I- INTRODUCTION GENERALE:

Découvrir de nouveaux médicaments de la manière la plus efficace et la moins coûteuse possible constitue un enjeu majeur pour les années à venir. Il est admis que, en moyenne, pour une molécule qui arrive sur le marché en tant que médicament innovant, 10 000 molécules sont synthétisées et testées. De plus, le développement d'un médicament demande généralement entre 10 et 15 ans de recherches. Il s'agit en effet de trouver une molécule qui doit à la fois présenter des propriétés thérapeutiques particulières, et posséder le minimum d'effets secondaires indésirables. Le prix de revient d'un médicament est essentiellement dû à ces synthèses longues et coûteuses.

Pour cette raison, l'industrie pharmaceutique s'oriente vers de nouvelles méthodes de recherche, qui consistent à prédire les propriétés et activités de molécules avant même que celles-ci ne soient synthétisées. Deux disciplines de la « chimie computationnelle » se sont développées en réponse à ce besoin : les relations structure-activité ou QSAR (Quantitative Structure-Activity Relationships) et les relations structure-propriété ou QSPR (Quantitative Structure-Property Relationships).

La découverte d'une telle relation permet de prédire les propriétés physiques et chimiques et l'activité biologique de composés, de développer de nouvelles théories ou de comprendre les phénomènes observés.

Elle permet également de guider les synthèses de nouvelles molécules, sans avoir à les réaliser, ou à analyser des familles entières de composés [1].

Les relations entre les structures des molécules et leurs activités sont généralement établies à l'aide de méthodes de modélisation par apprentissage statistique. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs, nombres réels mesurés ou calculés à partir des structures moléculaires. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée. Ces méthodes présentent cependant plusieurs inconvénients. Elles nécessitent en effet la sélection des descripteurs pertinents ainsi que leur calcul. De plus, les molécules sont des structures, qui peuvent être représentées par des graphes ; leur représentation par un vecteur de données induit donc une perte d'information, qui peut avoir une influence sur la qualité des modèles.

La première partie de ce mémoire présente une généralité sur les phénols ainsi que leurs toxicités, on présentant la méthodologie QSAR, nous rappelons les principaux types de descripteurs, ainsi que les méthodes expérimentales de mesure et les approches théoriques de calcules logP.

Une détermination expérimentale de logP n'étant pas toujours possible, par exemple pour les substances très hydrosolubles, et les substances très lipophiles , il est alors possible d'utiliser une valeur de logP déterminée par des méthodes théoriques. De nombreuses approches ont été et continuent d'être élaborées pour estimer logP. Trois programmes sur PC, disponibles dans le commerce (CLOGP, KOWWIN, DRAGON), sont fréquemment utilisés pour évaluer les risques en l'absence de données expérimentales.

Dans le premier chapitre nous allons utilisés deux méthodes différentes pour sélectionner les échantillons de calibration et de validation, la première est la division aléatoire, l'autre méthode se fait à l'aide d'algorithmes DULEX

Dans le chapitre deux de la deuxième partie après la construction des modèles par les deux descripteurs logP et pKa, sachons que les valeurs de logP sont calculées à l'aide des logiciels notés précédemment, on compare les paramètres statistiques obtenus par ces modèles.

I- Généralités sur les phénols:

Les phénols sont des alcools aryliques sur lesquels le groupe -OH est collé sur la partie hydrocarbure aromatique, le phénol est le plus simple de ces composés. La figure 1 montre quelques composés phénoliques importants. Les phénols ont des propriétés tout à fait différentes de celles des alcools aliphatiques et oléfiniques. Les composés phénoliques les plus importants ont les groupes nitro ($-NO_2$) et des halogènes (en particulier Cl) collés sur les anneaux aromatiques. Ces substituants peuvent affecter le comportement chimique et toxicologique [2].

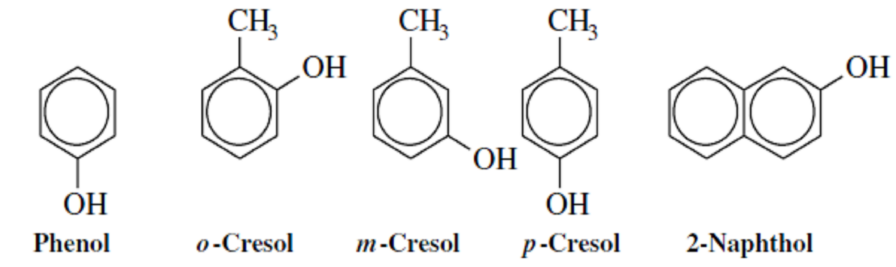
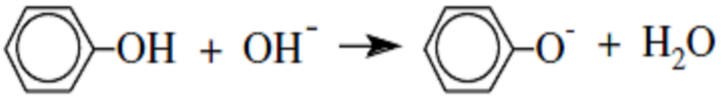


