1.5.2 Méthodes de recherche de médicaments

L'apparition de nouvelles technologies a bouleversé la recherche de nouveaux médicaments dans sa phase initiale. Celle-ci inclut tout d'abord la synthèse et l'isolement de nouvelles molécules puis leur essai sur des systèmes biologiques permettant de présupposer d'un intérêt thérapeutique éventuel. Cette phase était classiquement longue et pénible. La synthèse chimique relevait d'un art difficile ; au départ, le choix d'une structure de base se faisait sans guide. Les essais sur les animaux entiers ou les organes isolés étaient longs et complexes. Au total, malgré des progrès au fil des années, le processus relevait plus de la " pêche à la ligne " que de la démarche rationnelle.

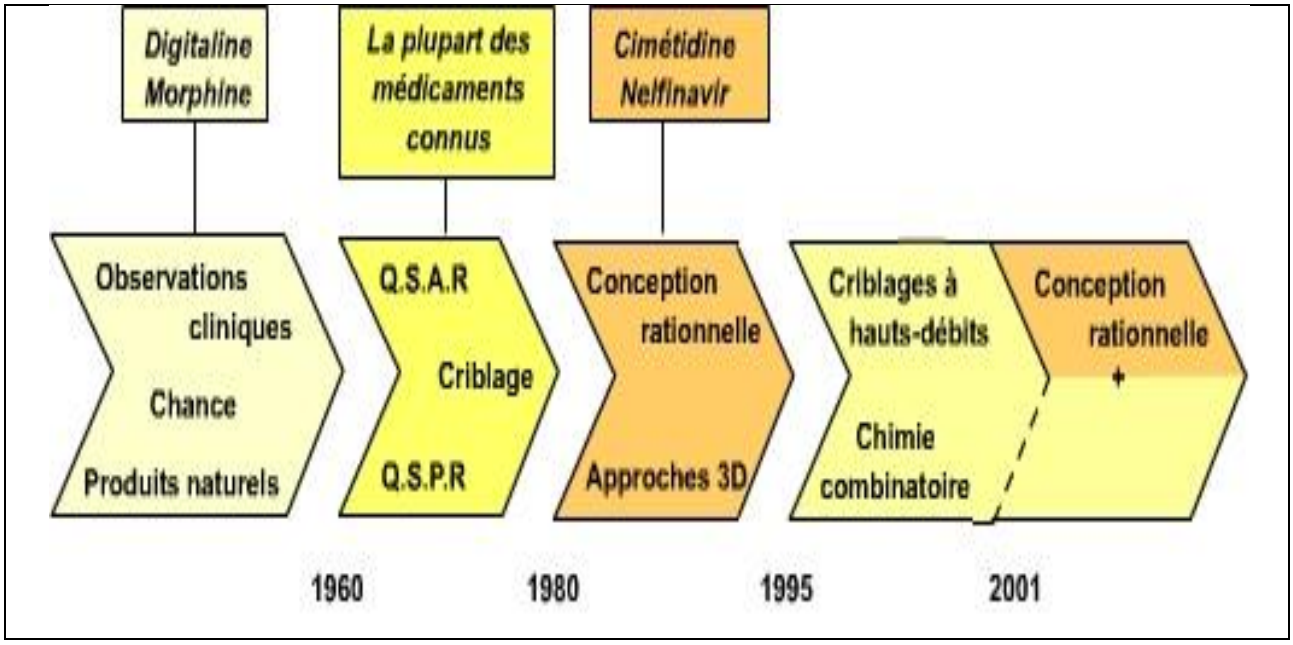

Fig.1.6 : Les approches de recherche des médicaments.

Trois approches ont profondément transformé cette recherche [13]:

1.5.2.1 Les techniques conformationnelles

La théorie des récepteurs postule que c'est l'union de la molécule de médicament avec une macromolécule qui est à l'origine de l'effet pharmacodynamique et plus généralement de la réponse thérapeutique. Cette union est fortement spécifique : seules quelques molécules privilégiées en sont capables. On dit que le médicament est comme " la clé dans la serrure ". On sait maintenant déterminer la conformation dans l'espace des protéines, notamment grâce à la radiocristallographie aux rayons X, donc celle des récepteurs. On peut donc prévoir quelles structures devront présenter les molécules pour pouvoir s'unir à eux. Cette recherche est aidée par les programmes informatiques qui permettent de visualiser les molécules et de les faire tourner dans l'espace (conception assistée par ordinateur).

