**République Algérienne Démocratique et Populaire**

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**Université Yahia FARES de MÉDÉA**

**Faculté des Sciences et de la Technologie**

**Mémoire pour l'obtention du diplôme de Magister en Génie des Procédés**

**Option : Pharmaceutique**

**Thème**

# MODÉLISATION DE LA PRÉDICTION DE L'EFFT THÉRAPEUTIQUE À PARTIR DE LA STRUCTURE MOLÉCULAIRE

**Présenté par :**

**Sarah REZAZI**

**Devant le jury composé de :**

| | | |
|---|---|---|
| **Dr. L. DEKAR** | Maître de Conférences (A), UYFM | **Président du Jury** |
| **Pr. Y. BAL** | Professeur, USDB | **Examinateur** |
| **Dr. M. N. BOUAZIZ** | Maître de Conférences (A), UYFM | **Examinateur** |
| **Pr. S. HANINI** | Professeur, UYFM | **Directeur de mémoire** |
| **Dr. C. SI-MOUSSA** | Maître de Conférences (B), UYFM | **Invité** |

**2010 – 2011**

# *Dédicace*

*J*e dédie ce modeste travail à mes très chers parents et je remercie Allah de les avoir protégés pour être témoin de ma réussite,

*À* mes frères et mes sœurs,

*À* toute ma famille,

*À* mes collègues et amies,

*À* toute la promotion de Génie des Procédés Pharmaceutiques,

*À* ceux qui me connaissent, et surtout qui m'aiment,

*A*vec mes profonds sentiments.

*Sarah*

# *Remerciements*

Tout d'abord je rends grâce à Allah, le tout puissant qui m'a donné la santé et la volonté pour arriver à ce stade.

J'exprime mes profonds remerciements à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Mes plus vifs remerciements iront également à Messieurs les membres de mon jury de mémoire, le Professeur **Y. Bal** et les Docteurs **L. Dekar** et **M.N. Bouaziz** pour avoir bien voulu consacrer une partie de leur temps à examiner et rapporter ce travail.

Ce travail a été réalisé sous la direction de Monsieur **S. Hanini**, Professeur à l'Université Dr. Yahia Farès de Médéa, qu'il trouve ainsi l'expression de notre profonde reconnaissance pour sa compétence, ses encouragements et ses précieux conseils durant le déroulement de ce travail.

Je souhaite également exprimer mes plus sincères remerciements et ma reconnaissance envers mon codirecteurs de mémoire, monsieur le Docteur **C. Si-Moussa** pour son aide précieuse, ses conseils éclairés, ses encouragements qu'ils n'ont cessé de me prodiguer afin de mener à bien cette tâche.

Je remercie tout particulièrement monsieur le Docteur **O. Rezazi** pour ses aimables contributions dans ce travail. Et je tiens également à adresser mes remerciements à **L. Khaouane, T. Omari, M. Medjeber** et **H. Mermouz** pour l'aide qu'ils m'ont apporté durant la réalisation de ce travail.

Enfin j'exprime mes profonds remerciements à tous les enseignants qui ont contribué à ma formation.

**ملخص: نمذجة التنبؤ بالنشاط العلاجي باستعمال التركيب الجزيئي**

من أجل وضع نماذج تسمح بالتنبؤ بالخصائص العلاجية للجزيئات ، قمنا بتطبيق الشبكة العصبونية على قاعدة من البيانات التجريبية ذات حجم {227،532} والمتمثلة في قيم الواصفات الجزيئية للجزيئات التي تنتمي إلى ستة مجموعات علاجية {مضادات الالتهابات ، المضادات الحيوية ، المسكنات ، مضادات الفيروسات ،مدرات البول والجلايكورتيكود} مأخوذة أساسا من المرجع كلمان و أنجل (2001) . الأساليب التقليدية المعتمدة للنمذجة تتميز بعدة عيوب تكمن في صعوبة تحديد الواصفات الجزيئية الفعالة. للتغلب على هذه المشاكل قمنا بتطوير تقنية إحصائية ( الإزالة التدريجية ) والتي تمكننا من تحديد العدد الأمثل للواصفات الجزيئية ، تم اعتماد الشبكات العصبونية ذات التغذية الأمامية و الانتشار الخلفي ، ومن أجل تحديد خصائص الشبكة المثالية قمنا بتطبيق إستراتيجية قائمة على التقويم بالتجربة و الخطأ مع اعتماد معيارين للتوقف وهما موافقة دالة خطأ مثبتة مسبقا وقدرة الشبكة العصبونية على التعميم ، و تم تعديل خصائص هذه الشبكة بهدف الحصول على أفضل نتائج . وأظهرت النماذج وجود علاقة جيدة من البيانات التجريبية التي تتجاوز 90 ٪ بالنسبة لجميع الأنشطة العلاجية كما أن دراسة قيمة الخطأ المتوسط النسبية المطلقة التي تم الحصول عليها خلال مرحلة الاختبار ومرحلة التعميم لا تتجاوز 0،875 ٪ .قارنا هذه النماذج مع النماذج التقليدية الأخرى المصممة على نفس قاعدة البيانات المستخدمة باستعمال الشبكة العصبونية. هذا النموذج التقليدي يتمثل في إجراء الانحدار الخطي المتعدد (RLM) مقترن بالاختيار المسبق للواصفات الجزيئية عن طريق تقنية تحليل المركبات الرئيسية (ACP). وجدنا أن النتائج المتحصل عليها باستعمال نماذج الشبكة العصبونية تعتبر أفضل بكثير من نتائج النماذج التقليدية ،كما أن الواصفات الجزيئية المنتقاة بطريقة إحصائية سمحت على نحو فعال بالربط الأمثل للخصائص العلاجية للجزيئات.

**كلمات مفاتيح:** الشبكة العصبونية الاصطناعية ، الواصفات الجزيئية ،تحليل المركبات الرئيسية، والانحدار الخطي المتعدد، الإزالة التدريجية.

**RÉSUMÉ: Modélisation de la prédiction de l'activité thérapeutique à partir de la structure moléculaire.**

Dans le but de développer des modèles pour la corrélation et la prédiction de l'activité thérapeutique des molécules nous avons appliqué la modélisation neuronale aux données expérimentales correspondant aux valeurs des descripteurs moléculaires des molécules appartenant à six classes thérapeutiques {anti-inflammatoire ; antibiotique ; analgésique ; antivirale ; diurétique et glucocorticoïde} collectées principalement de la référence Kleemann et Engel (2001), avec une base de données initiale de dimension {532, 227}. Les méthodes traditionnelles de modélisation présentent des principaux inconvénients résident dans la difficulté du choix des descripteurs pertinents. Pour s'affranchir de ces problèmes nous avons mis au point une technique statistique (élimination progressive) qui permet de déterminer le nombre optimal de descripteurs. Pour la recherche de la topologie du réseau de neurones optimal nous avons adopté la stratégie d'essai et erreur (trial and error), avec deux critères d'arrêt : vérification d'une fonction de performance (d'erreur) et la capacité de généralisation du réseau de neurones. Un ensemble de paramètres est modifié jusqu'à la satisfaction de la meilleure régression. Les modèles conçus ont montré une très bonne corrélation des données expérimentales qui dépasse 90% pour toutes les activités thérapeutiques étudiées ainsi que la valeur de l'erreur relative absolue moyenne obtenue lors de la phase de test et la phase de la généralisation ne dépasse pas 0,875% . Nous avons comparé ces modèles avec d'autres modèles traditionnels conçus sur la même base de données utilisées dans la modélisation neuronale. Ces modèles sont élaborés par la réalisation d'une régression linéaire multiple (RLM) précédé d'une étape d'analyse en composante principale (ACP) pour le choix des descripteurs moléculaires. Nous avons trouvé que les corrélations des modèles neuronaux sont nettement meilleures. Les résultats obtenus sont encourageants, ainsi les descripteurs moléculaires optimisés corrèlent efficacement l'activité thérapeutique des molécules.

**Mots clés :** Réseau de neurones artificiel, descripteurs moléculaires, analyse en composante principale, régression linéaire multiple, élimination progressive.

**ABSTRACT: Modeling the prediction of therapeutic activity from the molecular structure.**

In order to develop models for correlation and prediction of the therapeutic molecules we applied the neural modeling with experimental data corresponds to the values of molecular descriptors of the molecules belonging to six therapeutic classes {anti-inflammatory, antibiotic, analgesic, antiviral, diuretic and glucocorticoid} collected from the reference Kleemann and Engel (2001), with initial database size {532.227}. Traditional methods of modeling are the main disadvantages represented in the difficulty of selecting relevant descriptors. To overcome these problems we developed a statistical technique (progressive elimination) for determining the optimal number of descriptors. To search for the topology of the neural network we adopted the optimal strategy of trial and error (trial and error), with two stopping criteria: verification of a performance function (error) and the ability to generalization of the neural network. A set of parameters is modified to the satisfaction of the best regression. The models developed showed a good correlation of experimental data which exceeds 90% for all therapeutic activities studied and the value of the average absolute relative error obtained during the testing phase and the phase of generalization does not exceed 0.875 %. We compared these models with other traditional models designed on the same database used in modeling neuronal. These models are developed by conducting a multiple linear regression (MLR) preceded by a step of principal component analysis (PCA) for the selection of molecular descriptors. We found that the correlations of neural models are much better. The results are encouraging, and the optimized molecular descriptors correlate effectively the therapeutic molecules.

**Keywords:** Artificial Neural Network, molecular descriptors, principal component analysis, multiple linear regression, progressive elimination.

# Table des matières

Table des matières i
Liste des figures v
Liste des tableaux viii
Liste des acronymes x
Nomenclature et symboles xi

***Introduction*** 01

***Chapitre I: Travaux Antérieurs***

I.1. Introduction 05
I.2. Relation Quantitative De Structure-Activité (QSAR) 05
I.3. Revue des travaux réalisés sur la relation quantitative structure-activité 06
I.3.1. Analyse de Hammett 06
I.3.2. Analyse de Hunsch 10
I.4. Méthodes récente QSPR/QSAR 11
I.4.1 Modélisation par les SVM à partir de descripteurs traditionnels 11
I.4.2 Fonctions noyaux pour molécules 12
I.4.3 Autres méthodes de QSPR/QSAR 13
I.4.3.1. Méthode de contribution de groupes 13
I.4.3.2. Analyse comparative de champs moléculaires (CoMFA) 15
I.5. Techniques de modélisation à partir des structures des moléculaire 15
I.5.1. Régression linéaire multiple (RLM) 15
I.5.2. Réseaux de neurones 17
I.5.3. Méthode générale de sélection 18
I.6. Conclusion 20

***Chapitre II : Descripteurs moléculaires, calcul et méthodes statistiques de réduction***

II.1. Introduction 22
II.2. Descripteurs moléculaires 22
II.2.1. Différents types des descripteurs moléculaires 23
II.2.1.1. Descripteurs 1D 23
II.2.1.2. Descripteurs 2D 23
II.2.1.3. Descripteurs 3D 24
II.2.2. Sélection et application 26
II.2.2.1. Carte d'identité d'une molécule 26
II.2.2.2. Utilisation des descripteurs moléculaires 27
II.2.2.3. Sélection à l'aide de modèle QSAR 28
II.3. Méthodes de calculs des descripteurs 29
II.3.1. SMILES (Simplified Molecular Input Line Entry Specification) 30
II.3.1.1. Notion générale 30
II.3.1.2. Principes de SMILES et théorie des graphes 31

II.3.1.3. Conversion 32
II.3.2. Logiciels de calculs des descripteurs 33
II.3.2.1. Domaine privé 34
II.3.2.2. Domaine public 34
II.3.2.3. Domaine académiques 35
II.4. Réduction du nombre de descripteurs 36
II.4.1. Analyse en composantes principales 36
II.4.2. Méthode de régression des moindres carrés partiels 41
II.4.3. Sélection des variables pertinentes 41
II.5. conclusion 42

*Chapitre III: Réseaux de Neurones Artificiels (RNA)*
III.1 Introduction 44
III.2. Historique 44
III.3. Notions fondamentales sur les réseaux de neurones artificiels 46
III.3.1. Analogie entre neurone biologique et neurone artificiel 46
III.3.2. Concept du réseau de neurones de type feed forward 48
III.3.3. Fonctions d'activation 50
III.3.4. Architecture des réseaux de neurones artificiels 52
III.3.4.1. Réseaux de neurones à fonctions radiales de base 52
III.3.4.2. Réseaux compétitifs ou cartes auto organisatrices 54
III.3.4.3. Réseaux récurrents 55
III.3.5. Procédure de modélisation neuronale 56
III.3.5.1. Sélection des données 58
III.3.5.2. Collecte des données 58
III.3.5.3. Séparation des bases de données 58
III.3.5.4. Normalisation des données 59
III.3.5.5. Choix du type et paramètres du réseau de neurones 59
III.3.5.6. Choix de la fonction de transfert 60
III.3.5.7. Nombre d'itération 61
III.3.5.8. Estimation des paramètres ou apprentissage 61
III.3.5.9. Validation et généralisation 61
III.4. Conclusion 62

*Chapitre IV: Modélisation de la prédiction de l'activité thérapeutique*
IV.1. Introduction 64

**Partie 01: Élaboration de la base de données**
IV.1.1. Collection des données 66
IV.1.2. Dessin des molécules sur ChemDraw 67
IV.1.3. Génération de la notation SMILES par ChemDraw 67
IV.1.4. Élaboration du fichier texte de la molécule sous forme SMI 68
IV.1.5. Différentes étapes pour le calcul des descripteurs par DRAGON 69

IV.1.7. Codage des sorties 74
IV.1.8. Propriétés de la base finale 75

## Partie 02: Modélisation neuronale

IV.2.1. Introduction 77
IV.2.2. Réduction préliminaire 77
IV.2.3. Réduction par méthode statistique 77
IV.2.3.1. Méthode sélectionnée 77
IV.2.3.2. Procédure générale 77
IV.2.4. Optimisation de la prédiction de l'activité anti-inflammatoire 79
IV.2.4.1. Définition d'un anti-inflammatoire 79
IV.2.4.2. Liste des anti-inflammatoires utilisés 79
IV.2.4.3. Optimisation du nombre de descripteurs 81
IV.2.4.3.1. Méthode sélectionnée : élimination progressive 81
IV.2.4.3.2. Normalisation des données 81
IV.2.4.3.3. Répartition des données 82
IV.2.4.3.4. Architecture du réseau de neurone utilisée 82
IV.2.4.3.5. Résultats et discussion 82
IV.2.4.4.Optimisation du modèle neuronal pour la prédiction d'activité anti-inflammatoire 84
IV.2.4.5. Structure du réseau RNSO obtenu 94
IV.2.4.6. Résultats 94
IV.2.4.7. Analyse des résultats 96
IV.2.5. Optimisation de la prédiction de l'activité antibiotique 97
IV.2.5.1. Définition d'un antibiotique 97
IV.2.5.2. Liste des antibiotiques utilisés 97
IV.2.5.3. Optimisation du nombre de descripteurs 99
IV.2.5.4. Modélisation neuronale 99
IV.2.5.4.1. Type de réseau 99
IV.2.5.4.2. Nombre de couches cachées 100
IV.2.5.4.3. Algorithme d'apprentissage et la fonction d'activation 100
IV.2.5.5. Résultats 100
IV.2.5.6. Analyse des résultats 102
IV.2.5.7. Amélioration des capacités prédictives pour l'activité antibiotique 102
IV.2.5.8. Analyse des résultats 105
IV.2.6. Optimisation de la prédiction de l'activité analgésique 105
IV.2.6.1. Définition d'un analgésique 105
IV.2.6.2. Liste de molécules analgésiques utilisées 105
IV.2.6.3. Optimisation du nombre de descripteurs 107
IV.2.6.4. Modélisation neuronale 107
IV.2.6.5. Résultats 108
IV.2.6.6. Analyse des résultats 110
IV.2.7. Optimisation de la prédiction de l'activité antivirale 110
IV.2.7.1. Définition d'un antiviral 110

IV.2.7.2. Liste de molécules antivirales utilisées 110
IV.2.7.3. Optimisation du nombre de descripteurs 112
IV.2.7.4. Modélisation neuronale 112
IV.2.7.5. Résultats 113
IV.2.7.6. Analyse des résultats 115
IV.2.8. Optimisation de la prédiction de l'activité diurétique 115
IV.2.8.1. Définition d'un diurétique 115
IV.2.8.2. Liste de molécules diurétique utilisées 115
IV.2.8.3. Optimisation du nombre de descripteurs 116
IV.2.8.4. Modélisation neuronale 117
IV.2.8.5. Analyse des résultats 119
IV.2.9. Optimisation de la prédiction de l'activité glucocorticoïde 120
IV.2.9.1. Définition d'un glucocorticoïde 120
IV.2.9.2. Définition d'un glucocorticoïde 120
IV.2.9.3. Liste de molécules glucocorticoïdes utilisées 120
IV.2.9.4. Optimisation du nombre de descripteurs 121
IV.2.9.5. Modélisation neuronale 121
IV.2.9.6. Analyse des résultats 123
IV.2.9.7. Test de performance de prédiction du modèle construit 124
IV.2.10. Conclusion 125

**Partie 03: Analyse en composante principale & régression linéaire multiple**

IV.3.1. Introduction 128
IV.3.2. Présentation de la méthode d'analyse en composante principale 128
IV.3.3. Présentation de résultats 129
IV.3.3.1. Calcul des valeurs propres 129
IV.3.3.1.1. Discussion 130
IV.3.4.1.2. Représentation graphique des valeurs propres et du cumul% 131
IV.3.3.2. Interprétation des axes 131
IV.3.3.3.Interprétation des plans 134
IV.3.4 Régression linéaire multiple (RLM) 136
IV.3.4.1. Élaboration des modèles de régression multiple 136
IV.3.4.1.1. Présentation du modèle 136
IV.3.4.1.2. Forme matricielle 137
IV.3.4.2. Statistiques descriptives et résultats d'estimation des modèles 138
IV.3.4.2.1. Statistiques descriptives 138
IV.3.4.2.2. Estimation des modèles RLM 138
IV.3.5. Analyse comparative avec la modélisation neuronale 158
IV.3.6. Conclusion 159

***Conclusion générale*** 161
***Références bibliographiques*** 165
***Annexes*** 169

# Liste des figures


| | | |
|---|---|---|
| **LISTE DES FIGURES DU CHAPITRE I** | | |
| **Figure I.1** | Exemple de deux molécules, isomères de constitution | 14 |
| **LISTE DES FIGURES DU CHAPITRE II** | | |
| **Figure II.1** | Exemple de molécule convertie en graphe réduit. Le graphe ne garde que l'enchaînement de grands types d'éléments structuraux. ***Ar*** correspond à un groupe aromatique, L à un linker, F à un groupe fonctionnel | 25 |
| **Figure II.2** | Niveaux de description d'une molécule utilisés dans le criblage virtuel | 26 |
| **Figure II.3** | Exemple de dérivation de représentation SMILES | 30 |
| **Figure II.4** | Génération de la notation SMILES à partir de la structure moléculaire**.** | 32 |
| **Figure II.5** | Création de la structure moléculaire à partir de la notation SMILES. | 33 |
| **Figure II.6** | Différents logiciels utilisés dans le domaine privé | 34 |
| **Figure II.7** | Différents logiciels utilisés dans le domaine public | 34 |
| **Figure II.8** | Organigramme résume les différentes étapes d'une analyse ACP. | 39 |
| **Figure II.9** | Résultats d'ACP (espace des individus) | 40 |
| **Figure II.10** | Résultats d'ACP (espace des variables) | 40 |
| **LISTE DES FIGURES DU CHAPITRE III** | | |
| **Figure III.1** | Schéma d'un neurone biologique | 46 |
| **Figure III.2** | Schéma d'un neurone artificiel | 47 |
| **Figure III.3** | Schéma de comparaison entre un neurone biologique et un neurone artificiel | 48 |
| **Figure III.4** | Réseau de neurones artificiels de type feedforward | 49 |
| **Figure III.5** | Schéma d'un neurone de fonction radiale de base | 52 |
| **Figure III.6** | Schéma de l'implantation MATLAB® du réseau de neurones | 53 |
| **Figure III.7** | Schéma de l'implantation MATLAB® d'un réseau de neurones probabiliste | 54 |
| **Figure III.8** | Schéma de l'implantation MATLAB® d'un réseau de neurones à régression généralisée | 54 |
| **Figure III.9** | Architecture d'un réseau de neurones compétitifs ou réseau de Kohonen | 55 |
| **Figure III.10** | Schéma de l'implantation MATLAB® du réseau récurrent d'Elman | 56 |
| **LISTE DES FIGURES DU CHAPITRE IV** | | |
| **Partie 01** | | |
| **Figure IV.1.1** | Organigramme des déférentes étapes de réalisation de la base de données | 66 |
| **Figure IV.1.2** | Dessin des molécules sur ChemDraw. | 67 |
| **Figure IV.1.3** | Génération de la notation SMILES par ChemDraw | 68 |
| **Figure IV.1.4** | Elaboration du fichier texte de la molécule sous forme SMILES. | 68 |
| **Figure IV.1.5** | Première étapes pour le calcul des descripteurs par DRAGON | 69 |
| **Figure IV.1.6** | Deuxième étapes pour le calcul des descripteurs par DRAGON. | 69 |
| **Figure IV.1.7** | Troisième étape pour le calcul des descripteurs par DRAGON. | 70 |
| **Figure IV.1.8** | Quatrième étape pour le calcul des descripteurs par DRAGON. | 70 |
| **Figure IV.1.9** | Dernière étape pour le calcul des descripteurs par DRAGON | 71 |
| **Partie 02** | | |
| **Figure IV.2.1** | Organigramme représentant la procédure de réduction des descripteurs par la méthode élimination progressive | 78 |
| **Figure IV.2.2** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 83 |

| | | |
|---|---|---|
| **Figure VI.2.3** | Organigramme de la méthodologie de calcul RNSO | 85 |
| **Figure IV.2.4** | Effet de la fonction d'activation sur les performances du réseau de neurones. | 87 |
| **Figure IV.2.5** | Effet du type d'algorithme d'apprentissage sur la performance du réseau de neurones en termes d'ERAM. | 88 |
| **Figure IV.2.6** | Effet de la méthode de normalisation sur la performance du réseau de neurones en termes d'ERAM. | 90 |
| **Figure IV.2.7** | Performance du réseau de neurones à deux couches cachées en termes d'ERAM minimale | 91 |
| **Figure IV.2.8** | Performance du réseau de neurones à deux couches cachées en termes du coefficient de régression globale | 91 |
| **Figure IV.2.9** | Performance du réseau de neurones à deux couches cachées en termes d'ERAM minimale | 92 |
| **Figure IV.2.10** | Performance du réseau de neurones à deux couches cachées en termes du coefficient de régression globale. | 92 |
| **Figure IV.2.11** | Performance du réseau de neurones à trois couches cachées en termes d'ERAM minimale | 93 |
| **Figure IV.2.12** | Performance du réseau de neurones à trois couches cachées en termes du coefficient de régression globale. | 93 |
| **Figure IV.2.13** | Courbe de régression linéaire de l'activité anti-inflammatoire calculée par RNA optimisé avec l'activité anti-inflammatoire donnée expérimentalement pour le processus de test | 95 |
| **Figure IV.2.14** | Courbe de régression linéaire de l'activité anti-inflammatoire calculée par RNA optimisé avec l'activité anti-inflammatoire donnée expérimentalement pour le processus de généralisation | 95 |
| **Figure IV.2.15** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 99 |
| **Figure IV.2.16** | Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibiotique donnée expérimentalement pour le processus de test. | 101 |
| **Figure IV.2.17** | Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibiotique donnée expérimentalement pour le processus de généralisation. | 101 |
| **Figure IV.2.18** | Performance du RNA optimisé. | 104 |
| **Figure IV.2.19** | Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibactérienne donnée expérimentalement après réduction de la base de données. | 104 |
| **Figure IV.2.20** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 107 |
| **Figure IV.2.21** | Courbe de régression linéaire de l'activité analgésique calculée par RNA optimisé avec l'activité analgésique donnée expérimentalement pour le processus de test | 109 |
| **Figure IV.2.22** | Courbe de régression linéaire de l'activité analgésique calculée par RNA optimisé avec l'activité analgésique donnée expérimentalement pour le processus de généralisation | 109 |
| **Figure IV.2.23** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 112 |
| **Figure IV.2.24** | Courbe de régression linéaire de l'activité antivirale calculée par RNA optimisé avec l'activité antivirale donnée expérimentalement pour le processus de test | 114 |
| **Figure IV.2.25** | Courbe de régression linéaire de l'activité antivirale calculée par RNA optimisé avec l'activité antivirale donnée expérimentalement pour le processus de généralisation | 114 |
| **Figure IV.2.26** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 117 |
| **Figure IV.2.27** | Courbe de régression linéaire de l'activité diurétique calculée par RNA optimisé avec l'activité diurétique donnée expérimentalement pour le processus de test | 118 |
| **Figure IV.2.28** | Courbe de régression linéaire de l'activité diurétique calculée par RNA optimisé avec l'activité diurétique donnée expérimentalement pour le processus de généralisation | 118 |
| **Figure IV.2.29** | Effet de l'élimination progressive sur l'erreur relative absolue moyenne | 121 |

| | | |
|---|---|---|
| **Figure IV.2.30** | Courbe de régression linéaire de l'activité glucocorticoïde calculée par RNA optimisé avec l'activité glucocorticoïde donnée expérimentalement pour le processus de test | 122 |
| **Figure IV.2.31** | Courbe de régression linéaire de l'activité glucocorticoïde calculée par RNA optimisé avec l'activité glucocorticoïde donnée expérimentalement pour le processus de généralisation | 122 |
| **Figure IV.2.32** | Représentation schématique du réseau de neurones multicouches pour la prédiction de l'activité des glucocorticoïdes | 123 |
| **Partie 03** | | |
| **Figure IV.3.1** | Représentation graphique des valeurs propres et pourcentage de cumul | 131 |
| **Figure IV.3.2** | Cercle de corrélation du plan (1.2) | 134 |
| **Figure IV.3.3** | Cercle de corrélation du plan (1.3) | 135 |
| **Figure IV.3.4** | Courbe de régression pour l'activité anti-inflammatoire | 142 |
| **Figure IV.3.5** | Courbe de régression pour l'activité antibiotique | 147 |
| **Figure IV.3.6** | Courbe de régression pour l'activité analgésique | 149 |
| **Figure IV.3.7** | Courbe de régression pour l'activité antivirale | 152 |
| **Figure IV.3.8** | Courbe de régression pour l'activité diurétique | 154 |
| **Figure IV.3.9** | Courbe de régression pour l'activité glucocorticoïde | 156 |
| **LISTE DES FIGURES DE L'ANNEXES** | | |
| **Figure B.1** | Courbe de régression linéaire des valeurs prédites par RNA avec les valeurs données expérimentalement pour le processus de test pour la prédiction de l'activité des glucocorticoïdes. | 170 |
| **Figure B.2** | Courbe de régression linéaire des valeurs prédites par RNA avec les valeurs données expérimentalement pour le processus de généralisation pour la prédiction de l'activité des glucocorticoïdes. | 170 |