Figure -1: Les structures de quelques composés phénoliques

Les propriétés physiques des phénols schématisés dans la figure -1 sont résumées dans le tableau 1. Ces composés phénoliques sont des acides faibles qui s'ionisent sous forme d'ions phénolates en présence de base (eq 01) :



(eq 01)

Tableau-1 : Propriétés des phénols les plus importants :

Composé	**Propriété**		
		Temperature de fusion (°C)	Temperature d'ébullition(°C)
Phénol	Acide carboxilique; solide blanc; odeur caractéristique.	41	102
m-Cresol	Se trouve souvent en mélange avec l'ortho- et le para- ; liquide jaune clair.	11	203
o-Cresol	Solide.	31	191
p-Cresol	Cristal solide avec une odeur phénolique.	36	202
1-Naphtol	Alpha-naphtol; solide incolore.	96	282
2-Naphtol	Bêta-naphtol.	122	288

Les phénols sont extraits commercialement à partir du goudron sous forme de base aqueuse, comme les ions phénolates. L'utilisation commerciale principale du phénol réside dans la fabrication des résines de polymères phénoliques, généralement avec du formaldéhyde. Les phénols et les crésols sont employés comme des produits antiseptiques et désinfectants. Le phénol a été utilisé pour la première fois pour stériliser des blessures en chirurgie, (d'après le travail de Lord Lister en 1885). Le mélange utilisé était classiquement constitué de 3 mL de phénol à 50 %, de 2 mL d'eau et de 8 gouttes de savon et 8 gouttes d'huile de croton. En application cutanée, ce mélange permettait une dépigmentation. Dans ce type d'utilisation, il a été rapporté que plus de 30 % des adultes présentaient des *dysrythmies* (Morrison *et al.* 1991) [3]. Le seul cas publié dans la littérature correspond à l'utilisation d'un mélange de composition voisine (40 % de phénol, 0,8 % d'huile de croton dans du savon à base d'hexachlorophène et d'eau) chez un enfant âgé de 10 ans. Ce mélange a été appliqué sous anesthésie sur 1,9 % de la surface corporelle. Cinquante cinq minutes après le traitement, des extrasystoles ventriculaires polymorphes sont observées en l'absence de modification de la pression artérielle et des concentrations en sodium et potassium plasmatiques (Warner et Harper, 1985) [4].

Toxicologie des phénols

Généralement, les phénols possèdent les mêmes effets toxicologiques. Le phénol est un poison protoplasmique, il peut endommager toutes sortes de cellules. La dernière étude médicale a démontré que la désinfection avec le phénol cause un nombre étonnant d'intoxications [5].

Des doses mortelles de phénol peuvent être absorbées par la peau. Ses effets toxicologiques sont aigus principalement pour le système nerveux central. La mort peut se produire dans la demi-heure qui suit après l'exposition; les principaux organes endommagés par l'exposition périodique au phénol sont la rate, le pancréas, et les reins et peut être les poumons [6].

Méthodologie QSAR

Méthodologie QSAR :

I- Introduction:

Le besoin de mesurer l'impact des polluants sur l'environnement est en constante augmentation. La mesure de cet impact nécessite de connaître non seulement la toxicité des produits chimiques rejetés, mais aussi celle des molécules issues de leur dégradation. La quantité de mesures à effectuer pour mesurer ces toxicités est donc importante, ce qui augmente les coûts et les délais de développement de nouveaux produits chimiques. Une alternative à la mesure systématique de la toxicité de tels composés sur des animaux est le recours à un modèle, pour prédire l'activité de molécules appartenant à une famille donnée.

Une base de données relative à la toxicité a été compilée par l'Agence pour la Protection de l'Environnement Américaine (EPA) [07]. Cette base, nommée ECOTOX, recense les toxicités connues de molécules diverses sur la vie aquatique, les animaux terrestres et les plantes. L'OCDE a en particulier établi une base dans le but de développer un modèle capable de prédire l'activité toxique de molécules.

Deux méthodes de modélisation peuvent dès lors être mises en place. La plus directe consiste à établir un modèle global, valable pour toutes les classes de molécules. Il semble cependant plus approprié d'utiliser des modèles distincts pour modéliser des phénomènes ou des mécanismes différents.

L'approche qui consiste à établir un modèle pour chaque mode d'action conduit ainsi à des résultats plus précis. Elle nécessite cependant deux étapes : dans un premier temps, chaque molécule est affectée à une classe donnée ; sa toxicité est ensuite prédite grâce au modèle développé pour cette classe.

