_1 et a_2 ?" et d'autre part n_x questions "Cette molécule a-t-elle une valeur du descripteur comprise entre deux bornes x_1 et x_2 ?". La probabilité d'obtenir la réponse a (ou x) sera la fraction des molécules appartenant à une case (x,a) du tableau de contingences, rapportée au total des molécules appartenant à la même colonne (ou la même ligne).

Dans le logiciel OSCAR, ce calcul du degré de dépendance est mené après le test du χ^2. Il faut donc commencer par regarder si qualitativement activité et descripteur sont liés ou non avant d'en obtenir une mesure quantitative. Le logiciel fournit trois données, l'entropie totale des deux grandeurs, la fraction de redondance dans la connaissance de l'activité connaissant le descripteur et enfin le degré d'association des deux grandeurs (entre 0 et 1) équivalent au coefficient de Cramer. Comme nous l'indiquions ci-dessus, opter pour ce test ou pour celui de Kolmogorov dépend en définitive des valeurs de l'activité dont on dispose et de leur incertitude.

Il faut noter que si l'utilisateur peut faire lui-même le tri des molécules, c'est-à-dire choisir judicieusement les bornes qui conduisent à la meilleure séparation possible entre les classes, il est préférable qu'il laisse ce soin au logiciel OSCAR. En effet, il vaut mieux que les modalités du classement restent les mêmes pour comparer successivement plusieurs descripteurs géométriques. De plus, il existe plus d'une solution pour classer des molécules en 3, 4 ou 5 ensembles, surtout s'il faut imposer à chaque classe un nombre minimal d'espèces.

5.2 Reconnaissance de formes

Il existe encore peu d'applications analogues à la démarche de Mezey (paragraphe 4.5.3), c'est-à-dire fondées sur l'emploi de formes moléculaires "iso-propriété", la définition de critères de ressemblance pour les comparer puis la programmation de méthodes utilisant ces critères [82,83,84,85]. Comme nous l'avons vu, cela tient au fait que la reconnaissance de formes s'appuie sur des raisonnements inhabituels pour des chimistes. Il est d'ailleurs révélateur de constater que ce sont surtout des informaticiens ou des mathématiciens qui se sont intéressés à ces problèmes.

Le domaine le plus en pointe en chimie est celui des spectroscopies. En effet, la "forme" à analyser est ici un spectre, en deux dimensions seulement, et de frontières clairement délimitées. De plus, le dessin est lui-même décomposable en formes primitives qui ont une signification physique : en RMN, les pics sont dus à des protons et sont caractérisés par un déplacement chimique. Comparativement, nos surfaces isopotentielles en trois dimensions et de contours arbitraires ne présentent aucune caractéristique évidente et simple à analyser.

Nous nous proposons ici de définir nos molécules sous la forme de *vecteurs* dont les composantes seront des descripteurs mêlant indifféremment les aspects géométriques et électroniques. Nous avons à l'esprit deux idées :

- Comparer le potentiel calculé en des points correspondants de plusieurs molécules (paragraphe 4.5.2) apporte une information exacte mais parcellaire. Inversement, estimer la différence entre deux molécules par leur similarité (paragraphe 4.5.3) est une démarche intellectuellement plus satisfaisante mais très subjective, donc moins

fiable. Nous voudrions définir nous-même cette notion et substituer à une mesure numérique de la ressemblance entre deux molécules une mesure par classification tenant compte des activités pharmacologiques.

- Nous aimerions revenir à la démarche du paragraphe 4.4.3, c'est-à-dire imaginer des orientations successives de nos molécules par rapport à un même récepteur. A la notion de descripteurs d'une molécule, il s'agirait de préférer celle de descripteurs de configurations réactif - récepteur.

Commençons d'abord par revenir sur le principal aspect de la reconnaissance de formes, la définition d'une similarité entre objets, avant d'aborder les critères de classifications.

5.2.1 Distance et Similarité

Pour savoir à quel point deux objets sont proches l'un de l'autre dans l'espace - ou "ressemblants" quant à leur position - nous utilisons naturellement la distance d'Euclide. Plus généralement, pour des objets (vecteurs) dotés de caractéristiques (composantes), exprimées elles-mêmes sous forme numérique, la définition est due à Minkowski :

$$D_{A-B} = \left(\sum_i \left(x_{Ai} - x_{Bi} \right)^k \right)^{\frac{1}{k}}$$

Si les composantes sont simplement binaires (absence ou présence d'une propriété par exemple), la distance de Hamming, définie à l'aide d'opérations logiques et non plus numériques, est souvent utilisée :

$$D_{A-B} = \sum_i XOR \left(x_{Ai}\, x_{Bi} \right)^1 \tag{36}$$

Un troisième exemple [86] consiste à donner aux composantes une "base" de poids w comprise entre 0 et 1 et à les trier par importance décroissante. La distance entre deux espèces i et j est définie par :

$$d_{ij} = 1 - s_{ij} = 1 - \sum_{k=1}^{N} t_k\, w^k \tag{37}$$

où s_{ij} désigne la similarité des deux espèces, N, le nombre de composantes que t_k compare :

$$\begin{cases} t_k = 1 & \text{si } i_k = j_k \\ t_k = 0 & \text{si } i_k \neq j_k \end{cases}$$

Pour une base w de 0.5, la similarité varie donc entre 0 et un maximum qui tend vers 1 lorsque deux espèces sont identiques du point de vue d'un nombre N infini de propriétés. Inversement, il est possible de retenir une base optimale w_0 telle que

$$\sum_{k=1}^{N} \left(w_0 \right)^k = 1 \tag{38}$$

[1] XOR est le "ou" logique exclusif

et afin que la similarité entre deux espèces puisse atteindre 1 sans dépasser cette valeur (autrement dit ceci évite des distances négatives). L'avantage de ce développement en puissances entières est qu'à chaque valeur s_{ij}, il ne correspond qu'une seule décomposition $t = [t_k]$, c'est à dire une seule manière pour deux espèces de se distinguer.

Toutes ces définitions montrent que chacun peut définir la ressemblance comme il l'entend, en fonction, d'une part de la nature des objets à décrire sous forme de vecteurs, d'autre part des propriétés mathématiques voulues pour la similarité. Par exemple, la dernière définition semble adaptée à des molécules dont les composantes sont des grandeurs géométriques ou électroniques d'importance décroissante. En revanche, elle n'obéit pas forcément aux propriétés habituelles des distances dites "archimédiennes" :

$$d_{A-B} \succ 0 \text{ si } A \neq B \; ; \; d_{A-A} = 0$$
$$d_{A-C} \leq d_{A-B} + d_{B-C}$$

5.2.2 Matrices de classement

Intéressons-nous par exemple à la valeur du potentiel électrostatique dans N régions de l'espace et dans toutes nos molécules. Nous pouvons construire autant de vecteurs moléculaires dont les N composantes seront des réels exprimant ce potentiel. Pour comparer deux molécules i et j, définissons-en la similarité de la manière suivante :

$$s_{ij} = \sum_{k=1}^{N} w_k \, s_{ijk} \tag{39}$$

Dans cette expression les w_k désignent les poids que nous attribuons à chaque zone (leur "importance") et qui, dans un premier temps, seront tels que :

$$\sum_{k=1}^{N} w_k = 1 \tag{40}$$

Nous choisirons un nombre de zones suffisant pour bien décrire chaque molécule mais qui reste nettement inférieur à leur nombre total (généralement, le nombre d'espèces doit être au moins trois fois supérieur à celui des propriétés). Elles seront disposées près des substituants les plus importants et surtout, devront être bien séparées dans l'espace afin que les composantes s_{ijk} soient indépendantes les unes des autres. La ressemblance des molécules i et j du point de vue du potentiel mesuré dans la zone k peut s'écrire :

$$s_{ijk} = 1 - \frac{|F_j(k) - F_i(k)|}{F_{max}(k) - F_{min}(k)} \tag{41}$$

où F est un potentiel "corrigé" dont nous détaillerons l'expression un peu plus loin. Ici, la similarité varie encore entre 1 et 0 selon que les deux molécules ont le même potentiel ou au contraire les valeurs les plus différentes.

L'ensemble des éléments s_{ij} constitue une *matrice de similarité* $[S]$, cohérente mais dont le défaut est de considérer comme aussi dissemblables deux molécules qui diffèrent soit fortement pour une zone, soit faiblement sur l'ensemble des régions. Pour accentuer l'importance du premier effet par rapport au second, le mieux est de faire intervenir dans l'équation (39) la variance des s_{ijk} qui varie dans ces conditions entre 0.25 et 0 : [1]

$$var\,(s_{ijk}) \;=\; \frac{1}{N}\sum_{k=1}^{N}\left(s_{ijk} - \bar{s_{ij}}\right)^2 \tag{42}$$

$$s_{ij} \;=\; (1 - var\,(s_{ijk}))\sum_{k=1}^{N} w_k\; s_{ijk} \tag{43}$$

Pour classer ensuite nos molécules, nous pouvons introduire un degré de séparation b et indiquer que deux molécules i et j appartiennent à la même classe si $s_{ij} \geq b$. Nous obtiendrons autant de classes que de molécules si $b = 1$ et inversement une seule classe les contenant toutes si $b = 0$. L'avantage est que nous passons donc d'une matrice carrée d'indices de ressemblance à une classification plus sommaire mais plus compréhensible puisque la définition de la similarité y devient transparente pour l'utilisateur.

Une *matrice de classe* $[C]$ peut résumer l'information sur cette séparation des molécules en plusieurs ensembles. Chaque élément c_{ij} contient la plus forte ressemblance qui puisse être trouvée entre deux espèces u et v appartenant respectivement aux classes i et j :

$$c_{ij} \;=\; max\,(u \in i\,;\; v \in j)\; s_{uv} \tag{44}$$

5.2.3 Comparaison de deux classifications

Nous pouvons reprendre la même démarche pour classer les molécules selon une seule composante, leur activité pharmacologique. C'est d'ailleurs une tâche que nous avons déjà menée dans le paragraphe 5.1.1 pour établir les tableaux de contingence. La différence réside maintenant dans le fait que nous devons comparer, non plus des valeurs numériques (les activités aux descripteurs géométriques) mais des classifications exprimées sous une forme matricielle. De surcroît, nous devons tenir compte du nombre de composantes (ici neuf puis une) et du nombre de classes (de 2 à 5 généralement) qui peuvent très bien différer dans les deux cas.

Le critère le moins subjectif pour estimer la "distance" entre deux matrices de classement, est l'entropie. Nous l'avons déjà utilisée au paragraphe 5.1.2 à propos de molécules distribuées selon deux grandeurs numériques. Elle peut être étendue à des espèces dont nous connaissons la matrice de similarité :

$$h([S]) \;=\; -\sum_{i,j} s_{ij}\,\ln\,(s_{ij}) \;-\; \sum_{i,j}(1 - s_{ij})\,\ln\,(1 - s_{ij}) \tag{45}$$

[1] Plus rigoureusement, nous devrions diviser par $N - 1$ et non par N

Il apparaît là aussi que la masse d'information "contenue" dans une matrice de similarité, est maximale lorsque les indices de ressemblance sont tous égaux à 0.5 et minimale lorsqu'ils sont tous de 0 ou 1. Selon un raisonnement très analogue à celui du paragraphe 5.1.2, nous pouvons déduire *la distance entre deux matrices de similarité* A et F comme étant fonction du degré de redondance entre le contenu de chacune :

$$D(A-F) = -\sum_{i,j} a_{ij} \ln\left(\frac{a_{ij}}{f_{ij}}\right) - \sum_{i,j}(1-a_{ij}) \ln\left(\frac{1-a_{ij}}{1-f_{ij}}\right) \qquad (46)$$

Cette distance s'annule lorsque les deux classifications sont identiques.

5.2.4 Potentiel apparent

Nous avons donc l'intention de faire varier le poids de chaque zone de l'espace de manière à modifier la matrice de similarité F puis à minimiser la distance $D(A-F)$ qui la sépare du classement selon l'activité. Ainsi nous identifierons les régions de l'espace où le potentiel semble avoir le plus d'importance dans l'interaction site actif - site récepteur. Concrètement, nous modifierons les coefficients w_k de l'équation (43) tout en respectant la condition (40).

Revenons à présent sur la définition du terme $F_i(k)$ de l'équation (41) et donc sur la *forme* F de la molécule i dans la zone k. Nous pouvons exprimer ce terme comme dépendant du potentiel qui règne dans la région mais ne le mesurer qu'aux seuls points qui appartiennent aussi à la surface moléculaire. De plus, ces derniers n'ont pas tous la même importance sur cette enveloppe. Nous pouvons tenir compte de :

1. l'aire de la surface locale autour de chaque point. En effet, l'algorithme qui les calcule associe à chacun une petite facette. Plus elle est étendue, plus le point est "représentatif".

2. les rayons de courbure de chaque facette. Rappelons qu'il existe trois catégories de points, convexes, cols ou concaves, qui ne sont pas aussi facilement accessibles pour un récepteur situé "en face".

3. éventuellement, nous pourrions tenir compte de l'orientation du vecteur normal à la facette par rapport à une direction particulière de l'espace elle-même dirigée vers le récepteur. Cette variante supposerait des essais, donc des calculs itératifs puisque la topologie de ce dernier est inconnue.

Ceci nous amène à l'expression de la forme moléculaire i dans la zone k :

$$F_i(k) = \frac{1}{N_p} \sum_{p \in k} ar(p)\, co(p)\, \cos(\vec{dr}, \vec{t_p}) \sum_a V_a(p) \qquad (47)$$

dans laquelle p désigne un point de la zone, $ar(p)$ et $co(p)$, deux fonctions de l'aire et de la concavité de sa facette, $\vec{t_p}$ et $\vec{dr}$ étant respectivement le vecteur normal et le

vecteur directeur. Enfin, $V_a(p)$ est le potentiel exercé localement par l'atome a (noyau et intégrales biélectroniques monocentriques).

Comparativement aux autres descripteurs moléculaires, le terme F pondère le potentiel électrostatique par des considérations topologiques assez fines de la région. Il s'agit en quelque sorte d'un *potentiel apparent* pour un hypothétique récepteur.