# Liste des tableaux


| | | |
|---|---|---|
| **LISTE DES TABLEAUX DU CHAPITRE I** | | |
| **Tableau I.1** | Exemples pour des valeurs de ρ | 09 |
| **Tableau I.2** | Caractéristique de σ pour des substituants (para/méta) | 10 |
| **Tableau I.3** | Décomposition en groupes de deux isomères de constitution | 14 |
| **LISTE DES TABLEAUX DU CHAPITRE II** | | |
| **Tableau II.1** | Application de la notation SMILES sur des molécules de natures variées | 31 |
| **LISTE DES TABLEAUX DU CHAPITRE III** | | |
| **Partie 01** | | |
| **Tableau III.1** | Fonctions de transfert usuelles | 51 |
| **Tableau IV.1.1** | Résultats de calcul des descripteurs moléculaires. | 71 |
| **Tableau IV.1.2** | Signification des différents descripteurs moléculaires utilisés. | 72 |
| **Tableau IV.1.3** | Codage des sorties | 74 |
| **Tableau IV.1.4** | Caractéristiques de la base finale | 75 |
| **Partie 02** | | |
| **Tableau IV.2.1** | Molécules anti-inflammatoire utilisées dans la base de données initiale | 80 |
| **Tableau IV.2.2** | Descripteurs éliminés dans chaque passage d'élimination progressive | 83 |
| **Tableau IV.2.3** | Différentes fonctions d'activations utilisées dans la phase de conception du modèle neuronal. | 86 |
| **Tableau IV.2.4** | Différents algorithmes d'apprentissage utilisés lors de la phase de conception du modèle neuronal | 88 |
| **Tableau IV.2.5** | Différentes fonctions de normalisation utilisées | 89 |
| **Tableau IV.2.6** | Structure du réseau de neurones optimisé | 94 |
| **Tableau IV.2.7** | Molécules antibiotiques utilisées dans la base de données initiale | 98 |
| **Tableau IV.2.8** | Topologie du RNA optimisé pour l'activité antibiotique | 100 |
| **Tableau IV.2.9** | Topologie du RNA optimisé pour l'activité antibiotique après réduction. | 103 |
| **Tableau IV.2.10** | Molécules analgésiques utilisées dans la base de données initiale | 106 |
| **Tableau IV.2.11** | Topologie du RNA optimisé pour l'activité analgésique. | 108 |
| **Tableau IV.2.12** | Molécules antivirales utilisées dans la base de données initiale | 111 |
| **Tableau IV.2.13** | Topologie du RNA optimisé pour l'activité antivirale | 113 |
| **Tableau IV.2.14** | Molécules diurétiques utilisées dans la base de données initiale | 116 |
| **Tableau IV.2.15** | Topologie du RNA optimisé pour l'activité diurétique | 117 |
| **Tableau IV.2.16** | Molécules antihistaminiques utilisées dans la base de données initiale | 120 |
| **Tableau IV.2.17** | Topologie du RNA optimisé pour l'activité antihistaminique | 121 |
| **Tableau IV.2.18** | Résultats du calcul par le modèle RNA conçu pour la prédiction de l'activité des glucocorticoïdes. | 124 |
| **Partie 03** | | |
| **Tableau IV.3.1** | Valeurs propres calculées | 129 |
| **Tableau IV.3.2** | Matrice des coordonnées factorielles des variables | 132 |
| **Tableau IV.3.3** | Coordonnées factorielles des individus | 133 |
| **Tableau IV.3.5** | Résultats générales de la régression linéaire multiples | 138 |
| **Tableau IV.3.6** | Coefficients pour l'activité anti-inflammatoire | 139 |
| **Tableau IV.3.7** | Coefficients pour l'activité antibiotique | 139 |
| **Tableau IV.3.8** | Coefficients pour l'activité analgésique | 140 |
| **Tableau IV.3.9** | Coefficients pour l'activité antivirale | 140 |

| | | |
|---|---|---|
| **Tableau IV.3.10** | Coefficients pour l'activité diurétique | 140 |
| **Tableau IV.3.11** | Coefficients pour l'activité glucocorticoïde | 158 |
| **LISTE DES TABLEAUX DE L'ANNEXES** | | |
| **Tableau B.1** | Résultats du calcul du réseau de neurone artificiel, processus d'apprentissage. (Ensemble des molécules non antibiotiques et molécules antibiotique) | 171 |
| **Tableau B.2** | Résultats du calcul du réseau de neurone artificiel, processus de validation. (Ensemble des molécules non antibiotiques et molécules antibiotique) | 172 |
| **Tableau C.1** | Exemple de tableau des valeurs propres fourni par STAISTICA® pour la première base | 173 |
| **Tableau C.2** | Exemple de matrice des coordonnées factorielles des variables pour quelques molécules | 173 |
| **Tableau C.3** | Exemple de tableau de régression et résidus des anti-inflammatoires | 174 |
| **Tableau C.4** | Tableau de régression et résidus pour l'activité antibiotique | 175 |
| **Tableau C.5** | Exemple de Tableau de régression et résidus des analgésiques | 176 |
| **Tableau C.6** | Exemple de tableau de régression et résidus des antiviraux | 177 |
| **Tableau C.7** | Exemple de Tableau de régression et résidus des diurétiques | 178 |
| **Tableau C.8** | Exemple de Tableau de régression et résidus des glucocorticoïdes | 179 |

# Liste des Acronymes

| Acronyme | Signification |
|---|---|
| **ACP** | Analyse en Composantes Principales |
| **ADN** | Acide Désoxyribo Nucléique |
| **AN** | Artificial Neurone |
| **ASCII** | American Standard Code for Information Interchange |
| **CADD** | Computer Assisted Drug Design |
| **CoMFA** | Analyse Comparative de Champs Moléculaires |
| **EAM** | Erreur Absolue Moyenne |
| **ERAM** | Erreur Relative Absolue Moyenne |
| **InChl** | International Chemical Identifier |
| **IUPAC** | International Union of Pure and Applied Chemistry |
| **LCR** | Liquide Céphalo Rachidien |
| **LDA** | Analyse Discriminante Linéaire |
| **MLP** | Multi Layer Perceptron |
| **MoRSE** | Molecule Representation of Structures based on Electron diffraction |
| **MP** | McCulloch et Pitts |
| **OBD** | Optimal Brain DAMAGE |
| **OBS** | Optimal Brain SURGEON |
| **PDB** | Protein Data Bank |
| **PDD** | Diagrammes de Distribution Pharmacologique |
| **PLP** | Protéines Liant les Pénicillines |
| **QSAR** | Quantitative Structure Activity Relationship |
| **QSPR** | Quantitative Structure Propriety Relationship |
| **RLM** | Régression Linéaire Multiple |
| **RMN** | Résonance Magnétique Nucléaire |
| **RNA** | Réseau de Neurone Artificiel |
| **RSA** | Relation Structure Activity |
| **SDF** | Standard Delay Format |
| **SMILES** | Simplified Molecular Input Line Entry Specification |
| **SVM** | Machines à Vecteurs Supports |

# Nomenclature et symboles

| Lecture | Désignation |
|---|---|
| $\mathbf{b_i^*}$ | Coefficient du modèle de régression linéaire multiple |
| b | vecteur des coefficients $\mathbf{b_i^*}$ du modèle de régression linéaire multiple |
| $b_0$ | Ordonné à l'origine pour un modèle de régression linéaire multiple |
| C | Concentration. |
| D | dimension |
| e | Vecteur des résidus $e_i$ |
| $e_i$ | Résidus pour l'observation i |
| f | Fonction de régression |
| F | Test de Fisher |
| F | Axe factoriel |
| Fact | Factoriel |
| K | Constante de dissociation (d'équilibre) |
| K | Constantes d'équilibre pour les composés substitués. |
| $K_0$ | Constantes d'équilibre pour les composés non substitués |
| n | nombre d'observations |
| n | Nombre de variables |
| P | Coefficient de cloison du composé entre l'octanol et l'eau |
| $P_h$ | Coefficient de partage de l'hydrogène |
| $P_x$ | Coefficient de cloison de la substance X |
| t | Test de Student |
| val | valeur |
| w | Poids synaptique des connexions inter-neuronales |
| X | matrice des $\boldsymbol{X_i}$ |
| X | Nombre de substances |
| X | matrice des observations |
| $x_i$ | Variable explicative de l'observation i |
| $x_k$ | Ensemble de vecteurs de variables |
| Y | vecteur des $\boldsymbol{Y_i}$ |
| y | vecteur des valeurs de la grandeur à modéliser |
| $Y_i$ | Variable à expliquer de l'observation i |
| $\widehat{\boldsymbol{Y}}$ | Valeur prévue |

## Lettres Grecques

| Lecture | Désignation |
|---|---|
| α | Pente de la droite de regréssion linéaire |
| β | Ordonnée à l'origine |
| π | Notation de l'équation alternative de Hansch |
| ρ | Pente de la ligne de l'équation de l'énergie libre |
| σ | Paramètre de substituant de Hammett. |

## Exposants

| Lecture | Désignation |
|---|---|
| *i* | Constituent i d'un mélange ou élément i d'un vecteur |
| T | Transposé |
| T | Target |
| cal | Calculée |
| exp | Expérimentale |

## Indices

| Lecture | Désignation |
|---|---|
| c | Calculée |
| *i* | Élément i d'un vecteur |
| ij | Croisement de la ligne i avec la colonne j d'une matrice |
| j | Élément j d'un vecteur |
| k | Élément k d'un vecteur |
| mc | Minimum quadratique |
| t | Tabulé |

# *Introduction générale*

# INTRODUCTION GÉNÉRALE

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, 10000 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, coûteuses et finalement inutiles. 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). Elles consistent essentiellement en la recherche de similitudes entre molécules dans de grandes bases de données de molécules existantes dont les propriétés sont connues. 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.

D'autre part la révolution de l'informatique et le développement de nouvelles méthodes de calcul numérique basé sur l'intelligence artificielle et les sciences cognitives ont donné naissance durant les dernières décennies à de nouvelles méthodes de calcul et de modélisation telles que les réseaux de neurones artificiels (RNA), les systèmes experts, les algorithmes génétiques et la logique floue. L'efficacité, la fiabilité et la robustesse de ces méthodes ont été démontrées par leurs nombreuses applications dans les différentes disciplines de sciences et d'ingénierie.

## Problématique

Les relations entre les structures des molécules et leurs propriétés ou 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 résident principalement dans la sélection des descripteurs pertinents qui assurent une conservation de l'information et une bonne qualité des modèles.

Dans la littérature très peu de travaux ont été publiés sur l'application des méthodes neuronales pour la corrélation et/ou la prédiction des propriétés et activités thérapeutiques des molécules ainsi que leurs couplages aux méthodes statistiques. Les auteurs utilisent, généralement, un choix aléatoire des descripteurs moléculaires ce qui influe directement sur la qualité du modèle construit ou bien, utilisent des méthodes statistiques qui ont prouvé leurs limites. Nous citons par exemple, les travaux présentés par Miguel et al en 2003 et Korichi et al en 2006.

Dans le cadre de ce mémoire on va essayer de une trouver méthode de modélisation qui s'affranchit de la caractérisation des molécules par des descripteurs moléculaires, et établit directement des relations entre ces descripteurs moléculaires et l'activité thérapeutique. Le présent travail porte sur l'utilisation de la modélisation neuronale pour la prédiction de l'activité thérapeutique des molécules par l'exploitation des méthodes statistiques et la comparaison des capacités prédictives du modèle neuronal élaboré avec ceux de la méthode de régression linéaire multiple considérée parmi les méthodes traditionnelles de modélisation.

Le premier chapitre de ce mémoire présente les travaux antérieurs réalisés sur les méthodes traditionnelles et récentes de modélisation moléculaire QSAR/QSPR et la fiabilité de chaque méthode.

Le chapitre suivant introduit un rappel sur les principaux types de descripteurs, les problèmes liés à leur calcul et à leur sélection, ainsi que les principales techniques conventionnelles utilisées pour la réduction du nombre de descripteurs moléculaires.

Dans le troisième chapitre, les notions fondamentales des RNA ainsi que la stratégie utilisée dans ce travail pour le développement du RNA optimal seront présentées.

Le quatrième chapitre est consacré aux applications des différentes méthodes de modélisation pour la prédiction des propriétés et d'activités moléculaires. Nous y présentons les contributions de ce travail à l'application des méthodes statistiques pour la sélection des descripteurs pertinents pour la modélisation neuronale et la modélisation par régression linaire multiple de l'activité thérapeutique.

Les conclusions tirées à partir de ce travail et les études qui peuvent être envisagées dans le futur sont présentés à la fin du mémoire.

# *Chapitre I*

# *Travaux antérieurs*

# CHAPITRE I
# TRAVAUX ANTÉRIEURS

## I.1. Introduction

Les premiers essais de modélisation d'activités de molécules datent de la fin du 19ème siècle, lorsque Crum-Brown et Frazer [Crum-Brown., et Frazer, (1868)] 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 la notion générale de la relation quantitative structure-activité et les différentes travaux réalisés dans ce cadre toute en présentant les revue critiques des différentes théories de QSAR /QSPR et comment les molécules peuvent être représentées par des vecteurs de réels. Nous introduirons ensuite les outils de modélisation les plus utilisés, c'est-à-dire la régression linéaire multiple et la régression 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. Nous décrirons les méthodes de modélisation sous contraintes, ainsi que leurs applications en QSAR et QSPR. Enfin, de nouvelles méthodes de modélisation, telles que la méthode CoMFA, mises au point pour la modélisation d'activités biologiques, seront présentées.

## I.2. Relation Quantitative De Structure-Activité (QSAR)

La relation structure activité est depuis quelques années utilisée de façon systématique dans la conception de nouveaux médicaments et permet de prédire l'activité thérapeutique de molécules avant même de les avoir synthétisées. Cette prédiction, bien que non totale, a l'avantage d'économiser une partie de l'expérimentation initiale et la relation structure-activité est devenue partie intégrante d'un arbre de décision [Lhuguenot (2001)].

Une relation quantitative structure-activité (QSAR) relie les propriétés numériques de la structure moléculaire à son activité par un modèle mathématique. Le terme « relation quantitative de structure-propriété » (QSPR) est également employé, en particulier quand une certaine propriété autre que l'activité biologique est concernée.

De plus lorsque les propriétés ou structures physico-chimiques sont exprimées par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure à activité, entre les deux. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de la réponse biologique pour des structures similaires.

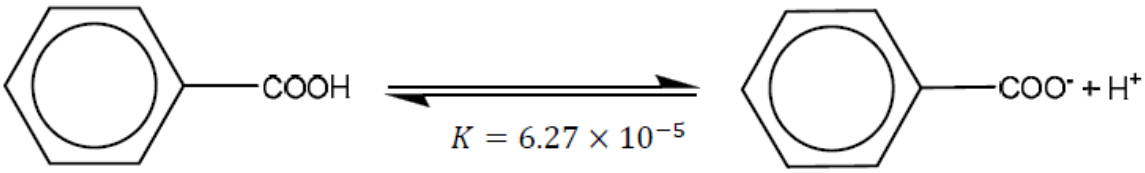
## I.3. Revue des travaux réalisés sur la relation quantitative structure-activité

Les premiers travaux réalisés sur QSAR Retour au 19eme siècle. En 1863, l'AFA Cros à l'Université de Strasbourg, a observé que la toxicité des alcools pour les mammifères plus que la solubilité dans l'eau des alcools diminués [Borman (1990)].

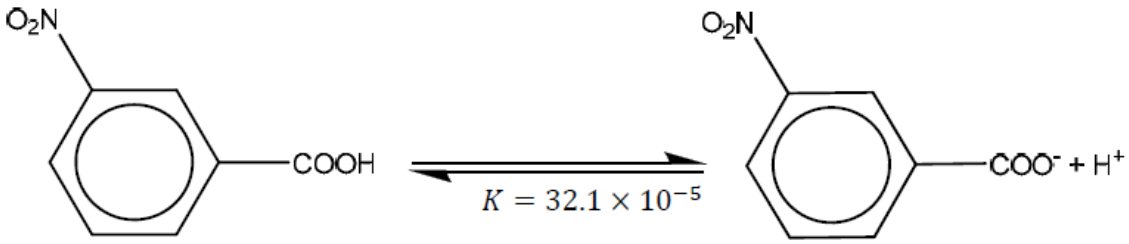
Dans les années 1890, Hans Horst Meyer de l'Université de Marburg et Charles Ernest Overton de l'Université de Zurich, travaillant indépendamment, a noté que la toxicité des composés organiques dépend de leur lipophilie [Borman (1990), Lipnick, et al. (1986)].

### I.3.1 Analyse de Hammett

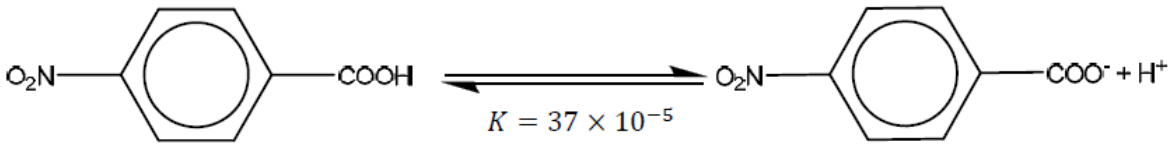
Peu de développement supplémentaire de QSAR eu lieu jusqu'à ce que le travail de Louis Hammett (1894-1987), qui en corrélation les propriétés électroniques des acides organiques et les bases de leurs constantes d'équilibre et de réactivité. Envisager la dissociation de l'acide benzoïque:



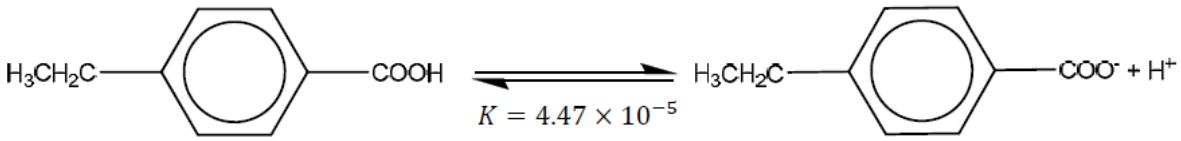

Hammett a observé que l'ajout de substituants au cycle aromatique de l'acide benzoïque a eu un effet quantitatif et ordonnée sur la constante de dissociation. Par exemple :




Un groupe nitro en position méta augmente la constante de dissociation, parce que le groupe nitro est électro attracteur, de manière à stabiliser la charge négative qui se développe. Considérons maintenant l'effet d'un groupe nitro en position para:

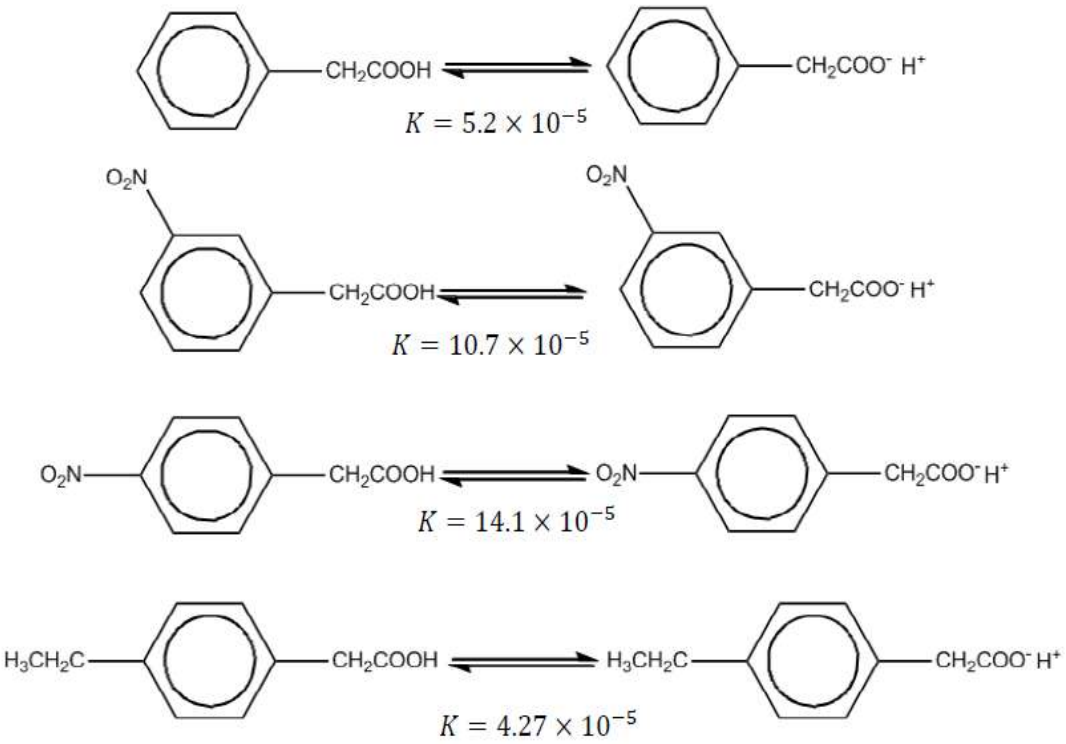



La constante d'équilibre est d'autant plus grande que pour le groupe nitro en position méta, indiquant encore plus d'électrons de retrait. Considérons maintenant le cas dans lequel un groupe éthyle est en position para:




Dans ce cas, la constante de dissociation est plus faible que pour le composé non substitué, indiquant que le groupe éthyle est donneur d'électrons, déstabilisant ainsi la charge négative qui se pose sur la dissociation.

Hammett a également observé que les substituants ont un effet similaire sur la dissociation des acides organiques et les bases. Il a envisagé la dissociation des acides phénylacétique:




Pour la dissociation du groupe nitro, l'effet étant moins pour le substituant méta que pour le substituant para, tout comme a été observée avec l'acide benzoïque.

Les relations élaborées par Hammett sont représentées par les équations I .1 et I.2

$$\boldsymbol{log\frac{K}{K_0} = \rho\, log\frac{K\prime}{K_0'}} \qquad \text{(I.1)}$$

Avec

- ρ : trait constant de proportionnalité à un équilibre donné.
- $K_0$ ou $K'_0$ représentent des constantes d'équilibre pour les composés non substitués
- K ou K' constantes d'équilibre pour les composés substitués.