Ces données ont fait l'objet de prédictions à l'aide d'autres méthodes, telles que la régression par les moindres carrés partiels [08], le logiciel ECOSAR [09], une méthode de partition floue adaptative (AFP) [10], ou les réseaux de neurones probabilistes [11]. Les descripteurs le plus souvent retenus sont le coefficient de partage octanol-eau, le pKa,...

Il est généralement admis que la toxicité de nombreuses substances, particulièrement les produits chimiques organiques industriels, est la conséquence de leur solubilité dans les lipides, alors que leurs caractéristiques moléculaires spécifiques ont peu ou pas d'influence. Leur mode d'action consisterait en la destruction des processus physiologiques associés aux membranes cellulaires.

II- Histoire Des QSAR :

Les premiers essais de modélisation d'activités de molécules datent de la fin du 19ème siècle, lorsque Crum-Brown et Frazer [12] postulèrent que l'activité biologique d'une molécule est une fonction de sa constitution chimique C (eq. 2).

$$\phi = f(C) \qquad \text{(eq. 02)}$$

Richet [13] a découvert que la toxicité des composés organiques suit inversement leur solubilité dans l'eau. Un tel rapport correspond à l'eq.2, où les $\Delta\phi$ représentent des différences entre les valeurs des activités biologiques, qui sont causées par les changements de ces propriétés chimiques et particulièrement les propriétés physico-chimiques, ΔC.

$$\Delta\phi = f(\Delta C) \qquad \text{(eq. 03)}$$

Aujourd'hui il n'y a aucune méthode qui applique l'eq. 02 pour traiter les données biologiques. Toutes les équations QSAR correspondent à l'eq. 03, parce que les différences dans l'activité biologique sont seulement quantitativement corrélées avec des changements de lipophilie et/ou d'autres propriétés physico-chimiques des composés.

On peut considérer l'année 1964 comme l'année de naissance de la méthodologie QSAR moderne. Deux articles ont été publiés, un par Hansch et Fujita intitulé " *method for the correlation of biological activity and chemical structure"* [14], l'autre par Free et Wilson portant pour titre *"A mathematical contribution to structure activity studies"* [15].
Les deux contributions ont commencé par élaborer deux nouvelles méthodes pour quantifier la relation Activité Biologique / Structure (QSAR) appelée "Hansch analysis" (linear free energy-related approach, extrathermodynamic approach) et Free "Wilson analysis", respectivement.

L' approche QSAR résulte de la combinaison de différents paramètres physico-chimiques de façon linéaire additive (eq. *04 ;* log (1/C) est le logarithme de l'inverse de la dose molaire qui produit ou empêche une certaine réponse biologique, logP est le logarithme du coefficient de partage de n-octanol/ eau). D'autres méthodes utilisent un paramètre connu sous le nom de paramètre de lipophilie calculée π (eq. 5), il est employé au lieu des valeurs mesurées de logP (comme les valeurs σ de Hammett sont employées au lieu des constantes d'équilibre des réactions organiques), et la formulation d'une équation parabolique pour la description quantitative non-linéaire des rapports lipophilie-activité (eq. 6) [16,17].

$$\log 1/C = a \text{ og } P + b\sigma + \ldots + \text{const.} \quad \text{(eq 04)}$$

$$\pi_{x} = \log P_{R-X} - \log \ddot{P}_{R-H} \quad \text{(eq 05)}$$

$$\log 1/C = a(\log P)^2 + b \log P + c\,\sigma + \ldots + \text{const} \quad \text{(eq 06)}$$

D'après la contribution signifiée par Fujita et Ban [18], le modèle libre de Wilson est défini par l'eq. 07, où a_{ij} est la contribution du groupe substituant $\mathbf{X_I}$ en position j, μ est la valeur de l'activité biologique (théorique) d'un composé référence dans la série ; toutes les contributions de groupe a_{ij} des différents substituants $\mathbf{X_I}$ se rapportent aux substituants correspondants (le plus souvent l'hydrogène) pour ce composé de référence.

$$\text{Log } 1 / c = \sum \text{ aij} + \mu \quad \text{(eq.07)}$$

III- Les Modèles QSAR/QSPR

Au cours des décennies passées, les Relations Quantitatives Structure- Activité/ Propriété (QSAR) sont devenues un puissant outil théorique, alternatif à la mécanique quantique, pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements. L'approche QSAR procède de l'hypothèse d'une correspondance univoque entre n'importe quelle propriété physique, affinité chimique, ou activité biologique d'un composé chimique et sa structure moléculaire [19]. Cette dernière peut être représentée par la composition chimique, la connectivité des atomes, la surface d'énergie potentielle, et la fonction d'onde électronique d'un composé. Différents descripteurs moléculaires physico- chimiques reflétant la structure peuvent être déterminés empiriquement ou en utilisant des méthodes théoriques et computationnelles de différentes complexités. Il est à souligner que la connaissance de la constitution chimique exacte et/ou de la structure moléculaire tridimensionnelle des composés chimiques étudiés est un pré- requis à l'application de l'approche QSAR.