Bien que hautement sophistiquée et évidemment plus ardue que ces quelques lignes pourraient le laisser croire, cette approche permet de ne plus s'en remettre au hasard dans la recherche des séries chimiques intéressantes. On voit, cependant, qu'il est indispensable de connaître au départ le récepteur pertinent, c'est-à-dire d'avoir une hypothèse physio-pathologique et d'avoir été capable d'identifier et d'isoler la protéine qui le porte. Là aussi, des progrès décisifs ont été faits dans l'isolement des protéines et, mieux encore, dans le repérage et le clonage des gènes qui commandent leur synthèse.

1.5.2.2 La chimie combinatoire

Il est désormais possible de synthétiser en une seule opération plusieurs centaines de molécules c'est ce que l'on appelle la chimie combinatoire. On part de la structure de base déterminée a priori comme il vient d'être dit et on génère systématiquement toutes les variations possibles en greffant des radicaux chimiques, des chaînes latérales, en modifiant le squelette, etc. Ceci se fait non plus étape par étape, mais en mettant en présence les réactifs nécessaires. On obtient ainsi d'un seul coup plusieurs centaines de molécules. Toutes les opérations, synthèse, isolement et identification, sont miniaturisées et robotisées. Le gain de temps et l'abaissement des

coûts sont considérables. On peut ainsi constituer une bibliothèque de plusieurs milliers de dérivés en quelques mois.

1.5.2.3 Le criblage à haut débit

Le problème est alors d'identifier parmi toutes ces molécules celles qui sont pourvues des propriétés biologiques les plus intéressantes. Le gain de temps et l'augmentation de la productivité apportés par la chimie combinatoire l'auraient été en vain si la productivité de cette phase de repérage appelée "criblage" n'avait pas été aussi améliorée. Aux essais longs et limités de la pharmacologie expérimentale classique, a succédé une technique qui permet d'essayer dans le minimum de temps des milliers de molécules.

Le test consiste à mettre en présence la substance à tester et un système biochimique (une enzyme par exemple) et de mesurer l'importance de la réaction éventuelle. L'essai peut être fait simultanément avec un grand nombre de systèmes, de significations très diverses. Tout dépend de ce que l'on met dans les tubes et, une fois de plus, on ne trouvera que ce que l'on cherche. Les systèmes biologiques testés ne sont pas indifférents : ce sont ceux dont on pense qu'ils interviennent de manière cruciale dans le déterminisme de la maladie. L'opération finale est celle du choix. La plupart du temps, toutes les molécules intéressantes ne peuvent pas passer en développement. Il faut donc sélectionner les plus prometteuses, compte tenu de leurs résultats aux tests (Figure 1.7).