Les valeurs de $\boldsymbol{log\frac{K\prime}{K}}$ sont toujours ceux de l'acide benzoïque et sont donnés le symbole σ. Par conséquent, nous pouvons écrire:

$$\boldsymbol{log\frac{K}{K_0} = \rho\sigma} \qquad \text{(I .2)}$$

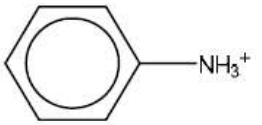
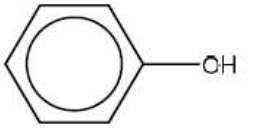
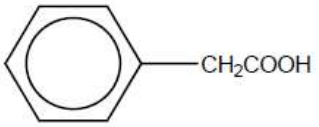
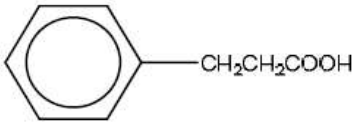
Elle concerne l'effet des substituants sur cet équilibre à l'effet de ces substituants sur l'équilibre acide benzoïque. C'est, si l'effet des substituants est proportionnellement plus élevé que sur l'équilibre acide benzoïque, puis ρ > 1, si l'effet est moindre que sur l'équilibre acide benzoïque, ρ <1. Par définition, ρ pour l'acide benzoïque est égal à 1.

σ Est un descripteur des substituants. L'ampleur des σ donne la force relative de l'électrondonneur de retirer ou de propriétés des substituants. σ est positif si le substituant est électronégative et si elle est donneur d'électrons.

QSAR fondée sur la relation de Hammett utilise des propriétés électroniques comme les descripteurs de structures. Des difficultés ont été rencontrées lors enquêteurs ont tenté d'appliquer les relations de type Hammett pour les systèmes biologiques, ce qui indique que d'autres descripteurs structuraux ont été nécessaires.

Pour développer une meilleure compréhension de ces relations, il est instructif d'examiner certaines valeurs de ρ et σ. Les valeurs de ρ sont fournies dans le tableau I.1:

**Tableau I.1 :** Exemples pour des valeurs de ρ [Hansch (1991)]..

| | |
|---|---|
|  | $\rho = 2.90$ |
|  | $\rho = 2.23$ |
|  | $\rho = 0.49$ |
|  | $\rho = 0.21$ |

Dans l'aniline et la molécule du phénol, les substituants sont en mesure d'exercer un effet plus important sur la dissociation de l'aniline et le phénol que dans l'acide benzoïque, donc la valeur de ρ >1.

Dans les acides phénylacétique et phénylpropionique, les substituent sont en mesure d'exercer un effet moindre sur l'équilibre que sur l'équilibre acide benzoïque donc la valeur de ρ<1.

Le tableau I.2 représente certaines valeurs caractéristiques de σ pour les substituants dans les positions méta et para

**Tableau I.2 :** Caractéristique de **σ** pour des substituants (para/méta) [Hansch (1991)].

| **Substituent** | $\sigma_m$ | $\sigma_p$ |
|---|---|---|
| **-H** | 0.0 | 0.0 |
| **-$NO_2$** | 0.71 | 0.78 |
| **-Cl** | 0.37 | 0.23 |
| **-$OCH_3$** | 0.12 | -0.27 |

### I.3.2 Analyse de Hunsch

La première utilisation de QSAR de rationaliser l'activité biologique [Hansch (1969)]. Hansch a développé les équations qui ont relié l'activité biologique aux caractéristiques électroniques et au hydrophobicité d'une molécule par exemple :

$$\boldsymbol{log\left(\frac{1}{C}\right) = k_1 logP - k_2(logP)^2 + k_3\sigma + k_4} \qquad \text{(I. 3)}$$

Où

- **C** : Concentration du composé exigé pour produire une réponse standard dans un temps donné.
- **Log P** : Logarithme du coefficient de cloison du composé entre l'octanol et l'eau, qui a été choisie par Hansch comme mesure appropriée de hydrophobicité relatif.
- **σ** : est le paramètre de substituant de Hammett.
- $\mathbf{k_1}$ ... $\mathbf{k_4}$ : sont des constantes.

Le composant hydrophobe a été considéré pour modeler la capacité des médicaments passé par des membranes de cellules.

Hansch a identifié une valeur optimale du hydrophobicité :

- ✓ Trop bas et le médicament ne diviserait pas dans la membrane de cellules ;
- ✓ Trop haut et le composé diviserait dans la membrane mais tendrait à rester là plutôt que de procéder à la cible réelle.

Ceci explique la dépendance parabolique de l'activité sur la manière alternative de logP. Un paramètre π est utilisé pour exprimer la notation de l'équation alternative de Hansch. C'est le logarithme du coefficient de partage d'un composé avec X substances relativement à un composé de parent dans lequel le substituant est hydrogène [Hansch (1991)].

$$\boldsymbol{\pi = log(\frac{P_x}{P_h})} \quad \text{(I.4)}$$

Ainsi

$$\boldsymbol{log\left(\frac{1}{C}\right) = k_1\pi - k_2\pi^2 + k_3\sigma + k_4} \quad \text{(I.5)}$$

Un énorme nombre d'équations de QSAR ont été rapportés dans la littérature, grand nombres ayant une forme fonctionnelle beaucoup plus compliquée que l'équation originale de Hansch. Différents paramètres ont été employés dans des équations de QSAR, conçues pour représenter les caractéristiques hydrophobes, électroniques ou stériques de la molécule. Les propriétés qui ont choisies pour l'inclusion dans l'équation de QSAR devraient être aussi non-corrélatives.

## I.4. Méthodes récentes QSPR/QSAR

Depuis leur développement, et plus particulièrement depuis 2001, l'outil informatique a été utilisé avec succès pour la modélisation de propriétés et d'activités moléculaires [Micheli et al. (2003)].

### I.4.1. Modélisation par les SVM à partir de descripteurs traditionnels

Considérons tout d'abord l'approche qui consiste à représenter les molécules par un vecteur de descripteurs traditionnels.

Les SVM (Support Vector Machines) constituent dans ce cas une méthode performante pour la recherche de la fonction de régression f reliant ces vecteurs de descripteurs aux grandeurs modélisées. Cette approche a montré son efficacité pour la classification de molécules selon leur activité pharmacologique [Byvatov et al. (2003), Muller et al. (2005)], mais aussi en régression. Les modélisations d'activités telles que la toxicité de phénols [Yao et al (2004)] ou l'activité inhibitrice de pyrimidines [Burbidge et al. (2001)], ainsi que de propriétés telles que la solubilité aqueuse [Lind et Maltseva (1855)], ont en effet conduit à de bons résultats, souvent supérieurs à ceux obtenus par d'autres méthodes sur les mêmes bases.

### I.4.2. Fonctions noyaux pour molécules

Parallèlement à cette approche se sont développées les fonctions noyaux pour molécules (Graph Kernels for Structures) [Gärtner (2003) ; Kashima et al. (2003)]. Cette méthode consiste à comparer des graphes représentant les liaisons covalentes des molécules, ou des ensembles d'atomes dans l'espace, à l'aide de fonctions noyaux. À chaque molécule est associé un vecteur qui recense certains sous-éléments prédéfinis (par exemple des séquences d'atomes) du graphe correspondant. La distance entre deux graphes peut ensuite être évaluée grâce à une fonction noyau, qui compare les vecteurs qui leur sont associés.

L'utilisation de fonctions noyaux est particulièrement appropriée en bioinformatique, domaine dans lequel les données sont souvent de grande dimension.

Les vecteurs associés aux molécules peuvent être considérés comme des vecteurs de descripteurs, calculés cette fois à partir des graphes moléculaires, par opposition aux descripteurs traditionnels précédemment introduits. Cette représentation pose le problème du choix des descripteurs, mais de façon différente : les graphes associés aux molécules peuvent avoir des structures variées, et le choix des sous-éléments caractéristiques dépend de ces structures. On distingue par exemple :

Les noyaux de convolution [Haussler et al. (1999)], obtenus par convolution de plusieurs noyaux. Il est ainsi possible de comparer des séquences de longueurs différentes, à l'aide de différents noyaux, et d'obtenir un noyau global par convolution.

Les noyaux de Fisher [Jaakkola et al. (2000)] et de spectres [Leslie *et al*. (2002)], notamment utilisés pour la classification de séquences biologiques, telles que les séquences d'acides aminés [Ding et Dubchak (2001), Vert et al. (2004)].

Les noyaux marginalisés [Mahé et al. (2005), Tsuda et al. (2002), Mahé et al. (2004)], permettant de comparer des données structurées plus variées, et pouvant être adaptés à la modélisation de propriétés de molécules, ou au traitement d'images.

La notion de distance permet de modéliser la propriété étudiée par le biais d'outils de modélisation (réseaux de neurones, machines à vecteurs supports…). Cette approche, décrite plus en détail dans l'article [Goulon et al. (2005)], présente l'avantage d'éviter le calcul et la sélection des descripteurs, mais demande en revanche de définir les sous-éléments

caractéristiques des graphes. De plus, les modèles les plus performants demandent des temps de calcul qui peuvent devenir très grands.

### I.4.3. Autres méthodes de QSPR/QSAR

La modélisation d'une propriété ou d'une activité moléculaire nécessite de disposer d'informations caractérisant les molécules, informations à partir desquelles la grandeur en question est prédite. Il peut s'agir de descripteurs, mais il existe des méthodes alternatives de caractérisation des molécules. Nous présenterons dans un premier temps la méthode de contribution de groupes, qui, bien que datant des débuts de la modélisation QSAR, est toujours utilisée pour des applications particulières.

Nous décrirons ensuite comment les techniques de calcul et de comparaison de champs moléculaires se révèlent particulièrement efficaces pour la modélisation d'activités biologiques. Nous introduirons finalement des méthodes récentes qui considèrent que les structures des molécules contiennent des informations dont il est possible de tirer directement profit pour la modélisation.

#### I.4.3.1. Méthode de contribution de groupes

Les méthodes de contribution de groupes consistent à évaluer une propriété en décomposant la molécule en un ensemble de groupes fonctionnels, et en sommant les contributions relatives à des fragments de molécules [Leo *et al*. (1975), Klopman *et al*. (1994)]. Ces contributions sont déterminées à partir d'une base d'exemples de molécules, dont les valeurs de la propriété sont connues. Plusieurs types de groupes fonctionnels peuvent être définis. Ils sont généralement organisés en un système hiérarchique :

- Les groupes d'ordre 0 sont des atomes, et le calcul d'une propriété est effectué en sommant les contributions de chacun des atomes de la molécule considérée.

- La décomposition en groupes d'ordre 1 consiste à découper la molécule en groupes d'atomes (tels que $-CH_2-$, $-CH_3$ ou -OH). Leurs contributions à une propriété donnée sont sommées sans que l'environnement de chacun des groupes dans la molécule ne soit pris en considération. Ainsi, le groupe $-CH_2-$ a une contribution fixe, qu'il soit relié à un carbone ou à un groupe oxygéné. Ces groupes sont assez souvent employés, car ils permettent d'estimer rapidement la valeur d'une propriété, avec une précision parfois suffisante (par exemple pour l'enthalpie de formation). Cependant, les résultats

obtenus, pour la température d'ébullition par exemple, ne sont pas toujours satisfaisants. De plus, certains isomères peuvent conduire à la même décomposition : il est alors impossible de les distinguer par cette méthode.

- Les groupes d'ordre 2 sont constitués des atomes centraux de la molécule (autres que H), accompagnés de leurs plus proches voisins, c'est-à-dire de tous les atomes auxquels ils sont reliés. Contrairement aux groupes d'ordre 1, ceux d'ordre 2 tiennent compte de l'environnement des atomes.

Le tableau I.3 présente une comparaison des décompositions des molécules de butan-2-ol et 2-méthylpropan-1-ol, représentées sur la Figure I.1.

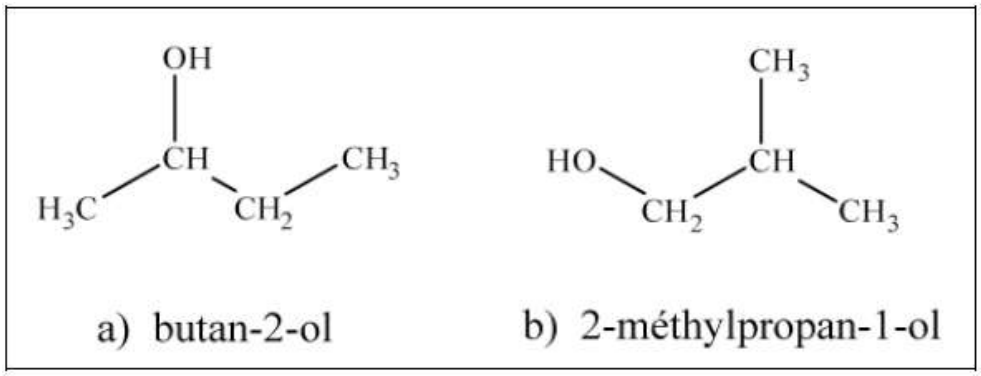



**Figure I.1:** Exemple de deux molécules, isomères de constitution [Goulon (2008)].

**Tableau I.3 :** Décomposition en groupes de deux isomères de constitution [Goulon (2008)].

| Méthode | Groupe | Nombre de groupes a) | Nombre de groupes b) |
|---|---|---|---|
| **Ordre 0** | C | 4 | 4 |
| | H | 10 | 10 |
| | O | 1 | 1 |
| **Ordre 1** | CH | 1 | 1 |
| | $-CH_2-$ | 1 | 1 |
| | $-CH_3$ | 2 | 2 |
| | -OH | 1 | 1 |
| **Ordre 2** | $C-(C)(H)_3$ | 2 | 2 |
| | $C-(C)_2(H)_2$ | 1 | 0 |
| | $C-(C)(O)(H)_2$ | 0 | 1 |
| | $C-(c)_3(H)$ | 0 | 1 |
| | $C-(C)_2(O)(H)$ | 1 | 0 |

### I.4.3.2 Analyse comparative de champs moléculaires (CoMFA)

La méthode CoMFA (pour Comparative Molecular Field Analysis) a été développée à partir de 1988 par Cramer. Elle est en particulier utilisée pour la modélisation d'activités biologiques, pour lesquelles les méthodes classiques se révèlent parfois peu performantes.

L'activité biologique d'une molécule dépend généralement de son interaction avec un récepteur donné. La modélisation de cette activité peut donc être réalisée en calculant les interactions de chaque molécule (ligand) avec ce récepteur, et en établissant une relation entre ces interactions et l'activité étudiée. Cette modélisation, repose sur le calcul de potentiels d'interactions moléculaires.

## I.5. Techniques de modélisation moléculaire

Ces dernières années ont vu l'apparition de nouvelles techniques de modélisation qui permet d'établir une relation directe entre les structures des molécules et leur propriétés. Ces techniques, reposent sur la représentation des molécules par des descripteurs moléculaires.

La modélisation par apprentissage consiste à trouver le jeu de paramètres qui conduit à la meilleure approximation possible de la fonction de régression, à partir des couples entrées / sorties constituant l'ensemble d'apprentissage. Le plus souvent, ces couples sont constitués d'un ensemble de vecteurs de variables (descripteurs dans le cas de molécules) $\left\{ \mathbf{x}_k, k = 1...n \right\}$, et d'un ensemble de mesures de la grandeur à modéliser $\left\{ y\left(\mathbf{x}_k\right), k = 1...n \right\}$. La détermination des valeurs de ces paramètres nécessite la mise en œuvre de méthodes d'optimisation qui diffèrent selon le type de modèle choisi. Nous allons tout d'abord présenter les principaux types de modèles faisant appel à l'optimisation paramétrique, sans contrainte, qui consiste à déterminer les paramètres optimaux par minimisation directe d'une fonction de coût par rapport aux paramètres du modèle.

### I.5.1. Régression linéaire multiple (RLM)

La régression linéaire multiple est la méthode la plus simple de modélisation. Elle consiste à rechercher une équation linéaire par rapport à ses paramètres, reliant la variable à modéliser $\mathbf{y}$ au vecteur d'entrées $\mathrm{n}\mathbf{x} = x_k \left\{ k = 1...q \right\}$.

Ces entrées peuvent être soit les variables primaires, soit des fonctions non paramétrées, ou à paramètres fixés, de ces variables (par exemple des monômes). L'équation linéaire recherchée est de la forme :

$$\boldsymbol{g(X,\theta) = \sum_{k=1}^{q} \theta_k x_k = X\theta} \quad \text{(I .6)}$$

Ou

$\boldsymbol{\theta = \{\theta_k, k = 1 \ldots q\}}$ est le vecteur des paramètres ; **X**, matrice des observations de taille ($N$, $q$), est définie comme la matrice dont les éléments de la colonne $k$ prennent pour valeurs les $N$ mesures de la variable $k$. Pour chaque élément $i$ de la base d'apprentissage, le résidu est défini comme la différence entre la valeur de la grandeur à modéliser pour cet élément $i$ et l'estimation du modèle :

$$\boldsymbol{R_i = y^i - g(x^i, \theta)} \quad \text{(I .7)}$$

L'apprentissage est réalisé par minimisation de la fonction de coût des moindres carrés, qui mesure l'ajustement du modèle $g$ aux données d'apprentissage :

$$\boldsymbol{J(\theta) = \sum_{i=1}^{N} (R_i)^2 = \sum_{i=1}^{N} \left[y^i - g(X^i, \theta)\right]^2 \|y - X\theta\|^2} \quad \text{(I .8)}$$

La fonction $j(\boldsymbol{\theta})$ est une fonction positive quadratique en $\boldsymbol{\theta}$ son minimum est unique. Il est donné par :

$$\boldsymbol{\theta_{mc} = \left(X^T X\right)^{-1} X^T y} \quad \text{(I .9)}$$

Les paramètres $\boldsymbol{\theta_k}$ obtenus sont appelés coefficients de régression partielle ; chacun d'eux mesure l'effet de la variable explicative $x_k$ concernée sur la propriété modélisée lorsque les autres variables explicatives sont maintenues constantes. La régression linéaire est facile à mettre en œuvre, et les coefficients $\boldsymbol{\theta_k}$ obtenus peuvent être interprétés : ils mesurent l'influence de chacune des variables sur la grandeur étudiée. Cependant, il est souvent nécessaire d'avoir recours à des modèles de plus grande complexité [Goulon (2008)].

### I.5.2. Réseaux de neurones

Les réseaux de neurones formels [Dreyfus *et al*. (2004)] étaient, à l'origine, une tentative de modélisation mathématique des systèmes nerveux, initiée dès 1943 par McCulloch et Pitts. Les réseaux de neurones ont pour but de modéliser des processus, à partir d'exemples de couples entrées / sorties. Ils ont la propriété d'approximation universelle : un réseau de neurones comportant un nombre fini de neurones cachés, de même fonction d'activation, et un neurone de sortie linéaire, est capable d'approcher uniformément, avec une précision arbitraire, toute fonction bornée suffisamment régulière, sur un domaine fini de l'espace de ses variables. De plus, il s'agit d'approximateurs parcimonieux : une approximation par un réseau de neurones nécessite en général moins de paramètres que les approximateurs usuels (toutes les notions concernant cette méthode de modélisation seront détaillées dans le chapitre III)

Notons ici que très peu de travaux ont été publiés sur l'application des RNA pour la corrélation et/ou la prédiction des propriétés thérapeutiques des molécules. Nous allons exposer chronologiquement les principaux travaux antérieurs effectués sur la modélisation moléculaire par les réseaux de neurones :

- **Miguel et al (2003)** ont étudié la combinaison des réseaux de neurones artificiels de type MLP (perceptron multicouches) et l'analyse discriminante linéaire (LDA) pour la prédiction de l'activité antibactérienne par l'utilisation d'un ensemble de descripteurs moléculaires sélectionnés aléatoirement pour discriminer entre les médicaments antibactériens et non-antibactériens. Les résultats ont confirmé la capacité discriminante des descripteurs topologiques proposés et l'efficacité de l'utilisation combinée de l'LDA et MLP dans la sélection de nouvelles structures ayant une activité antibactérienne. Cet article peut être considéré comme le seul travail qui a une relation directe avec l'objectif de notre étude.

- **Adina et al (2005)** ont conçu une méthode de réseaux de neurones pour prédire l'activité biologique qui se repose sur une représentation locale du ligand. Un ensemble de 42 composés, tel chaque composé sélectionné est représenté par un vecteur de quatre propriétés : le volume, logP, le moment dipolaire et un simple paramètre «stérique» relatif à la forme du composé, Cette représentation ligand a été testée en utilisant les réseaux de neurones sur un ensemble de dérivés urée-cycliques,

inhibant le VIH protéase. Le modèle neuronal conçu permet d'obtenir un coefficient de corrélation entre la prévision et l'expérimentation de l'ordre de 0,76 pour la base global et 0,88 lorsque trois valeurs aberrantes ont été éliminées.

- **Guerra et al (2009)** ont conçu un modèle neuronal basé sur une représentation numérique moléculaire en utilisant le programme CODES, de prévoir l'absorption par voie orale. Ce modèle prédit l'absorption des deux types de composés (de haute et basse absorption) avec un niveau de précision globale de 74%. Ils ont prouvé que la méthode CODES / RNA est non seulement utilisé comme moyen de criblage à haut débit et l'amélioration des capacités des procédures d'absorptions, mais aussi l'amélioration en corrélation in vitro–in vivo pourrait être abordée.

### I.5.3. Méthode générale de sélection du modèle

Lors de la modélisation, nous recherchons à la fois la complexité la mieux adaptée au problème étudié, et les paramètres du modèle correspondant. Le modèle doit en effet être de complexité suffisante pour rendre compte de la relation entre les variables explicatives et la grandeur modélisée, mais il ne doit pas être trop complexe, afin de ne pas être trop sensible au bruit présent dans les données. En effet, lorsque la complexité du modèle envisagé augmente, on observe généralement une diminution du coût d'apprentissage, car le modèle s'ajuste de plus en plus précisément aux données d'apprentissage. En revanche, l'erreur de généralisation diminue pour atteindre un minimum, puis augmente avec la complexité du modèle, car celui-ci est alors sur-ajusté aux données d'apprentissage, et au bruit présent dans ces données. Ce compromis est également connu sous le nom de dilemme biais-variance : le biais caractérise l'écart entre les estimations et les mesures, et la variance reflète l'influence du choix de la base d'apprentissage sur le modèle [Geman et al. (1992)].

De plus, lorsque les modèles envisagés ne sont pas linéaires par rapport à leurs paramètres, la sélection du modèle optimal ne consiste pas uniquement à choisir la meilleure famille de fonctions, comme c'est le cas pour des modèles linéaires (modèles polynomiaux ou combinaisons linéaires de fonctions non paramétrées). La sélection s'effectue donc souvent en plusieurs étapes.

- Pour les modèles non linéaires en leurs paramètres, la première sélection s'effectue au sein d'une famille de fonctions donnée. En effet, la fonction de coût n'étant pas

quadratique en les paramètres du modèle, elle ne possède pas un minimum unique. L'optimisation de cette fonction peut donc conduire à des minima différents, donc à différents modèles de même complexité. Il est donc nécessaire de réaliser plusieurs apprentissages, pour une complexité donnée, et de choisir celui qui est susceptible de posséder les meilleures propriétés de généralisation.

- Il faut ensuite sélectionner, parmi les meilleurs modèles de complexités différentes, celui qui présente les meilleures capacités de généralisation.

L'erreur de généralisation ne peut pas être évaluée exactement ; la base de données disponible est en effet de taille limitée, et la distribution de probabilité des données, dont dépend cette erreur, est généralement inconnue. Il est donc nécessaire de trouver une approximation de cette erreur de généralisation. Nous allons ainsi présenter les méthodes les plus utilisées pour effectuer ces sélections.

La méthode la plus commune, dite méthode de validation simple ou hold-out, consiste à construire, à partir de l'ensemble des données, deux ensembles : les paramètres du modèle sont ajustés sur la base d'apprentissage, et l'erreur de généralisation est évaluée sur la base de validation. L'erreur en validation est définie de façon analogue au coût d'apprentissage :

$$\boldsymbol{E_V = \sqrt{\frac{1}{N_V}\sum_{i=1}^{N_V}(R_i)^2}} \qquad \text{(I.10)}$$

Où $N_V$ est le nombre d'exemples dans la base de validation.

Pour des modèles de complexité donnée, la méthode de sélection consiste à effectuer plusieurs apprentissages à partir de différentes initialisations des paramètres, lorsque l'on a affaire à un modèle non linéaire en ses paramètres. Il faut alors choisir, parmi les modèles obtenus, celui qui fournit la plus faible erreur en validation. Il s'agit d'une méthode très rapide, mais d'efficacité limitée.

De plus, les exemples de la base de validation ne sont pas utilisés pour l'estimation des paramètres du modèle, et peuvent faire défaut : le modèle n'étant pas ajusté à certains types d'exemples, ses capacités de généralisation sont limitées. Par conséquent, l'erreur de généralisation est plus souvent estimée à l'aide d'autres méthodes, telles que la validation croisée ou le leave-one-out [Geman et al. (1992)].