Le succès de l'approche QSAR dépend de façon critique de la définition précise et de l'utilisation appropriée des descripteurs moléculaires. On distingue, arbitrairement, **les descripteurs moléculaires empiriques** des **descripteurs moléculaires théoriques**.

Les descripteurs empiriques peuvent être divisés en deux classes générales (tableau 1), la première reflète les interactions électroniques intramoléculaires **(descripteurs structurels)** alors que la seconde tient compte des interactions intermoléculaires dans les milieux condensés tels que les liquides et les solutions **(descripteurs de solvatation)**.

Tableau -2 : Classification d'ensemble des descripteurs moléculaires empiriques

classe	Sous- classe
Descripteurs structurels	- Constantes d'induction - Constantes de résonance - Constantes stériques
Descripteurs de solvatation	- Echelles de polarité - Echelles de polarisabilité - Echelles d'acidité - Echelles de basicité - Echelles mixtes

Les descripteurs structurels les plus répandus ont été définis pour quantifier les propriétés d'induction, l'effet mésomère ou de résonance, et les effets stériques des composés chimiques. Les descripteurs de solvatation reflètent les interactions du soluté avec la masse du solvant environnant **(effets de solvant macroscopiques** ou **non spécifiques)**, et les liaisons spécifiques, souvent des liaisons hydrogène entre le soluté et les molécules individuelles de solvant **(effets de solvant spécifiques** ou **microscopiques)**. Les effets de solvant macroscopiques sont quantifiés en utilisant diverses échelles de polarité et de polarisabilité. Les descripteurs des effets de solvant microscopiques impliquent les échelles générales d'acidité et de basicité. Certaines échelles empiriques d'effets de solvant (échelles mixtes) peuvent impliquer en même temps ces deux effets macroscopique et microscopique. Le coefficient de partage octanol/ eau, log P, est le représentant typique de tels descripteurs.

Les descripteurs moléculaires théoriques peuvent, conventionnellement, être répartis en un certain nombre de classes, selon leur complexité ou leur méthode de calcul. Les descripteurs théoriques les plus simples sont des **descripteurs constitutionnels** qui peuvent être construits à partir de l'information sur la composition chimique du composé considéré. Les nombres, absolus et relatifs, des différents types d'atomes et de liaisons chimiques, la masse molaire, et le nombre de différents cycles dans le composé représentent quelques descripteurs constitutionnels typiques. **Les descripteurs**, **ou indices**, **topologiques** décrivent la connectivité des atomes dans la molécule. On a avancé [19] que les indices topologiques pouvaient encoder des interactions moléculaires subtiles et non pas seulement renseigner sur le degré de ramification des liaisons chimiques ou la distribution de la masse spécifique dans la molécule. **Les descripteurs géométriques** sont obtenus à partir de la structure tri-

dimensionnelle des molécules définie par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée, par exemple, par le rayon atomique de Van der Waals. Les molécules de la plupart des composés chimiques possèdent une certaine flexibilité conformationnelle et les surfaces de potentiels moléculaires respectives possèdent de multiples minima locaux. Selon la structure de la molécule, le nombre de ces minima peut être très grand et, par conséquent, il est plutôt difficile de trouver le minimum d'énergie global pour des conditions expérimentales établies.

Evidemment, les descripteurs géométriques peuvent varier de façon significative selon les conformations utilisées dans le calcul de ces descripteurs. Dans une certaine mesure, **les descripteurs théoriques liés à la distribution de charge** peuvent également dépendre de la conformation. Ces descripteurs sont basés sur la structure tri- dimensionnelle et la distribution des charges dans la molécule. Ces dernières peuvent se présenter comme charges atomiques partielles obtenues à partir d'un schéma empirique ou en utilisant des fonctions plus sophistiquées basées sur la fonction d'onde de la molécule calculée par la chimie quantique.

Un certain nombre de **descripteurs quanto- chimiques basés sur les OM** ont été employés dans le développement d'équations QSAR. Les plus utilisés sont les énergies des OM frontières, c'est-à-dire, l'énergie calculée de la plus basse orbitale moléculaire inoccupée ($\varepsilon_{LUMO)}$, et l'énergie de la plus haute orbitale moléculaire occupée (ε_{HOMO}), et la différence entre ces énergies. De même, différents indices de réactivité déduits de la théorie de la superdélocalisabilité de Fukui ou d'autres constructions théoriques ont gagné en popularité parmi les chercheurs.