* * *

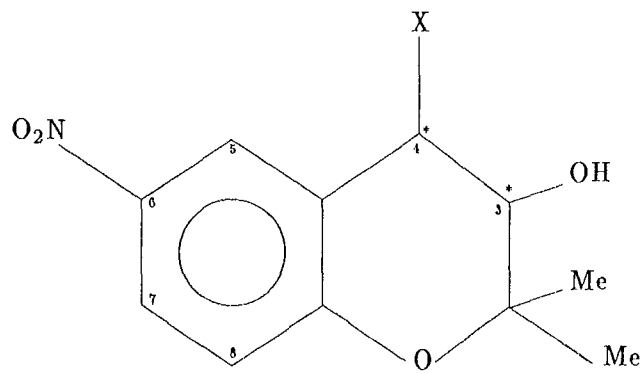
6 Applications

Nous évoquons maintenant quelques applications du logiciel OSCAR à l'étude d'une famille de composés analogues. Notre idée est de reprendre pas à pas des résultats publiés par ailleurs et d'examiner de quelle manière notre programme peut les affiner, les confirmer ou peut-être même les contredire.

6.1 Les benzopyranes

A l'occasion de l'étude de benzopyranes, des auteurs ont constaté que des dérivés substitués du 3,4-dihydro 2,2-diméthyle trans-4-(isopropylamino) 6-nitro 2*H*-1-benzopyrane 3-ol avaient une activité d'*anti-hypertenseur* à condition que les substituants des carbones 3 et 4 soient disposés en trans l'un par rapport à l'autre (figure 15) [58].

Figure 15: Benzopyranes substitués. Première série



Ces composés ne diffèrent que par le substituant X qui, dans un premier temps, est soit une chaîne aliphatique, soit un cycle $(CH_2)_n$. Ils présentent donc l'avantage de poser peu de problèmes de conformations, ce qui réduit les difficultés de leur étude. Nous nous proposons de chercher les raisons de la différence entre leur activité respective.

— Tests pharmacologiques —

Chaque composé est administré à plusieurs rats rendus préalablement hypertendus [58]. Après absorption, la pression systolique est mesurée à intervalles réguliers pour

définir ainsi l'activité comme le rapport entre la chute maximale de la pression sanguine et le nombre de millimoles de composé (tenant compte lui-même du poids de chaque cobaye). Le tableau 12 regroupe ces résultats.

Tableau 12: Activité des benzopyranes. Composés aliphatiques et cyclisés

composé	substituant X	dose en mg / kg pondéral	chute max. de la pression sanguine	barre d'erreur (mm Hg)
1	$NHCHMe_2$	100	36	4
		300	69	14
2	NH_2	30	23	1
		100	45	5
3	NHMe	30	21	8
		100	43	10
4	NMe_2	100	29	9
5	NEt_2	30	34	7
		100	54	3
6	$NHCMe_3$	100	7	3
7	NHcPr	10	31	8
		100	95	9
8	$NH(CH_2)cPr$	100	13	6
9	$NH(CH_2)_2OH$	100	27	7
10	$NH(CH_2)_3OH$	100	28	2
11	$NH(CH_2)_3Cl$	0.3	3	14
		1	41	11
		3	85	13
		10	81	12
12	$NH(CH_2)_2cNC_4H_8$	100	23	6
13	cNC_4H_8	1	52	7
		3	79	3
		10	98	6
14	cNC_5H_{10}	1	28	10
		3	73	13
		10	105	15
15	cNC_6H_{12}	10	17	2
		100	96	8
16	cNC_7H_{14}	3	34	10
		10	53	23
		30	82	15
17	bipipéridine	10	20	10
18	$cN(CHMeCH_2)_2$	30	26	9
19	$cN(C_2H_4)_2CHMe$	10	28	8
20	$cN(C_2H_4)_2O$	10	23	4
		30	55	7
		100	70	6
21	$cN(C_2H_4)_2N-\varphi$	10	10	11

Les valeurs de l'activité y sont données avec des barres d'erreur parfois très importantes en dépit du nombre de rats utilisés et de mesures faites. Pour certains composés, ces dernières ont pu être relevées pour plusieurs doses inférieures à 30 mg. Nous avons donc vérifié si l'activité variait de manière identique d'un composé à l'autre en fonction du dosage (figure 16), ce qui semble être le cas [1]. Un tel test est indispensable pour savoir si la comparaison des valeurs numériques est possible. De plus, il permet d'introduire une activité "moyenne", plus facile à manipuler dans la plupart des calculs. Dans la pratique, OSCAR permet toujours de travailler soit avec cette moyenne, soit à dose équivalente.

Compte tenu de la différence d'ordres de grandeurs entre les valeurs du tableau 12, nous utiliserons plus souvent le logarithme, népérien ou décimal, de l'activité. De plus, l'importance des barres d'erreurs nous incitera à faire abstraction dans les autres graphes des faibles écarts qui peuvent séparer certains composés.

— Premiers constats —

Au début du tableau 12, le substituant X est linéaire, plus ou moins long et plus ou moins symétrique. Les activités sont faibles et peu différentes les unes des autres, même dans le cas de **2**. La présence d'un groupement cyclopropane ne change apparemment rien, contrairement à des substituants OH ou Cl à l'extrémité de la chaîne.

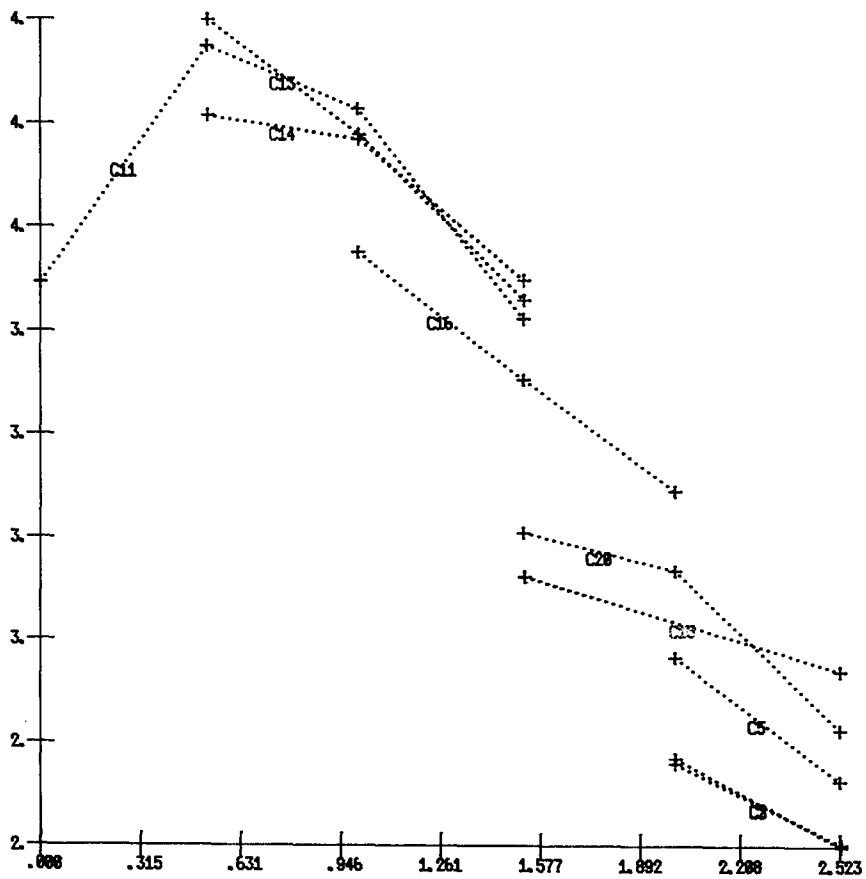
Ces derniers composés sont délicats à considérer. **7** et **8** sont certainement peu stables et donc susceptibles d'intervenir dans des réactions différentes de celles des autres espèces. **9**, **10** et **11** soulèvent deux problèmes. Le premier est celui de leur conformation à plusieurs degrés de liberté et qui rendra un peu arbitraire la géométrie optimisée que nous leur assignerons. Le second tient à leur lipophilie qui risque de les distinguer, là encore, des autres molécules.

La seconde partie du tableau contient des composés dont le substituant X est cyclique, incluant un nombre croissant de (CH_2). L'activité est plus grande dans le cas de **13** et **14** que pour les suivants. Les variations sont alors plus faibles, ces espèces étant toutes plus actives que les premières du tableau.

— Grandeurs globales —

Avant de nous focaliser sur la taille du substituant X, nous avons jugé utile de soumettre chaque molécule à des calculs de grandeurs d'ordre général. D'un point de vue thermodynamique, l'énergie totale et la chaleur de formation confirment la plus grande stabilité des composés cyclisés par rapport aux aliphatiques (de -67000 à -96000

[1] la première valeur du composé **11** semble aberrante.



Activite mesuree pour differentes doses. Premiere serie.

kcal / mol). Comme prévu, **7** et **8** sont les seuls à avoir une chaleur de formation positive. **13** et **14** sont moins stables que les autres cyclisés, de même que **11**.

Le moment dipolaire varie de 3.8 à près de 7 debyes, le maximum étant atteint pour les plus gros substituants cycliques. Les autres descripteurs beaucoup plus liés à la structure électronique - les polarisabilités ou les délocalisabilités - confirment encore une nette séparation entre les premières et les dernières espèces.

Nous avons enfin calculé le volume et la surface moléculaire qui varient entre 209 et 324 $\AA^3$ d'une part, et entre 222 et 328 $\AA^2$ d'autre part. La forme - c'est-à-dire l'indice d'écart à la sphéricité du volume moléculaire - varie entre 0.23 et 0.36, montrant ainsi qu'un gros substituant rend le benzopyrane plus compact, davantage qu'il ne le "déforme".

6.1.1 Encombrement stérique

Si la plupart des grandeurs précédentes distinguent nettement les composés aliphatiques des composés cyclisés, il faut aussi comprendre pourquoi **13** et **14** sont bien plus actifs que leurs proches voisins. Visiblement, il ne peut s'agir que de l'encombrement stérique, en dépit d'une grande similitude de leurs formes.

Nous avons demandé au logiciel OSCAR de superposer ces molécules et de leur imposer un même système de coordonnées défini dans **17** par les atomes 11, 29 et 30. Nous avons mesuré l'éloignement des contours de chaque composé dans plusieurs directions divergentes à partir de cette origine. Les graphes obtenus n'ont pas montré de relation significative avec l'activité, donc de conditions intéressantes comme par exemple, "l'activité exige un faible encombrement stérique dans cette région" ou au contraire "la forme de chaque composé actif présente une protubérance dans cette direction".

Mais contrairement à ce que le dessin sur papier permettrait de le supposer, la géométrie optimisée est bien différente d'un composé à l'autre, y compris pour **13** et **14**. Il est donc prudent de déterminer les conformères possibles, c'est-à-dire de calculer la variation de l'énergie totale ΔE_t de chaque composé en fonction de l'angle de rotation α de X. Si cette variation est faible, le composé aura toute latitude pour orienter son substituant de manière adéquate. Inversement, s'il existe une "barrière" énergétique, cette condition ne pourra être remplie. Le tableau 13 mentionne ces résultats pour les substituants les plus caractéristiques.

La lecture de ce tableau permet de confirmer la facilité du substituant à tourner dans les composés **13** et **14**. En particulier, il n'y a qu'un seul petit domaine de variation de l'angle dièdre qui soit "interdit" comme conformation (revoir le tableau 6 page 38 à propos de la répartition statistique des molécules). Par comparaison, les derniers

Tableau 13: Stabilités des conformères des benzopyranes

13		**14**		**15**		**16**		**18**	
ΔE_t kcal/m	α degrés	ΔE_t kcal/m	α degrés	ΔE_t kcal/m	α degrés	ΔE_t kcal/m	α degrés	ΔE_t kcal/m	α degrés
0.	98	1.5	85	0.3	89	3.5	86	3.0	83
5.7	52	0.	42	8.4	41	2.8	46	3.9	49
6.6	5	0.1	-6	6.1	-2	3.0	-2	3.2	3
1.1	-39	0.2	-54	1.5	-49	3.7	-55	0.3	-50
1.1	-82	6.4	-101	0.	-91	2.4	-98	0.	-93
2.6	-125	3.3	-148	2.8	-133	13.0	-150	0.4	-137
1.3	-165	0.3	170	1.7	-174	4.5	167	4.2	-176
2.3	149	1.9	125	1.9	138	0.	128	3.5	134

composés sont pratiquement figés dans une géométrie rigide.

Nous avons donc repris le calcul des éloignements en ne considérant pas seulement la conformation la plus stable d'un composé mais toutes celles susceptibles d'être adoptées par une proportion minimale de ses molécules. Les directions préférentielles étaient les mêmes que précédemment, tracées à partir du carbone 11 dans le composé bipipéridine **17**. Le tableau 14 regroupe les distances obtenues, les carbones de **17** qui définissent chaque direction étant rappelés.

En lisant ce tableau par colonne, il apparaît que pour chaque direction de calcul, il existe toujours un conformère de **13**, stable, dont la surface de contact est particulièrement proche du cycle benzopyrane et de l'atome d'azote. Il en est de même pour **14** mais uniquement pour les trois directions tracées vers les atomes du substituant les plus proches du cycle aromatique. Ces résultats montrent donc qu'il est possible de trouver un critère géométrique assez parlant qui singularise les deux composés les plus actifs par rapport aux suivants et même qui peut confirmer la plus forte activité de **13** par rapport à **14**.

Quels sont les avantages et les limites de ce genre de calculs ? Nous avons déjà indiqué que ces paramètres étaient fiables et très rapides à obtenir. De plus, il n'est pas coûteux (en temps) d'affiner les résultats en augmentant le nombre de conformères et celui des directions [1]. Autrement dit, il est possible de connaître avec précision l'encombrement de nos composés et de dresser ainsi une sorte de *carte*, sur laquelle nous reporterons des zones de l'espace qui doivent ou ne doivent pas être occupées par le composé idéal. Nous reviendrons sur ce point ultérieurement.

Inversement, la restriction majeure de cette méthode est qu'elle ne prouve en rien

[1] En fait, les tableaux qui précèdent sont un résumé d'une étude plus complète, qui fournit bien plus de résultats numériques...