## I.6. Conclusion

Le passage en revue de la littérature réalisé dans ce travail fait sortir principalement les points suivants :

- QSAR est un facteur principal dans le développement de la technologie pharmaceutique.
- Un énorme nombre d'équations de QSAR ont été rapportées dans la littérature. Différents paramètres ont été employés dans des équations de QSAR, conçues pour représenter les caractéristiques hydrophobes, électroniques ou stériques de la molécule.
- La modélisation de l'activité biologique peut être réalisée en calculant les interactions de chaque molécule avec son récepteur, et en établissant une relation entre ces interactions et l'activité étudiée, déférentes méthodes sont utilisées pour établir ce type de relations, tel que la méthode de contribution de groupes te l'analyse comparative des champs moléculaires
- Les méthodes récentes considèrent que les structures des molécules contiennent des informations dont il est possible de tirer directement profit pour la modélisation.
- La modélisation vise à fournir un modèle qui soit non seulement ajusté aux données d'apprentissage, mais aussi capable de prédire la valeur de la sortie sur de nouveaux exemples, ce modèle doit rendre compte de la relation entre les variables explicatives et la grandeur modélisée avec une complexité minimale, et une bonne capacité de généralisation. Tel que la régression linéaire multiple et la modélisation par réseaux de neurones artificiels qui peuvent être considérées comme méthodes prometteuses très utilisées dans la modélisation moléculaire

# *Chapitre II*

# *Descripteurs moléculaires, calcul et méthodes statistiques de réduction*

**CHAPITRE II**

**DESCRIPTEURS MOLÉCULAIRES, CALCUL ET MÉTHODES STATISTIQUES DE RÉDUCTION**

## II.1. Introduction

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, 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. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels [Goulon (2008)].

Dans ce chapitre nous présentant tout d'abord les notions générales sur les descripteurs moléculaires et la classification des différentes types des descripteurs existant ainsi que leur méthodes de calculs par divers logiciels, notre études sur l'application de ces descripteurs seras dans le domaine de la modélisation moléculaire QSAR, Nous présenterons le problème de la sélection de ces descripteurs pour des meilleures performances du modèle, ainsi que les stratégies les plus efficaces pour le résoudre toute en utilisant les méthodes statistiques.

## II.2. Descripteurs moléculaires

On en dénombre aujourd'hui plus de 3000 types de descripteurs moléculaires, 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 [Hansch *et al*. (1995)], 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 celles qui sont les plus pertinentes pour la modélisation.

## II.2.1. Différents types des 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.

### II.2.1.1. 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 [Goulon (2008)].

### II.2.1.2. 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 [Wiener (1947)], l'indice de Randić [Randić (1975)], l'indice de connectivité de valence de Kier-Hall [Kier *et al*. (1976)] et l'indice de Balaban [Balaban (1982)]. 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 Randić

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,. 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.

### II.2.1.3. 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, 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 :

- **Descripteurs géométriques** les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.

- **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 [Heritage *et al*. (1998)] est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type MoRSE [Schuur et *al*. (1996)] *(*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.

Par ailleurs, le calcul de certains descripteurs demande une étape préliminaire d'alignement des molécules, pour les placer dans une orientation commune. Nous reviendrons sur ces descripteurs, qui sont souvent spécifiques à certaines méthodes (CoMFA, CoMSIA...).

Nous citerons un dernier type de descripteurs utile pour la recherche par similarité et donc la diversité : les graphes réduits [Monge (2006)] Ils codent l'enchaînement de groupements prédéfinis. Il existe différents types de graphes réduits. Un exemple est présenté Figure II.1

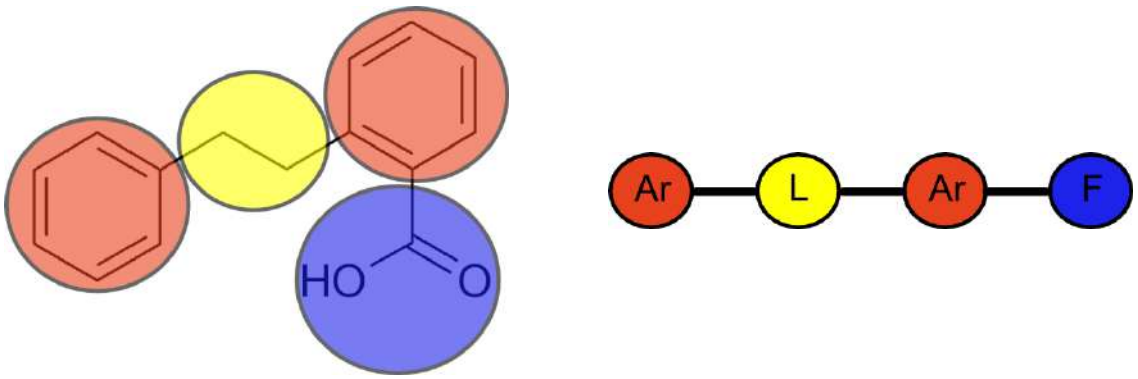



**Figure II.1 :** Exemple de molécule convertie en graphe réduit. Le graphe ne garde que l'enchaînement de grands types d'éléments structuraux. ***Ar*** correspond à un groupe aromatique, ***L*** à un linker, ***F*** à un groupe fonctionnel [Monge (2006)]

On notera que les graphes réduits ne gardent pas les types atomiques, mais codent les molécules en fonction de types de groupements très généralistes. Ces descripteurs peuvent donc être utilisés pour chercher, à partir d'un composé actif, d'autres composés actifs de familles chimiques très différentes.

## II.2.2. Sélection et application

### II.2.2.1. Carte d'identité d'une molécule

Afin d'établir des modèles de prédiction du comportement biopharmaceutique des molécules (absorption, métabolisme, élimination, toxicité, etc.). Des outils de chémo-informatique permettent une meilleure manipulation de l'ensemble des données bien collectés.

Cette modélisation repose sur l'utilisation des descripteurs moléculaires (de topologie, de géométrie, d'interactions électroniques et/ou lipophiles, de propriétés quantiques ou graphes moléculaires), véritable carte d'identité d'une molécule.

La Figure II.2 résume les différents niveaux de description d'une molécule en commençant par les plus simples, le nom et la représentation chimique, jusqu'aux descripteurs 3D faisant appel à des volumes et des surfaces moléculaires (4D).

Les descripteurs 1D sont déduits de la formule brute (nombres d'atomes, nombre de C, O, N …)., les descripteurs 2D de la structure plane de la molécule (donneurs/accepteurs d'hydrogène, nombre de cycles…). Les descripteurs 3D sont déduits de la structure 3D de la molécule (volume, surface accessible au solvant…). La notion de 4D peut être employée quand on explore l'espace autour de la molécule (surfaces moléculaires).

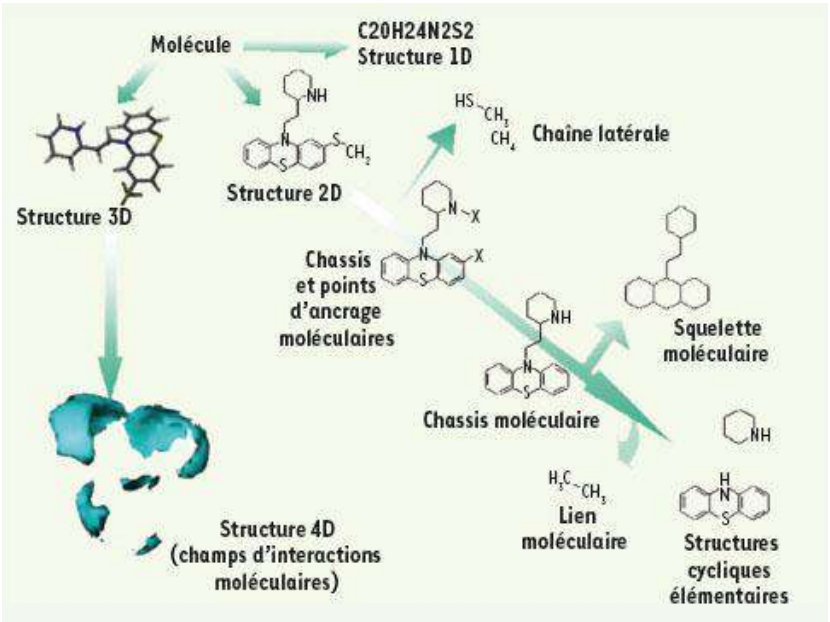



**Figure II.2 :** Niveaux de description d'une molécule utilisés dans le criblage virtuel [Philippe et *al*. (2009)].

### II.2.2.2. Utilisation des descripteurs moléculaires

Ces descripteurs sont utilisés pour filtrer des bases de structures chimiques permettant d'éliminer des composés dont les caractéristiques sont trop éloignées de celles d'un médicament.

Les règles dites de Lipinsky utilisent, par exemple, une combinaison de propriétés physicochimiques essentielles dans l'absorption orale. Et on peut également vouloir s'affranchir de certains fragments chimiques comme :

- les fonctions réactives susceptibles de former des liaisons covalentes [Opera (2000)] ;
- les fonctions connues comme générant des faux positifs pour des raisons physico-chimiques du type agents chélatants, inhibiteurs suicides, composés polyanioniques, etc. [Richton (2003)] ;
- les fonctions chimiques susceptibles de réaliser des agrégats (promiscuous aggregating inhibitors) et faussant le résultat biologique [Serdler *et al*. (2003)].

Ces approches sont souvent utilisées dans des processus de focalisation de bases. Elles permettent, dans le travail de Monge et al. [Monge (2006)], de réduire une base chimique de 173 803 composés à 10 000 molécules ayant des propriétés compatibles avec celles d'un futur médicament.

Mais elles ont également été utilisées avec succès pour enrichir des chémiothèques avec des apports externes tant sur la base du potentiel médicamenteux que d'un point de vue diversité chimique.

### II.2.2.3. Sélection à l'aide de modèle QSAR

Ensuite on s'engage la construction du modèle, par des approches QSAR (quantitative structure activity relationship) le plus souvent [Sperandio (2008)]. Ces techniques permettent de relier par une relation mathématique les descripteurs chimiques, soit à l'activité biologique dans une approche de relation structure-activité quantitative, soit à une propriété (physicochimique ou pharmacocinétique) dans le cas des relations structure-propriété quantitative (QSPR pour quantitative structure property relationship). De tels modèles QSAR ou QSPR, basés sur des descripteurs 2D ou 3D, sont le point de départ de nombreux processus de sélection de molécules. Ces modèles sont tous construits sur le même mode mettant en œuvre un jeu de référence appelé jeu d'apprentissage permettant de sélectionner le(s) descripteur(s) le(s) plus adapté(s) dans la construction d'un modèle. Les modes de construction de ces modèles peuvent être simples (régression linéaire ou multilinéaire) ou plus sophistiqués permettant de rechercher la meilleure combinaison de descripteurs (algorithmes génétiques, réseaux de neurones, forêts aléatoires, séparateurs à vaste marge).

Elle sera suivie d'une étape de validation appelée validation externe et réalisée avec un nouveau jeu de données, souvent généré postérieurement, et qui va permettre de bien mesurer la pertinence du modèle et de définir le domaine d'application.

Le pouvoir prédictif de ces modèles et leur domaine d'application pour de futures séries chimiques dépendent de la composition du jeu d'apprentissage utilisé pour construire le modèle. Il faut se rappeler qu'un modèle QSAR parfait ne pourra donc pas permettre une prédiction pour toute sorte de substances (molécules trop différentes de celles du jeu d'apprentissage) et avoir un niveau de précision supérieur à celui de la méthode expérimentale qui a permis de générer les données du jeu d'apprentissage.

## II.3. Méthodes de calculs des descripteurs

Afin de calculer les descripteurs moléculaires d'une molécule donnée, on doit passer par plusieurs étapes avant d'établir la base de données contenant les valeurs de ces descripteurs, la première de ces étapes consiste à représenter la structure moléculaire dans forme bien déterminée après avoir dessiné sa structure moléculaire.

Différents formats de fichiers ont été décrits pour représenter la structure des molécules. Les plus connus sont SMILES, SDF, mol2 (Tripos) et PDB (Protein Data-Bank). Le format SMILES (Simplified Molecular Input Line Entry System) permet de représenter les molécules sur une seule ligne de caractères. Les tirets simples ou doubles représentent des liaisons de même nature, et un système de parenthèses permet de repérer les embranchements au niveau de la molécule. Les chiffres permettent d'identifier les connections de cycle. Le problème de ce code est sa dégénérescence. En effet, à une même molécule peuvent correspondre plusieurs codes SMILES.

Le code SMILES canonique a permis de résoudre ce problème. L'avantage principal de ce format est qu'il est léger, permettant le stockage de bases de molécules de grande taille dans un espace mémoire restreint. Cependant, d'autres informations sont souvent nécessaires, notamment la structure 3D de la molécule. Des formats comme PDB, SDF ou mol2 contiennent les coordonnées cartésiennes de chaque atome, et éventuellement leur charge partielle (mol2), ainsi que toutes les liaisons formant la molécule. OpenBabel permet la conversion d'un format 3D vers un autre format 3D.

La conversion d'un format 2D vers un format 3D nécessite de générer de l'information et est rendue possible grâce à des outils comme Omega, Corina Catalyst (www.accelrys.com), Concord (www.tripos.com) ou Frog (bioserv.rpbs.jussieu.fr). Selon la molécule, différents états d'ionisation et stéréo-isomères peuvent être générés et il est important de le vérifier avant [Toshio (1995)].

## II.3.1. SMILES (Simplified Molecular Input Line Entry Specification)

### II.3.1.1. Notion générale

SMILES est un langage symbolique de description de la structure des molécules chimiques sous forme de courtes chaînes de caractères. Les chaînes SMILES peuvent être importées dans la plupart des éditeurs de structure moléculaire pour être reconverties en représentation 2D ou en modèle 3D, et par conséquent, permettent de construire simplement des bases de données informatiques, la figure II.3 représente un exemple de génération de la notation SMILES pour la molécule de ciprofloxacine.

**A :** Structure de la molécule, la Ciprofloxacine.

**B :** Suppression des hydrogènes, ouverture des cycles et numérotation.

**C :** Arbre couvrant la structure dé-cyclisée.

**D :** Chaîne SMILES correspondante, le code couleur correspond à la couleur des branches sur le panneau (www.opensmiles.org).

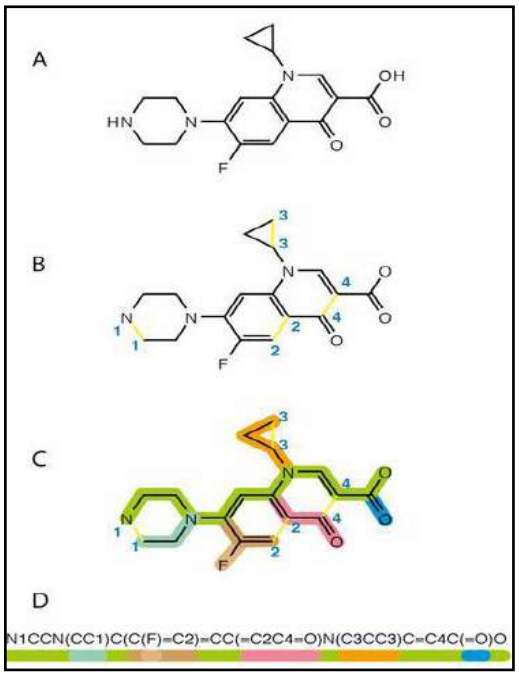



**Figure II.3 :** Exemple de dérivation de représentation SMILES

La définition du langage SMILES a été initialement introduite par David Weininger [Weininger (1988)]. Plusieurs extensions et modifications ont été ajoutées depuis, notamment par la société Daylight Chemical Information Systems, dont Weininger est le fondateur. Il existe d'autres langages à ligne de texte pour représenter les molécules, dont notamment InChl, introduit par l'IUPAC. SMILES est cependant plus répandu, car plus facile à lire par l'utilisateur humain et surtout parce qu'il est implémenté dans de nombreux logiciels de chimie moléculaire.

### II.3.1.2. Principes de SMILES et théorie des graphes

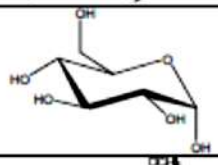
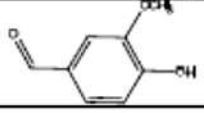
SMILES est basé sur la représentation d'une molécule chimique sous forme d'un graphe, au sens mathématique du terme. Les atomes sont les sommets du graphe, et les liaisons chimiques en sont les arêtes. Chaque sommet du graphe est étiqueté avec le symbole de l'atome correspondant (C, N, O…).

La chaîne de caractère SMILES correspondant à une molécule donnée est obtenue en parcourant progressivement le graphe suivant un algorithme de parcours en profondeur et en imprimant successivement le symbole de chaque sommet traversé.

Le graphe de la molécule est préalablement simplifié en supprimant tous les atomes d'hydrogène (notation implicite) et en ouvrant les cycles pour transformer le graphe en arbre couvrant (spanning tree). À chaque fois qu'un cycle est ouvert, un suffixe numérique est ajouté pour indiquer la connexion des sommets correspondant à la liaison chimique supprimée. Des parenthèses sont utilisées pour indiquer les points de branchement sur l'arbre.

De par la nature de ce langage, il existe, pour une même molécule organique constituée de plusieurs atomes autres que l'hydrogène, plusieurs écritures possibles, selon l'atome de départ et l'ordre de parcoure de la structure moléculaire. Ainsi, CCO, OCC, C(C)O et C(O)C sont toutes quatre des représentations valides de la molécule d'éthanol CH3–CH2–OH [Weininger (1988)]. Le tableau II.1 présente la notation SMILES pour quelques molécules.

**Tableau II.1 :** Application de la notation SMILES sur des molécules de natures variées [Weininger (1988)].

| Molécules | Structure | Notation Smiles |
|---|---|---|
| Diazote |  | N#N |
| Isocianate de méthyle |  | CN=C=O |
| Glucose |  | O[C@H]([C@@H](CO)O[C@H](O)[C@@H]1O)[C@@H]1O |
| Vanilline |  | O=CC1=CC(OC)=C(O)C=C1 |

### II.3.1.3. Conversion

Il existe des algorithmes informatiques (par exemple le logiciel ACD/ChemSketch,) utilisés pour transformer automatiquement le schéma d'une molécule donnée en chaîne SMILES. Il n'existe en général pas de description unique d'une molécule et que plusieurs SMILES peuvent décrire la même structure chimique. Ceci dépend de l'atome dont on part pour décrire la structure, de l'ordre dans lequel on décrit les branches du graphe moléculaire et de la manière dont on ouvre les cycles, la figure II.4 représente la méthode de génération de la notation SMILES à partir de la structure moléculaire en utilisant le ChemSketch, ce dernier peut assurer aussi le cas inverse représenté dans la figure II .5.

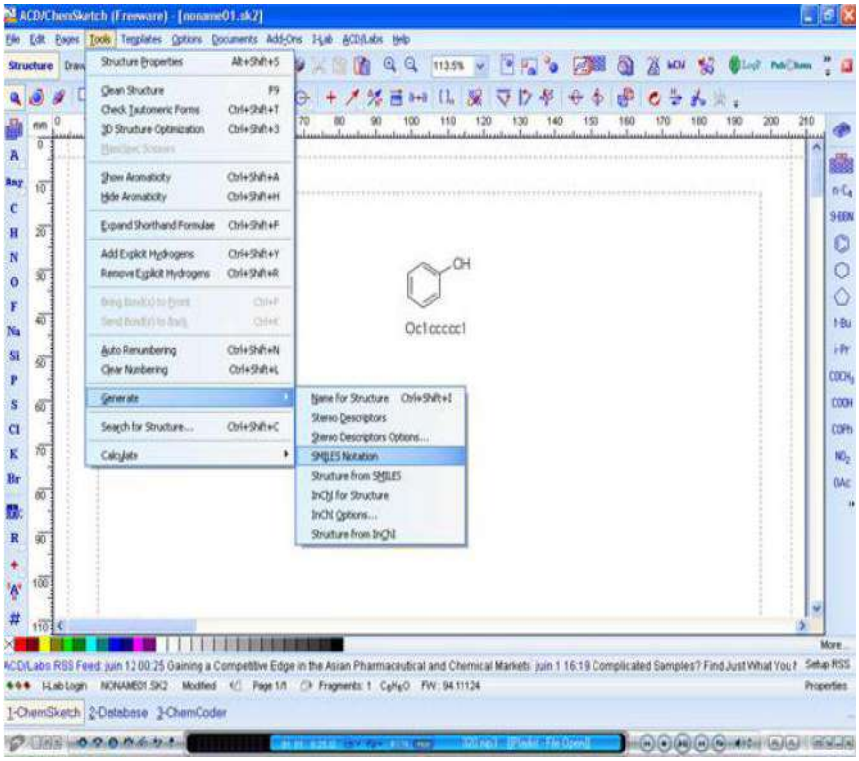



**Figure II.4 :** Génération de la notation SMILES à partir de la structure moléculaire**.**

Réciproquement, la description SMILES d'une molécule peut être reconvertie en représentation chimique 2D classique, au moyen d'algorithmes de représentation de graphe

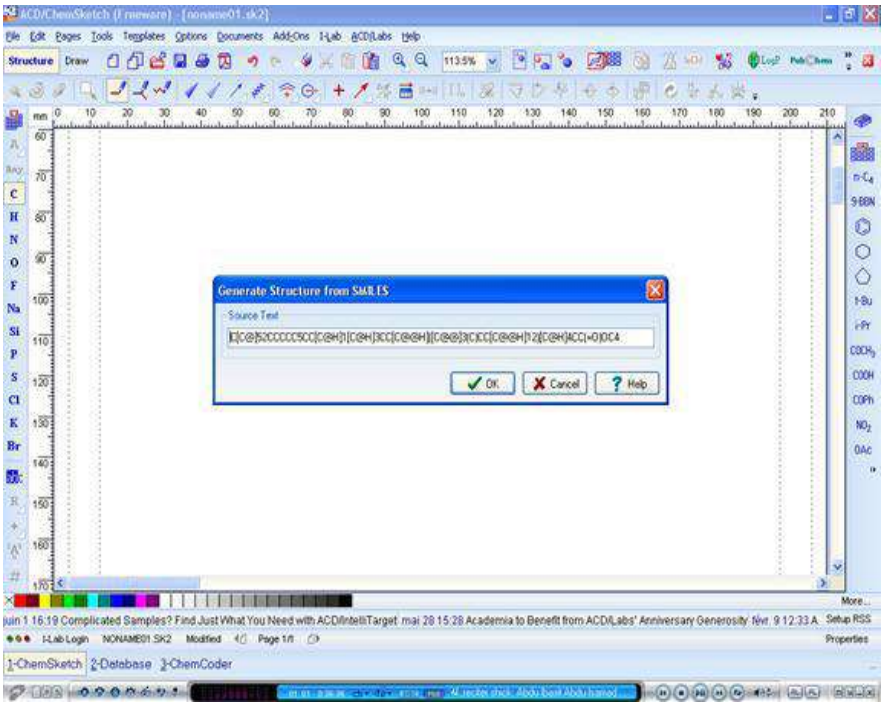



**Figure II.5 :** Création de la structure moléculaire à partir de la notation SMILES.

## II.3.2. Logiciels de calculs des descripteurs

La méthode SMILES permet d'échanger numériquement des descripteurs de molécules arbitrairement complexes par le biais d'un fichier texte classique .ces descripteurs peuvent être calculées par d'autres logiciels tel que DRAGON® en premier lieu (c'est une application utilisée pour le calcul de descripteurs moléculaires à l'origine développé par la Chimio-métrie Milano et QSAR Research Group).

Ces descripteurs peuvent être utilisés pour évaluer la structure moléculaire-activité ou structure-propriété des relations, ainsi que pour l'analyse de similitude et de dépistage.