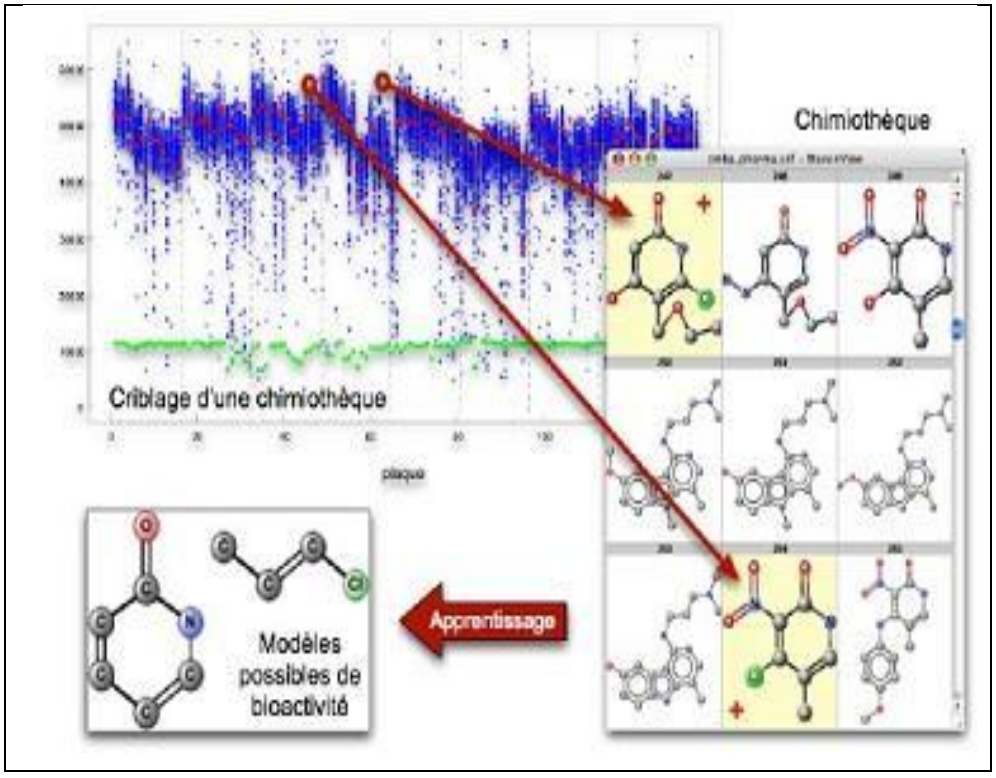

Fig.1.7 : Criblage à haut débit.

1.6 Quelques logiciels

Voici une petite sélection de logiciels pour DOS, Windows et Linux [14]:

1.6.1 Logiciels utilisés pour la représentation moléculaire

- **ChemSketch:** Editeur de formules en 2D et 3D. Optimisation géométrique par mécanique moléculaire.

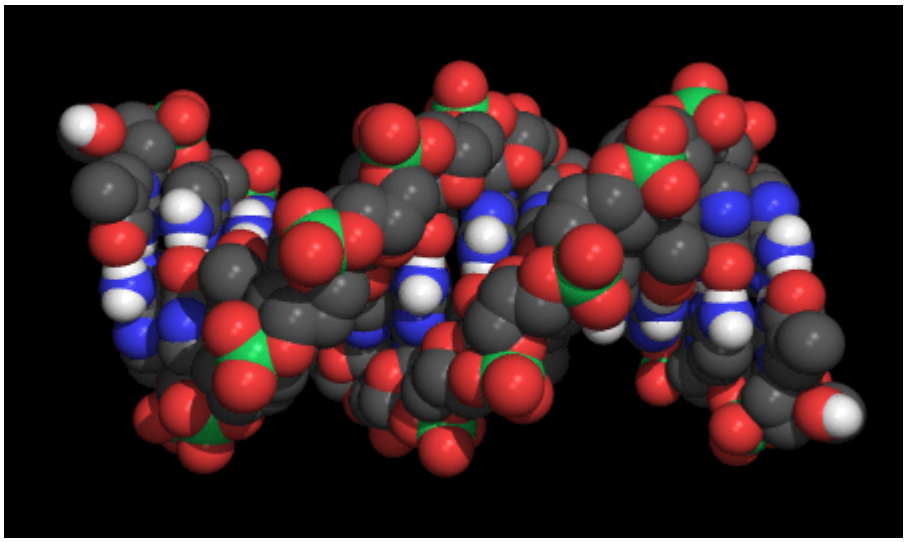

Fig.1.8 : Représentation moléculaire.

- **RasMol**: Visualiseur de molécules.
- **PovChem** : Permet de créer des images de molécules en 3D par la technique du "tracé de rayons" à l'aide du logiciel POV-Ray.
- **Logiciels pour la spectroscopie.**
- **Spartan**: Un logiciel de modélisation en trois dimensions; calcule, entre autres, des énergies, des états de transition, des conformations, offre de nombreuses possibilités de visualisation.

1.6.2 Logiciels utilisés pour les calculs de propriétés

- **MOPAC** : Calculs quantiques semi-empiriques.
- **Open Babel:** Convertisseur de fichiers de coordonnées moléculaires.
- **Tinker** : Ensemble très complet de programmes pour la mécanique et la dynamique moléculaires.
- **Vega** : Ce programme permet de calculer un grand nombre de propriétés moléculaires (volume, lipophilie...), d'analyser des trajectoires de dynamique moléculaire et de réaliser divers traitements sur les fichiers de coordonnées. (Note : la version Windows de Vega contient une copie de MOPAC, permettant de réaliser des calculs quantiques).
- **VMD**: Visual Molecular Dynamics (VMD). Visualiseur de trajectoires de dynamique moléculaire. Le site présente en outre une très belle collection d'images et d'animations.
- **Gaussian 98**: Calculateur très puissant
- **Molden**: Logiciel de visualisation de géométrie, de fréquences harmoniques, d'orbitales moléculaires, permet la représentation de résultats issus de différents logiciels de calculs de chimie.