Tableau 14: Distances cycle benzopyrane - surface moléculaire (angströms)

composé	α	azote		pont près de OH			pont près de φ			
	degrés	29	30	32	36	35	33	37	34	31
13	98	3.31	4.37	2.29	2.42	2.23	4.22	5.26	4.12	4.42
	-39	4.27	5.42	4.80	5.33	4.06	5.59	3.37	2.91	4.03
	-165	3.23	2.66	3.92	4.15	4.75	2.88	4.07	5.62	5.07
14	42	6.29	2.66	2.30	4.96	4.89	5.39	5.30	5.17	4.34
	-6	6.34	2.96	2.30	2.42	4.81	4.85	5.51	5.00	5.76
	-54	6.40	4.77	3.40	2.42	2.30	5.58	5.19	5.54	5.36
	170	6.48	5.48	3.27	5.22	5.00	5.55	4.71	5.37	4.30
15	-91	5.79	5.82	3.94	3.04	4.12	6.22	5.63	5.52	5.40
	-174	6.26	4.91	3.12	4.87	4.93	6.02	5.70	4.77	5.18
	89	6.25	5.43	2.29	4.81	2.23	5.78	5.33	5.58	5.71
16	128	3.61	4.27	5.24	6.11	6.49	4.54	2.89	4.45	4.34
17		4.86	5.01	5.42	5.67	5.29	5.48	6.20	5.46	5.20
18	-50	3.35	5.71	4.79	5.44	4.57	4.79	5.22	4.04	4.44
	-93	5.05	4.76	4.70	5.11	4.96	4.97	2.89	5.11	5.02
	-137	5.03	5.19	5.48	4.65	5.16	4.85	4.87	5.42	5.41

son utilité... S'il est certain que la taille du substituant est importante (sinon comment expliquer que **14** soit plus actif que **15** ?), il est possible qu'elle intervienne de manière beaucoup moins directe.

6.1.2 Surfaces de contact locales

Imaginons que l'activité provienne, non pas d'une taille moindre du substituant (**13**) ou d'une conformation relativement proche du cycle aromatique (**14**), mais des caractéristiques propres aux hétéro-atomes de la région. Pourtant, un rapide calcul de charges montre que rien ne distingue les composés de ce point de vue. Les carbones du substituant sont tous légèrement électronégatifs ce qui est un résultat attendu en série aliphatique. Les charges des atomes de carbone, d'azote et d'oxygène voisins ne varient pas davantage.

Ces derniers atomes jouent cependant un rôle puisqu'une condition nécessaire à l'activité est qu'ils soient placés de part et d'autre du plan moyen du cycle pyrane. Le point clé doit donc résider dans leur *accessibilité*, exprimée comme étant la portion de la surface moléculaire directement attenante à chacun. Le tableau 15 regroupe les valeurs obtenues pour ces hétéro-atomes.

Tableau 15: Accessibilités locales ($\AA^2$)

composé	α	substituant	azote	hydrogène	OH	cis	trans
	degrés	X	29	19		$\overbrace{n_{29} + h}$	$\overbrace{h_{19} + oh}$
13	98	87.1	2.1	7.4	20.3	13.2	27.7
	-39	87.9	1.8	10.4	19.4	-	29.8
	-165	84.8	-	9.5	20.9	-	30.4
14	42	100.5	-	8.4	20.0	-	28.4
	-6	102.4	-	7.9	19.0	-	26.9
	-54	100.0	1.8	9.2	20.4	11.3	29.6
	170	99.6	0.5	10.1	19.9	-	30.0
15	-91	117.7	-	8.0	19.6	-	27.6
	-174	108.4	-	9.4	20.6	7.4	30.0
	89	122.2	-	8.4	19.9	-	28.3
16	128	127.5	-	6.9	17.4	-	27.3
17		131.7	-	8.2	17.5	-	25.7
18	-50	111.6	-	8.4	17.7	-	26.1
	-93	111.1	-	5.9	16.7	-	22.6
	-137	111.6	2.0	4.9	19.7	-	24.6

Rappelons brièvement les modalités du calcul. Nous avons demandé à OSCAR de considérer la liste exhaustive des voisins de l'atome d'azote sauf du côté du carbone 11. Ces deux derniers étant définis comme atomes correspondants, ceci suffit pour que le logiciel identifie dans chaque composé tous les atomes du substituant sans qu'eux-mêmes soient nécessairement correspondants. La surface moléculaire au sens de Connolly est calculée dans la foulée puis ses points sont triés pour ne retenir que ceux dont l'atome le plus proche appartient à X. Les valeurs du tableau 15 correspondent à la sommation des éléments de surface dont chaque point est représentatif.

La surface globale varie de 87 $\AA^2$ pour **13** à plus de 130 $\AA^2$ pour le bipipéridine. Les variations sont peu sensibles d'un conformère à l'autre (à l'exception de 15) mais importantes entre les composés. Ces résultats sont donc significatifs et seraient déterminants si la forme du récepteur devenait mieux approchée par ailleurs. Dans ce cas, nous aurions un instrument de comparaison portant sur une échelle de quelques dizaines d'angstroms donc moins ciblé que la palette des descripteurs dont nous disposons déjà.

La surface mesurée au niveau de l'azote est nulle pour la plupart des composés, ce qui signifie que l'encombrement stérique est tel qu'aucun point de la surface n'est plus proche de cet atome que d'un autre. Là encore, **13** et **14** se singularisent par une accessibilité appréciable, avec la même hiérarchie entre eux. Nous avons donc là une seconde indication qui explique leur activité respective. **18** est aussi accessible mais nous avons vu qu'il était figé dans une conformation précise, manifestement défavorable.

Par comparaison, nous avons mesuré l'accessibilité de l'hydrogène quaternaire proche de l'azote. Les valeurs sont beaucoup plus élevées, avec toujours la même tendance favorable aux composés actifs. Elles confirment surtout que "la voie d'accès" à l'atome d'azote ne peut se situer que de ce côté. Les dernières colonnes permettent de comparer l'accessibilité de part et d'autre du plan moyen du cycle furane. Elles confirment par le calcul que le substituant alcool doit impérativement être le plus éloigné de l'azote pour en dégager l'accès.

6.2 Substitutions aromatiques

La suite de l'étude des benzopyranes consistait à retenir les deux composés les plus actifs, **13** et **14**, et à essayer différents substituants du premier cycle aromatique. Une trentaine de composés ont donc été synthétisés avant d'être testés dans les mêmes conditions que précédemment. Les tableaux 16 et 17 en donnent la liste en rappelant en sixième colonne le nombre de groupes $(CH_2)_n$ du substituant X.

Il existait a priori un troisième candidat possible, **20**, dont les caractéristiques géométriques sont très proches de **14** et qui peut facilement tourner autour de l'axe C11 - X. L'unique différence réside dans l'atome d'oxygène situé à son extrémité mais qui n'est probablement guère plus actif dans sa position qu'un groupe méthylène. A priori, il n'était donc pas évident que l'activité de ce composé soit beaucoup plus faible que celle des deux autres. OSCAR nous en a fourni une explication immédiate lors des calculs précédents. En effet, nous avons constaté que **20** était le seul composé à présenter un potentiel électrique de signe négatif à l'extrémité de X. Nous en déduisons donc - en première approximation - que le récepteur doit présenter à nos molécules une région de même potentiel.

Avant d'entreprendre une étude systématique de ces nouveaux composés, nous nous sommes encore assuré que l'activité variait de la même manière selon le dosage. Seules les valeurs de **39** et **40** semblent sujettes à caution (figure 17).

Tous les composés de cette seconde série sont moins actifs que **13** et **14** mais davantage que les premières espèces cyclisées. Nous ne relevons que trois exceptions ; la première lorsque le groupe nitro est décalé d'une position sur le cycle aromatique (**25**), les autres lorsqu'il est remplacé par un groupe cyano (**28** et **27**). Les positions 5 et 8 (se reporter à la figure 15 de la page 72), plus proches du cycle pyrane, sont peu favorables même avec ces substituants (**23** et **26** puis **30**). Cette forte sensibilité de l'activité au site de substitution n'apparaît pas évidente. Ainsi **24**, homologue de **25** mais dérivé de **14** et non de **13** est très peu actif, et même le remplacement du groupe nitro par un cyano ne change rien (**29**).

Tableau 16: Benzopyranes (seconde série). Composés 22 à 34

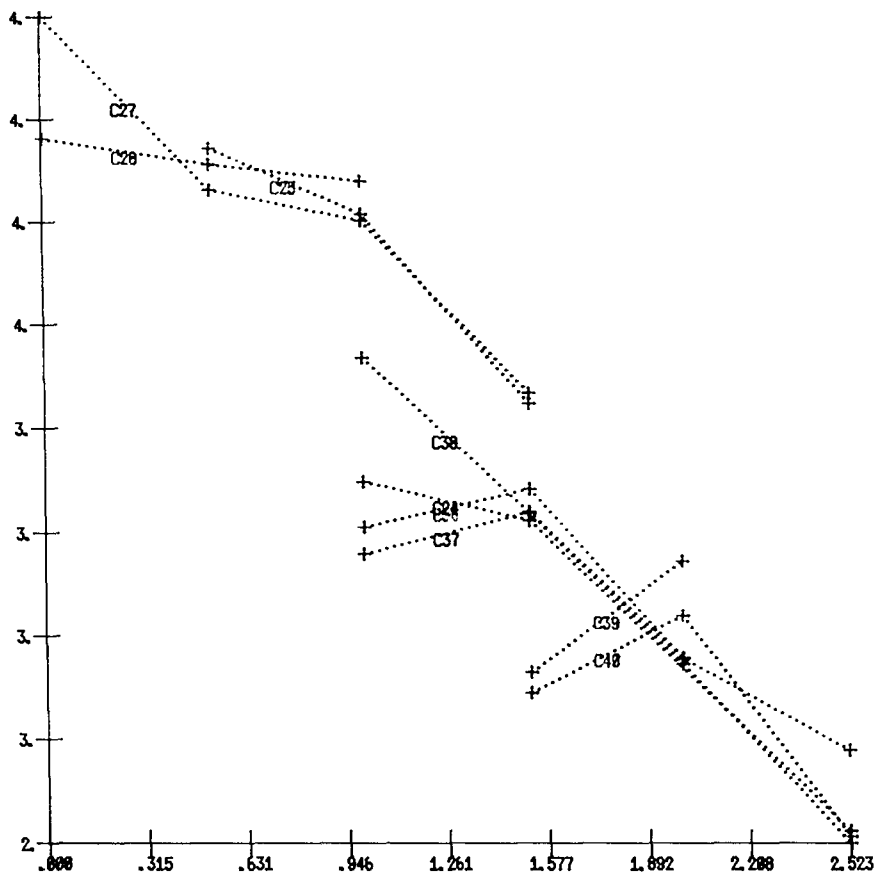
composé	R_5	R_6	R_7	R_8	n	dose en mg / kg pondéral	variation de la pression sanguine	
22					4	30	10	6
						100	46	17
23	NO_2				5	100	12	9
24			NO_2		5	3	18	13
						10	48	8
						100	79	13
25			NO_2		4	1	44	9
						3	90	11
						10	99	4
26	CN				5	10	13	4
						100	33	5
27		CN			5	0.3	29	6
						1	35	7
						3	88	8
						10	107	3
28		CN			4	0.3	15	4
						1	43	7
						3	117	2
29			CN		5	30	7	3
						100	33	16
30				CN	5	10	5	16
						100	33	16
31		Cl			5	30	7	2
						100	69	5
32		Me			5	10	34	24
						30	44	11
						100	43	5
33		Me			4	100	9	5
34		OMe			4	30	29	6
						100	17	10

Il apparaît visiblement une sorte de "couplage" entre le substituant du cycle aromatique et la conformation du cycle X, résultat qui ne s'impose pas d'emblée puisque les deux sites sont tout de même assez éloignés l'un de l'autre (plus de 5 angströms dans le composé le plus actif). Les dernières molécules seront utiles pour le préciser puisqu'il s'agit essentiellement de substitutions variées en position 6, sans conséquences notables sur l'activité.

Tableau 17: Benzopyranes (seconde série). Composés 35 à 47

composé	R_5	R_6	R_7	R_8	n	dose en mg / kg pondéral	variation de la pression sanguine	
35		F			5	10	16	3
						100	26	1
36		COOMe			4	3	14	5
						10	58	5
						30	65	6
37		COMe			5	3	12	8
						10	51	7
						30	65	15
						100	127	3
38		COMe			4	3	39	6
						10	53	13
						30	66	13
						100	78	10
39		CH(OH)Me			5	10	20	5
						30	115	11
40		C(NOH)Me			5	10	17	4
						30	80	19
						100	74	5
41		$CONH_2$			5	10	9	4
						100	30	6
42		$CONH_2$			4	100	21	8
43		NHCOMe			5	10	6	8
						100	14	6
44			NO_2	Me	5	100	39	2
45	φ	φ			5	100	4	4
46			φ	φ	4	100	4	5
47		COOH			4	10	2	5
48		NH_2			4	100	6	8

Du point de vue des grandeurs portant sur l'ensemble de chaque molécule, aucune tendance ne ressort clairement. Qu'il s'agisse de stabilité thermodynamique, de lipophilie, de forme géométrique ou de caractéristiques électroniques, rien ne distingue les composés selon l'activité. Nous ne parlerons donc que de critères locaux.



Activite mesuree pour differentes doses. Seconde serie.

6.2.1 Charges locales

En examinant les charges totales de chaque carbone du cycle aromatique, nous avons constaté des variations assez prévisibles en fonction des substituants (tableau 18) mais qui ne se traduisent pas par une distorsion significative sur l'ensemble du cycle [1].

Tableau 18: Charges des carbones du cycle aromatique

position	minimum		maximum	
10	5.83	**13** et **14**	5.95	**44**
5	5.97	**44**	6.17	**13** et **14**
6	5.95	**29**	6.39	**44**
7	5.94	**34**	6.45	**13** et **14**
8	5.91	**26**	6.32	**23**

Des observations similaires peuvent être faites pour les substituants. Dans le cas d'un groupe nitro, il apparaît évidemment un report de charge de l'azote sur les deux oxygènes. Mais globalement, l'effet mésomère donneur l'emporte de peu sur l'effet inductif attracteur, et le carbone du cycle aromatique reste lui-même assez négatif (comme pour - Cl par opposition à - F). C'est le cas inverse pour les groupes cyano qui, dans nos composés, sont les seuls à toujours présenter un excès de charge d'environ 0.15. Ceci peut donc expliquer un comportement assez différent en dépit d'une électronégativité a priori voisine.