On peut subdiviser les types de logiciels disponibles en trois domaines :

## II.3.2.1. Domaine privé

La figure II.6. Présente les différents logiciels utilisés dans le domaine privé

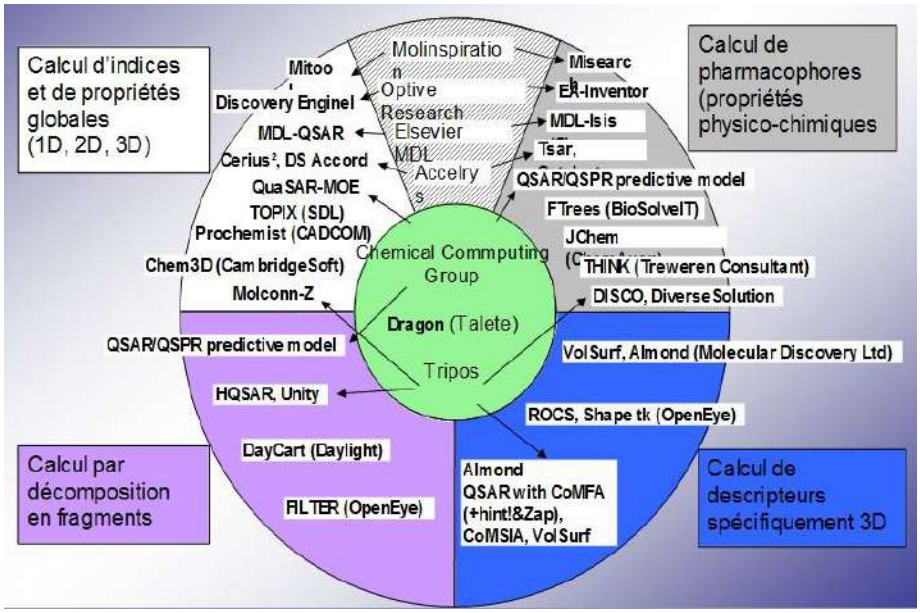



**Figure II.6 :** Différents logiciels utilisés dans le domaine privé (http://www.daylight.com, 2006; http://www.epa.gov, 2006)

## I.3.2.2. Domaine public

La figure II.7. Présente les différents logiciels utilisés dans le domaine public

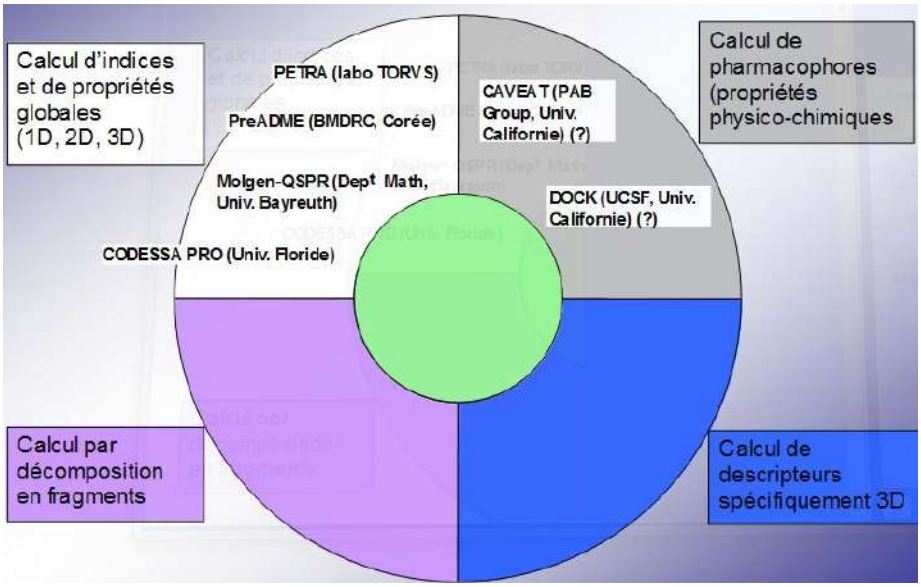



**Figure II.7 :** Différents logiciels utilisés dans le domaine public (http://www.daylight.com, 2006; http://www.epa.gov, 2006).

### II.3.2.3. Domaine académiques

Peu de logiciels freeware sont disponible même pour les académiques, donc on peut citer les logiciels les plus utilisés suivants :

- **Academic package**
  - ChemAxon (Marvin, jchem): calcul de descripteurs topologiques
  - OpenEye Sci. Soft. (FILTER)
  - Treweren Consultant (THINK)

- **Calcul en ligne**
  - E-Dragon (logiciel Dragon en ligne)
  - Molgen
  - PreADME
  - CORINA (autorise la transformation de structure 2D→3D de 1000 molécules gratuitement)

## II.4. Réduction du nombre de descripteurs

Comme nous l'avons rappelé, 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 sur-ajusté à ces exemples, et incapable de prédire la grandeur modélisée sur de nouvelles observations.

Il est donc nécessaire de réduire la dimensionnalité 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.

### II.4.1. Analyse en composantes principales

L'analyse en composantes principales (ou ACP) [Jolliffe (2002)], est une technique d'analyse de données utilisée pour réduire la dimension de l'espace de représentation des données, pour réaliser une analyse en composante principale on doit passer par les étapes suivantes :

1. **Quantifier la variabilité contenue dans un tableau de données**

On appelle inertie la quantité d'information contenue dans un tableau de données.

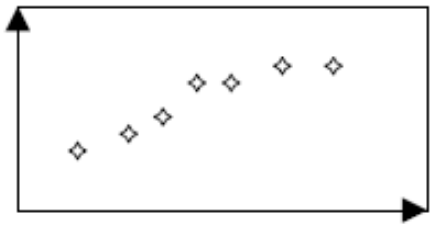
- Une inertie nulle signifie que tous les individus sont presque identiques.
- L'inertie du nuage sera égale à la somme des variances des j caractères.
- Si les j caractères sont centrés-réduits, l'inertie sera égale à j.

**2. Projeter sur un plan un tableau de données à j dimensions**

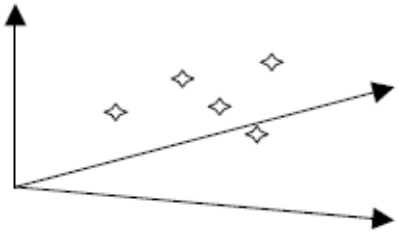
- L'ACP est une méthode descriptive.
- Son objectif est de représenter sous forme graphique l'essentiel de l'information contenue dans un tableau de données quantitatif.
- Dans un tableau de données à j variables, les individus se trouvent dans un espace à j dimensions.

**3. Représentation graphique**

- Lorsqu'il n'y a que deux dimensions (largeur et longueur par exemple), il est facile de représenter les données sur un plan :



- Avec trois dimensions (largeur, hauteur et profondeur par ex.), c'est déjà plus difficile



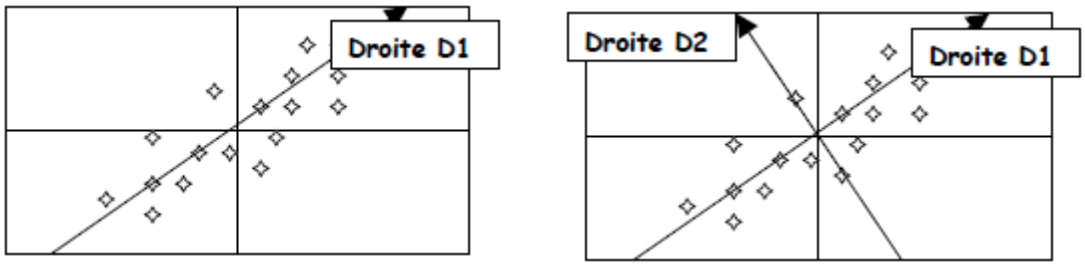
- Mais au-delà de 3 dimensions, il est impossible de représenter les données sur un plan ou même de les visualiser mentalement.

**4. Projeter la réalité sur un plan**

- Nous avons l'habitude de dessiner ou photographier la réalité.
- Nous naturellement passons d'un espace à 3 dimensions à un espace à 2 dimensions.
- Selon le point de vue, l'information retenue ne sera pas la même.
- L'ACP nous propose un point de vue permettant de voir au mieux les individus d'un tableau.

**5. Résumer les données**

- Lorsqu'on projette les données sur un plan, on obtient un graphique déformé de la réalité.
- Le rôle de l'ACP est de trouver des espaces de dimensions plus petites minimisant ces déformations.
- On utilise un espace à 2 dimensions (un plan). Ce plan est appelé le plan principal. Il est constitué de deux droites perpendiculaires.
- La méthode consiste à calculer la première droite D1 de façon à maximiser les carrés des distances de projection des points sur la droite.




- Puis une 2éme droite D2 perpendiculaire à la première

**6. Composantes principales**

- Les droites D1 et D2 sont des caractères synthétiques obtenus par des combinaisons linéaires avec les variables d'origines.
- Ces droites sont appelées composantes principales, ou axes principaux.
- La première composante principale doit "capturer" le maximum d'inertie du tableau des données. La variance des individus doit être maximale.
- Il reste un résidu non expliqué par cette première composante. C'est sur ce résidu qu'est calculée la deuxième composante principale.

## 7. Caractères des composantes principales

- La première composante principale "capture" le maximum d'inertie du tableau des données.
- La deuxième composante principale est un complément, une correction de la première.
- La deuxième composante principale doit avoir une corrélation linéaire nulle avec la première (orthogonalité).
- Il n'y a pas de redondance d'information entre deux composantes principales.
- On calcule les autres composantes de la même manière.

## 8. ACP en résumé

La figure II.8 représente l'organigramme qui résume les différentes étapes d'une analyse ACP

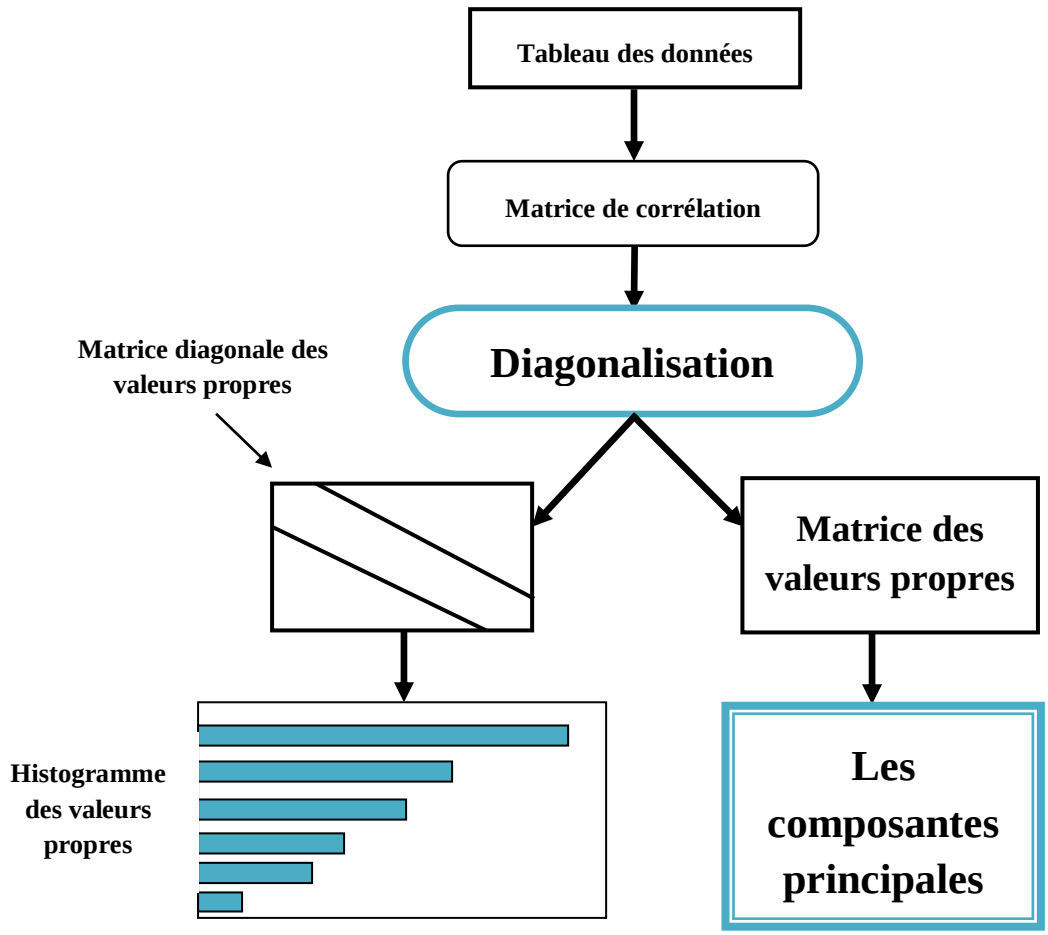



**Figure II.8 :** Organigramme résume les différentes étapes d'une analyse ACP.

## 9. Exemple de résultats d'ACP

Les figures II.9 et II.10 représentent les résultats obtenus après la réalisation d'une analyse en composantes principales en termes d'espace des individus et espaces des variables aussi [Bouchier (2006)].

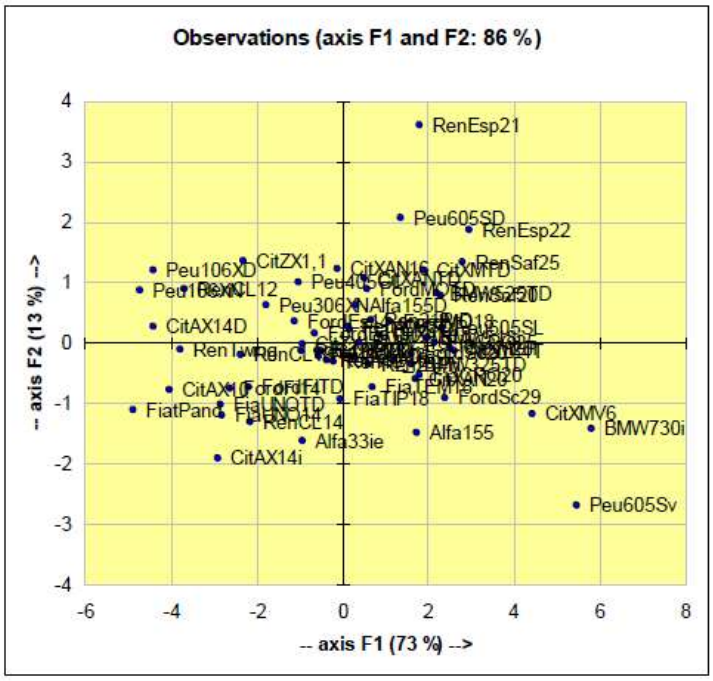



**Figure II.9 :** Résultats d'ACP (espace des individus) [Bouchier (2006)].

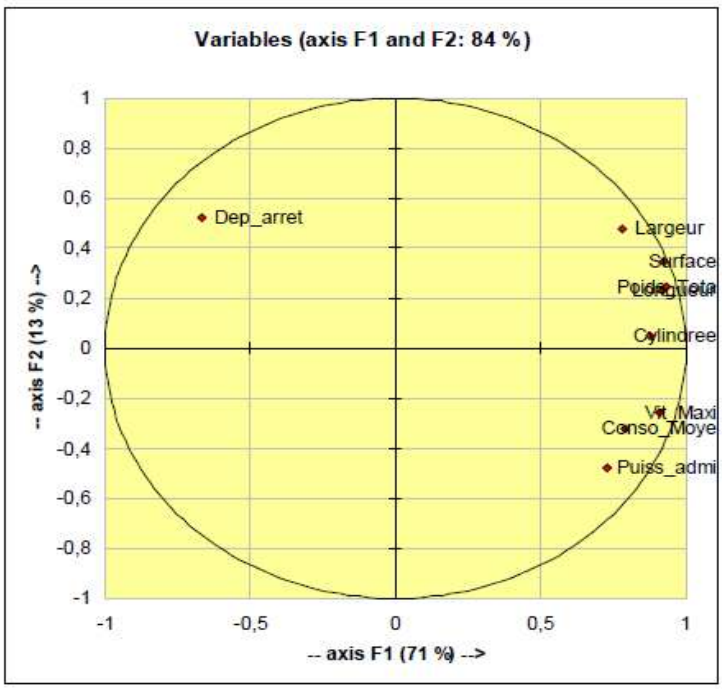



**Figure II.10 :** Résultats d'ACP (espace des variables) [Bouchier (2006)].

### II.4.2. Méthode de régression des moindres carrés partiels

La régression des moindres carrés partiels [Wold (1966)] (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.

### II.4.3. Sélection des variables pertinentes

L'analyse en composantes principales a pour but de réduire les corrélations entre les variables, mais cette étape de réduction est indépendante de la grandeur modélisée. Or, les variables calculées n'ont pas nécessairement une influence sur cette grandeur. Il est nécessaire d'éliminer celles dont l'influence est inférieure à celle du bruit, et de sélectionner uniquement les plus pertinentes d'entre elles. Il est possible, à partir de $p$ descripteurs, de former $2_p$ sous-ensembles de variables, donc $2_p$ modèles, dont il faudrait comparer les performances pour déterminer le meilleur jeu de descripteurs. Cette méthode permet d'envisager toutes les combinaisons possibles de descripteurs, mais sa mise en œuvre est très lourde. Des approches alternatives ont donc été développées.

- **Sélection progressive :** consiste à incorporer les variables au modèle une à une, en sélectionnant, à chaque étape, la variable dont la corrélation partielle avec la grandeur modélisée est la plus élevée. À l'inverse, lors de l'élimination progressive, on débute la modélisation avec l'ensemble des descripteurs, en les éliminant un par un jusqu'à obtenir le meilleur jeu de composantes.

- **Sélection pas à pas** : est une combinaison des deux méthodes évoquées précédemment. Les variables sont incorporées une à une dans le modèle, par sélection progressive. Cependant, à chaque étape, on vérifie que les corrélations partielles des variables précédemment introduites sont encore significatives.

Une méthode plus efficace est celle du **descripteur sonde** [Stoppiglia *et al*. (2003), Stoppiglia (1997)]. Cette sélection est généralement effectuée à partir d'un modèle linéaire en ses paramètres, par exemple un modèle polynomial. En effet, la pertinence d'une variable est indépendante du modèle considéré, et les variables ainsi sélectionnées peuvent être utilisées comme variables de modèles plus complexes. La méthode consiste à ajouter un descripteur aléatoire, appelé « descripteur sonde », à la liste des variables candidates. Ces variables sont alors classées selon leur pertinence par une méthode d'orthogonalisation. On se fixe le risque que l'on est prêt à accepter de conserver une variable candidate alors qu'elle est moins bien classée que le descripteur sonde, et l'on en déduit le rang, dans le classement, au-delà duquel les variables sont éliminées.

## II.5. Conclusion

Ces dernières années ont vu l'apparition de nouvelles techniques de modélisation qui permet d'établir une relation directe entre les structures des molécules et leur propriétés. Ces techniques, tout comme la méthode que nous avons adoptée, reposent sur la représentation des molécules par des descripteurs moléculaires.

Nous avons vu comment peut-on traiter les molécules chimiques par les outils informatiques dans le but de déterminer ces descripteurs moléculaires qui représentent les différentes propriétés chimiques, physiques et thérapeutiques donc on a besoin d'une représentation de ces molécules ainsi que des bases de données. La notation SMILES est l'une des méthodes les plus utilisées. Elle permet d'échanger numériquement des descriptions de molécules arbitrairement complexes. Par le biais d'un fichier texte classique la représentation SMILES étant très compacte, elle permet de stocker simplement des fichiers de plusieurs centaines de milliers de molécules (chimiothèques).

# *Chapitre III*

# *Réseaux de Neurones Artificiels (RNA)*

# CHAPITRE III
# RÉSEAUX DE NEURONES ARTIFICIELS (RNA)

## III.1. Introduction

Un réseau de neurones artificiel (RNA) est un processeur massivement distribué en parallèle, composé d'unités de traitement simples, qui a une propension naturelle pour stocker la connaissance empirique et de la rendre disponible à l'usage. Il ressemble au cerveau humain sur deux aspects:

- La connaissance est acquise par le réseau de son environnement à travers un processus d'apprentissage.
- Les forces de liaison inter-neuronale, connues sous le nom de poids synaptiques, sont employées pour stocker la connaissance acquise.

En abordant ce chapitre par cette définition d'un RNA, reproduite du livre de Haykin (1999), on peut alors, imaginer les vastes domaines d'application potentielle des RNA, concevoir l'intérêt qui leur a été réservé durant les deux dernière décennies, traduit par le nombre énorme de publications, et enfin comprendre pourquoi ils sont devenus un outil robuste de calcul scientifique dans plusieurs disciplines de sciences et d'ingénierie.

En raison de la disponibilité et la richesse bibliographique qui traite les fondements et les applications des RNA il est inutile de présenter avec détails toutes les notions théoriques des RNA,

## III.2. Historique

Pour les réseaux de neurones artificiels, tout commence en 1943, lorsque deux scientifique : le neuropsychiatre Warren S. McCulloch (1898-1972) et le mathématicien Walter Pitts (né en 1924) de l'Université de Chicago dans l'Illinois, conçoivent le premier modèle du neurone biologique, baptisé neurone formel ou automate à seuil [McCulloch et Pitts (1943)].

Un peu plus tard, en 1949, le neurophysiologiste Donald O. Hebb (1904-1985) publia ses travaux sur la « plasticité synaptique » qui sont à la base des mécanismes d'apprentissage. Il présente dans son ouvrage (The Organization of Behavior) une règle d'apprentissage qui porte encore son nom.

En 1957, Frank Rosenblatt développa le Perceptron au Laboratoire d'Aéronautique de l'Université de Cornell dans le but de comprendre la mémoire humaine, l'apprentissage et les processus cognitifs.

Le perceptron est un réseau de neurones artificiels inspiré du système visuel, possédant une couche de neurones "perceptive" et une couche de neurones "décisionnelle" [Rosenblatt (1962)]. Sa mise au point (en 1962) fut le départ des premières machines adaptatives. Ce réseau, qui parvient à apprendre à identifier des formes simples et à calculer certaines fonctions logiques, constitue le premier système artificiel exhibant une faculté jusque-là réservé au vivant, la capacité d'apprendre par l'expérience (l'apprentissage).

Dans la même période, en 1960, Bernards Widrow, de Stanford, conçut l'Adaline (Adaptative Linear Element), une machine basée sur le principe des composants électriques et utilisée pour le traitement du signal. Ce modèle sera par la suite le modèle de base des réseaux multicouche.

En 1969, Marvin Minsky et Seymour Papert, du Massachusetts Institute of Technology, publièrent « Perceptrons », une analyse très critique démontrant les limites des machines de type Perceptron et du connexionniste comme modèle de la cognition. Ils montrèrent que les réseaux de neurones comme le Perceptron n'étaient même pas capables de comprendre la fonction logique XOR (le « ou » exclusif). La neuro-informatique vit fondre ses crédits de recherche au profit d'autres domaines comme l'étude des systèmes experts.

Il faut attendre le début des années 80 pour assister à la renaissance de la recherche dans le domaine des RNA. John J. Hopfield, du Californie Institute de la Technologie, démontra en 1982 dans une publication de l'Académie des sciences [Hopfield (1982)] l'utilité des réseaux complètement connectés basés sur la règle de Hebb, qui ont permis de définir les notions d'attracteurs et de mémoire associative. En1984, le Finlandais Tuevo Kohonen proposa un modèle de « cartes autoorganisatrices » reposant sur le principe des cartes cérébrales.

En 1986, D.E.Rumelhart décrit dans son ouvrage "Parallel distributed processing" [Rumelhart., et al. (1986)] l'algorithme d'apprentissage dit de rétropropagation de l'erreur qui fournit un moyen simple d'entraîner les neurones des couches cachées. Vers la fin des années 80, les réseaux de neurones sont devenus un domaine de recherche attractif, qui réunit autour

d'une thématique commune des mathématiciens, des psychologues, des physiciens, des biologistes, des ingénieurs et des informaticiens.

## III.3. Notions fondamentales sur les réseaux de neurones artificiels

### III.3.1. Analogie entre neurone biologique et neurone artificiel

L'idée de réseaux neuraux artificiels a été inspirée de la manière par laquelle les neurones biologiques traitent les informations. Ce concept est utilisé pour mettre en œuvre des logiciels de simulations des processus massivement parallèles qui impliquent des éléments de traitement connectés dans une architecture de réseau. L'apprentissage dans le cerveau humain se fait à travers un réseau des neurones qui sont connectés par l'axone, les synapses et des dendrites (figure III.1).

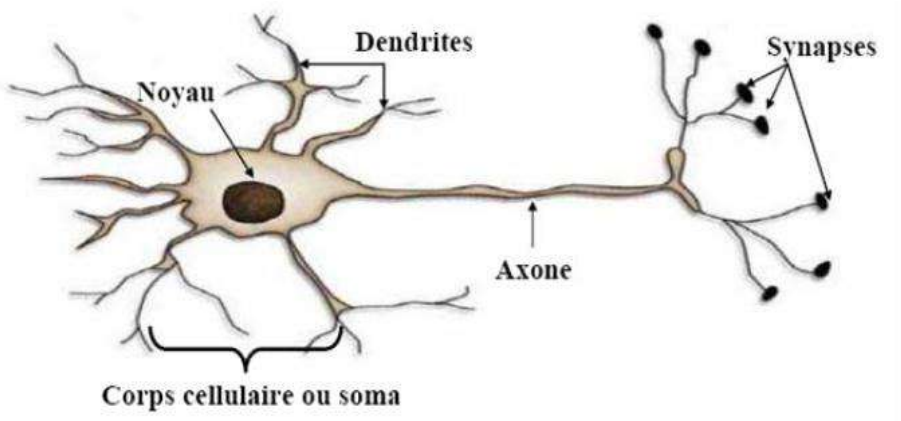



**Figure III.1 :** Schéma d'un neurone biologique [Bouzidi, (2006)].

Le neurone biologique est donc une cellule composée d'un corps cellulaire et d'un noyau. Le corps cellulaire se ramifie pour former ce que l'on nomme les dendrites. C'est par les dendrites que l'information est acheminée de l'extérieur vers le soma, corps du neurone. L'information traitée par le neurone chemine ensuite le long de l'axone (unique) pour être transmise aux autres neurones. La jonction entre deux neurones est appelée la synapse. Une résistance synaptique variable affecte le flux d'informations entre deux neurones biologiques. Les dendrites, recueillent, via les synapses, les signaux électrochimiques ou les potentiels d'action venant d'autres cellules ou de l'extérieur et les achemine vers le corps cellulaire ou le soma. Le soma, outre son rôle concernant le métabolisme de la cellule, collecte et concentre les informations et en fait une sommation dite "spatiotemporelle".