Tous les descripteurs théoriques ne peuvent être strictement classés selon le schéma présenté dans le tableau 2. Par exemple, les indices topographiques sont déduits de l'information contenant à la fois la topologie et la géométrie des molécules. **Les indices électrotopologiques** sont fondés sur la topologie et la distribution de charge alors que les aires de surfaces partielles chargées sont des descripteurs qui encodent à la fois la distribution de charge et la géométrie des molécules. De tels descripteurs peuvent être classés comme **descripteurs moléculaires mixtes ou combinés**.

Les descripteurs moléculaires peuvent être définis pour tout le système moléculaire étudié ou pour n'importe laquelle de ses parties (fragments). Par exemple, la majorité des descripteurs empiriques structurels sont reliés à des fragments moléculaires appelés substituants. En conséquence, les molécules d'une série congénère de composés chimiques sont divisées formellement en deux ou plusieurs fragments qui correspondent à une unité

structurale constante Y (c'est-à-dire le centre de réaction) et à des unités structurales variables Xi (les substituants). Les relations QSAR/QSPR sont ainsi présentées comme suit :

$$P = P_0^{(Y)} + \sum_i \sum_k \alpha_{ik}^{(Y)} D_{ik}^{(X)} \qquad \text{(eq. 08)}$$

Où $P_0^{(Y)}$ est l'ordonnée à l'origine correspondant au fragment moléculaire constant Y, les $D_{ik}^{(X)}$ sont les descripteurs moléculaires de type k pour les fragments variables X_i, et les $\alpha_{ik}^{(Y)}$ sont les coefficients de développement caractéristiques d'une série donnée de composés $X_i Y$.

Tableau -3 : Classification générale des descripteurs moléculaires théoriques

classe	Sous- classe
Descripteurs constitutionnels	- Dénombrement des atomes ou des liaisons. - Descripteurs basés sur les masses atomiques.
Descripteurs topologiques	- Indices topologiques (connectivité). - Descripteurs théoriques d'information. - Descripteurs topochimiques.
Descripteurs géométriques	- Descripteurs liés à la distance. - Descripteurs liés à l'aire de la surface. - Descripteurs liés au volume. - Descripteurs du champ stérique moléculaire.
Descripteurs liés à la distribution de charge	- Charges atomiques partielles. - Moments électriques moléculaires - Polarisabilités moléculaires. - Descripteurs du champ électrique moléculaire.
Descripteurs liés aux orbitales moléculaires	- Energie des OM frontières - Ordres de liaison - Indices de réactivité de Fukui.
Descripteurs température dépendants	- Fonctions thermodynamiques. - Descripteurs facteurs de Boltzmann pondérés.
Descripteurs de solvatation	- Energie électrostatique de solvatation. - Energie de dispersion de solvatation. - Enthalpie libre de formation de cavité. - Descripteurs de liaison hydrogène. - Entropie de solvatation. - Descripteurs d'énergie de solvatation linéaire théorique.
Descripteurs mixtes	- Descripteurs topographiques. - Descripteurs électrotopologiques. - Descripteurs de la charge partielle de l'aire de la surface.

La plupart des descripteurs théoriques qui apparaissent dans le tableau 2 peuvent être calculés soit pour la molécule entière soit pour un fragment moléculaire pré- défini.

IV- Méthodes utilisées pour le développement de modèles QSAR/QSPR

IV-1- Introduction

L'application pratique des gammes des descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche aisée [19]. Tout d'abord, un très grand nombre (>3000) de descripteurs moléculaires, de différentes complexités et de conceptions diverses ont été imaginés et proposés au cours des (60 dernières) années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie, ni même proposée, pour la sélection de descripteurs adaptés parmi la myriade de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition.

Une autre difficulté dans la sélection des descripteurs QSAR découle de la non standardisation des gammes de descripteurs. Les gammes empiriques des constantes d'induction, de résonance et d'effet stérique des constituants, ou les échelles empiriques d'effets de solvant comportent des erreurs intrinsèques liées aux erreurs respectives des mesures expérimentales. Par ailleurs, les méthodes quanto- mécaniques appliquées aux calculs des descripteurs moléculaires et aux distributions de charges liés aux OM sont souvent basées sur différents paramètres semi- empiriques, ou l'utilisation de différents ensembles de base dans les calculs ab- initio. Naturellement, un descripteur construit à l'aide de différentes méthodes expérimentales ou théoriques, pour divers composés, ne peut être utilisé pour le calcul d'un modèle QSAR unique. Une approche systématique pour la sélection de gammes de descripteurs pour le calcul de modèles QSAR est basée sur la discrimination statistique entre de larges ensembles de descripteurs.