1.7 Applications de chemoinformatique

1.7.1 Commentaires généraux

L'étendue des applications de chemoinformatique est très large ; en effet, n'importe quel champ de chimie peut profiter de ses méthodes. Nous avons dit que la chemoinformatique est l'application des méthodes d'informatique pour résoudre des problèmes chimiques.

Quel, alors, sont les problèmes principaux considérés par un chimiste ?

Il doit réaliser que la tâche principale de la chimie n'est pas tellement de produire des produits chimiques mais de produire des propriétés, les propriétés qui s'avèrent justement être attachées aux produits chimiques. La société a besoin d'une **variété de propriétés**, par exemple, pour les maladies traitantes, pour des voitures de coloration, pour construire les maisons stables, pour

coller des matériaux ensemble, pour les visages embellissant, pour des vêtements de nettoyage, etc. La première question qu'un chimiste doit répondre est :

Quelle structure j'ai besoin pour obtenir la propriété désirée ?

C'est le secteur de la **structure des propriétés ou** structure des **rapports d'activité**. Une fois qu'on a obtenu une idée sur quelle structure portera la propriété désirée, on doit répondre à la prochaine question :

Comment peux-je synthétiser cette structure ?

C'est le secteur de la **conception de synthèse**. Tout à fait une variété de problèmes doivent être résolues en concevant des synthèses efficaces, problèmes qui concerne le développement court de stratégie de synthèse et prévoir le cours des réactions chimiques.

Une fois une réaction dans un arrangement de synthèse a été exécutée, on doit répondre à la prochaine question :

Ce qui est le produit de la réaction que j'ai exécutée ?

C'est le secteur de **l'élucidation de structure**. Notre connaissance de chimie n'est pas encore assez profonde que nous pouvons toujours être sûrs que la réaction que nous exécutons prend le cours désiré. L'utilisation d'information spectroscopique doit être faite pour élucider la structure du produit de réaction. Tous ces problèmes sont trop complexes pour être résolus par des calculs basés sur les premiers principes. Ils ont tous besoin de beaucoup d'information d'être traité et profondément la connaissance chimique. À ce sens, cependant, les méthodes de chemoinformatique peuvent aider à répondre à ces trois questions fondamentales [13].

1.7.2 Applications de chemoinformatique par secteurs de chimie

Quelques applications typiques de chemoinformatique dans différents secteurs de chimie sont énumérées ci-dessous. Il doit être souligné que cette liste est loin d'être complet [15]!

1.7.2.1 L'information chimique

- Stockage et récupération des structures chimiques et des données associées pour contrôler la pléthore de données.
- Diffusion des données sur l'Internet.
- Édition absolue des données à l'information.

1.7.2.2 Tous les domaines de chimie

- Prévision des propriétés physiques, chimiques, ou biologiques des composés.

1.7.2.3 Chimie analytique

- Analyse des données de la chimie analytique pour faire des prévisions sur la qualité, l'origine, et l'âge des objets étudiés.
- Elucidation de la structure d'un composé basé sur des données spectroscopiques.

1.7.2.4 Chimie organique

- Prévision du cours et des produits des réactions organiques.
- Conception des synthèses organiques.

1.7.2.5 Conception de médicament

- Etablissement de la relation structure-activité.
- Comparaison des bibliothèques chimiques.
- Définition et analyse de diversité structurale.
- Planification des bibliothèques chimiques.
- analyse des voies biochimiques.

Le domaine de chemoinformatique est loin entièrement d'être développé. Il y a beaucoup de secteurs et problèmes qui peuvent encore tirer bénéfice de l'application des méthodes de chemoinformatique. Il y a beaucoup d'espace pour l'innovation en cherchant de nouvelles applications et en développant de nouvelles méthodes.