Les autres descripteurs plus théoriques que sont les autopolarisabilités atomiques et les délocalisabilités électrophiles ou nucléophiles n'apportent pas d'informations plus pertinentes. Mais nous avons constaté qu'ils singularisaient presque systématiquement les composés **33**, **34** et **48** substitués en position 6 respectivement par Me, OMe et NH_2. De même avons nous noté l'auto-polarisabilité exceptionnelle du composé le plus actif, **28**, en position 6. Nous n'avons pas trouvé d'explication satisfaisante à ce résultat. Peut-être ne s'agit-il que d'un (curieux) artefact...

6.2.2 Formes de potentiel

Les remarques précédentes nous incitent à nous focaliser sur la région située entre le cycle aromatique et le substituant X, notamment pour expliquer l'activité très différente de **24** et **25**. Nous avons préalablement vérifié que les données du tableau 13 sur la prépondérance de certains conformères par rapport à d'autres restaient valables. Le

[1] les valeurs obtenues dépendent beaucoup de la méthode semi-empirique, en particulier les halogènes. Toutes proviennent ici d'un calcul PM3

seul changement de **13** à **25** est en effet le déplacement d'un substituant éloigné de plusieurs angströms.

Si **25** est actif tandis que **24** ne l'est pas, cela est dû à l'existence d'un domaine de variation de l'angle dièdre cycle X - structure benzopyrane, pour lequel le premier a un conformère stable, tandis que le second n'en a pas. Le tableau 13 nous indique qu'il s'agit des angles compris entre -70 et -130 degrés. Nous avons donc sélectionné deux conformères de **24** (géométries optimisées à -82 et -111 degrés) afin d'y repérer des caractéristiques originales que le second composé ne pourrait pas offrir.

Nous n'avons découvert aucun critère de nature purement géométrique susceptible d'apporter une réponse. La mesure des distances entre atomes correspondants, celle du volume de la zone concave comprise entre le cycle X et les sites de substitution ou encore le calcul des éloignements montrent souvent des résultats intéressants sur quelques composés (jusqu'à une dizaine) mais aussi suffisamment d'exceptions pour mettre en doute la validité de ces descripteurs.

Les surfaces isopotentielles apportent en revanche une information d'autant plus intéressante que leur allure est largement imprévisible, contrairement aux courbes d'isodensité électronique. Les figures 5 et 6 mettent ainsi en évidence des contours asymétriques et très changeants. L'inconvénient est qu'une étude systématique sur un grand nombre de composés demande beaucoup de temps, d'autant qu'il faut réitérer les calculs sur plusieurs plans de référence (paragraphe 4.5.1) afin de ne rien "oublier". Après de multiples essais, la conclusion est qu'il est fréquent d'observer des débuts de tendances, c'est-à-dire des formes relativement spécifiques aux composés les plus actifs, y compris dans le cas de **24** et **25**. Mais il n'en reste pas moins vrai que leurs fluctuations ne permettraient pas d'utiliser ces données pour concevoir un nouveau composé actif en conséquence.

6.2.3 Traitement par classifications

Déçu de ne pouvoir trouver un bon critère portant sur la similitude de ces surfaces, nous avons décidé de lancer des calculs de reconnaissance de formes plus systématiques (et plus longs) selon les modalités exposées au paragraphe 5.2.

Nous avons pris **22** comme composé de référence et subdivisé la surface moléculaire en cinq régions distinctes, situées entre les cycles aromatique et pyrolidinique. Les quatre premières sont proches des sites de substitution tandis que la dernière encercle l'atome d'azote et ses deux voisins immédiats sur X. Chaque zone est délimitée grâce à une distance limite entre chaque point qu'elle peut contenir et un (ou plusieurs) atome(s) correspondant(s). Ces derniers permettent donc de déduire dans les autres molécules

les zones correspondantes : toutes sont adjacentes mais bien distinctes [1]. En chaque point d'une zone, OSCAR calcule alors le potentiel en le pondérant par l'accessibilité locale, c'est-à-dire par la concavité et l'aire de l'élément de surface.

Pour chaque molécule, nous obtenons finalement cinq valeurs d'un potentiel "moyen" qui tient compte de l'accessibilité de la zone tout en étant peu sensible aux rotations des groupes substituants, NH_2 ou COOMe par exemple. Ce sont donc de "bonnes" valeurs du potentiel exercé localement.

Ces valeurs numériques sont autant de composantes indépendantes les unes des autres et qui peuvent être comparées entre molécules pour construire une matrice de similarité. Cette dernière sert à son tour à une classification des molécules en 2, 3, 4 ou 5 ensembles, comme pour l'activité. Pour clarifier ce schéma détaillé au paragraphe 5.2, nous donnons ici un exemple de traitement mené avec les zones définies précédemment.

Le premier tableau est un simple classement des molécules selon l'activité ou plus exactement selon son logarithme décimal. Les numéros supérieurs à 50 sont ceux de composés supplémentaires [59] que nous avons inclus dans le calcul afin d'augmenter le nombre d'espèces par rapport à celui des propriétés (ici 52 sur 5). Le second tableau est la classification analogue selon le potentiel. Il est suivi de la matrice de classes dont chaque élément contient le plus fort indice de ressemblance qui puisse être trouvé entre deux molécules appartenant à deux classes différentes. Elle montre surtout la qualité de la classification [2] dans la mesure où les dissemblances sont généralement assez fortes (jusqu'à 0.67) et en donne la hiérarchie : **40** et **41** sont plus proches de la plupart des composés que de **30** et **58**. Enfin, la distance entre matrices de similarité est une estimation de l'écart qui sépare les deux classifications. Elle n'est significative que dans des comparaisons.

Dans ce calcul, nous avons exclu deux composés, 7 et 8. Ils se démarquent tellement des autres par leurs valeurs de potentiel qu'ils constituent chacun leur propre classe (leur potentiel est constamment positif dans toutes les zones, y compris près du cycle X). La classification précédente est alors un peu faussée comme le confirme d'ailleurs une valeur beaucoup plus élevée pour la distance entre matrices de similarité (-871), ainsi que pour l'entropie de la matrice des classes. Nous avons déjà noté leur particularité structurale qui peut expliquer un tel comportement.

— Zones importantes —

Pour classer nos molécules, nous n'avons fait qu'une seule correction qui est celle de la variance des indices de similarité. Rappelons qu'elle consiste simplement à exagérer les dissemblances de potentiel fortes et locales par rapport à celles qui sont plus faibles

[1] deux régions peuvent donner l'impression de s'interpénétrer sur les photographies mais c'est à cause de la profondeur...

[2] l'entropie en est une "note"

Tableau 19: Classement de formes

en 5 ensembles selon l'activité

classe	composés
1	56
2	53, 55, 57, 58
3	17, 19, 20, 24, 32, 36, 39, 40
4	11, 13, 14, 16, 25, 27, 28, 37
	38, 51, 52, 54, 59
5	les autres (26)

en 5 ensembles selon le potentiel

classe	composés
1	40, 41
2	30, 58
3	25, 29, 39, 51, 52, 53, 54
4	23, 26, 43
5	les autres (38)

matrice de classes du potentiel

$$\begin{bmatrix} 1.00 & 0.67 & 0.67 & 0.78 & 0.90 \\ 0.67 & 1.00 & 0.79 & 0.76 & 0.81 \\ 0.67 & 0.79 & 1.00 & 0.83 & 0.92 \\ 0.78 & 0.76 & 0.83 & 1.00 & 0.82 \\ 0.90 & 0.81 & 0.92 & 0.82 & 1.00 \end{bmatrix}$$

entropie de la matrice de classes : 4.86

poids attribué à chaque zone : 0.25 chacune

distance entre matrices de similarité : -516

et plus étendues. Mais nous avons aussi la possibilité d'attribuer à chaque zone un poids différent pour modifier la matrice de similarité, et par conséquent sa ressemblance avec la matrice analogue obtenue avec l'activité. Partant de cinq zones d'un poids de 0.20 chacune, un calcul itératif pour optimiser ces poids nous a fourni le quintuplet (0.16 ; 0.23 ; 0.00 ; 0.30 ; 0.31) qui abaisse la distance entre matrices à -491.

Tableau 20: Benzopyranes (seconde série). Composés 51 à 60

composé	R_6	R_7	n	dose en mg / kg pondéral	variation de la pression sanguine	
51	NHCOMe	NO_2	4	1.0	25	4
				10.	99	3
52	NHCOMe	NO_2	5	1.0	44	6
				10.	89	13
53	NH_2	NO_2	4	0.1	22	8
				0.3	83	5
				1.0	71	3
54	NH_2	NO_2	5	0.3	42	10
				1.0	67	2
55	NO_2	NHCOMe	5	0.1	78	6
				0.3	108	6
56	NO_2	NH_2	4	0.1	122	7
57	NO_2	NH_2	5	0.1	110	4
				0.3	111	4
58	NC	NHCOMe	4	0.1	36	8
				0.3	149	5
59	H_2NCO	NHCOMe	5	0.3	34	6
				1.0	50	23
60	NC	NH_2	4	0.1	54	12
				0.3	115	20

Ce résultat conduit à plusieurs remarques. La première est qu'apparemment la valeur du potentiel dans la troisième zone, c'est-à-dire près de la position 6 de substitution n'est pas un des facteurs qui déterminent l'activité de nos composés. En effet, au fur et à mesure que OSCAR attribue à ce site un poids décroissant, le classement des molécules ressemble un petit peu plus à celui que donne l'activité. Ce résultat n'est pas si surprenant, d'une part parce qu'il confirme notre échec lors de l'observation des surfaces isopotentielles, d'autre part parce que beaucoup de nos composés (seconde partie du tableau 17) sont inactifs mais très différents dans le potentiel qu'ils exercent localement.

Les positions 8 et 7 ont une importance secondaire, sauf si elles font l'objet simultanément de substitutions. Ceci confirme les faibles différences d'activité entre les composés du tableau 17 et le fait que ceux du tableau 20 soient plus actifs qu'eux. Les deux dernières zones semblent être le plus susceptibles de distinguer les molécules. Pour la quatrième, nous ne disposons malheureusement que de deux composés, **23** et **26** et pour lesquels l'encombrement stérique devient prédominant (cycle pipéridinique). D'autres composés substitués au même endroit auraient été très utiles pour approfondir ce point.

6.2.4 A l'approche d'un récepteur...

Nous voudrions enfin donner un aperçu d'une autre possibilité offerte par le logiciel OSCAR et dont le principe est exposé au paragraphe 4.4.3. Il s'agit de comparer l'encombrement stérique de plusieurs composés en prenant successivement des directions de référence que l'on imagine orientées vers le récepteur. De cette manière, il est possible de regarder si au fur et à mesure de leur "progression" le long de cet axe, certaines molécules ne sont pas rapidement génées par l'encombrement d'un de leurs substituants.

Dans la pratique ce type d'étude présente quelques inconvénients. Comme il s'agit d'une observation très fine du volume réellement occupé par une molécule qui se déplace dans l'espace, il est difficile d'en comparer plus de 3 ou 4 simultanément (OSCAR permet d'en traiter 6). De plus, pour tenir compte des conformations possibles (imaginons par exemple **11** se déplaçant selon la direction C11 - X), le logiciel doit offrir une possibilité d'animation à l'écran. En bref, il s'agit d'un outil puissant, essentiellement limité par les capacités de l'utilisateur à appréhender des formes dans l'espace...

De telles observations sont illustrées par les trois photos qui suivent. Elles ont été prises dans une direction définie d'un côté par l'oxygène du cycle benzopyrane, de l'autre par le milieu du segment NO_2X. Nous imaginons donc que le récepteur se trouve dans cette direction et qu'il présente d'une part une zone de potentiel positif face à la fonction cyano ou nitro, d'autre part une partie concave face au substituant X. Pour alléger l'affichage, nous n'avons dessiné que les points de l'enveloppe moléculaire de deux composés, **13** et **17** en rappelant les numéros des atomes de la région (13, 27 et 28 sont les atomes du groupe NO_2dans les deux molécules, les autres sont les carbones ou les hydrogènes du cycle X). La première photographie est prise relativement "loin" du centre des deux molécules : elle indique qu'à cette distance, leur encombrement stérique diffère peu (points rouges et verts en nombre comparable et dans les mêmes régions). En revanche, les photographies suivantes confirment un encombrement stérique de plus en plus gênant pour le second composé du côté du substituant X.

Conclusions

A l'issue de ce travail, nous voudrions faire quelques observations :

La principale difficulté des études en QSAR reste le manque considérable d'informations sur les processus qui interviennent dans l'organisme humain, après administration d'un médicament. Dans la pratique, les seules données dont nous disposons sont des valeurs numériques incertaines, probablement dénaturées, et de toute manière, terriblement réductrices d'une succession complexe de processus chimiques. Quant à l'exploitation de ces données, notre logiciel n'offre pas de solution miracle : OSCAR "fait avec ce qu'il a", s'efforçant de proposer des outils simples et adéquats pour comparer l'activité respective de plusieurs molécules.

A ce sujet, le lecteur aura sans doute remarqué une certaine méfiance à l'égard des statistiques. Au cours de ce travail, nous avons été assez surpris du recours fréquent dans les publications, à un outil qui, par nature, est incapable d'expliquer les relations entre structure et activité de manière rationnelle ou méthodique. En définitive, cet instrument est-il réellement intéressant et combien de médicaments sont-ils issus de l'application d'une relation statistique établie auparavant sur une série de composés analogues ? A cette question volontairement provocatrice, nous pourrions d'ailleurs en ajouter une autre : n'y a-t-il pas dans bien des travaux, une légère tendance à faire dire aux statistiques beaucoup plus que ce qu'elles peuvent mathématiquement démontrer ?