Dans ce qui suit nous passerons en revue diverses approches utilisées pour le développement des " meilleures" équations QSPR dans de grands espaces de descripteurs.

En dernier ressort, les modèles QSAR peuvent être développés selon des modèles mathématiques différents, généralement en relation avec l'analyse statistique multivariée. Le premier modèle, et le plus largement utilisé, consiste en une équation (multi) linéaire obtenue par régression des données expérimentales en fonction d'un ensemble de descripteurs pré-sélectionnés (ou d'un seul), en utilisant la méthode des moindres carrés ordinaires (MCO). Dans quelques cas, les modèles physiques ou chimiques connus du phénomène étudié laissent prévoir certaines formes mathématiques non linéaires (exponentielles ou logarithmiques) de la dépendance entre les données expérimentales et les descripteurs moléculaires. Les modèles

QSAR peuvent alors être établis à l'aide de la technique de régression par les moindres carrés non linéaires. D'autres modèles ont été développés en utilisant l'analyse factorielle ou l'analyse en composantes principales. L'intérêt de ces méthodes est qu'elles évacuent le problème de multicolinéarité inhérent aux méthodes de régression linéaires. Cependant, l'interprétation des équations QSAR est alors entravée par la nature formelle des facteurs ou des composantes principales. Une alternative aux méthodes très classiques de régression linéaire multiple (RLM) et d'analyse en composantes principales (ACP) est la technique de régression par les moindres carrés partiels (MCP ou PLS) [20-36].

IV-2- Méthodes de régressions linéaire et multilinéaire

IV-2-1 Aperçu général

Comme signalé auparavant, l'investigateur choisit dans chaque cas un ou plusieurs descripteurs supposé(s) refléter les interactions physiques ou chimiques à la base de la propriété moléculaire ou de la caractéristique du phénomène étudié. Ce choix, encore une fois, est habituellement fondé sur l'intuition chimique, la tradition, ou simplement la disponibilité du descripteur. Néanmoins, cinq principes peuvent aider à la sélection de descripteurs moléculaires convenables pour l'établissement de modèles QSAR. Ce sont:

a) Un nombre maximal de données expérimentales (de préférence toutes) doivent être caractérisées par des valeurs de descripteurs originaux complémentaires.
b) Les valeurs des descripteurs doivent être obtenues de la même source et, de préférence, mesurées selon le même protocole expérimental ou calculées en utilisant le même logiciel.
c) Le nombre de descripteurs dans les modèles de régression multiples doit être minimisé, sans perte d'information, ce que mettent en évidence les critères statistiques (valeurs des tests t et F…).
d) Dans les modèles RLM, les descripteurs utilisés doivent être statistiquement orthogonaux.
e) Pourvu que les autres critères soient similaires, la nature physique ou chimique du descripteur sélectionné doit être la plus proche de la propriété ou du phénomène étudié.

En réalité, il est difficile de se conformer pratiquement aux 5 principes énoncés. Cependant, la négligence de plusieurs d'entre eux peut conduire à des équations inutiles sans aucun pouvoir prédictif sinon très limité.

II-2-2- Evaluation préliminaire des données

Avant d'entamer le développement effectif des équations de régression QSAR et QSPR, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes).

On distingue habituellement dans un tel pré- traitement des données les analyses univariées des analyses bivariées [31-36].

Dans l'analyse univariée, il est recommandé de vérifier la conformité des données à la distribution normale. Une précaution particulière doit être prise lors de la procédure de régression subséquente si les valeurs de la propriété étudiée, ou d'un descripteur, ne suivent pas la loi de Laplace- Gauss.

Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bivariée, c'est-à-dire de calculer le coefficient de corrélation linéaire R entre chacune des paires de l'ensemble des descripteurs. Si R est statistiquement significatif (R>0,95), ces deux descripteurs ne peuvent être utilisés simultanément lors de l'analyse par RLM.

II-2-3- Régression linéaire multiple

Un modèle de régression linéaire multiple entre une variable expliquée Y et p variables explicatives X_1,……, X_p, s'écrit pour tout i=1,……, n :

$$yi= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon \qquad \text{(eq.09)}$$

ou les y_i, x_{i1}, x_{i2},…, x_{ip} sont des données respectivement relatives aux variables Y, X_1,……., X_p.

Les estimateurs β_j sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ε_i représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires e_i, différence entre les valeurs observées y_i et les valeurs estimées $\hat{y}_i$.