Si le potentiel somatique dépasse un certain seuil, il y a émission d'un potentiel d'action, qui correspond à une oscillation électrique très brève. L'axone est un prolongement cellulaire chargé de faire passer l'information nerveuse à d'autres cellules à travers les synapses grâce à son arborisation terminale. Les synapses se trouvent au point de contact entre neurones, elles convertissent les impulsions nerveuses en signaux chimiques c'est-à-dire en molécules de neurotransmetteur. En effet, l'arrivée d'un potentiel d'action à l'extrémité d'un axone entraîne le passage d'une quantité de neurotransmetteurs dans l'intervalle synaptique. Ils permettent alors un passage pondéré de l'influx nerveux. Le nombre de synapses par neurone varie considérablement de plusieurs centaines à une dizaine de milliers.

Le neurone artificiel (figure III.2) reçoit les entrées qui sont analogues aux impulsions électrochimiques que les dendrites de neurones biologiques reçoivent d'autres neurones. Il est donc un opérateur mathématique très simple qui traduit la représentation mathématique d'un neurone biologique. Selon, Dreyfus (2005) "un neurone est une fonction algébrique non linéaire, paramétrée, à valeur bornée". Il possède des entrées qui peuvent être les sorties d'autres neurones, ou des entrées de signaux extérieurs, et une sortie. La valeur de la sortie résulte du calcul de la somme des entrées, pondérées par des coefficients (dits poids synaptiques) et du calcul d'une fonction non linéaire bornée de cette somme pondérée. Cette fonction (dite fonction d'activation) est le plus souvent une fonction sigmoïde, mais peut également être une gaussienne, un échelon.

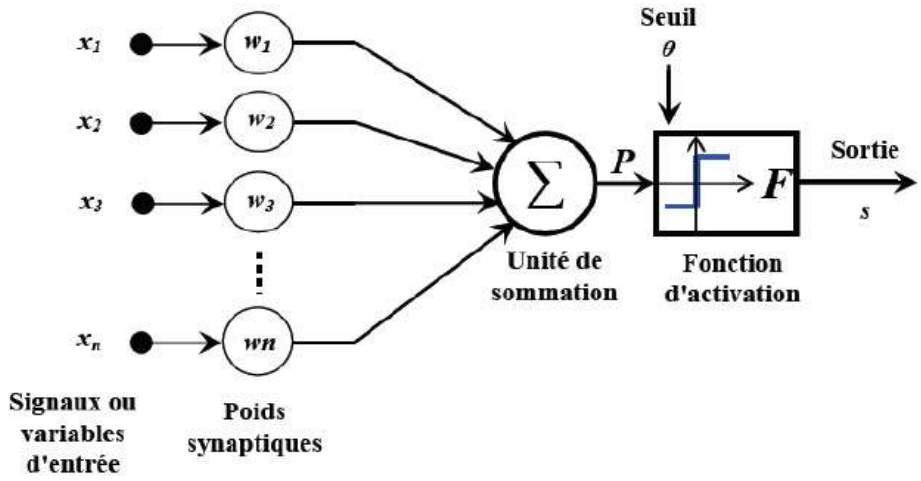



**Figure III.2 :** Schéma d'un neurone artificiel [Bouzidi (2006)].

La figure III.3 résume la correspondance entre le neurone artificiel et le neurone biologique. On remarque dans cette description schématique, que chaque composant ou constituant du neurone artificiel trouve son inspiration biologique.

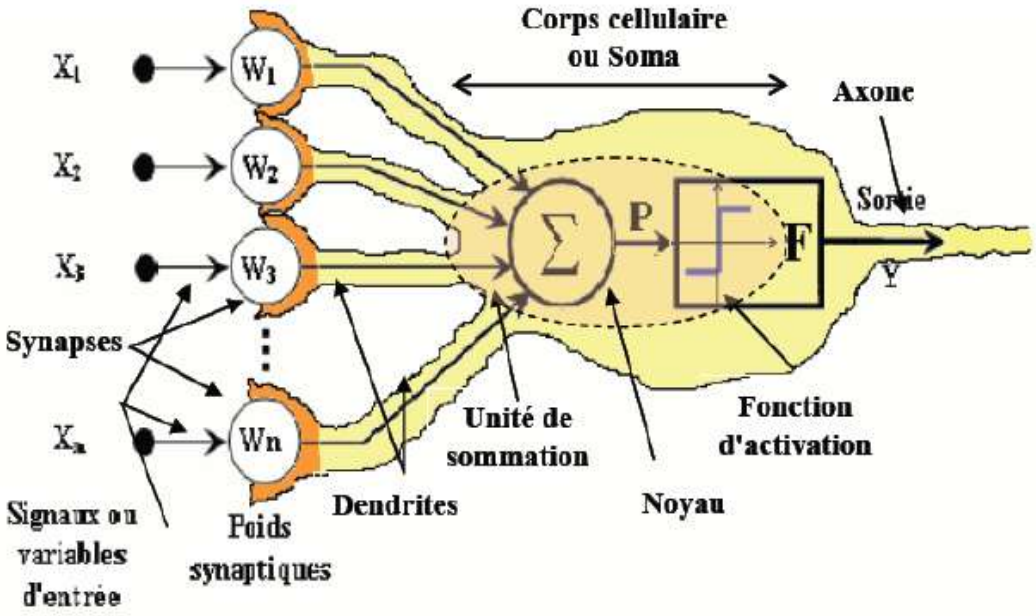



**Figure III.3 :** Schéma de comparaison entre un neurone biologique et un neurone artificiel [Bouzidi (2006)].

### III.3.2. Concept du réseau de neurones de type feedforward

Les neurones artificiels sont arrangés en couches (figure III.4), où la couche d'entrée reçoit des signaux d'entrée du monde réel et chaque couche successive reçoit les sorties pondérées de la couche précédente comme son entrée résultant ainsi en un RNA de type feedforward (cas de perceptrons multicouches), dans lequel chaque sortie d'une couche est alimentée en avant à la couche qui la succède où elle est traitée. Les sorties de la dernière couche constituent les sorties au monde réel.

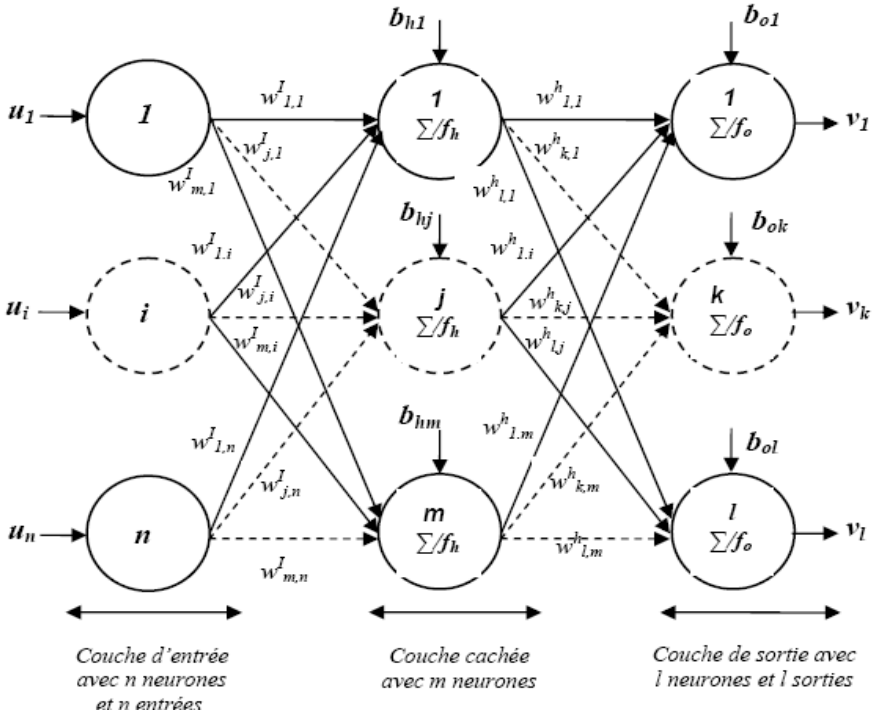



**Figure III.4:** Réseau de neurones artificiels de type feedforward [Si-moussa (2009)].

Dans un tel RNA de type feedforward un neurone, dans une couche cachée ou une couche de sortie, a deux tâches :

Il additionne le biais aux entrées pondérés de plusieurs connexions et applique ensuite une fonction d'activation (dite aussi fonction de transfert ou encore fonction de seuillage) à cette somme comme donnée par l'équation III.1 (pour le neurone *j* de la couche cachée de la figure (III.4) :

$$\boldsymbol{Z_i = f_h\left(\sum_{i=1}^{n} w_{ji}^{I} u_i + b_{hj}\right);\ \ j = 1, 2, \dots, m} \qquad \text{(III.1)}$$

Il propage la valeur résultante par les connexions partantes aux neurones de la couche suivante où elle subit le même processus tel que donné par l'équation III.2 (par exemple les sorties de la couche cachée alimentée au neurone k de la couche de sortie donnent la Sortie $V_k$) :

$$\boldsymbol{v_k = f_0\left(\sum_{j=1}^{m} w_{kj}^{h} z_j + b_{0k}\right);\ \ k = 1, 2, \dots, I} \qquad \text{(III.2)}$$

En combinant des équations III.1 et III.2 on obtient la relation entre la sortie $V_k$ et les entrées du RN:

$$\boldsymbol{v_k = f_0\left(\sum_{j=1}^{m} w_{kj}^{h} f_h\left(\sum_{i=1}^{n} w_{ji}^{I} u_i + b_{hj}\right) + b_{0k}\right);\ k = 1, 2, \dots, l} \qquad \text{(III.3)}$$

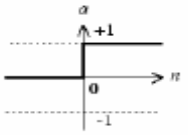
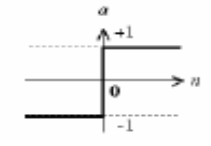
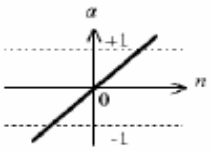
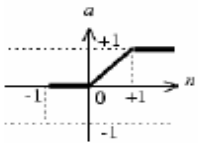
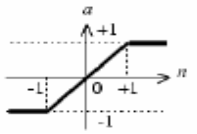
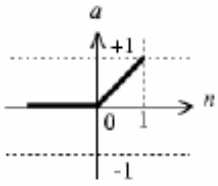
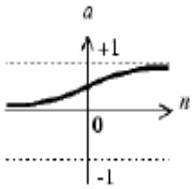
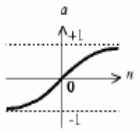
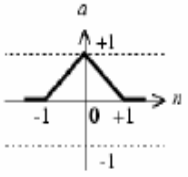
Le nombre de neurones dans la couche d'entrée et la couche de sortie est fixé par le nombre de variables indépendantes et dépendantes respectivement. Le modélisateur définit le nombre de couches cachées et le nombre de neurones dans chaque couche cachée. Le développement du modèle est réalisé par un processus d'apprentissage dans laquelle un ensemble de données expérimentales des variables indépendantes est présenté à la couche d'entrée du réseau.

Les sorties de la couche de sortie comprennent une prédiction des variables dépendantes. Le réseau apprend les relations entre les variables indépendantes et dépendantes par la comparaison itérative des sorties calculées et des sorties expérimentales et l'ajustement postérieur de la matrice de poids et le vecteur des biais de chaque couche par un algorithme d'apprentissage de rétro propagation du gradient de l'erreur. Ainsi le réseau développe un modèle neuronal capable de prédire avec précision acceptable les variables de sortie comprises dans le domaine des données d'apprentissage. Par conséquent, l'objectif de la modélisation neuronale est de réduire au minimum les erreurs de prédiction de l'ensemble des données de validation présentées au réseau à la fin de l'étape d'apprentissage.

### III.3.3 Fonctions d'activation

La fonction d'activation calcule la valeur de l'état du neurone qui est transmise aux neurones avals. La fonction d'activation sert aussi à introduire une non-linéarité dans le fonctionnement du neurone. À la différence des neurones biologiques dont l'état est binaire, la plupart des fonctions de transfert sont continues, offrant une infinité de valeurs possibles comprises dans l'intervalle [0, +1] ou [-1, +1]. Il existe de nombreuses formes possibles pour la fonction de transfert. Le tableau 3.1 résume quelques fonctions de transfert utilisées dans le programme de calcul techniques MATLAB® [Hagan et al (1996)]. En outre, du fait que la continuité et la différenciabilité des fonctions d'activation à tous les points sont les caractéristiques requises par les algorithmes d'optimisation actuels, les fonctions d'activation typiques qui accomplissent ces exigences sont la fonction linéaire, la fonction tangente hyperbolique et la fonction sigmoïde logarithmique. Selon Samarasinghe (2006) la fonction linéaire est plus appropriée pour les neurones de sorties alors que les fonctions sigmoïdes sont plus appropriées pour les neurones des couches cachées.

**Tableau III.1 :** Fonctions de transfert usuelles [Si-moussa (2009)].

| Nom de la fonction | Relation entrée/sortie | Graphe de la fonction | Fonction Matlab correspondante |
|---|---|---|---|
| **Seuil** | **a = 0 si n < 0**<br>**a = 1 si n ≥ 0** |  | **hardlim** |
| **Seuil symétrique** | **a = -1 si n < 0**<br>**a = 1 si n ≥ 0** |  | **hardlims** |
| **Linéaire** | **a = n** |  | **purelin** |
| **Linéaire saturée** | **a = 0 si n < 0**<br>**a = n si 0 ≤ n ≤ 1**<br>**a = 1 si n > 1** |  | **satlin** |
| **Linéaire saturée symétrique** | **a = -1 si n < -1**<br>**a = n si -1≤ n ≤ 1**<br>**a = 1 si n > 1** |  | **satlins** |
| **Linéaire positive** | **a = 0 si n < 0**<br>**a = n si n ≥ 0** |  | **Poslin** |
| **Logistique (sigmoïde logarithmique)** | $a = \frac{1}{1+e^{-n}}$ |  | **logsig** |
| **Tangente hyperbolique** | $a = \frac{e^{+n}-e^{-n}}{e^{+n}+e^{-n}}$ |  | **Tansig** |
| **Triangulaire de base** | **a = 1- $\|n\|$, -1≤ n ≤ +1**<br>**a = 0, autrement** |  | **Tribas** |

### III.3.4. Architecture des réseaux de neurones artificiels

Nous avons introduit plus haut le concept de RNA de type feedforward, il existe cependant d'autres types de RNA. Une description succincte, basée sur le guide d'utilisateur de la boite à outils des RN de MATLAB® [Demuth et Beale (2000)], des principaux types de RNA est donné dans ce qui suit:

#### III.3.4.1. Réseaux de neurones à fonctions radiales de base

Il s'agit d'une "spécialisation" d'un perceptron multicouche. Les neurones de ces réseaux sont actifs pour une zone restreinte de l'espace d'entrée qui est définie par des fonctions radiales de base (RBF, Radial Basis Function Network). Cette technique consiste à utiliser la non linéarité d'une fonction pour ponctuer la réponse des neurones des couches cachées. La figure III.5 montre l'implantation MATLAB® d'un neurone RBF. Ici l'entrée nette de la fonction de transfert de radbas (une gaussienne) est la distance vectorielle entre son vecteur de poids w et le vecteur d'entrée p, multiplié par le biais b. Le signe ||dist|| dans cette figure signifie que la fonction accepte le vecteur d'entrée p et une ligne la matrice des poids et calcule leur produit.

Entrées Neurone de fonction radiale de base Graphe de la fonction

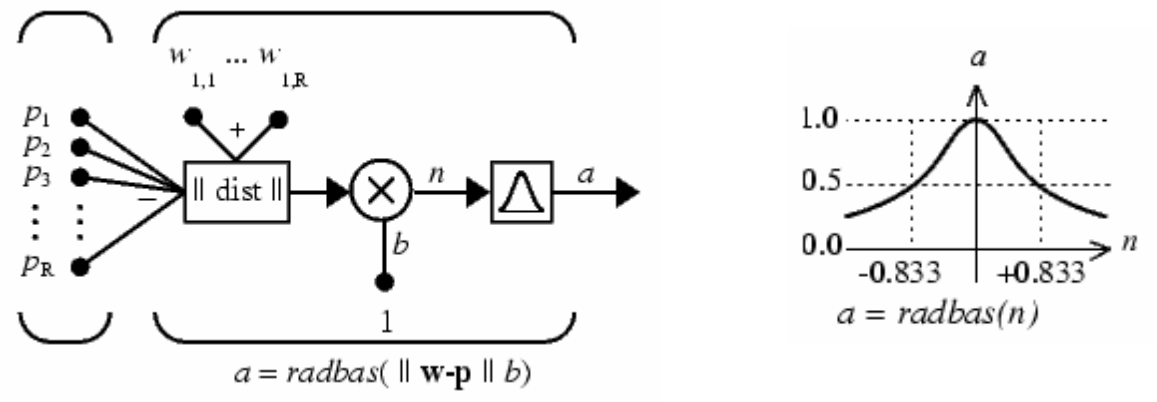



**Figure III.5 :** Schéma d'un neurone de fonction radiale de base [Si-moussa (2009)]

La fonction radiale de base à un maximum de 1 quand son entrée est 0. Lorsque la distance entre w et p diminue la sortie augmente. Ainsi, un neurone RBF agit comme un détecteur qui produit 1 chaque fois que l'entrée p est identique à son vecteur de poids w. Le biais b permet à la sensibilité du neurone d'être ajustée. Par exemple, si un neurone avait un

biais de 0.1 il donnerait une sortie de 0.5 pour n'importe quel vecteur d'entrée p à une distance vectorielle de 8.326 (0.8326/b) de son vecteur de poids w.

On distingue différents types de RN RBF selon le processus que subit la sortie de la seule couche cachée de neurones de fonction radiale de base. Les figures III.6, III.7 et III.8 montrent, respectivement, les structures des implantations MATLAB® des RN de fonction radiale de base, RN probabilistes (Probabilistic Neural Networks) et RN à régression généralisée (Generalized Neural Networks). Il faut noter qu'ici, la fonction de transfert est une exponentielle et l'opérateur sommation disparaît au profit de l'opération multiplication (élément par élément des matrices). Les réseaux RBF nécessitent beaucoup plus de neurones qu'un réseau feedforward.

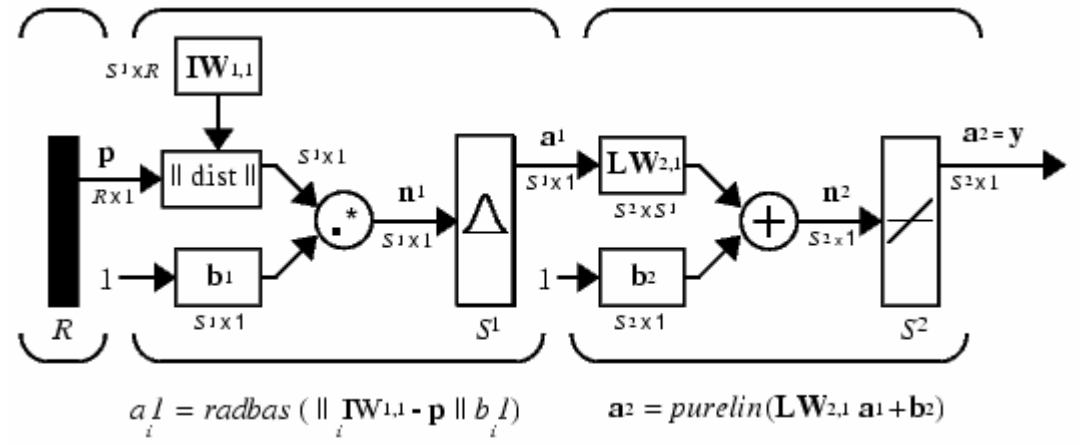



**Figure III.6 :** Schéma de l'implantation MATLAB® du réseau de neurones RBF [Si-moussa, (2009)].

Les réseaux neurones probabilistes sont généralement utilisés pour des problèmes de classification. La première couche qui est un RBF, donne une information sur la ressemblance entre la donnée d'entrée et le jeu de données utilisé lors de l'apprentissage. La deuxième couche produit comme sortie un vecteur de probabilité. Finalement, une fonction de transfert compétitive produit 1 ou 0.

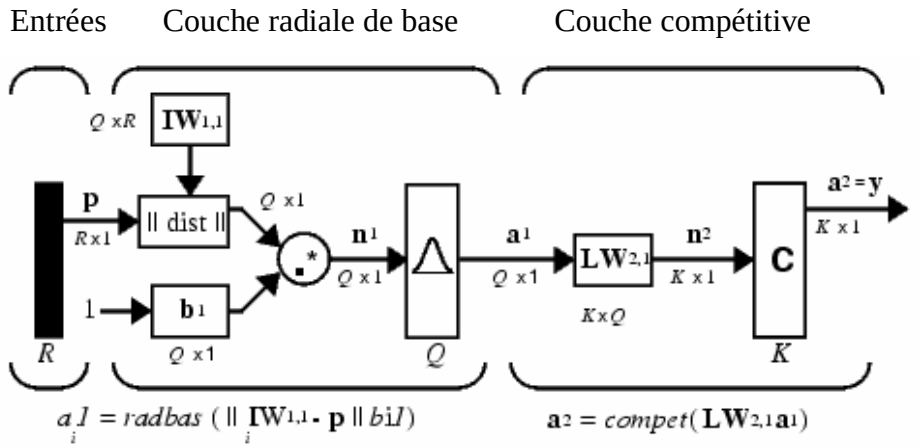



**Figure III.7:** Schéma de l'implantation MATLAB® d'un réseau de neurones probabiliste [Si-moussa, (2009)].

Dans un réseau à régression généralisée, il y a un réseau RBF auquel on ajoute une couche de sortie constituée d'une fonction de transfert linéaire. Le signe nprod signifie une multiplication élément par élément, moralisée par la somme des éléments de a. Ces réseaux sont aussi utilisés en tant qu'approximateur de fonction, mais sont plus lourds d'utilisation que les MLP.

### III.3.4.2. Réseaux compétitifs ou cartes auto organisatrices

Ces réseaux peuvent détecter des régularités et des corrélations dans les données d'entrées et peuvent adapter leurs réponses en fonction de ces entrées. Les neurones de ces réseaux peuvent reconnaître des groupes dans les vecteurs d'entrée. Les cartes auto organisatrices peuvent apprendre à reconnaître des groupes de vecteurs similaires, ce qui entraîne que les neurones physiquement proches les uns des autres dans la couche de neurones à répondre à ces vecteurs d'entrée similaires.

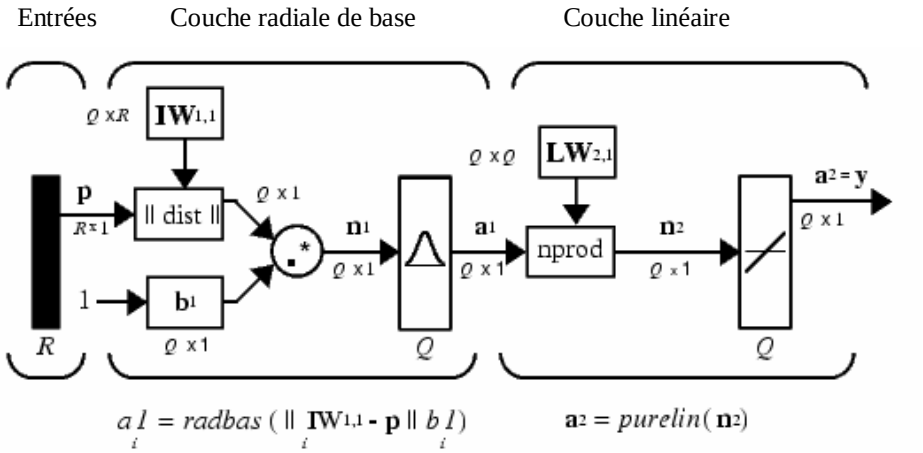



**Figure III.8 :** Schéma de l'implantation MATLAB® d'un réseau de neurones à régression généralisée [Si-moussa, (2009)].

La figure III.9 montre l'architecture d'un réseau de neurones compétitif ou réseau de Kohonen. Le signe ||ndist|| dans cette figure signifie la multiplication du vecteur d'entrées p et la matrice de poids d'entrées IW1,1 et produit un vecteur ayant S1 éléments. Les éléments sont le négatif des distances entre le vecteur d'entrées p et les vecteurs iIW1,1 formés des lignes de la matrice de poids d'entrées. Il Calcule l'entrée nette n1 d'une couche compétitive en trouvant la distance négative entre le vecteur d'entrées p et les vecteurs de poids et ajoutant les biais b. Si tous les biais sont nuls, la valeur maximale de l'entrée d'un neurone est 0. Cela arrive quand le vecteur d'entrées p est égal au vecteur de poids de ce neurone.