Notre logiciel inclut des méthodes de classifications, inspirées de notions d'intelligence artificielle. Nous tenons à démystifier cette expression qui suscite souvent des regards sceptiques ou ironiques. C'est une question de crédibilité, car il faut bien avouer que bien des applications fondées sur ces techniques ont conduit à des échecs. Quoique assez arbitraires, les critères entropiques utilisés dans OSCAR, ne sont ni mauvais, ni trop subjectifs. Ils participent à ce souci de tirer le meilleur parti de données pharmacologiques insuffisantes, en conduisant à de bonnes classifications des molécules selon leurs caractéristiques structurales. Par exemple, étudier l'influence relative du potentiel électrique dans différentes régions de l'espace peut apporter une information utile ("ce n'est pas la peine de mettre tel substituant en tel site car cela ne changera rien à l'activité") et certainement originale par rapport aux autres logiciels de modélisation moléculaire.

La reconnaissance de formes est une autre de ces expressions à interpréter plus prosaïquement, ne serait-ce que parce que le mot "forme" peut concerner des notions très différentes. Avec OSCAR, nous nous sommes efforcés de considérer l'interaction réactif

- récepteur d'un point de vue original, simultanément électrique et topologique. Nous pensons que c'est la voie la plus sérieuse à explorer pour mieux comprendre les relations structure - activité, parce qu'elle mèle deux aspects que l'on traite souvent séparément. De ce point de vue, il reste là un travail considérable à mener. Une idée intéressante et très ambitieuse consisterait d'ailleurs à envisager de manière systématique des récepteurs possibles...

* * *

Le travail exposé dans les chapitres qui précèdent appelle d'autres remarques :

L'étude des benzopyranes montre clairement l'absence de méthode toute faite pour étudier précisément leur activité. L'examen de ces composés a souvent nécessité la modification du code source de notre programme pour y ajouter de nouvelles options ou développer de nouveaux outils adaptés aux questions qui se posaient. Dans un laboratoire comme le nôtre, riche de logiciels en modélisation moléculaire, nous avons souvent constaté que ces derniers n'avaient pas la souplesse nécessaire pour répondre à des questions ponctuelles, mêmes simples. A posteriori, nous pensons avoir eu raison de faire d'OSCAR non pas une boite noire qui pourrait susciter, à juste titre, une certaine méfiance, mais un outil de recherche particulièrement adapté aux besoins sans cesse renouvelés des pharmacologues.

Le choix initial de la famille des benzopyranes s'explique par le fait que cette cinquantaine de composés offrait un bon condensé des difficultés à résoudre en programmation : multiples conformères, activités différant parfois de plus d'un ordre de grandeur, problèmes d'encombrement stérique, délocalisation dans un cycle aromatique, substituants variés, etc. Lors de nos essais, nous avons également utilisé d'autres types de molécules en fonction de leurs caractéristiques les plus marquantes (comme les benzodiazépines pour leurs propriétés électroniques ou des espèces fortement ramifiées pour tester les algorithmes de parcours). De ce fait, les résultats que nous avons présentés au dernier chapitre laisseront probablement des pharmaciens sur leur faim, puisque nous cherchions beaucoup plus à tester la viabilité de nos descripteurs de la relation structure - activité, qu'à rechercher de nouveaux antihypertenseurs...

Du premier point de vue, nous pouvons conclure que parmi la masse d'informations que recèle une structure moléculaire, il est possible de mener un tri pour ne retenir qu'un petit nombre de grandeurs remarquables. Définis rigoureusement, ces descripteurs que sont par exemple l'aire d'une surface de contact, l'accessibilité d'un atome, l'éloignement de deux substituants, un excès de charge locale ou bien encore une différence de potentiel, sont suffisamment fiables pour qu'il soit raisonnable de chercher une explication de l'activité à partir de leur comportement. De plus, seules de telles grandeurs permettent d'imaginer comment se font l'approche et la reconnaissance mutuelle de deux espèces chimiques qui, en cas de succès, les conduiront à une réaction indispensable pour obtenir finalement un effet thérapeutique.

Ce travail nous amène enfin à une dernière réflexion. Trop de molécules sont laissées de côté dans les études de QSAR du fait de leur inactivité : il y a là une perte considérable d'informations qui sont pourtant parfaitement exploitables et peut-être même riches d'enseignements. Il nous semble donc que des progrès plus significatifs dans les relations structure - activité dépendront largement de la capacité des futurs logiciels de modélisation moléculaire à traiter et comparer simultanément un grand nombre de structures moléculaires, mêmes différentes.

Incluant des notions aussi variées que la chimie quantique, le graphisme moléculaire, les méthodes de classifications, les statistiques et la topologie des molécules, conçu tout particulièrement pour traiter simultanément un grand nombre de molécules - y compris en arrière-plan - OSCAR nous paraît donc une bonne base de départ pour des études beaucoup plus systématiques dans les relations structure - activité.

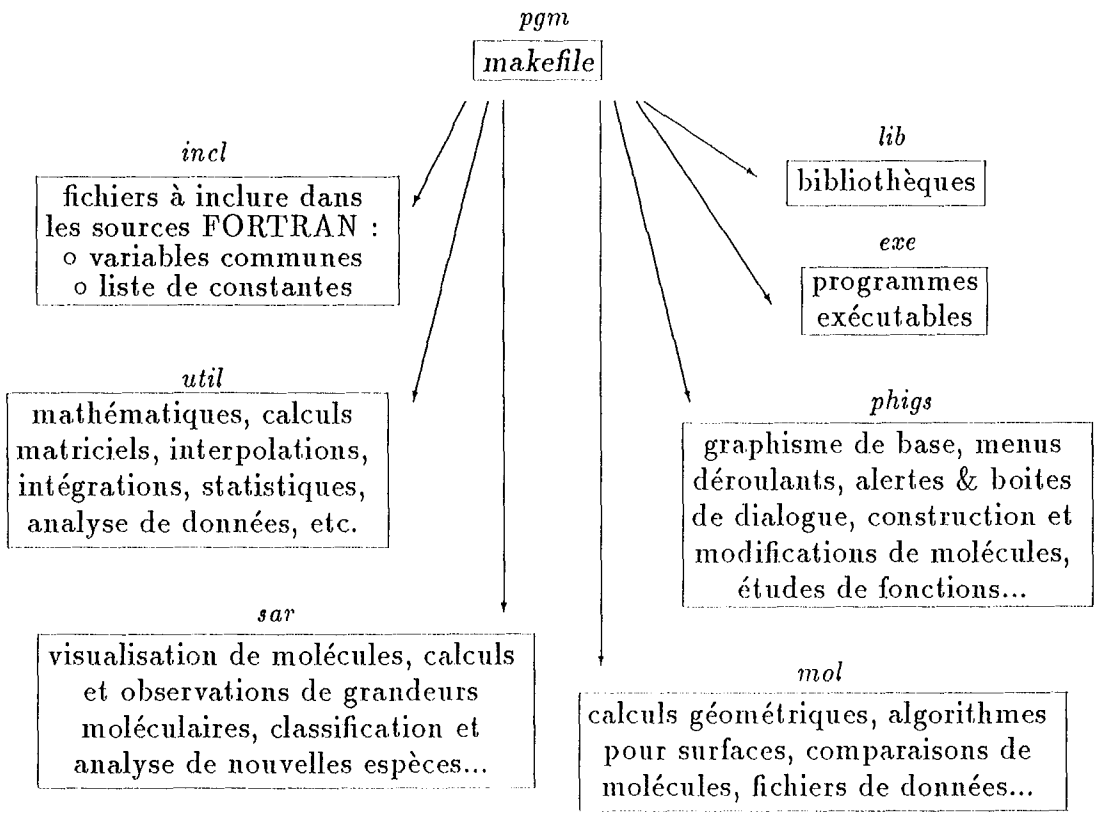
* * *

ANNEXES

A Programmation FORTRAN

Cette annexe est destinée aux programmeurs désireux de profiter de la partie informatique de notre travail. Toute la programmation écrite pour cette thèse est regroupée sous un répertoire identifié par la variable d'environnement shell $PGM. Les procédures sont ensuite réparties dans les *directories* ci-dessous :

Figure 18: Arborescence des algorithmes sous UNIX



Quelques conventions ont été adoptées lors de l'écriture des *routines* pour en faciliter ultérieurement la compréhension. On notera surtout que :

o les variables sont explicitement déclarées, que ce soit leur type (entières, réelles, etc.), leur allocation (statique ou dynamique) ou leur taille (en nombre d'octets). Les constantes débutent par la lettre k et inversement aucune variable ne débute par cette lettre,

o les variables communes sont consommées avec modération. Il n'en est fait usage que dans les programmes eux-mêmes et non dans les bibliothèques (à l'exception de *phigs.a*, où elles contiennent quelques données sur les stations de travail initialisées par les routines *gbsows* et *gbinws*). Elles sont regroupées par fichiers à inclure, ce qui évite les changements de noms intempestifs.

o de manière générale, on appréciera l'écriture du source, abondamment commenté et dont la présentation est digne d'un maniaque. Nous avons nous-même tellement perdu de temps à comprendre le source d'autres personnes pour faire attention à cet aspect sordide mais capital du travail.

Tableau 21: Taille du source FORTRAN

subdirectory	*util*	*mol*	*phigs*	*sar*
taille des fichiers sources (ko)	157	432	374	217
taille du code (ko) après optimisation	141	211	523	1051
place en mémoire [a]	105	856	904	3502

[a] variables statiques uniquement

Une caractéristique d'OSCAR est de tenir relativement peu de place en mémoire. Cela tient au faible nombre de données (liste des atomes, résultats de calculs, etc.) dont le programme a réellement besoin en permanence et à une organisation très modulaire qui fait appel à des variables purement locales (non comptabilisées dans le tableau ci-dessus).

En revanche, il peut être gourmand quant à la place nécessaire sur disque, essentiellement à cause des fichiers de résultats de chaque molécule, contenant, entre autres, la fonction d'onde multiélectronique. A cela, il faut ajouter les fichiers contenant les points de chaque enveloppe moléculaire (compter une centaine de kilo-octets pour une molécule de 40 atomes). A titre d'exemple, l'étude d'une centaine de composés peut demander une trentaine de mégaoctets. C'est pourquoi, toutes nos procédures de lecture de fichiers permettent de lire directement des fichiers préalablement compressés par une commande du système d'exploitation. Dans ce cas, le gain de place sur disque peut atteindre les deux tiers, mais au prix d'une lecture plus lente.

— Utilitaires —

Nous indiquons page suivante, les *routines* les plus importantes de la bibliothèque *util.a*. Elles sont fréquemment "appelées" dans les autres bibliothèques, ainsi que dans mes autres programmes.

Tableau 22: Procédures utilitaires

utcabp	changement d'angle entre points de l'espace
utccc	centrage de coordonnées cartésiennes
utcdap	changement d'angle dièdre entre points de l'espace
utcdbp	changement de distance entre points de l'espace
utcddp	discrimination entre un descripteur et une propriété
utcdg	construction d'un arbre de décision
utcdsm	calculs de différences entre matrices de similarité
utcf	factorielle d'un entier
utcfsem	ouverture & et fermeture de fichiers ; messages d'erreur
utcgf	*gamma function*
utcibf	*incomplete beta function*
utcisl	intersections entre une sphère et une droite
utclr	régressions linéaires
utcpn	calcul de polynômes
utcqks	*qks function* pour degré d'association de deux distributions
utcsm	calculs de matrices de similarité
utcsdf	calcul de dérivées secondes pour interpolations
utdcm	matrice de passage entre systèmes de coordonnées
utdcs	définition d'un système local de coordonnées
utdcs2	variante simplifiée (pour une simple droite)
utdrm	définition d'une matrice de rotations
utdsm	définition d'une matrice d'homothéties
utdt	définition du tenseur d'une propriété donnée
utdtm	définition d'une matrice de translations
utfdl	*fit* de données sur une droite
utfdl2	estimation de l'erreur commise sur un *fit*
utfdle	*fit* de données sur une expression polynomiale
utgcp	barycentre de points pour une propriété donnée
utge	fonction erreur préfé finie
utgivs	volume d'intersections entre sphères
utgmdd	*moments* d'une distribution de données
utgpn	décomposition de chaînes de caractères
uticsf	interpolation par une fonction spline
utipf	interpolation par une fonction polynômiale
utirm	intégration par la méthode de Romberg
utirpf	interpolation par un quotient de polynômes
utism	intégration par la méthode de Simpson
utitr	intégration par la méthode des trapèzes
utlcd	corrélations linéaires entre distributions de données
utmm	produit de deux matrices
utopf	décompression et ouverture de fichiers

Tableau 23: Procédures utilitaires (suite)

utqce	classifications d'espèces
utqcm	système local de coordonnées et matrices de passage
utqec	calcul des coefficients d'un polynôme
utqeg	diagonalisation de matrices ; valeurs et vecteurs propres
utqep	équation d'un plan passant par trois points
utqim	inversion de matrices
utqipl	intersection d'un plan et d'une droite
utqmlr	calculs de régressions multi-linéaires
utqv	traitement de vecteurs...
utrcd	corrélations entre distributions nominales (Kendall, etc.)
utrfp	lecture de fichiers de points
uts...	résolutions d'équations diverses...
utsda	tri de tableaux de réels (nombreuses variantes)
utssc	séparation en classes d'espèces
utst	*Student's t test* de deux distributions
uttdcs	test du χ^2 entre distributions nominales
uttde	calcul entropique du degré d'association
uttdks	test de Kolmogorov et Smirnov
uttdks2	variante avec comparaison avec une fonction prédéfinie
uttp	transformation d'un point par une opération géométrique
. . .	

* * *

B OSCAR le guide...

B.1 Fichiers de molécules

Pour créer un premier fichier de données pour une molécule, une solution consiste à utiliser le programme GEOMHELP [87] dans lequel l'utilisateur précise la position de chaque atome en coordonnées internes (distances, angles et angles dièdres). Le mieux est d'ajouter les atomes un à un dans le fichier et de relancer OSCAR à chaque fois pour construire graphiquement la molécule. Le logiciel indiquera si certaines distances interatomiques sont trop faibles ou mettra en évidence d'éventuelles distorsions. Cette méthode est aussi rapide que d'utiliser le *builder* (outil de construction de molécules) d'un autre logiciel de modélisation moléculaire [1].