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses :

a) Les résidus (**e**) ont une espérance mathématique nulle :

$$E(\mathbf{e}) = \mathbf{0} \qquad \text{(eq.10)}$$

b) Le modèle choisi est correct (aucune variable explicative n'a été omise).

c) Les résidus sont indépendants entre eux :

$$E(e_i, e_j) = 0 \qquad \text{si } i \neq j \qquad \text{(eq. 11)}$$

leurs covariances sont nulles.

d) Les résidus ont tous même variance σ^2 (propriété d'homoscédasticité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que :

e) Les résidus suivent une distribution normale (de Laplace- Gauss).

L'analyse des résidus présente un intérêt à plusieurs égards. Elle permet en effet de vérifier, a posteriori, la validité du modèle utilisé, en ce qui concerne, d'une part la forme de celui- ci (linéarité ou non linéarité de la relation, par exemple) et d'autre part, certaines hypothèses plus spécifiques, telles que l'égalité des variances résiduelles, la normalité des résidus ou l'absence d'auto- corrélation.

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 4 à 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degré de liberté final (n-p-1) doit être [37] tel que :

$$n- p- 1 \geq 10 \qquad \text{(eq. 12)}$$

n étant la dimension de l'échantillon, et p le nombre de variables explicatives entrant dans la construction du modèle.

Pour les modèles à plus de deux descripteurs, de faibles coefficients de corrélation croisés n'assurent pas forcément l'orthogonalité des descripteurs. Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux obéissent [37] à la condition FIV< 5.

Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \text{(eq. 13)}$$

Où $\bar{y}$ est la valeur moyenne des valeurs observées pour l'ensemble de calibrage.

- La racine de l'écart quadratique moyen de calcul :

$$\sigma_N = EQMC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}} \qquad \text{(eq. 14)}$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen de prédiction (EQMP), et celle calculée sur l'ensemble de validation externe (EQMPext) :

$$EQMP = \sqrt{\frac{1}{n} \sum_{1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2} = \sqrt{\frac{PRESS}{n}} \qquad \text{(eq. 15)}$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i \right)^2}{n_{ext}}} \qquad \text{(eq. 16)}$$

La validation croisée par « leave – one - out » (LOO) [38] consiste à recalculer le modèle sur (n-1) observations, et à utiliser le modèle ainsi obtenu pour calculer la grandeur d'intérêt du composé écarté, notée $\hat{y}_{(i)}$. On répète le procédé pour chacune des grandeurs

d'intérêt. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (eq.15), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction [38] :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad \text{(eq. 17)}$$

Contrairement à R^2 qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [39].

Si de petites valeurs de Q^2_{LOO} indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q^2_{LOO} est une condition nécessaire de robustesse et d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante, et peut conduire à une surestimation de la capacité prédictive du modèle lorsqu'il est appliqué à des composés réellement externes.

Evidemment, on peut être amené à écarter 2, 3 ou un plus grand nombre d'éléments à la fois, ce qui conduit aux procédures LMO (leave – many- out).

Dans le cas où on a suffisamment de données qui n'ont pas servi dans la création du modèle ou après collecte de nouvelles, on peut ou on doit procéder à la validation de ce dernier, c'est la validation externe. La statistique se rapportant à ce procédé, notée Q^2_{ext}, est calculée comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_{(i)}\right)^2 \Big/ n_{ext}}{\sum_{i=1}^{n_{tr}} \left(y_i - \bar{y}_{tr}\right)^2 \Big/ n_{tr}} \qquad \text{(eq. 18)}$$

Pour une grande valeur de Q^2_{LOO}, une valeur élevée de Q^2_{ext} permet de présager d'une bonne capacité prédictive du modèle.

La validation interne peut être également réalisée en utilisant la technique du bootstrap : Q^2_{boot} (bootstrapping). Elle consiste à simuler m échantillons de même taille n que l'échantillon initial. Ils sont obtenus par tirage au hasard avec remise parmi les n individus

observés au départ, ceux-ci ayant tous la même probabilité 1/n d'être choisis [38,40]. Contrairement aux validations croisées par LOO et LMO, les méthodes de bootstrap sont plus efficaces et plus stables.

VI- 3- Sélection des deux ensembles (validation et calibration)

Il existe différentes méthodes de sélection des échantillons de calibration et de validation, la plus simple est la division aléatoire en deux ensemble disjoints. Une autre méthode de partage des données se fait à l'aide d'algorithmes (DULEX, CADEX, OPIST).
Dans ce travail la méthode aléatoire a été confrontée à l'éclatement par l'algorithme DUPLEX, que nous avons écrit.

Notre écriture du programme pour l'algorithme DUPLEX s'appuie sur sa description faite par R.D. Snee [41] qui mit au crédit de R.W. Kennard son développement. L'édition et l'exécution de ce programme utilisent l'environnement du logiciel MATLAB [42].