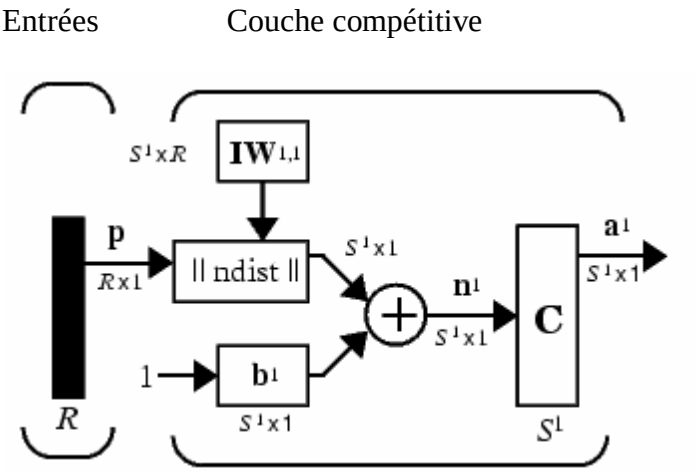



**Figure III.9 :** Architecture d'un réseau de neurones compétitifs ou réseau de Kohonen [Si-moussa, (2009)].

### III.3.4.3. Réseaux récurrents

Un réseau de neurone est dit bouclé ou récurrent lorsque des sorties de neurones d'une couche sont connectées aux entrées d'une couche d'indice inférieur. Autrement dit, lorsqu'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ. La sortie d'un neurone du réseau peut donc être fonction d'elle-même ; cela n'est évidemment concevable que si la notion de temps est explicitement prise en considération. Raison pour laquelle ce type de réseaux sont appelés dynamique. Ainsi, à chaque connexion d'un réseau de neurones bouclé est attaché, outre un poids comme pour les réseaux non bouclé, un retard, multiple entier de l'unité de temps choisie. Une grandeur, à un instant donné, ne pouvant pas être fonction de sa propre valeur au même instant, tout cycle du graphe du réseau doit avoir un retard non nul.

Le réseau d'Elman (figure III.10) à des neurones tansig dans sa couche cachée récurrente et un neurone linéaire dans sa couche de sortie. Ce type de réseau peut aussi approximer n'importe qu'elle type de fonction pourvu que la couche cachée ait assez de neurones. Puisque ce réseau peut enregistrer de l'information pour une référence future, il est capable d'apprendre des associations aussi bien temporelles que spatiales. Ces réseaux sont des réseaux récursifs, un peu plus complexes que les MLP. Chaque cellule est connectée à toutes les autres et les changements de valeurs de cellules s'enchaînent en cascade jusqu'à un état stable. Ces réseaux sont bien adaptés à la reconnaissance de formes.

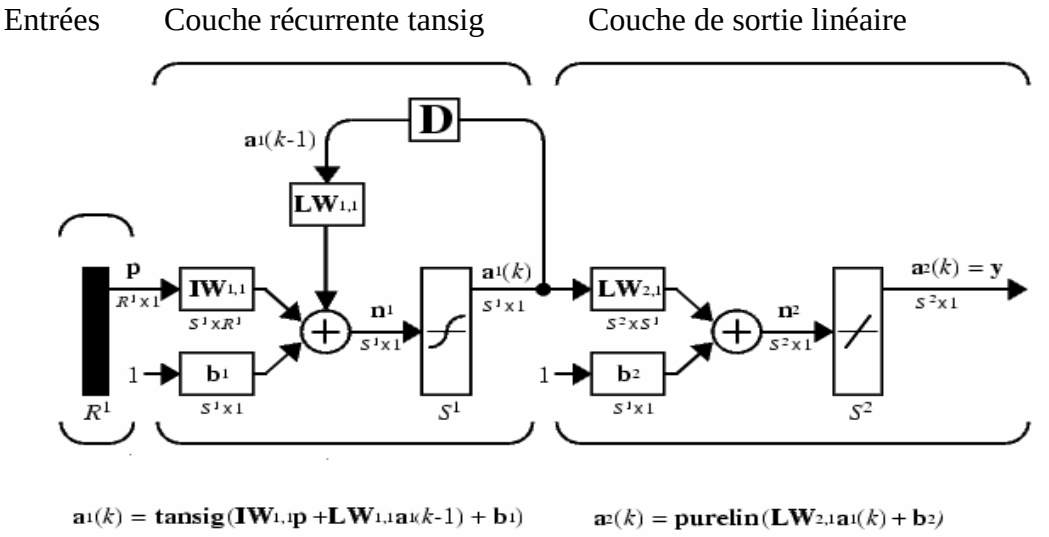



**Figure III.10 :** Schéma de l'implantation MATLAB® du réseau récurrent d'Elman [Si-moussa, (2009)].

### III.3.5. Procédure de modélisation neuronale

Le modèle boite noire constitue la forme la plus primitive de modèle mathématique : il est réalisé uniquement à partir de données expérimentales ou observations ; il peut avoir une valeur prédictive, dans un certain domaine de validité, mais il n'a aucune valeur explicative [Dreyfus et al. (2002)]. A l'opposé, un modèle de connaissance, que l'on pourrait qualifier de boite blanche, est issu de l'analyse des phénomènes physiques, chimiques, biologiques, etc., dont résulte la grandeur que l'on chercher à modéliser. Ces phénomènes sont mis en équation à l'aide des connaissances théoriques disponibles au moment de l'élaboration du modèle.

La démarche scientifique s'efforce, dans la mesure du possible, de construire des modèles de connaissance : l'élaboration d'un modèle de connaissance nécessite l'existence d'une théorie, alors qu'un modèle boite noire requiert seulement la disponibilité de mesures. Entre la boite noire et le modèle de connaissance se situe le modèle semi- empirique, ou modèle boite grise, qui contient à la fois des équations résultant d'une théorie, et des équations purement empiriques, résultant d'une modélisation de type boite noire [Dreyfus et al. (2002)].

Les réseaux de neurones constituent le plus souvent des modèles boite noire, néanmoins dans certaines conditions, ils peuvent être utilisés avec profit dans le cadre de modèles boite grises. Un modèle statique est régi exclusivement par des équations algébriques ; un modèle dynamique, en revanche, obéit à des équations différentielles et à des équations algébriques [Dreyfus et al. (2002)]. Nous examinerons en bref, exclusivement, dans ce qui suit la conception d'un modèle statique type boite noire, réalisé à l'aide de réseaux de neurones bouclés (MLP).

L'objectif de la modélisation boite noire est de trouver, à partir des données ou des mesures disponibles, une relation déterministe, si elle existe, entre les variables d'entrées du modèle x et la grandeur à modéliser y.

La procédure de développement d'un réseau de neurone qui répond à une modélisation statique peut se résumée aux points suivants [Bouzidi (2006)]:

1. Sélection des données
2. Collecte des données
3. Séparation des bases de données,
4. Normalisation des données,
5. Choix du type de réseau de neurones,
6. Choix du nombre de neurones dans la couche cachée
7. Choix de la fonction de transfert
8. Estimation des paramètres ou apprentissage,
9. Validation et généralisation.

### III.3.5.1. Sélection des données

Lorsqu'on conçoit un modèle à partir de mesures, il est indispensable que l'ensemble des variables d'entrée soit aussi réduit que possible ; en effet, à chaque entrée du modèle correspond un nombre de paramètres égal au nombre de neurones cachés. Il est donc nécessaire.

- Que la représentation des entrées soit aussi compacte que possible,
- Que la représentation des entrées contienne tous les facteurs pertinents par rapport à la grandeur à modéliser, mais seulement ceux-ci. En effet, la présence de variables qui ont, sur la grandeur à modéliser, moins d'influence que les perturbations ou le bruit, entraîne celle de paramètres superflus et introduit des variations non significatives sur les variables d'entrées, qui constituent autant de sources d'erreur pour le modèle [Dreyfus et al. (2002)].

Avant la conception d'un modèle neuronale, il convient donc de procéder à une sélection des entrées. Cette sélection revêt deux aspects :

- Réduction de la dimension de l'espace de représentation des variables du modèle,
- Élimination des entrées qui n'ont qu'une faible influence sur la sortie.

### III.3.5.2. Collecte des données

Une fois les entrées pertinentes du modèle sont bien connues, l'objectif de cette étape est de les recueillir, à la fois pour développer (l'apprentissage) le réseau de neurones et pour le tester. La base de donnée reconstituée doit être représentative des données susceptibles d'intervenir en phase d'utilisation du système neuronal.

### III.3.5.3. Séparation des bases de données

Afin de développer un modèle neuronal, il est nécessaire de disposer d'au moins deux bases de données : une base pour effectuer l'apprentissage et une autre pour valider le réseau obtenu et déterminer ses performances. On peut utiliser une troisième base de données, si elle est disponible, pour tester le réseau. Il n'y a pas de règle précise pour déterminer le partage de manière quantitative des données. Il résulte souvent d'un compromis tenant compte du nombre de données dont on dispose, l'essentiel, est que la base de données consacrées à

l'apprentissage doit être la plus importante. Il important de souligner que les sous-ensembles formés à partir de la base de données de départ devront être représentatifs de celle-ci [Omari (2009)].

### III.3.5.4. Normalisation des données

Avant tout apprentissage, il est indispensable de normaliser et de centrer toutes les variables d'entrée. En effet, si des entrées ont des grandeurs très différentes, celles qui sont petites n'ont pas d'influence sur l'apprentissage. Il existe plusieurs approches de normalisation des données dont le choix dépend de la nature des données et de la fonction à approximer. Cependant, les deux méthodes suivantes sont les plus utilisées :

La normalisation qui a pour but d'avoir données ayant une distribution normale de moyenne égale à zéro et un écart type égal à un. Cette méthode a l'équation de normalisation suivante:

$$\boldsymbol{Valeur\ normalisée} = \frac{\boldsymbol{Valeur\ réelle - Valeur\ moyenne}}{\boldsymbol{Ecart\ Type}}$$

La fonction ***prestd*** de MATLAB® réalise la normalisation selon l'équation 3.4, la fonction inverse de dénormalisation est réalisée par la fonction ***poststd***.

La normalisation qui a pour but d'avoir les données normalisées comprises entre -1 et +1 en utilisant l'équation de normalisation suivante:

$$\boldsymbol{Valeur\ normalisée} = \frac{\boldsymbol{2(Valeur\ réelle - Minimum)}}{\boldsymbol{Maximum - Minimum}} - \boldsymbol{1}$$

La fonction premnmx de MATLAB® réalise la normalisation selon l'équation 3.5, la fonction inverse de dénormalisation est réalisée par la fonction posmnmx. Dans ce travail nous avons utilisé ces deux méthodes de normalisation [si-moussa (2009)].

### III.3.5.5. Choix du type et paramètres du réseau de neurones

- **Type du réseau**

Le réseau de type feedforward est le réseau le plus utilisé pour l'approximation des fonctions des données statiques. Il est d'ailleurs le réseau par défaut pour les interfaces graphiques de réseaux de neurones, ***nntool*** et ***nftool***, de MATLAB®. Par conséquent, le réseau de type ***feedforward***, fonction ***newff*** de MATLAB®, a été notre choix.

- **Nombre de couches cachées**

Il n'y a pas de règles générales qui fixent le nombre de couches cachées d'un MLP pour un problème donné. Par conséquent la stratégie d'essai et erreur (trial and error) est la plus raisonnable [Curry et Morgan (2006)]. Le critère d'arrêt est la capacité de généralisation (phase de validation du réseau) et une fonction de performance donnée [Plumb et al. (2005)]. La capacité de généralisation est évaluée en calculant un vecteur du meilleur ajustement linéaire [α, β, R] de la sortie calculée par le RNA en fonction de la sortie cible pour un ensemble de données (égal à un tiers de la base de données et bien distribuées sur l'espace de données) non utilisées dans la phase d'apprentissage. Ce vecteur est constitué de la pente de la droite de régression linéaire, α, de l'ordonné à l'origine, β, et du coefficient de corrélation, R. Idéalement ce vecteur doit être égal à [α, β, R] = [1, 0, 1]. Ce vecteur est aisément obtenu à l'aide de la fonction postreg de MATLAB®. Pour la fonction de performance nous avons utilisé l'erreur relative absolue moyenne (ERAM) ou l'erreur absolue moyenne (EAM), entre la sortie calculée et la sortie cible, à fixer pour chaque sortie.

La meilleure approche neuronale envisageable est celle des MLP à une seule couche cachée, donc en commençant par une seule couche et on fait varier le nombre de neurones en commençant par un nombre réduit (égal au nombre des neurones de la couche d'entrée) et en ajoutant chaque fois un neurone jusqu'à un nombre maximum (égal à cinq fois le nombre des neurones de la couche d'entrée). Cette variation se fait à l'intérieur d'une boucle externe qui fait varier l'algorithme d'apprentissage, elle-même à l'intérieur d'une autre boucle externe qui fait varier la fonction d'activation.

Si aucune topologie à une seule couche, qui vérifie le critère d'arrêt, n'est pas obtenue on considère alors l'utilisation d'une deuxième couche cachée [Si-moussa (2009)].

### III.3.5.6. Choix de la fonction de transfert

La fonction de transfert neuronale, que l'on appelle aussi fonction d'activation, est la fonction mathématique appliquée à la sortie de chaque neurone. Là encore, il n'existe pas de méthode permettant de prévoir a priori quelle sera la fonction qui donnera les meilleurs résultats. Mais pour les MLP, il faut des fonctions de type sigmoïde pour les neurones de la couche cachée et généralement une fonction linéaire pour le neurone de la couche de sortie [Omari (2009)].

### III.3.5.7. Nombre d'itération

Le nombre d'itération est lié à un autre facteur, très souvent négligé par certains, qui est l'initialisation des poids synaptiques et les biais des neurones cachés et neurones de sortie. L'utilisation de différentes initialisation aide à éviter le problème se convergence vers le minimum local. Dans ce travail nous avons opté pour l'initialisation aléatoire par défaut de MATLAB® et l'exécution de l'apprentissage de la même topologie d'un réseau avec un nombre suffisant (20 à 30) de différentes initialisations.

### III.3.5.8. Estimation des paramètres ou apprentissage

Pour les modélisations statiques au moyen des MLP, il s'agit de faire un apprentissage supervisé. Le critère d'arrêt de l'apprentissage est souvent calculé à partir d'une fonction de coût, caractérisant l'écart entre les valeurs de sortie obtenues et les valeurs désirées. Ce qui permet un arrêt adéquat de l'apprentissage pour obtenir de bonnes performances de généralisation. L'obtention des paramètres architecturaux optimaux se fera par comparaison des performances obtenues pour différentes architectures de réseaux de neurones [Elie (1997)].

### III.3.5.9. Validation et généralisation

Après l'obtention de la meilleure structure du RN qui vérifie la fonction de performance choisie pour chaque sortie, on présente au réseau un ensemble de données de validation pour tester sa capacité de généralisation.

Donc la validation consiste à tester le réseau après ou au cours de l'apprentissage avec des données proches des données utilisées lors de la phase d'apprentissage. Les résultats de ce test permettront au modèle d'éviter les situations de sous ou de sur apprentissage. En revanche la généralisation s'effectue avec des données tout à fait nouvelles et différentes de la base de données de départ. [Omari (2009)]. L'évaluation est réalisée par le calcul de vecteur du meilleur ajustement linéaire cité plus haut est une évaluation visuelle de la distribution des sortie cible par rapport à la droite de l'ajustement linéaire idéal.

## III.4. Conclusion

Dans 65 ans, les réseaux de neurones ont bien pu trouver leur place au sein de l'industrie et de recherche. Mais la puissance de ces réseaux, calqués à partir du modèle biologique, inspire de grandes applications, non encore découverts.

Tout au long de ce chapitre, nous avons donné un maximum d'informations sur les réseaux de neurones d'une façon générale tout en mettant l'accent sur les réseaux de neurones de type feedforward, vraisemblablement les plus utilisés dans les modélisations statiques du fait qu'ils sont doués de la propriété d'approximation universelle.

Nous avons présenté un algorithme d'apprentissage très utilisé dans la communauté connexionniste: l'algorithme de rétro propagation du gradient. Plusieurs travaux lui ont été consacrés et un certain nombre de modifications ont été proposées pour accélérer le temps d'apprentissage, réduire les besoins en mémoire de stockage et améliorer la capacité de généralisation ainsi que les performances du réseau obtenu. Un résumé des caractéristiques des algorithmes d'apprentissage de rétro propagation implantés dans MATLAB® a été présenté.

Vers la fin, nous avons également présenté la stratégie implantée dans les programmes élaborés pour la modélisation neuronale qui sera utilisée dans la prédiction de l'activité thérapeutique à partir de la structure moléculaire.

# *Chapitre IV*

# *Modélisation de la prédiction de l'activité thérapeutique*

# CHAPITRE IV
# MODÉLISATION DE LA PRÉDICTION DE L'ACTIVITÉ THÉRAPEUTIQUE

## IV.1. Introduction

Les outils classiques de l'analyse de données en sciences humaines sont connus. Ils appartiennent à une même famille : analyse de variance, analyse typologique, analyse en composantes principales, etc. L'analyse révèle des causalités, des typologies, des caractéristiques qui facilitent la spécification de modèles explicatifs. L'approche classique consiste à formuler explicitement un modèle et à l'estimer. L'explication d'une réalité complexe peut toutefois buter sur les difficultés d'estimation des modèles complexes ou sur les limites (et les biais) des modèles plus facilement estimables mais exagérément simplifiés par rapport à la réalité.

Une bonne prédiction sur les données résulte généralement d'un bon modèle explicatif mais on sait qu'il est possible de prévoir avec succès sans pour autant expliquer. Le choix de la forme fonctionnelle se fait alors sans référence à un cadre conceptuel à priori cohérent avec le phénomène analysé.

C'est dans ce contexte que le présent travail a été mené. Il a pour but d'illustrer d'une part les capacités des réseaux de neurones et de montrer d'autre part qu'un modèle neuronal précédé d'une optimisation statistique de ses entrées permet d'améliorer sensiblement les capacités prédictives de ce modèle.

On expose en première partie, les différentes étapes permettant d'obtenir une base de données qui décrit numériquement les structures moléculaires des médicaments appartenant à différentes classes thérapeutiques par leurs descripteurs moléculaires. La deuxième partie est consacrée à la modélisation neuronale des données optimisées statistiquement, la troisième partie est une application de l'une des méthodes traditionnelles de prédiction les plus utilisée en QSAR qui est l'analyse des données ACP suivie d'une régression linéaire multiple (RLM). Une conclusion permet de dégager les principales capacités de telles méthodes pour la prédiction de l'activité thérapeutique.

# Partie 1

# Élaboration de la base de données

## IV.1.1. Collection des données

Dans cette étape nous avons sélectionné 800 molécules, ces molécules appartiennent à 6 principales classes de médicaments : anti-inflammatoires, antibiotiques, analgésiques, antiviraux, diurétiques et glucocorticoïdes, ainsi d'autres activités ont été utilisées dans le but d'obtenir une généralisation du modèle neuronale.

La figure IV.1 présente un organigramme qui résume les différentes étapes de réalisation de la base de données :

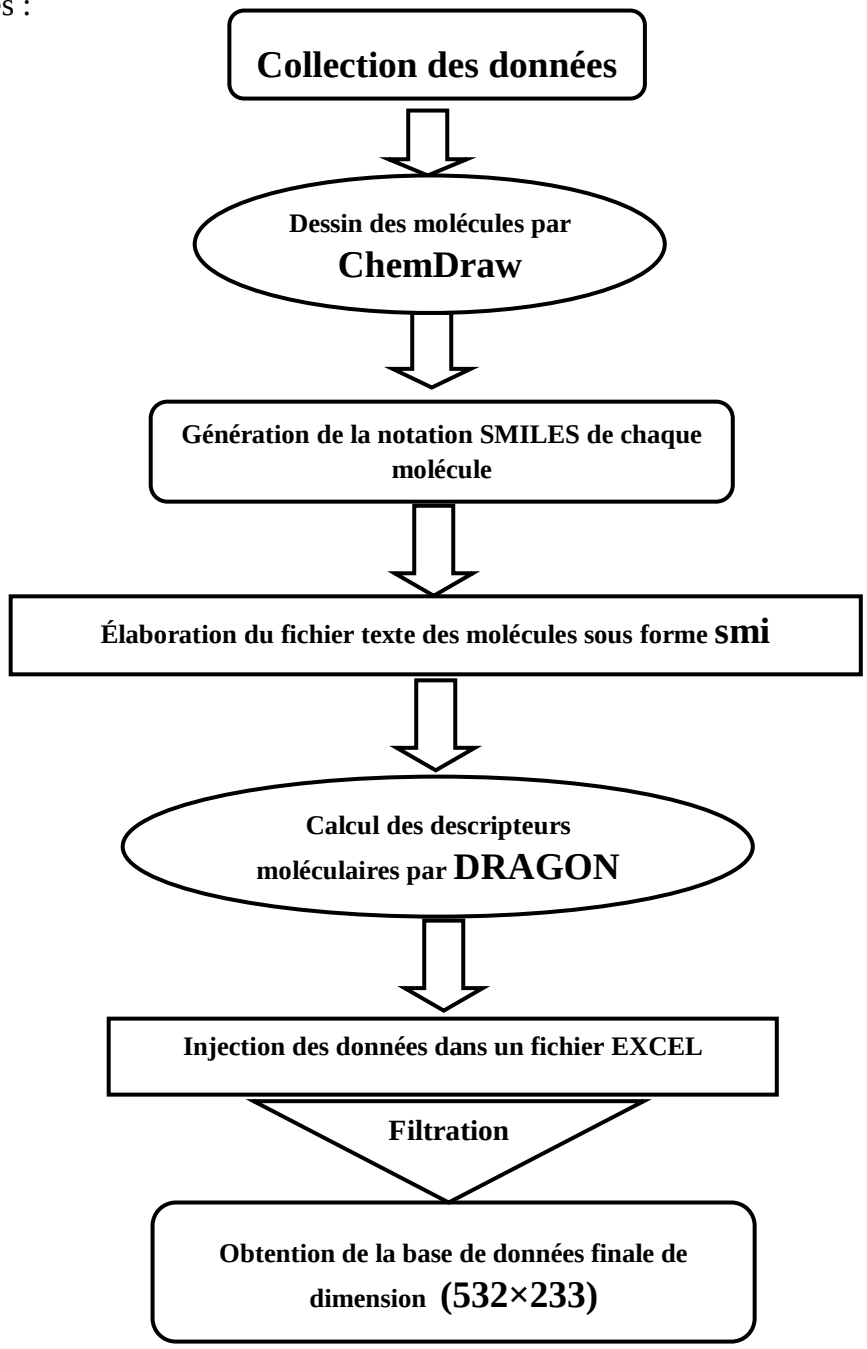



**Figure IV.1.1 :** Organigramme des déférentes étapes de réalisation de la base de données

## IV.1.2. Dessin des molécules sur ChemDraw

Pour réaliser cette tâche, on a utilisé le Chembio office comme un outil de dessin (figure IV.1.2), ce dernier nous permet d'effectuer une première filtration où on a éliminé les molécules qui possèdent une structure moléculaire compliquée ou bien qui possèdent un poids moléculaire important.

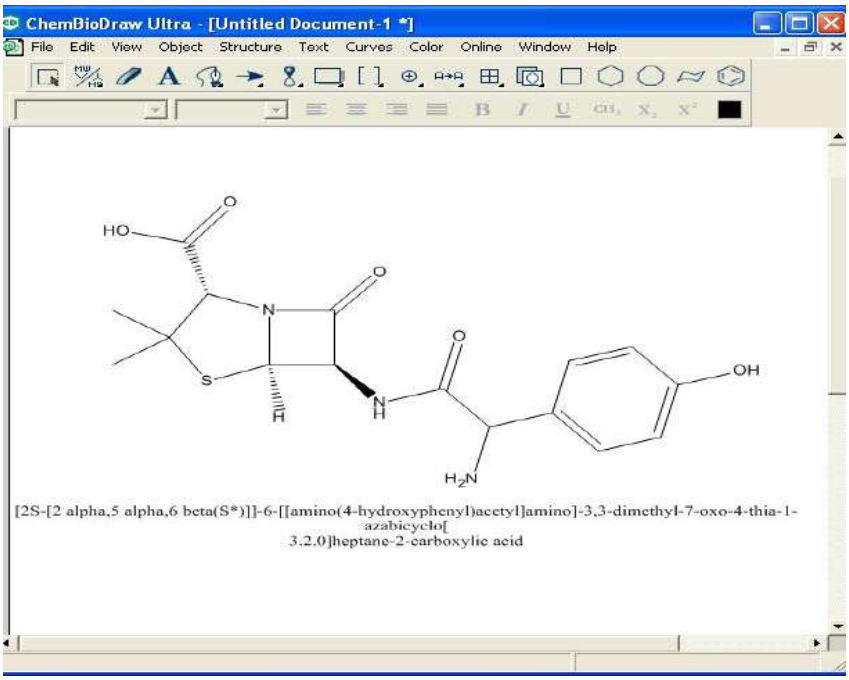



**Figure IV.1.2 :** Dessin des molécules sur ChemDraw.

## IV.1.3. Génération de la notation SMILES par ChemDraw

En utilisant le même logiciel (Chembio office) on peut générer les notations SMILES correspondantes à chaque molécule (figure IV.1.3), cette étape nous permet aussi de réaliser une deuxième filtration des données qui a pour objectif d'éliminer les molécules cationiques et anioniques et également celles qui sont composées.