1.8 Applications

1.8.1 L'utilisation de QSAR et de méthodes informatiques dans la conception de médicament

Le travail de **[16]** décrit la base de QSAR moderne dans la découverte de médicament et présente quelques défis et demandes courants de découverte et d'optimisation des candidats de médicament. Les modèles de QSAR tiennent compte du calcul des propriétés physico-chimiques (par exemple, lipophilicité), de la prédiction de l'activité biologique (ou de la toxicité), aussi bien que l'évaluation de l'absorption, de la distribution, du métabolisme, et de l'excrétion (ADME). Dans la recherche pharmaceutique, QSAR a un intérêt particulier pour les étapes préclinique de la découverte de médicament pour remplacer l'expérimentation pénible et coûteuse, de filtrer de grandes bases de données chimiques, et de choisir des candidats de médicament. Cependant, pour faire partie de stratégies de découverte et de développement de médicament, le besoin de QSARs de répondre à différents critères (par exemple, predictivité suffisant). Ce travail décrit la base de QSAR moderne dans la découverte de médicament et présente quelques défis et demandes courants de découverte et d'optimisation des candidats de médicament

1.8.2 Prévisions confiantes de ségrégation des propriétés des produits chimiques pour le criblage virtuel des médicaments

Le travail de **[17]** présente une méthodologie pour évaluer la confiance en prédiction d'une propriété physico-chimique ou biologique. L'identification des prédictions incertaines de composés est cruciale pour le procédé moderne de découverte de médicament. Cette tâche est accomplie par la combinaison de la méthode de prédiction avec une carte à organisation automatique. De cette façon, la méthode peut isoler des prédictions incertaines aussi bien que des prédictions confiantes. La méthode à quatre ensembles de données différents à été appliqué, et des différences significatives dans les prévisions moyennes ont été obtenu. Cette approche constitue une nouvelle manière pour évaluer la confiance, puisqu'elle recherche non seulement des situations d'extrapolation mais également elle identifie des problèmes d'interpolation.

1.8.3 Classification de composée chimique avec les modèles de structure extraits automatiquement

Le travail de **[18]** propose de nouvelles méthodes de classification de structure chimique basée sur l'intégration de l'exploitation de la base de données de graphe et l'exploitation de données et des fonctions de noyau de graphe de la machine d'apprentissage. Dans cette méthode, ils ont d'abord identifié un ensemble de modèles généraux de graphe dans des données de structure chimique. Ces modèles sont alors employés pour augmenter une fonction de noyau de graphe qui calcule par paires la similitude entre les molécules. La matrice de similitude obtenue est employée comme entrée pour classifier les composés chimiques par l'intermédiaire des machines à noyau telles que la machine à vecteur de support (SVM). Les résultats obtenus indiquent que l'utilisation d'une approche basée-modèle pour la similarité de graphe rapporte des profils d'exécution, et parfois excédant cela des approches existantes de situation actuelle.

1.8.4 Méthode d'arbres à sortie noyau pour la prédiction de sorties structurées et l'apprentissage de noyau

Le travail présenté dans **[19]** propose une **extension** des méthodes d'arbres pour la prédiction de sorties structurées et à l'apprentissage supervisé d'un noyau. Cette extension est basée sur l'utilisation d'un noyau sur la sortie de ces méthodes qui leur permet de construire un arbre à la seule condition qu'un noyau puisse être défini sur l'espace de sortie.

Cet algorithme, appelé OK3 (pour "output kernel trees"), généralise les arbres de classification et de régression ainsi que les méthodes d'ensemble d'arbres. Il hérite de plusieurs caractéristiques de ces méthodes telles que l'interprétabilité, la robustesse aux variables non pertinentes et la résistance à l'échelle sur le nombre d'entrées. Cet algorithme donne de bons résultats sur deux problèmes de nature très différente : un problème de complétion de motif and un problème d'inférence de graphe.

1.8.5 Approche multi-classes de représentation des molécules pour la conception des produits-procédés assistées par ordinateur

La Conception de Produits Assistée par Ordinateur (CPAO) est largement utilisée dans le domaine « Process System Engineering » (PSE), comme un outil puissant pour la recherche de nouveaux produits chimiques. Les étapes cruciales de la CPAO sont la génération des molécules et **l'estimation des propriétés**, particulièrement quand les structures moléculaires complexes comme les arômes sont recherchés.

Le travail présenté dans **[20]** présente une approche multi-classes de représentation des molécules basée sur les graphes moléculaires et la connaissance chimique.

Trois catégories de groupes fonctionnels sont proposées : groupes élémentaires, groupes de base et groupes composés. Ces derniers servent à générer quatre classes de représentation qui peuvent être utiles pour la **prédiction des propriétés** et dans la conception des molécules (CAMD). La méthodologie est utile pour intégrer des méthodes de contribution de groupes dans les simulateurs où certaines molécules ne sont pas référencées.

Cette méthodologie peut être aussi utile pour le développement de méthodes de contribution de groupes se basant sur une décomposition automatique.