S'il faut définir toute une série d'espèces analogues, commencer par construire une pseudo molécule de base et optimiser sa géométrie. Ensuite, pour chaque molécule, ajouter à ce noyau ses substituants. Tenir compte des conformères possibles, surtout si les différences d'énergie risquent d'être significatives en fonction de l'orientation de chaque groupement. Garder à l'esprit le fait qu'a priori, chaque molécule sera considérée dans une géométrie idéale et rigide...

Il sera possible de lire dans OSCAR les fichiers de données de GEOMO/S et de CHIMISTE, qui fournissent les coordonnées cartésiennes de chaque atome. Pour un autre format (MAD, etc.), il suffirait d'écrire une procédure de lecture puis de l'inclure dans le *link* du programme. Tous les fichiers concernant une même molécule doivent, de préférence, se situer dans un seul *directory* et avoir des noms qui ne diffèrent que par le suffixe.

B.2 Lancement d'OSCAR

Le logiciel fonctionne actuellement sur IBM RT PC et IBM RISC 6000. Dans les deux cas, taper simplement oscar et répondre à la question qui demande la station de travail souhaitée. Pour cela, un *PROFILE* - fichier qui précise certaines caractéristiques graphiques propres au programme - doit se trouver dans le *working directory*. L'affichage peut être sensiblement différent sur les deux types d'ordinateurs, de même que les touches utilisées au clavier...

[1] Je conseille aussi à l'utilisateur mon propre programme *builder*...

Il est possible de charger en mémoire dès le début de la session toute une série de molécules. Créer un fichier "liste", y inscrire le nom de chaque fichier à lire puis taper oscar liste. Cette facilité s'avère pratiquement indispensable dès qu'il s'agit de travailler plusieurs jours sur la même série de composés.

Une fois le programme lancé, il affiche une liste d'options disponibles en bas de l'écran. Il suffit de taper sur la fonction désirée, sachant que *Cancel* [1] permet toujours de revenir en arrière. Si OSCAR affiche une alerte, une boîte de dialogue ou tout autre message, la même touche annule la requête en cours. Par ailleurs, il sera fréquemment demandé à l'utilisateur de préciser une valeur numérique, un nom de fichier, ou plus généralement une chaîne de caractères. Dans ce cas, la saisie sera validée avec *Enter* [2], ou annulée avec *Cancel*. La touche *End* [3] arrête immédiatement l'exécution du programme.

B.3 Calculs des orbitales moléculaires

Une fois obtenus les fichiers de base, l'étape suivante consiste à lancer le programme GEOMO/S [17] et obtenir ainsi autant de fichiers de résultats. Conserver si possible les mêmes options de calcul, en particulier la méthode semi empirique (se reporter au paragraphe 2.1 quant à son choix). Certaines procédures de OSCAR refusent de comparer des molécules calculées par des méthodes différentes.

Les données lues en définitive dans les fichiers de résultats sont les coordonnées atomiques après optimisation de géométrie, le nombre d'orbitales moléculaires (comme combinaisons linéaires d'orbitales atomiques de Slater), l'énergie et les vecteurs propres de chacune, l'énergie totale et celle des liaisons, la chaleur de formation et le moment dipolaire. Si on souhaite s'intéresser aux polarisabilités globale et locales, utiliser l'option ad hoc de GEOMO/S puis le programme POL pour obtenir le fichier correspondant de résultats.

Noter enfin que la procédure de lecture tient compte du format actuel des fichiers de résultats de GEOMO/S : en cas de modification de ce dernier, le source d'OSCAR devra être corrigé. Une amélioration ultérieure consistera d'ailleurs à inclure directement les calculs SCF dans le logiciel.

---

[1] touche Scroll Lock sur station X, touches Esc ou Alt - Delete sur RT PC

[2] touches Return ou Action selon les stations de travail

[3] Shift - Esc sur RT PC

B.4 Prise en compte de l'activité

Pour inclure d'autres données en mémoire, OSCAR utilise des fichiers suffixés ".val". Ils doivent contenir notamment l'activité de la molécule et l'incertitude sur sa mesure (le programmeur a tenu compte du fait que ces valeurs varient parfois de plus d'un ordre de grandeur...). Ne pas oublier les conversions nécessaires pour rapporter le résultat d'un test biologique donné (dans un de nos exemples, la chute de la pression sanguine d'un rat de un kilo) au nombre de molécules de composé actif. Spécifier une incertitude de 1. si elle n'est pas connue.

Le même fichier peut inclure la liste des atomes définis en tant qu'*atomes de base* (ou atomes *correspondants*). Ce sont ceux qui occupent les mêmes positions dans l'espace, d'une molécule à l'autre, indépendamment de leur nature ou de leur numérotation. Veiller à ce qu'ils soient définis dans le même ordre d'une molécule à l'autre, le logiciel ne pouvant vérifier lui-même si cela est fait correctement.

B.5 Molécules en mémoire et à l'écran

Il est possible de définir des listes de molécules *en cours d'étude*. Par défaut, chaque molécule est ajoutée à la liste numéro 1, qui est celle utilisée au début du programme (liste dite courante). Les options du menu *Study* permettent ensuite d'ajouter ou de retrancher une molécule de la liste courante ou de changer cette dernière. Ce système est pratique pour sélectionner parmi les molécules en mémoire celles qu'il faut soumettre à un calcul donné. Noter que l'option *Auto* remplit elle-même les listes 2, 3 et 4 avec respectivement les molécules les plus actives, les molécules les moins actives et un panachage des deux catégories.

Durant toute la session, faire la distinction entre les molécules présentes en mémoire et celles qui, de surcroît, sont construites graphiquement. Si le premier nombre peut facilement être modifié (modifier la constante *knm* du programme et le recompiler), le second est fonction du nombre maximal de vues PHIGS disponibles sur la station de travail. Les limites actuelles sont respectivement de 99 et 30 molécules.

Les options d'OSCAR agissent soit :

- sur la molécule courante, première molécule visible à l'écran,
- sur la molécule dont l'utilisateur précise les premières lettres du nom ou bien le numéro en mémoire (0 désignant par extension toutes les molécules)
- sur chaque molécule de la liste d'étude (menu *Study* uniquement).

Si appuyer sur une touche de fonction reste sans effet, c'est donc qu'aucune molécule n'est présente en mémoire ou bien qu'aucune n'est construite graphiquement. Il peut

arriver aussi qu'il y ait confusion entre le numéro de la molécule en mémoire (affecté "chronologiquement" par OSCAR) et le numéro que l'utilisateur attribue lui-même à ses composés.

B.6 Outil de manipulation des vues 3D

Dans la partie droite de l'écran apparaissent des cases qui donnent accès - après désignation par l'utilisateur - à autant de fonctions. Constituant un utilitaire intégré mais indépendant du logiciel, on les retrouvera aussi dans d'autres programmes graphiques que j'ai écris. Ces fonctions agissent sur la première vue 3D [1] dont le numéro est inscrit tout en haut. Si l'affichage semble incorrect sur une station IBM RISC 6000, il peut être nécessaire de rafraîchir l'écran sous *X-Windows* (option *refresh*) ou bien de cliquer deux fois de suite sur la case *Act*.

next,last Renvoie la première vue au dernier plan ou inversement amène la dernière vue au premier plan. Le numéro de la nouvelle première vue est alors inscrit juste au-dessus. Cette option permet le passage d'une molécule à l'autre.

Shd La première vue 3D est rendue alternativement opaque ou transparente. Ceci permet les comparaisons et les superpositions. Utiliser la coloration des liaisons pour mieux distinguer les molécules si besoin.

Act La première vue 3D est rendue alternativement visible ou non, sans qu'elle change d'ordre d'affichage. Le titre de la molécule est alors réécrit en italique pour rappeler que la première molécule visible n'est pas la molécule courante.

Bkg Changement de la couleur de fond de la première vue 3D (si elle est opaque). En appuyant sur le troisième bouton de la souris, on peut indiquer directement un index de couleur.

Del Affichage d'une alerte de confirmation pour détruire la première vue 3D. Utiliser cette option pour détruire une molécule graphiquement sans la supprimer en mémoire.

Hdc Zoom de l'image réduite aux seules vues 3D pour faciliter les copies d'écran à l'aide d'une imprimante. Quitter en appuyant sur une touche de fonction.

Plt Dialogue pour redessiner l'image sur une feuille de papier via une table traçante. Préciser les numéros des vues à imprimer, les proportions du dessin dans la largeur d'une feuille de format A4 et une légende facultative à inscrire en bas. Après clic sur *Ok*, un fichier suffixé .gdf est créé dans le *working directory*. Taper ensuite sur AIX la commande print -gdf fichier pour l'imprimer.

[1] j'ai distingué en programmation les vues de présentation de celles qui contiennent des objets en trois dimensions. Nous ne parlons ici que des secondes.

Zbf [1] Activation ou désactivation du *z-buffer*. Si oui, l'image donne une impression de relief, mais sa mise à jour est ralentie... Noter que cette option n'a pas d'influence sur la qualité des copies d'écran sur une imprimante IBM 5087.

Tsp [1] Activation ou désactivation du mode de transparence partielle. Cette option permet de distinguer l'éclairage des facettes avant et arrière dans chaque molécule, sans rendre leur vue 3D transparente pour autant.

Lsr [1] Dialogue permettant de définir une ou plusieurs sources de lumière (couleur et type). Ralentit nettement la durée de rafraichissement de l'image. L'utilisateur a le choix entre des sources de lumières ambiantes, positionnelles ou directionnelles. Permet d'obtenir de belles photos...

para Modifie le mode de projection de la première vue 3D sur l'écran. En parallèle, chaque point est projeté sur un plan de vue suivant un axe normal à ce dernier. En perspective, la projection se fait à partir d'un point de fuite où se situe l'œil de l'observateur. Le mode parallèle est à préférer pour toute comparaison, le mode pespective donne une image plus réaliste.

with Flag de solidarisation du mode de projection. En position *on*, chaque changement de visualisation imposé à la première vue 3D est reproduit identiquement sur toutes les autres.

view Déplace le plan de vue où est projetée l'image. Les unités utilisées sont "équivalentes" à des centimètres en profondeur.

prp Déplacement du *projection reference point*, c'est-à-dire de l'œil de l'observateur dans le cas d'une projection de la première vue 3D en perspective.

near,far Positions des plans de clipping proche et lointain. Cette option permet d'effacer les portions de la première vue 3D qui ne sont pas dans le champ visuel de l'observateur. Mêmes unités que pour *view*. Si le dessin semble incomplet, c'est souvent parce que ces plans ont été trop rapprochés du plan de vue.

lat,ver Angles latéral et vertical (en degrés) d'observation de la première vue 3D.

Init 2D Réinitialise le mode de visualisation de la première vue 3D avec les mêmes options que celles des vues de présentation (en parallèle).

sx,sy,sz,zoom Facteurs d'homothétie appliqués le long des axes Ox, Oy et Oz de la première vue 3D. Ne pas confondre ces axes avec ceux, spécifiques, de la molécule associée. Le troisième bouton permet d'indiquer des valeurs numériques absolues. Ces zooms étant progressifs, on préfèrera souvent l'option de mise à l'échelle dans une ou plusieurs molécules (ci-dessous).

tx,ty Translations absolues (en centimètres) le long des axes de l'écran imposées à la première vue 3D.

notrl,norot Suppression des translations ou des rotations imposées à la première vue 3D.

[1] option sans effet si la station de travail n'a pas de coprocesseur graphique 3D

rx,ry,rz Rotations absolues (en degrés) imposées par rapport aux axes de l'écran et à son axe normal. Le troisième bouton permet d'indiquer directement des valeurs numériques.

with Flag de solidarisation. En position *on*, chaque mouvement 3D imposé à la première vue 3D est appliqué identiquement aux autres. A utiliser donc avec prudence.

B.7 Outil de construction de molécules

La vue de présentation de chaque molécule se compose de trois zones :

- un titre que l'utilisateur peut modifier après un clic dessus. Il est parfois utilisé dans différentes options de *design*. Ne pas le confondre avec le nom de la molécule, chaîne de caractères lues dans les fichiers de données au même titre que les coordonnées atomiques.
- une zone de commentaires juste au-dessus sur fond noir ou gris (elle peut être vidée en cliquant dessus). C'est là que OSCAR inscrit les résultats acquis en cours de session. Il est donc conseillé de construire chaque molécule de la liste de travail courante.
- une case mentionnant le numéro de la vue 3D associée. Pour la molécule courante (première molécule visible à l'écran), ce numéro coïncide avec celui qui est inscrit en haut de la zone droite de l'écran. Utile lors du *plot* de la molécule.

Cliquer sur cette dernière case avec le premier bouton de la souris donne accès à divers utilitaires qui agissent sur la molécule courante. L'autre bouton donne accès aux options de *design* qui, elles, portent sur toutes les molécules construites. Voici la liste des deux ensembles de fonctions :

internal coordinate Taper en bas de l'écran le numéro de deux, trois ou quatre atomes. OSCAR inscrira dans la zone de commentaires la distance (en angströms), l'angle ou l'angle dièdre (en degrés) entre eux.

distance atom - plane Indiquer de la même manière les numéros de quatre atomes. Le plan est défini à partir des trois derniers, A, B et C de la manière suivante : un premier vecteur directeur unit A et B, le second lui est perpendiculaire, colinéaire et de même sens que le vecteur $\vec{AC}$.

scale in Å by cm Par défaut, une échelle de 0. permet de "remplir" la vue 3D le mieux possible avec la molécule. Cette option est complémentaire du zoom (action sur la structure et non sur la vue).

rotation about axis L'utilisateur indique les numéros de deux atomes puis l'angle (en degrés) de rotation de la molécule courante autour de l'axe qu'ils définissent. Cette option est complémentaire des rotations précédentes (repère local à la molécule et non propre à la vue).

translation about axis Idem, mais pour une translation algébrique (donnée en centimètres) le long de l'axe des deux atomes.

spheres update [1] Remise à jour des sphères atomiques de covalence ou de Van der Waals après rotation de la molécule courante autour des axes Ox ou Oy de sa vue 3D (ce sont en fait des disques 2D).

bond coloration Changement de la couleur des liaisons dans la molécule courante. On peut préciser -2 pour une coloration uniforme (ce qui permet de bien distinguer deux molécules superposées), -3 pour qu'elle soit fonction de l'ordre de chaque liaison (par défaut), ou tout autre index positif.

clean Permet de "nettoyer" le dessin d'indications diverses (axes, enveloppe de contact, points, etc.)

lighting Dialogue pour choix des conditions d'éclairage de la molécule courante. Sont paramétrables, entre autres, les coefficients de diffusion et de réflexion de la lumière sur les facettes avant ou arrière de chaque atome. Le mieux est de faire des essais...