L'algorithme DUPLEX commence par une transformation des données (variables explicatives seulement) qui est faite comme suit :

1. Standardisation des données par la formule : $z_{ij} = \dfrac{x_{ij} - \overline{x}_j}{s_j \sqrt{(n-1)}}$

   s_j : Ecart-type du *j ème* régresseur.

   $\overline{x}_j$: Moyenne du *j ème* régresseur

   x_{ij} : Valeur du régresseur *j* pour la *i ème* observation.

   n : Nombre d'observations.

   Pour les i = 1, 2, 3… n et j=1, 2, 3… k ; les éléments de la matrice $\boldsymbol{Z}$ sont ainsi calculés.
2. On calcul la matrice symétrique définie positive $\boldsymbol{A = Z'Z}$.
3. La factorisation de Cholesky peut être maintenant faite pour le calcul de la matrice $\boldsymbol{B}$ telle que $\boldsymbol{A = B'B}$.
4. Les nouvelles coordonnées des n observations c'est-à-dire les $\boldsymbol{w}$ sont obtenues par le calcul de la matrice $\boldsymbol{W = ZB^{-1}}$

Une fois la transformation faite on utilise les points orthonormalisés pour calculer les distances euclidiennes entre les paires possibles de points. La distribution des points ou leur

éclatement en un ensemble de prédiction et un second d'estimation se fait en utilisant ces distances de la manière suivante :

i. La distance la plus grande correspond à la paire de points que l'algorithme classe comme points d'estimations, ces points (E1 et E2) sont ensuite éliminés.

ii. Les deux points (P1 et P2) les plus éloignés, dans les $(n-2)$ restants, sont placés dans l'ensemble de prédiction puis automatiquement éliminés des $(n-2)$ points.

iii. L'algorithme tient compte seulement des distances des $(n-4)$ points par rapport aux points d'estimation préalablement choisis (E1 et E2).Chaque point M à deux distances une par rapport à E1, ME1, l'autre, ME2, est celle à E2 ; mais M sera caractérisé par la plus petite distance entre elles. La plus grande distance entre ces $(n-4)$ petites distances, qui caractérisent ces $(n-4)$ points, identifie le point E3 qui est par conséquent le plus éloigné de la paire (E1, E2) et est, subséquemment à cela, classé avec eux dans l'ensemble d'estimation et éliminé des $(n-4)$ point.

iv. Pour le choix du point P3 de l'ensemble de prédiction, qui sera éliminé des $(n-5)$ observations avant de procéder à l'étape (v); l'algorithme continue et refait l'étape (iii) mais en utilisant cette fois la paire (P1 et P2) comme référence pour les $(n-5)$ points restants.

v. Le processus computationnel se poursuit en plaçant les points alternativement dans un ensemble ou dans l'autre. Chaque point (Ei ou Pi) assigné à un ensemble sera éliminé et contribuera au choix du point suivant, car la distance de ce point (Ei ou Pi) au ($\left(n-(i+3)\right)$ ou $\left(n-\left(i+4\right)\right)$ respectivement) points restants sera prise en compte pour caractériser ces derniers. Par exemple ; le point P3 jouera un rôle, avec P1 et P2 bien sure, pour trouver le point P4 parmi les $(n-7)$ points restants.

Les données en main peuvent être ainsi fractionnées, et dans n'importe quel ratio, en un ensemble de prédiction et un autre d'estimation en spécifiant le nombre de point requis que nous jugerons convenable à notre étude.

Pour Snee des précautions particulières doivent être prises avant de scinder les données par cette procédure :

- Le nombre d'observations n doit être supérieur ou égale à $(2k+26)$ si l'ont veut des ensembles d'égales dimensions.

- L'ensemble de prédiction doit contenir au moins 15 observations pour un contrôle rigoureux du pouvoir prédictif du modèle par les statistiques usuelles (Q^2_{ext}, $SDEP_{ext}$).
- Des points répliques d'autres points, ou des points étant des proches voisins, doivent en être purgée nos données d'origines avant tout traitement par DUPLEX.

V- Calcul des descripteurs moléculaires

La représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation QSAR/QSPR. Les performances du modèle élaboré et la précision des résultats sont étroitement liées au mode de détermination de ces descripteurs.

Nous avons utilisé le logiciel de modélisation moléculaire Hyperchem 6.03 [43] pour représenter les molécules puis, à l'aide de la méthode semi-empirique AM1, obtenir les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF (pour restricted Hartree-Fock ou formalisme de Hartree-Fock avec contrainte de spin) sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,1 kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique DRAGON [44] pour le calcul de AlogP, MlogP.

Nous avons aussi utilisé les deux logiciels Ultra CLOGP CampridgeSoft [45], et KOWWIN [46], pour les calculs des descripteurs ClogP, et $\log K_{owwin}$ successivement.