1.8.6 Relation structure moléculaire-Odeur (Utilisation des Réseaux de Neurones pour l'estimation de l'Odeur Balsamique)

Le travail de **[21]** présente une approche de **prédiction de l'odeur** des molécules **basée sur les descripteurs moléculaires.** Les techniques d'analyse en composantes principales (ACP) et d'analyse de colinéarité permettent d'identifier les descripteurs les plus pertinents. Un réseau de neurones supervisé à deux couches (cachée et sortie) est employé pour corréler la structure moléculaire à l'odeur. Un ensemble de paramètres est modifié jusqu'à la satisfaction de la meilleure régression. Le réseau neurologique corrèle d'une manière satisfaisante les molécules avec leur odeur assignée, basée sur des descripteurs moléculaires suffisamment nombreux et divers. Mais il ne peut pas prévoir l'odeur balsamique et ses sous-notes.

1.8.7 Une distance d'écoulement de réseau entre les graphes étiquetés

Le travail présenté dans **[22]** propose une mesure **de similarité originale** entre les graphes étiqueté **qui a des** applications à l'analyse de donnée structurée, par exemple : chemical informatics, web document clustering, etc. Les métriques exactes sur des graphes basés sur de sous-graphe isomorphisme ont été proposés plus tôt mais en raison du manque d'un efficace algorithme, ils ne peuvent pas être appliqués sur de grandes données. La métrique proposé basé sur les graphes **exploite** la similitude de contexte de sommet **et calcule** des points assortis globaux dans le temps polynômial dans la taille des graphes en utilisant une formulation d'écoulement de réseau du problème. Cette métrique est employée, dans un cadre distinctif **pour prévoir les propriétés** chimiques comme la cancérogénicité et la mutagénicité des molécules et les examiner sur des ensembles de données de PTC et de MUTAG. Les grains définis positifs construit en utilisant cette métrique présente une exécution améliorée de manière significative des grains existants finis pour des graphes sur la plupart des ensembles de données démontrant l'efficacité de la technique.

1.8.8 La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique

Le problème auquel s'intéresse le travail présenté dans **[23]** est la découverte de nouvelles familles de réactions chimiques à partir de bases de données de réactions, et montre en quoi ce problème peut se reformuler en un problème particulier de fouille de graphes. La découverte de nouvelles réactions présente un grand intérêt pour la synthèse en chimie organique, discipline dont le but est la conception de molécules complexes à partir de composants chimiques usuels et de réactions. En effet, plus un expert de la synthèse a de réactions à sa disposition, plus il peut créer de nouveaux produits à partir d'un ensemble donné de molécules et plus il peut optimiser le plan de synthèse d'une molécule cible donnée.

1.8.9 Regrouper des molécules : Influence des mesures de similitude

Le travail présenté dans [24] présente les résultats d'une étude expérimentale pour analyser l'effet de diverses mesures de similitude (ou distance) sur la qualité de regroupement d'un ensemble de molécules.il se concentre principalement sur les approches de regroupement capables de traiter directement la représentation 2D des molécules (c.-à-d., des graphes). Dans un tel contexte, il semble approprié d'employer une approche basée sur des mesures asymétriques de similitude.

Plusieurs d'autres travaux ont été énoncés dans [25].

1.9 Conclusion

Nous avons analysé le concept de chemoinformatique sous l'angle de sa modélisation. Il nous a été donné de constater que dans ce domaine, d'énorme volume d'information produite par la recherche en chimie ne peut être traitée et analysée que par les moyens informatiques.

Dans ce chapitre nous avons présenté la définition de la chemoinformatique en tant que domaine de la science qui consiste en l'application de l'informatique aux problèmes relatifs à la chimie. Par la suite nous avons présenté les concepts de bases de la chimie, ainsi qu'une vue d'ensemble de chemoinformatique, soulignant les problèmes et les solutions communs aux divers sous-domaines plus spécialisés.

Ensuite des Méthodes de la modélisation moléculaire qui ont pour but de prévoir la structure et la réactivité des molécules ou des systèmes de molécules ont été présentées, ainsi qu'une petite sélection de logiciels utilisés pour la représentation moléculaire et les calculs de propriétés.
En fin quelques applications typiques de chemoinformatique dans différents secteurs de chimie ont été énumérées.
Nous montrerons dans le chapitre suivant les techniques de modélisation par apprentissage qui ont permis la mise en place de nombreuses méthodes ; elles reposent pour la plupart sur la recherche d'une relation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriété ou l'activité que l'on souhaite prédire.