La case à cocher dessinée en bas du menu permet - pour certaines des options qui précèdent - de reproduire l'action sur toutes les molécules construites et pas seulement la "courante". Ce sont les options *scale*, *bond coloration* et *lighting*. Les fonctions qui suivent sont accessibles après clic sur le deuxième bouton de la souris.

no atoms Options de *design* permettant de ne voir que certains types d'atomes. Utile pour alléger un affichage trop riche d'informations. Remarque : le symbole "f" désigne les atomes fictifs (numéro atomique : 0)

no symbols Options portant sur les annotations inscrites au droit de chaque atome. *NB* : les coordonnées sont celles, centrées, de la molécule : a priori, elles ne coïncident pas avec celles du fichier. Les polarisabilités atomiques doivent être préalablement connues en mémoire, donc lues dans un fichier ".pol". Les indications HOMO (etc.) indiquent les poids relatifs de chaque atome dans les orbitales moléculaires frontières. Ces différentes données ne peuvent être affichées qu'une fois calculées (les fonctions correspondantes sont indiquées plus loin).

no bonds Options permettant de cacher certains types de liaisons, en particulier les liaisons "faibles" ou celles qui joignent un atome d'hydrogène.

bond symbols Indications de la distance (en angströms), de l'angle ou de l'angle dièdre (en degrés) entre deux atomes. Certaines valeurs sont négatives par pure convention. L'angle indiqué entre deux atomes A et B (le numéro de B étant plus petit

[1] uniquement sur IBM RT PC

que celui de A) est $\widehat{ABC}$, C étant le voisin de B dont le numéro est le plus petit. Même convention pour les angles dièdres.

informations Il est possible de faire apparaître un trièdre pour rappeler les axes de la vue, les composantes du moment dipolaire ou de la polarisabilité totale de la molécule. Ces options sont mutuellement exclusives, de même que le tracé des points appartenant à la surface externe de la molécule.

autres informations Les dernières options permettent l'affichage de données propres au dessin : axes de translation ou de rotation, système de coordonnées local à la molécule, etc.

B.8 Calculs sur les graphes y = fct (x)

Comme pour les molécules, la vue de présentation du graphe se compose de trois zones :

- un titre que l'utilisateur peut modifier en cliquant dessus. A priori, il est fonction du type de tracé demandé : activité en fonction d'un paramètre géométrique, grandeur moléculaire le long d'une direction de référence, etc.
- une zone de commentaires juste au-dessus sur fond noir ou gris (elle peut être vidée en cliquant dessus). C'est là que sont inscrits les résultats numériques des calculs présentés ci-dessous.
- une case mentionnant le numéro de la vue 3D associée. Ce numéro coïncide toujours avec celui qui est inscrit en haut de la zone droite de l'écran. A mentionner dans le dialogue qui permet un *plot* du graphe.

Noter que les options *next* et *last* qui permutent les vues 3D n'agissent que sur les molécules. Le graphe est visualisé avec une priorité plus élevée ce qui explique qu'il apparait toujours devant. Pour quitter le graphe, utiliser l'option *Del* pour détruire la vue 3D ou bien demander l'affichage d'un autre graphe.

Les options qui suivent constituant un utilitaire indépendant du logiciel OSCAR proprement dit, on comprendra que certaines d'entre elles servent peu pour l'étude des relations structure - activité. Comme pour les molécules, elles sont réparties en deux ensembles, accessibles chacun par un bouton de la souris.

scale Changement d'échelle par rapport aux deux axes de l'écran. Les unités à préciser ne sont pas des centimètres, mais celles des deux grandeurs du graphe. Une échelle de 0. correspond à un bon "remplissage" de l'écran.

operating curve Chaque option agit sur une courbe dite "courante". Par défaut, il s'agit de la courbe numéro 1. Utiliser cette option pour la changer.

exclusion Il est possible d'"exclure" momentanément des points de la courbe courante, c'est-à-dire de procéder exactement comme s'ils n'existaient pas. Option pratique pour les corrélations ou pour les interpolations, notamment pour éliminer des points aberrants. Dans les dernières options, les points inclus et exclus constituent deux distributions qui peuvent faire l'objet de comparaisons. Visuellement, les points exclus sont redessinés avec des annotations en italique. Pour exclure ou ré-inclure un point dans la courbe courante, indiquer simplement son numéro (les points sont numérotés par abscisse croissante). Pour en exclure plusieurs, indiquer les numéros des deux points extrêmes. Utiliser les variantes "by z limits" pour faire le même tri par rapport à une ordonnée-seuil.

inclusion Option inverse de la précédente. Permet de ré-inclure des points dans la courbe courante.

linear correlation Corrélation linéaire dans la courbe courante. Elle ne porte que sur les points "non exclus". Cette option ne fournit pas de droite (faire une corrélation polynomiale de degré 1 pour cela) mais simplement la statistique du χ^2 et le coefficient de Pearson.

polynomial correlation Recherche de l'unique polynôme de degré n passant par $n-1$ points. Les coefficients du polynôme sont indiqués en commentaires jusqu'au degré 6. Au-dela, la corrélation est déconseillée, compte tenu de l'"instabilité" d'une telle fonction mathématique. Le χ^2 équivaut à l'habituel coefficient de régression linéaire.

clear correlations effacement des courbes précédentes.

interpolations Recherche d'un polynôme d'interpolation entre les points non exclus de la courbe courante. Il résulte de l'application de la méthode de Lagrange (ou méthode des différences moindres). L'option suivante permet de rechercher un polynôme qui s'exprime comme le quotient de deux fonctions polynomiales. Plus adéquat pour le tracé de certaines courbes à forme elliptique. Le tracé peut faire apparaître des "pôles" qui correspondent à des valeurs nulles du dénominateur.

cubic spline Interpolation par une fonction spline cubique. Le plus commode puisque la fonction d'interpolation ne passe pas forcément par chaque point. Les dérivées secondes sont imposées aux deux points extrêmes puis déduites entre les deux.

clear interpolations Effacement des courbes précédentes.

save Option de sauvegarde d'une ou plusieurs des courbes affichées. L'utilisateur doit indiquer le nom d'un fichier de sauvegarde et le numéro de chaque courbe à sauver (0 désignant toutes les courbes). On peut préciser une ligne de commentaires. Le format adopté permet une relecture de ces fichiers dans certaines options de OSCAR. Il est prévu également une option de rechargement direct de courbe.

statistics Calcul de la moyenne, de la variance et de l'écart type sur les abscisses et sur les ordonnées des points non exclus de la courbe courante.

compare Compare, dans la courbe courante, les deux distributions de points, exclus et non exclus. Le premier test est le *Student's t* qui compare les moyennes de deux distributions dont on suppose qu'elles ont la même variance. Le second est celui de Kolmogorov qui calcule l'écart maximal qui sépare les deux distributions : il est d'autant plus fiable que les fluctuations sont faibles entre ordonnées. Dans les deux cas, une valeur numérique (la *significance*) indique à quel point le test est fiable : ces valeurs doivent être, si possible, inférieures à 0.005.

Les options qui suivent agissent sur le *design* des courbes. Elles sont accessibles après un clic sur le deuxième bouton de la souris.

curves Permet de sélectionner la ou les courbes à visualiser. Si le graphe semble un peu "vide", c'est probablement qu'une courbe est restée invisible... (l'invisibilité est conservée lors du tracé d'un graphe à un autre).

points Permet de ne voir les annotations que de quelques points. Utile pour alléger l'affichage dans le cas de courbes à grand nombre de points.

design Rend visible ou invisible (suivant le bouton de la souris) l'annotation de chaque point : coordonnée x ou z, numéro du point, pic, autre annotation, ligne pointillée entre points. D'autres options permettent l'affichage ou la disparition des axes, de leurs graduations ou encore des jonctions qui unissent les points de même numéro d'une courbe à une autre.

Une précision s'impose quant à la numérotation des points. Pour tout graphe affiché à l'écran, il y a deux numérotations. La première les numérote par abscisse croissante. Elle correspond à l'option de design *identifier* et est utilisée dans tous les calculs. La seconde est la numérotation antérieure au tri par abscisse croissante. Elle correspond à l'option de design *other*.

* * *

C Notions de mécanique moléculaire

L'*énergie de déformation* d'une molécule pour l'amener d'une géométrie optimisée à une certaine conformation, peut être définie de façon empirique comme la somme de plusieurs contributions. Nous en donnons ici une expression parmi d'autres [88], étant entendu que chaque auteur peut l'affiner à son gré...

$$E_{def} = \sum_{liaisons} E_b + \sum_{angles} E_\theta + \sum E_{nb} + \sum E_{op} + \sum_{dièdres} E_{tor} + \sum E_{dd}$$

E_b désigne l'énergie d'élongation ou de contraction d'une liaison :

$$E_b = f_d\ (r - r_0)^2$$

relation dans laquelle f_d et r_0 sont les paramètres empiriques à ajuster (r_0 provenant généralement des tables) et r, la distance entre les deux atomes.

E_θ représente l'énergie qu'il faut dépenser pour faire varier un angle de valence :

$$E_\theta = f_\theta\ (\Delta\theta^2 - \alpha\Delta\theta^3),\ \Delta\theta = \mid \theta - \theta_0 \mid$$

où f_θ et θ_0 sont les paramètres analogues aux précédents et α une constante.

E_{nb} rend compte des interactions entre atomes non liés selon l'expression :

$$E_{nb} = fe^{-gr} - (e/r^6)$$

r désignant la distance qui les sépare. Il peut être complété par un terme propre aux liaisons hydrogènes.

E_{op} est un terme plus fin qui exprime l'énergie à dépenser pour déplacer un carbone sp_2 hors du plan défini par ses trois voisins :

$$E_{op} = f\ D^2$$

D étant la distance perpendiculaire entre ce plan et la position de l'atome.

E_{tor} est le terme de torsion d'un angle dièdre :

$$E_{tor} = \sum_n \nu_n\ (1 \pm \cos\ n\ \varphi)$$

sommation dans laquelle chaque ν est un paramètre empirique et φ l'angle de torsion.

Enfin, chaque E_{dd} représente l'énergie d'interaction électrostatique définie par :

$$E_{dd} = K q_i \, q_j \, / \, (D \, r_{ij})$$

q_i et q_j désignant les charges respectives des atomes i et j séparés par une distance r_{ij} et D jouant un rôle analogue à celui d'une constante diélectrique.

Ces expressions font donc intervenir des paramètres plus ou moins empiriques dont certains sont tabulés dans la littérature. Pour nous, l'essentiel est moins d'obtenir de bonnes valeurs de l'énergie pour deux conformations que d'évaluer correctement l'écart qui les sépare. En effet, nous ne recherchons pas la grandeur thermodynamique en elle-même, mais souhaitons simplement identifier les répartition statistique des conformères d'une molécule.

En définitive, nous avons intégré dans le logiciel OSCAR l'essentiel de la méthode MM2, une des plus connues en mécanique moléculaire. Elle sera commode et assez rapide à utiliser pour le chimiste. Pour de plus amples détails sur son principe, nous invitons le lecteur à se reporter aux travaux de Burkert [89] et Allinger [90].

* * *

Bibliographie

[1] C. Hansch, *Relations Structure - Activité.* Edité par la société de chimie thérapeutique, Paris (1974).

[2] K.B. Lipkowitz, D.B. Boyd, *Reviews in computational chemistry.* VCH publishers (1990).

[3] J.K. Seydel, *QSAR and strategies in the design of bioactive compounds.* VCH, Weinheim (FRG) (1984).

[4] I. Lukovits, A. Lopata, *J. Med. Chem.*, **23**, 449-59 (1980). Decomposition of pharmacological activity indices into mutually independent components using principal component analysis.

[5] R. Rozot, *Facteurs électroniques et reconnaissance de formes en CAO de molécules pharmacologiquement actives. Application à la famille des benzodiazépines.* Thèse présentée à l'Université de Nancy I (1988).

[6] G. Klopman, *J. Am. Chem. Soc.*, **106**, 7315-21 (1984). Artificial intelligence approach to structure - activity studies. Computer automated structure evaluation of biological activity of organic molecules.

[7] G. Klopman, A.N. Kalos, *J. Comput. Chem.*, 6, 492-506 (1984). Causality in structure - activity studies

[8] C.K. Hancock, E.A. Meyers, B.J. Yager, *J. Am. Chem. Soc.*, **83**, 4211 (1961).

[9] R.W. Taft, *Steric effects in organic chemistry.* vol 3, M.S. Newman ed., New York, Wiley (1956).

[10] L. Hammett, *Physical organic chemistry.* Mc Graw-Hill, New York (1970).

[11] M. Born, J.R. Oppenheimer, *Ann. Physik*, **84**, 457 (1927).

[12] J.A. Pople, D.A. Beveridge, *Approximate molecular orbital theory.* McGraw-Hill, New York. (1970).

[13] J.L. Rivail, *Eléments de Chimie quantique à l'usage des chimistes.* InterEditions / Editions du CNRS, 291-5 (1989).

[14] M.J.S. Dewar, W. Thiel, *J. Am. Chem. Soc.*, **99**, 4899 (1977). Ground states of molecules. 38. The MNDO method. Approximation and parameters.

[15] J.J.P. Stewart, *J. Comput. Chem.*, **10**, 209 (1989). Optimization of parameters for semi-empirical methods. I. Method.

[16] J.J.P. Stewart, *J. Comput. Chem.*, **10**, 221 (1989). Optimization of parameters for semi-empirical methods. II. Applications.

[17] D. Rinaldi, P.E.J. Hoggan, A. Cartier, *QCPE*, 9, 128 (1989).