La chemoinformatique est fondamentalement basée sur la modélisation, cette dernière est donc une activité essentielle en vue d'effectué automatiquement la classification, la prédiction,…etc.

Les premiers essais de modélisation d'activités de molécules datent de la fin du 19ème siècle, lorsque Crum-Brown et Frazer **[26]** postulèrent que l'activité biologique d'une molécule est une fonction de sa constitution chimique. Mais ce n'est qu'en 1964 que furent développés les modèles de "contribution de groupes", qui constituent les réels débuts de la modélisation QSAR. Depuis, l'essor de nouvelles techniques de modélisation par apprentissage, linéaires d'abord, puis non linéaires, ont permis la mise en place de nombreuses méthodes ; elles reposent pour la plupart sur la recherche d'une relation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriété ou l'activité que l'on souhaite prédire. Nous montrerons tout d'abord comment les molécules peuvent être représentées par des vecteurs de réels, et comment ces descripteurs sont sélectionnés. Nous introduirons ensuite les outils de modélisation sans contrainte les plus utilisés, c'est-à-dire la régression linéaire multiple et la régression non linéaire à l'aide de réseaux de neurones, qui sont fondés sur le calcul de descripteurs. Nous présenterons le problème de la sélection de modèle, ainsi que les stratégies les plus efficaces pour le résoudre. Enfin, d'autres méthodes de modélisation, telles que la méthode CoMFA, mises au -point pour la modélisation d'activités biologiques, seront présentées.

2.1 Les descripteurs

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules, et ces structures elles-mêmes, en un ensemble de nombres réels appelés descripteurs ; une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété ou activité moléculaire, à l'aide d'outils de modélisation classiques. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels. On en dénombre aujourd'hui plus de 3000 types, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. Ils peuvent être obtenus de manière empirique ou non-empirique, mais les descripteurs calculés, et non

mesurés, sont à privilégier : ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est un des objectifs de la modélisation. Il existe cependant quelques descripteurs mesurés : il s'agit généralement de données expérimentales plus faciles à mesurer que la propriété ou l'activité à prédire (coefficient de partage eau-octanol **[27]**, polarisabilité, ou potentiel d'ionisation).

Avant toute modélisation, il est nécessaire de calculer ou de mesurer un grand nombre de descripteurs différents, car les mécanismes qui déterminent l'activité d'une molécule ou une de ses propriétés sont fréquemment mal connus. Il faut ensuite sélectionner parmi ces variables elles qui sont les plus pertinentes pour la modélisation.

2.1.1 Les descripteurs moléculaires

Nous allons présenter les descripteurs moléculaires les plus courants, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, mais véhiculent peu d'informations. Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

Les descripteurs 1D sont accessibles à partir de la formule brute de la molécule (par exemple C_6H_6O pour le phénol), et décrivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est-à-dire les atomes qui le constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution.

Les descripteurs 2D sont calculés à partir de la formule développée de la molécule. Ils peuvent être de plusieurs types.

- Les ***indices constitutionnels*** caractérisent les différents composants de la molécule.

Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles…

- Les ***indices topologiques*** peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Les plus fréquemment utilisés sont l'indice de Wiener **[28]**, l'indice de Randié **[29]**, l'indice de connectivité de valence de Kier-Hall **[30]** et l'indice de Balaban **[31]**. L'indice de Wiener permet de caractériser le volume moléculaire et la ramification d'une molécule : si l'on appelle distance topologique entre deux atomes le plus petit nombre de liaisons séparant ces deux atomes, l'indice de Wiener est égal à la somme de toutes les distances topologiques entre les différentes paires d'atomes de la

molécule. L'indice de Randié est un des descripteurs les plus utilisés ; il peut être interprété comme une mesure de l'aire de la molécule accessible au solvant.

Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités, telles que les activités biologiques. Des descripteurs, accessibles à partir de la structure 3D des molécules, ont pu être calculés grâce au développement des techniques instrumentales et de nouvelles méthodes théoriques.

Les descripteurs 3D d'une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes; leurs calculs nécessitent donc de connaître, le plus souvent par modélisation moléculaire empirique ou *ab initio*, la géométrie 3D de la molécule. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

- Les ***descripteurs géométriques*** les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.
- Les ***descripteurs électroniques*** permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique de molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée et de la plus basse vacante sont des descripteurs fréquemment sélectionnés.

Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.

− ***Descripteurs spectroscopiques*** : les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemples par leurs fonctions d'onde vibrationnelles.

En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci ; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. Les spectres infrarouges peuvent être obtenus soit de manière expérimentale, soit par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Le

descripteur EVA [32] est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type *MoRSE* [33] *(Molecule Representation of Structures based on Electron diffraction)* sont calculés à partir d'une simulation du spectre infrarouge ; ils font appel au calcul des intensités théoriques de diffraction d'électrons.

2.1.2 Réduction du nombre de variables

Un grand nombre de descripteurs différents sont collectés pour la modélisation d'une grandeur donnée, car les facteurs déterminants du processus étudié ne sont a priori pas connus. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur la grandeur modélisée, et les variables ne sont pas toujours mutuellement indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension du vecteur d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre d'exemples de la base d'apprentissage, le modèle risque d'être surajusté à ces exemples, et incapable de prédire la grandeur modélisée sur de nouvelles observations.
Il est donc nécessaire de réduire la dimension des variables d'entrée. Plusieurs approches sont possibles pour résoudre ce problème :

- réduire la dimension de l'espace des entrées ;
- remplacer les variables corrélées par de nouvelles variables synthétiques, obtenues à partir de leurs combinaisons ;
- sélectionner les variables les plus pertinentes.

Nous allons maintenant décrire les méthodes les plus fréquemment utilisées.

2.1.2.1 L'analyse en composantes principales

L'analyse en composantes principales (ou ACP) [34], est une technique d'analyse de données utilisée pour réduire la dimension de l'espace de représentation des données.
Contrairement à d'autres méthodes de sélection, celle-ci porte uniquement sur les variables, indépendamment des grandeurs que l'on cherche à modéliser. Les variables initiales sont remplacées par de nouvelles variables, appelées composantes principales, deux à deux non corrélées, et telles que les projections des données sur ces composantes soient de variance maximale. Elles peuvent être classées par ordre d'importance.
Considérons un ensemble de n observations, représentées chacune par p données. Ces observations forment un nuage de n points dans R^p.

Le principe de l'ACP est d'obtenir une représentation approchée des variables dans un sous-espace de dimension *k* plus faible, par projection sur des axes bien choisis ; ces axes principaux sont ceux qui maximisent l'inertie du nuage projeté, c'est-à-dire la moyenne pondérée des carrés des distances des points projetés à leur centre de gravité. La maximisation de l'inertie permet de préserver au mieux la répartition des points. Dès lors, les *n* composantes principales peuvent être représentés dans l'espace sous-tendu par ces axes, par une projection orthogonale des *n* vecteurs d'observations sur les *k* axes principaux. Puisque les composantes principales sont des combinaisons linéaires des variables initiales, l'interprétation du rôle de chacune de ces composantes reste possible. Il suffit en effet de déterminer quels descripteurs d'origine leur sont le plus fortement corrélés.

Les variables obtenues peuvent ensuite être utilisées en tant que nouvelles variables du modèle. Par exemple, la régression sur composantes principales **[35]** (ou PCR) est une méthode de modélisation dont la première étape est une analyse en composantes principales, suivie d'une régression linéaire multiple.

2.1.2.2 La méthode de régression des moindres carrés partiels

La régression des moindres carrés partiels **[36,37]** (MCP, ou PLS) est également une méthode statistique utilisée pour construire des modèles prédictifs lorsque le nombre de variables est élevé et que celles-ci sont fortement corrélées. Cette méthode utilise à la fois des principes de l'analyse en composantes principales et de la régression multilinéaire. Elle consiste à remplacer l'espace initial des variables par un espace de plus faible dimension, sous-tendu par un petit nombre de variables appelées « variable latentes », construites de façon itérative. Les variables retenues sont orthogonales (non corrélées), et sont des combinaisons linéaires des variables initiales. Les variables latentes sont obtenues à partir des variables initiales, mais en tenant compte de leur corrélation avec la variable modélisée, contrairement aux variables résultant de l'analyse en composantes principales. Elles doivent ainsi expliquer le mieux possible la covariance entre les entrées et la sortie. Elles sont alors les nouvelles variables explicatives d'un modèle de régression classique, telles que la régression linéaire multiple.