[18] M.T.C. Martins Costa, *CHIMISTE. Un ensemble de logiciels de modélisation moléculaire quantique.* Thèse soutenue à l'Université de Nancy I (1988).

[19] D. Rinaldi, *Applications des méthodes semi-empiriques à l'étude des propriétés électriques moléculaires et des interactions entre molécules.* Thèse présentée à l'Université de Nancy I (12/12/1975).

[20] L.B. Kier, *Molecular orbital theory in drug research.* Academic press, New York (1971).

[21] A.D. Buckingham, *Intermolecular forces.* Advances in chemical physics, vol 12, edited by J.O. Hirschfelder, Intersciences Publishers (1967). Permanent and induced molecular moments and long-range intermolecular forces.

[22] K. Fukui, *Molecular orbitals in chemistry.* P.O. Lowdin, B. Pullman eds., Academic Press, New York (1964).

[23] J.L. Rivail, *Eléments de Chimie quantique à l'usage des chimistes.* InterEditions / Editions du CNRS, 385-7,395-8 (1989).

[24] J. Mathieu, R. Panico, *Mécanismes réactionnels en chimie organique.* Hermann, Paris (1972).

[25] D. Rinaldi, J.L. Rivail, *Theoret. Chim. Acta (Berlin),* **32**, 243-51 (1974). Calcul théorique des polarisabilités électroniques moléculaires. Comparaison des différentes méthodes.

[26] A. Cartier, J.L. Rivail, *Chemometrics and intelligent laboratory systems,* **1**, 335-47 (1987). Electronic descriptors in quantitative structure - activity relationships.

[27] C. Edmiston, K. Ruedenberg, *Rev. Mod. Phys.,* **32**, 457 (1963).

[28] J.M. Foster, S.F. Boys, *Rev. Mod. Phys.,* **32**, 300 (1960).

[29] J.E. Ash, P.A. Chubbs, S.E. Ward, S.M. Welford and P. Willett, *Communication, storage and retrieval of chemical information.* Ellis Horwood. Chichester, England (1975).

[30] D. Weininger and A. Weininger, *J. Chem. Inf. Comput. Sci.,* **28**, 31 (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.

[31] D. Weininger, A. Weininger and J.L. Weininger, *J. Chem. Inf. Comput. Sci.,* **29**, 97 (1989). SMILES, 2. Algorithm for generation of unique SMILES notation.

[32] A.T. Balaban, *Chemical applications of graph theory.* Academic Press, New York (1976).

[33] R.C. Read and D.G. Corneil, *J. Graph Theory,* **1**, 363 (1977). The graph isomorphism disease.

[34] R.E. Tarjan, *Am. Chem. Soc. Symp. Ser.,* **46**, 1 (1977). Graph algorithms for chemical computation.

[35] W.T. Wipke and T.M. Dyott, *J. Am. Chem. Soc.,* **96**, 4834 (1974). Stereochemically unique naming algorithm.

[36] H.L. Morgan, *J. Chem. Doc.,* **5**, 107 (1965). The generation of a unique machine description for chemical structures: a technique developped at Chemical Abstracts Service.

[37] P.G. Dittmar, R.E. Stobaugh and C.E. Watson, *J. Chem. Inf. Comput. Sci.,* **16**, 111 (1976). The Chemical Abstracts Service Chemical Registry System. I. General design.

[38] G.W. Adamson, J. Cowell, M.F. Lynch, A.H.W. McLure, W.G. Town and A.M. Yapp, *J. Chem. Doc.,* **13**, 153 (1973). Strategic considerations in the design of a screening system for substructure searches of chemical structure files.

[39] L. Hodes, *J. Chem. Inf. Comput. Sci.*, **16**, 88 (1976). Selection of descriptors according to discrimination and redundancy. Application to chemical substructure searching.

[40] P. Willett, *J. Chem. Inf. Comput. Sci.*, **19**, 159 (1979). A screen set generation algorithm.

[41] J. Figueras, *J. Chem. Doc.*, **12**, 237 (1972). Substructure search by set reduction.

[42] A. von Scholley, *J. Chem. Inf. Comput. Sci.*, **25**, 235 (1985). A relaxation algorithm for generic chemical structure screening.

[43] E.H. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1965). A graph-theoretic algorithm for matching chemical structures.

[44] W.T. Wipke, G.I. Ouchi, S. Krishnan, *Artificial Intelligence*, **11**, 173 (1978). Simulation and Evaluation of Chemical Synthesis, SECS.

[45] J.M. Barnard, M.F. Lynch, S.M. Welford, *J. Chem. Doc.*, **21**, 151 (1981). Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL, a formal language for the description of generic chemical structures.

[46] F.H. Allen, O. Kennard and R. Taylor, *Acc. Chem. Res.*, **16**, 146 (1983). Systematic analysis of structural data as a research tool in organic chemistry.

[47] F.H. Allen, O. Kennard, D.G. Watson, L. Brammer, A.G. Orpen and R. Taylor, *J. Chem. Soc., Perkin Trans.*, **2**, S1 (1987). Tables of bond lengths determined by x-ray and neutron diffraction. Part 1. Bond lengths in organic compounds.

[48] A.G. Orpen, L. Brammer, F.H. Allen, O. Kennard, D.G. Watson and R. Taylor, *J. Chem. Soc., Dalton Trans.*, **2**, S1 (1989). Tables of bond lengths determined by x-ray and neutron diffraction. Organometallic compounds and coordination complexes of the d- and f- block metals.

[49] M. Randic, *J. Am. Chem. Soc.*, **97**, 6609-15 (1975). On characterization of molecular branching.

[50] D.H. Rouvray, *J. Comput. Chem.*, **8**, 470 (1987). The modeling of chemical phenomena using topological indices.

[51] L.B. Kier and L. Hall, *Molecular connectivity in structure-activity analysis.* Wiley, New York (1986).

[52] A.T. Balaban, A. Chiriac, I. Motov and Z. Simon, *Lecture Notes in Chemistry*, **15**, Springer-Verlag, Berlin (1980). Steric fit in quantitative structure-activity relations.

[53] W.J. Murray, L.H. Hall, and L.B. Kier, *J. Pharm. Sci.*, **64**, 1978 (1975). Molecular connectivity III: relationship to partition coefficient.

[54] C.W. Crandell, D.H. Smith, *J. Chem. Inf. Comput. Sci.*, **23**, 186 (1983). Computer-assisted examination of compounds for common three-dimensional substructures.

[55] T.H. Varkony, Y. Shiloach, D.H. Smith, *J. Chem. Inf. Comput. Sci.*, **19**, 104 (1979). Computer-assisted examination of chemical compounds for structural similarities.

[56] D.J. Danzinger, P.M. Dean, *J. Theor. Biol.*, **116**, 215 (1985). The search for functional correspondences in molecular structure between two dissimilar molecules.

[57] R.P. Sheridan, R. Nilakantan, J.S. Dixon, R. Venkataraghavan, *J. Med. Chem.*, **29**, 899 (1986). The ensemble approach to distance geometry: applications to the nicotinic pharmacophore.

[58] J.M. Evans, C.S. Fake, T.C. Hamilton, R.H. Poyser, E.A. Watts, *J. Med. Chem.*, **26**, 1582-9 (1983). Synthesis and antihypertensive activity of substitued *trans*-4-amino-3,4-dihydro-2,2-diméthyl-2H-1-benzopyran-3-ols.

[59] J.M. Evans, C.S. Fake, T.C. Hamilton, R.H. Poyser, G.A. Showell, *J. Med. Chem.*, **27**, 1127-31 (1984). Synthesis and antihypertensive activity of 6,7-disubstitued *trans*-4-amino-3,4-dihydro-2,2-diméthyl-2H-1-benzopyran-3-ols.

[60] F.M. Richards, *Ann. Rev. Biphys. Bioeng.*, 6, 151-76 (1977).

[61] K.D. Gibson, H.A. Scheraga, *Mol. Phys.*, **62**, 1247 (1987). Exact calculation of the volume and surface-area of fused hard-sphere molecules with unequal atomic radii.

[62] M.L. Connolly, *QCPE 75.* (1981). Molecular surface.

[63] J.L. Pascual-Ahuir, E. Silla, *J. Comput. Chem.*, **11**, 1047-60 (1990). GEPOL: An improved description of molecular surfaces. I. Building the spherical surface set.

[64] E. Silla, I. Tunon, J.L. Pascual-Ahuir, *J. Comput. Chem.*, **12**, 1077-88 (1991). GEPOL: An improved description of molecular surfaces. II. Computing the molecular area and volume.

[65] A. Verloop, W. Hoogenstraaten, J. Tipker, *Drug design*, **7**, 165 (1976).

[66] A. Verloop, in *QSAR and strategies in the design of bioactive compounds.* J.K. Seydel ed., VCH, Weinheim (FRG) (1984). The sterimol approach. Possible contribution to receptor mapping.

[67] E. Bernardy, P. de Donato, J.M. Cases, *Projet ESPRIT no 2263.* rapport no 6 (1991). Ultra high media for magnetic storage.

[68] P.G. Mezey, *Int. J. Quantum. Chem., Quantum. Biol. Symp.*, **12**, 113 (1986). Group theory of electrostatic potentials: a tool for quantum chemical drug design.

[69] P.G Mezey, *J. Math. Chem.*, **2**, 299 (1988). Shape group studies of molecular similarity: shape groups ans shape graphs of molecular contour surfaces.

[70] P.G Mezey, *J. Math. Chem.*, **2**, 325 (1988). Global and local relative convexity and oriented relative convexity: application to molecular shapes in external fields.

[71] P.G. Mezey in *Computational chemical graph theory and combinatorics.* D.H. Rouvray, Ed., Nova Publications, New York (1989). The topology of molecular surfaces and shape graphs.

[72] G.A. Arteca, V.B. Jammal, P.G. Mezey, J.S. Yadav, M.A. Hermsmeier, T.M. Gund, *J. Mol. Graphics*, **6**, 45 (1988). Shape group studies of molecular similarity: relative shapes of Van der Waals and electrostatic potential surfaces of nicotinic agonists.

[73] G.A. Arteca, V.B. Jammal, P.G. Mezey, *J. Comput. Chem.*, **9**, 608 (1988). Shape group studies of molecular similarity and regioselectivity in chemical reactions.

[74] G.A. Arteca, P.G. Mezey, *Int. J. Quantum. Chem., Biol. symp.*, **14**, 133 (1987). A method for the characterization of molecular conformations.

[75] P.G. Mezey in *Concepts and applications of molecular similarity.* G.M. Maggiora, M.A. Johnson, Eds., Wiley-Interscience (1990). Three-dimensional topological aspects of molecular similarity.

[76] M.A. Johnson, *J. Math. Chem.*, **3**, 117 (1989). A review and examination of the mathematical spaces underlying molecular similarity analysis.

[77] R. Von Mises, *Mathematical theory of probability and statistics.* Academic Press, New York, chapitres IX (C) et IX (E) (1986).

[78] W. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes.* Cambridge University Press (1986).

[79] J.P. Fénelon, *Qu'est-ce que l'analyse des données ?.* Lefonen, Paris (1981).

[80] J.C. Simon, *La reconnaissance des formes par algorithmes.* Editions Masson, Paris (1984).

[81] C.E. Shannon, W. Weaver, *The mathematical theory of communication.* The University of Illinois Press, Urbana (1949).

[82] T.E. Klein, C. Huang, T.E. Ferrin, R. Langridge, C. Hansch, *Computer assisted drug receptor mapping analysis.* T.H. Pierce, B.A. Hohne eds. (1986).

[83] F.A. Momany, R. Pitha, V.J. Klimkowski, C.M. Venkatachalam, *Drug design using a protein pseudoreceptor.* B.A. Hohne, T.H. Pierce eds. p 82-91 (1989).

[84] A.J. Stuper, W.E. Brugger, P.C. Jurs, *Computer assisted studies of chemical structure and biological evaluation.* Wiley, New York (1979).

[85] D.E. Walters, A.J. Hopfinger, *J. Mol. Struct. (Theochem),* **134**, 317-23 (1986). Case studies of molecular shape analysis to elucidate drug actions.

[86] O. Iordache, J.P. Corriou, D. Tondeur, *Hungarian Journal of Industrial Chemistry Veszprém,* **199**, 265-74 (1991). Neural network for system classification and process fault detection.

[87] A. Cartier, *QCPE*, **584**. Conversational interface to GEOMO/S program.

[88] D.J. Duchamp, *Am. Chem. Soc. Symp. Ser.,* **112**, E.C. Olson, R.E. Christoffersen Eds., Washingston DC, 79-102 (1979). Molecular mechanics and crystal structure analysis in drug design.

[89] U. Burkert and N.L. Allinger, *Molecular mechanics.* ACS monograph 177, Americal chemical society, Washington DC, 1982.

[90] N.L. Allinger, *J. Am. Chem. Soc.,* **99**, 8127 (1977). Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms.

Table des Matières

* * *

Figures

* * *

Tableaux

* * *

UNIVERSITE DE NANCY I

NOM DE L'ETUDIANT : Monsieur MATHIS Hervé

NATURE DE LA THESE : DOCTORAT DE L'UNIVERSITE DE NANCY I
en CHIMIE (chimie informatique et théorique)

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 28 OCT. 1992 n° 496

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

UNIVERSITE NANCY I LE PRESIDENT

M. BOULANGÉ

RESUME

Cette thèse présente un logiciel qui a pour but de mieux comprendre les relations entre caractéristiques structurales et propriétés thérapeutiques de molécules envisagées comme médicaments.

L'idée majeure est de soumettre une famille de composés, d'une part à des calculs de chimie quantique, d'autre part à des méthodes de reconnaissance de formes, afin d'observer si certaines propriétés moléculaires sont discriminantes vis-à-vis d'une activité pharmacologique mesurée par ailleurs. Des résultats sont détaillés, comparant les aptitudes respectives de quelques benzopyranes comme antihypertenseurs.

Outre un rappel des notions de base de la chimie quantique, les principaux aspects traités ici relèvent de la programmation (graphisme moléculaire notamment) et de l'analyse de données (classifications et statistiques).

MOTS CLES

analyse discriminante - benzopyranes - chimie quantique - classification - entropie informationnelle - graphisme moléculaire - intelligence artificielle - QSAR - reconnaissance de formes