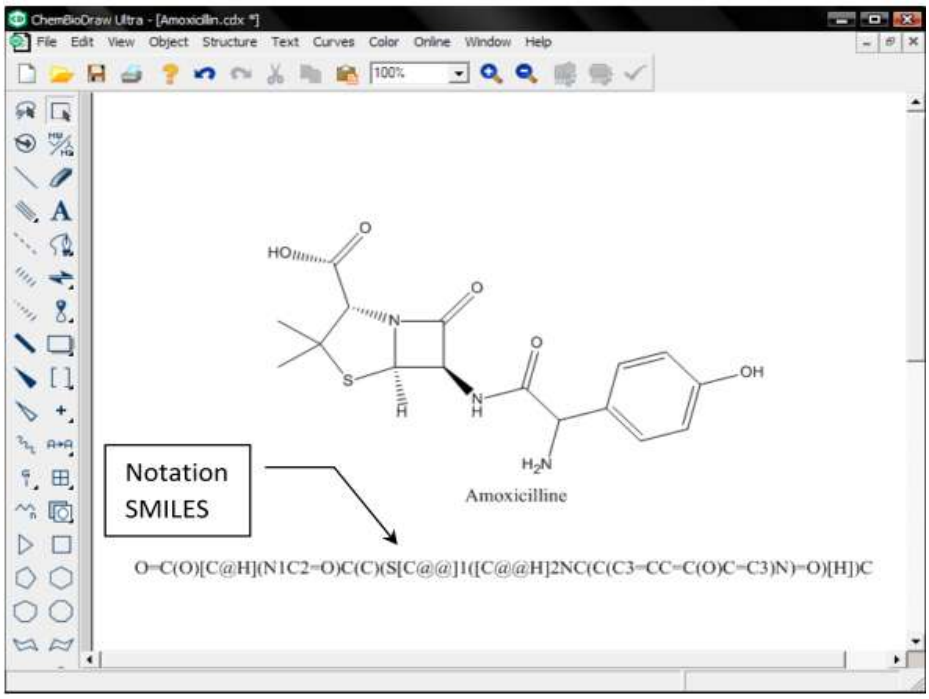



**Figure IV.1.3 :** Génération de la notation SMILES par ChemDraw.

## IV.1.4. Élaboration du fichier texte de la molécule sous forme smi

Les notations générées précédemment par Chembio office sont collectées par la suite dans un fichier texte sous forme smi, et qui seront utilisées par la suite dans le calcul de descripteurs moléculaires, la figure IV.1.4 illustre une partie des données représentées sous forme smi.

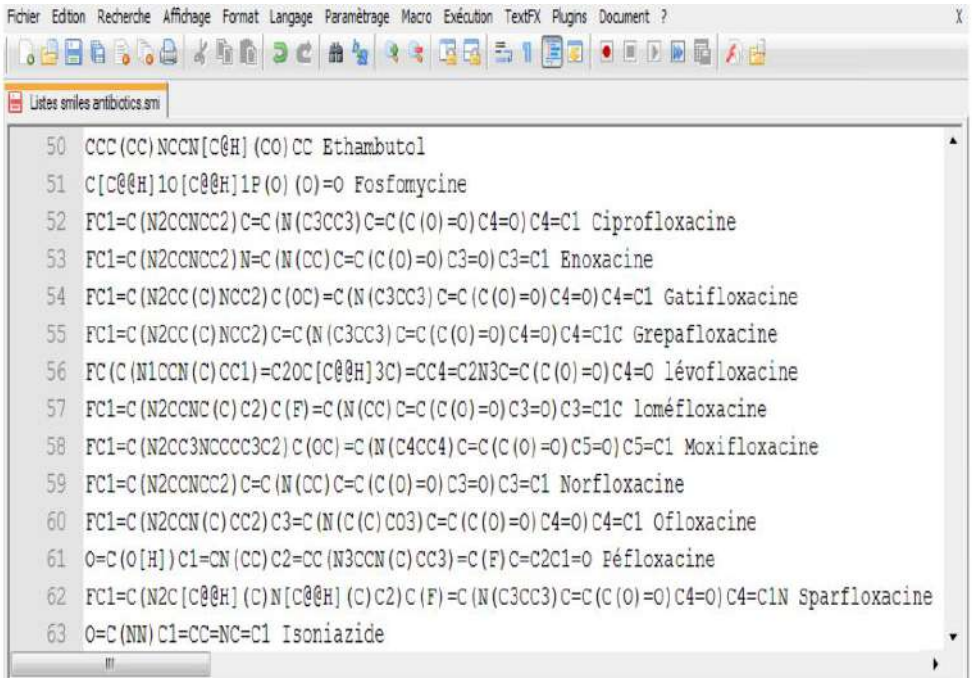



**Figure IV.1.4 :** Elaboration du fichier texte de la molécule sous forme SMILES.

## IV.1.5 . Différentes étapes pour le calcul des descripteurs par DRAGON

Les blocs des descripteurs moléculaires choisis et la méthodologie de calcul des différents descripteurs (constitutionnels, topologiques, géométriques, fragments atome-centrés et comptes fonctionnels de groupe) par le logiciel DRAGON sont résumés dans les figures IV.1.5, IV.1.6, IV.1.7 et IV.1.8

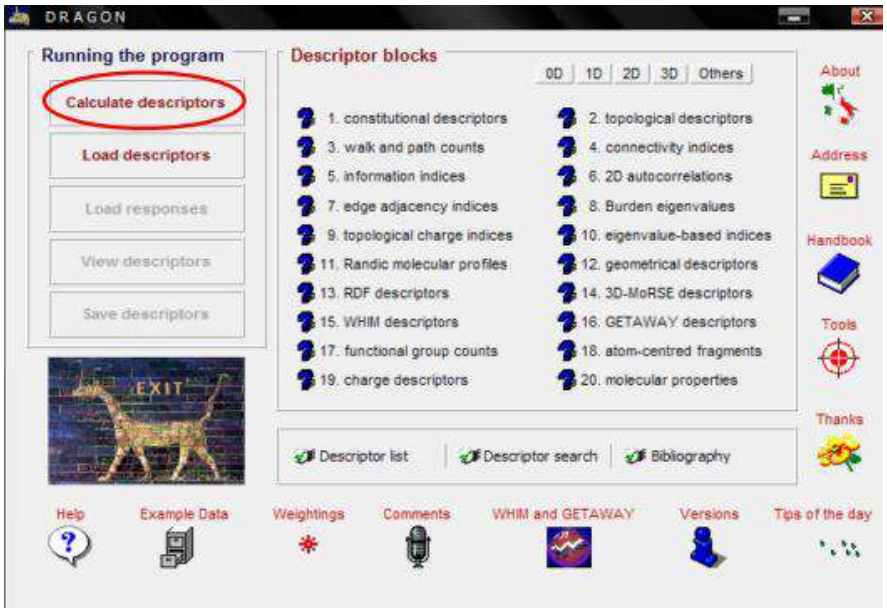



**Figure IV.1.5 :** Première étapes pour le calcul des descripteurs par DRAGON

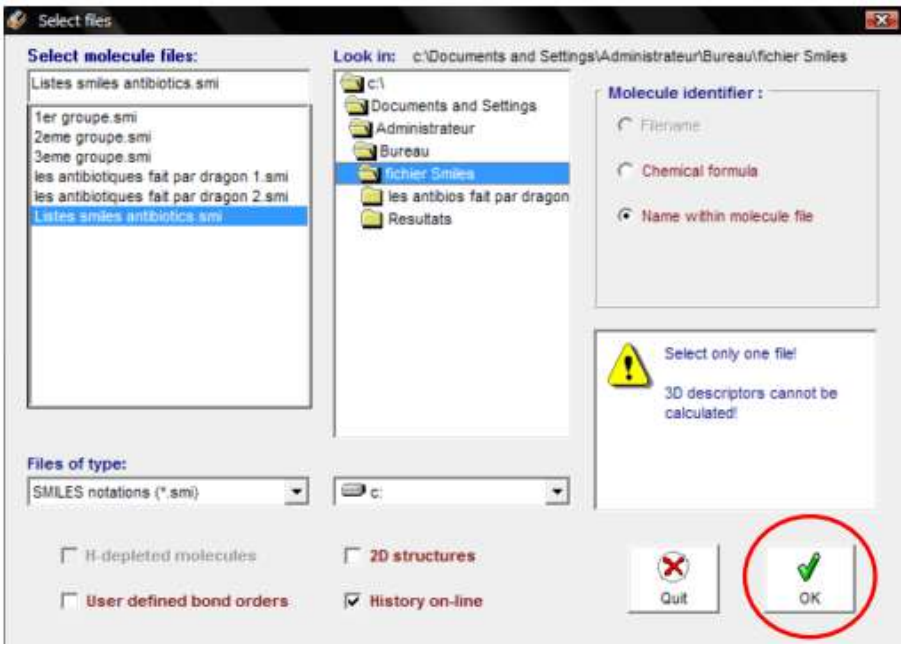



**Figure IV.1.6:** Deuxième étapes pour le calcul des descripteurs par DRAGON.

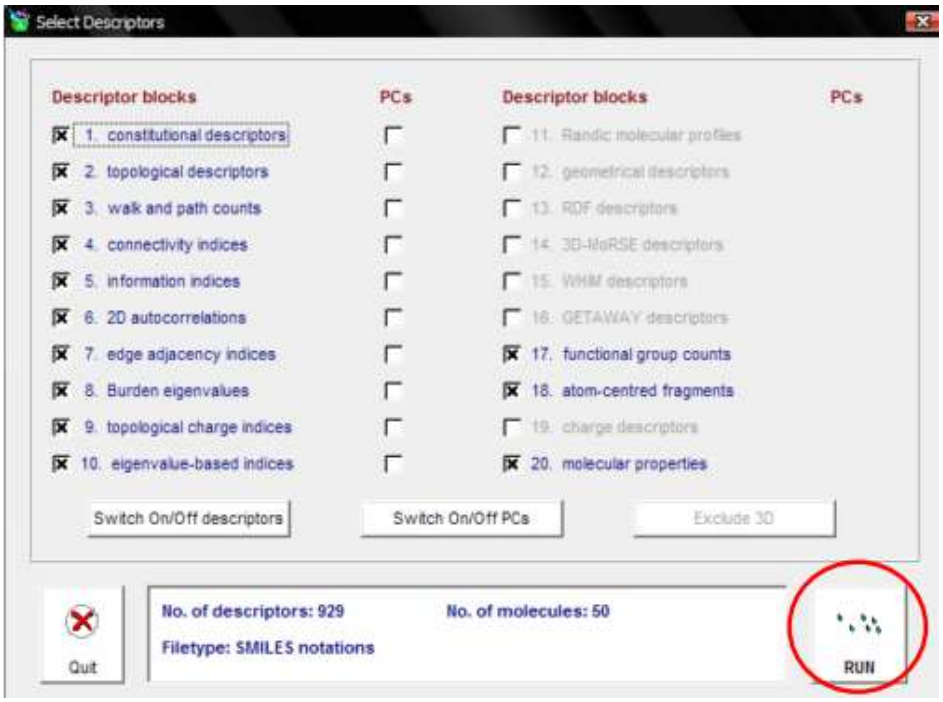



**Figure IV.1.7 :** Troisième étape pour le calcul des descripteurs par DRAGON.

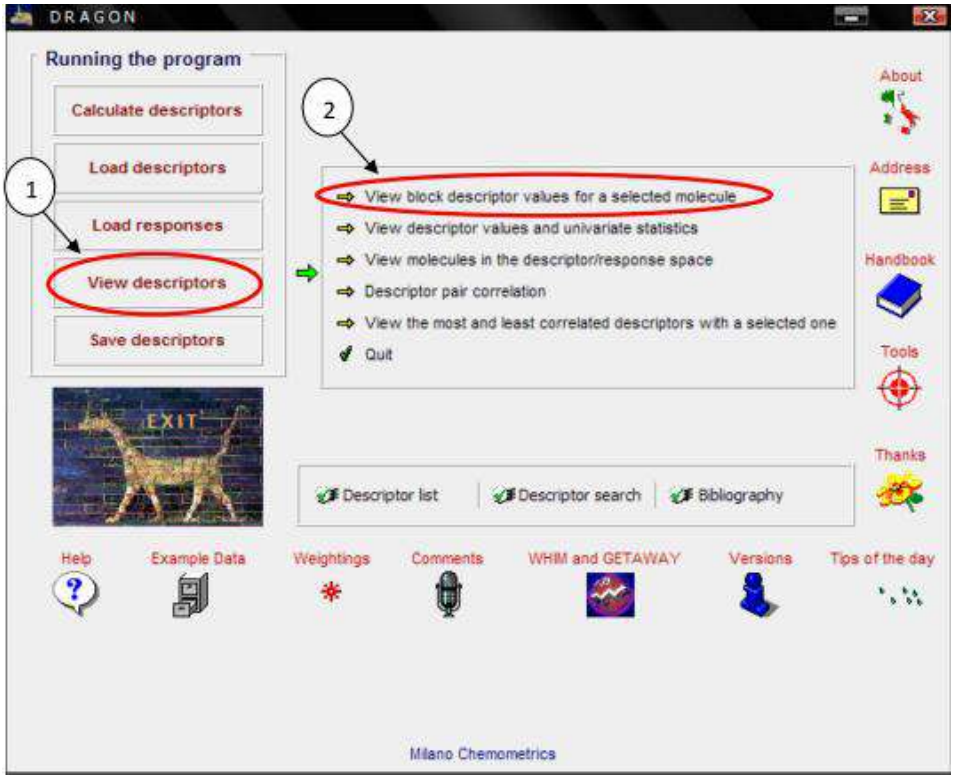
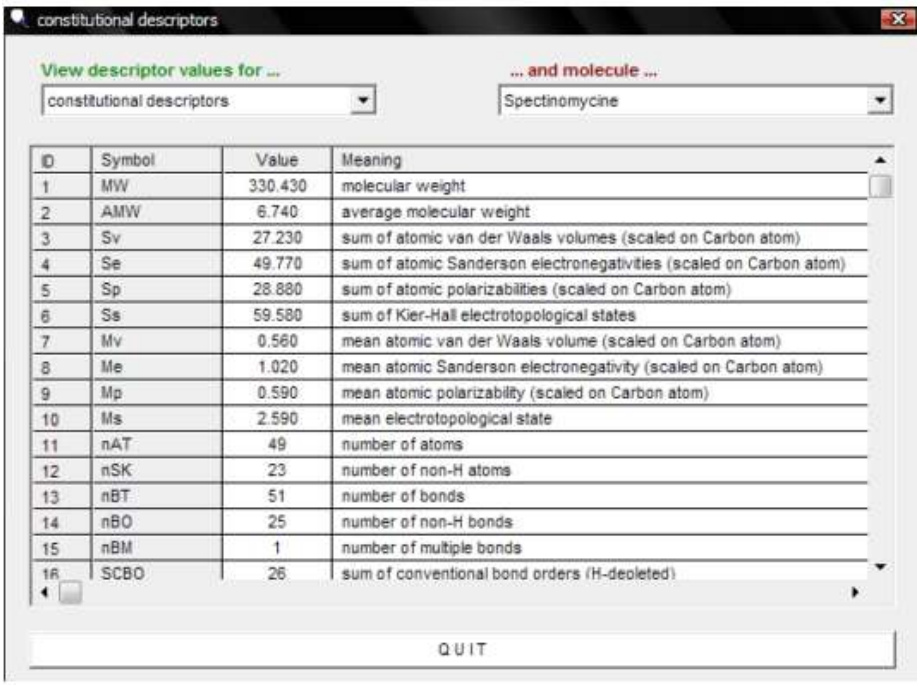



**Figure IV.1.8 :** Quatrième étape pour le calcul des descripteurs par DRAGON.




| ID | Symbol | Value | Meaning |
|---|---|---|---|
| 1 | MW | 330.430 | molecular weight |
| 2 | AMW | 6.740 | average molecular weight |
| 3 | Sv | 27.230 | sum of atomic van der Waals volumes (scaled on Carbon atom) |
| 4 | Se | 49.770 | sum of atomic Sanderson electronegativities (scaled on Carbon atom) |
| 5 | Sp | 28.880 | sum of atomic polarizabilities (scaled on Carbon atom) |
| 6 | Ss | 59.580 | sum of Kier-Hall electrotopological states |
| 7 | Mv | 0.560 | mean atomic van der Waals volume (scaled on Carbon atom) |
| 8 | Me | 1.020 | mean atomic Sanderson electronegativity (scaled on Carbon atom) |
| 9 | Mp | 0.590 | mean atomic polarizability (scaled on Carbon atom) |
| 10 | Ms | 2.590 | mean electrotopological state |
| 11 | nAT | 49 | number of atoms |
| 12 | nSK | 23 | number of non-H atoms |
| 13 | nBT | 51 | number of bonds |
| 14 | nBO | 25 | number of non-H bonds |
| 15 | nBM | 1 | number of multiple bonds |
| 16 | SCBO | 26 | sum of conventional bond orders (H-deoleted) |

**Figure IV.1.9 :** Dernière étape pour le calcul des descripteurs par DRAGON.

Le résultat de calcul des descripteurs moléculaires par DRAGON est représenté dans le tableau IV.1.1 :

**Tableau IV.1.1:** Résultats de calcul des descripteurs moléculaires.

| Les blocs des descripteurs | Le nombre des descripteurs dans le bloc | Le nombre des descripteurs après la réduction |
|---|---|---|
| Descripteurs constitutionnels | 50 | 27 |
| Descripteurs Topologiques | 115 | 105 |
| Descripteurs géométriques | 110 | 38 |
| Fragments atome-centrés | 67 | 17 |
| Comptes de promenade et de chemin | 56 | 14 |
| Propriétés moléculaires | 29 | 12 |

- ✓ **Signification des descripteurs moléculaires utilisés**

Le tableau Tableau IV.1.2 donne la signification des différents descripteurs utilsés après la réduction du nombre.

**Tableau IV.1.2 :** Signification des principaux descripteurs moléculaires utilisés.

| | Indice | Description de l'indice |
|---|---|---|
| **Descripteurs constitutionnels** | AW | poids moléculaire |
| | AMW | poids moléculaire moyen |
| | Sv | somme de volumes atomiques de Van Der Walls (mesurés sur l'atome de carbone) |
| | Se | somme atomique d'électronégativités de Sanderson (mesurés sur l'atome de carbone) |
| | Sp | somme de polarisabilités atomiques (mesurées sur l'atome de carbone) |
| | Ss | somme d'états électro-topologique de Kier-Hall |
| | Me | moyen d'électronégativité atomique de Sanderson (mesuré sur l'atome de carbone) |
| | Mp | moyen de polarisabilités atomique (mesuré sur l'atome de carbone) |
| | Ms | état électro-topologique moyen |
| | nAT | nombre d'atomes |
| | nSK | nombre d'atomes de non-H |
| | nBT | nombre de liens |
| | nBO | nombre de liens de non-H |
| | nBM | nombre de liens multiples |
| | SCBO | somme d'ordres en esclavage conventionnels (H-épuisés) |
| | nCIC | nombre d'anneaux |
| | nCIR | nombre de circuits |
| | RBN | nombre de liens rotatifs |
| | nDB | nombre de liens doubles |
| | nAB | nombre de liens aromatiques |
| | nH | nombre d'atomes d'hydrogène |
| | nC | nombre d'atomes de carbone |
| | nN | nombre d'atomes d'azote |
| | nO | nombre d'atomes d'oxygène |
| | nCL | nombre d'atomes de chlore |
| | nX | nombre d'atomes d'halogène |
| | nF | nombre d'atomes de fluor |
| | nBR | nombre d'atomes de brome |
| | nI | nombre d'atomes d'iode |
| | nB | nombre d'atomes de bore |
| | nHM | nombre d'atomes lourds |
| | nBnz | nombre de benzène-comme anneaux |
| | nR06 | nombre de 6 anneaux membrés |
| **Descripteurs Topologiques** | ZM1 | premier indice M1 de Zagreb |
| | ZM1V | premier indice de Zagreb par des degrés de sommet de valence |
| | ZM2 | deuxième indice M2 de Zagreb |
| | ZM2V | deuxième indice de Zagreb par des degrés de sommet de valence |
| | S1K | Indice de 1-chemin de forme alpha-modifié par Kier |
| | S2K | Indice de 2-chemins de forme alpha-modifié par Kier |
| | S3K | Indice de 3-chemins de forme alpha-modifié par Kier |
| | PHI | Indice de flexibilité de Kier |
| | BLI | Indice de benzène-likeliness de Kie |
| | PW2 | path/walk 2 – indice de forme de Randi |
| | PW3 | path/walk 3 – indice de forme de Randi |
| | PW4 | path/walk 4 – indice de forme de Randi |
| | PW5 | path/walk 5 – indice de forme de Randi |
| | PJI2 | Indice de forme 2D de Petitjean |
| | UNIP | unipolaire |
| | Lop | Indice central d'élagage |
| | Dz | Indice de Pogliani |
| | Ram | Indice de ramification |

| | | |
|---|---|---|
| | Pol | nombre de polarité |
| | SMTI | Indice Topologique Moléculaire De Schultz (MTI) |
| | SMTIV | Schultz MTI par des degrés de sommet de valence |
| | GMTIV | Gutman MTI par des degrés de sommet de valence |
| | W | Indice de la saucisse W |
| | HyDp | indice de hyper-distance-chemin |
| | W | Indice de détour |
| | Ww | Indice de hyper-détour |
| | Wap | Indice de saucisse de tout-chemin |
| | CSI | Indice excentrique de connectivité |
| | ECC | excentricité |
| | UNIP | uni polarité |
| | CENT | centralisation |
| | VAR | variation |
| | BAC | Indice central de Balaban |
| | T(N..N) | somme de distances topologiques entre N..N |
| | T(N..O) | somme de distances topologiques entre N..O |
| | (N..Cl) | somme de distances topologiques entre N..Cl |
| | T(O..O) | somme de distances topologiques entre O..O |
| | T(O..Cl) | somme de distances topologiques entre O..Cl |
| **Descripteurs géométriques** | MEcc | excentricité moléculaire |
| | L/Bw | rapport de longueur-à-largeur par WHIM |
| | G(N..N) | somme de distances géométriques entre N..N |
| | G(N..O) | somme de distances géométriques entre N..O |
| | G(N..S) | somme de distances géométriques entre N..S |
| | G(N..F) | somme de distances géométriques entre N..F |
| | G(N..Cl) | somme de distances géométriques entre N..Cl |
| | G(N..I) | somme de distances géométriques entre N..I |
| | G(O..O) | somme de distances géométriques entre O..O |
| | G(O..S) | somme de distances géométriques entre O..S |
| | G(O..F) | somme de distances géométriques entre O..F |
| | G(O..Cl) | somme de distances géométriques entre O..Cl |
| | G(O..I) | somme de distances géométriques entre O..I |
| | G(S..S) | somme de distances géométriques entre S..S |
| | G(S..F) | somme de distances géométriques entre S..F |
| | G(S..Cl) | somme de distances géométriques entre S..Cl |
| | G(S..I) | somme de distances géométriques entre S..I |
| | G(F..F) | somme de distances géométriques entre F..F |
| | G(F..Cl) | somme de distances géométriques entre F..Cl |
| | G(Cl..Cl) | somme de distances géométriques entre Cl..Cl |
| | G(Cl..I) | somme de distances géométriques entre Cl..I |
| **Fragments atome-centrés** | C-001 | CH3R/CH4 |
| | C-002 | CH2R2 |
| | C-024 | R--CH—R |
| | C-027 | R--CH—X |
| | C-038 | Al-C(=X)-Al |
| | H-046 | H attaché C0 (sp3) à aucun X a attaché à prochain C |
| | H-047 | H attaché C0 (sp3) à 1X a attaché à prochain C |
| | H-048 | H attaché à C2 (sp3)/C1(sp2)/C0(sp) |
| | H-049 | H attaché à C3 (sp3)/C2(sp2)/C3(sp2)/C3(sp) |
| | H-050 | H attaché au heteroatom |

| | | |
|---|---|---|
| | H-051 | H attaché à l'alpha-C |
| | H-052 | H attaché à C0(sp3) avec le 1X a attaché à prochain C |
| | H-053 | H attaché à C0(sp3) avec 2X a attaché à prochain C |
| | O-057 | phénol/enol/OH carboxylique |
| | Cl-089 | Cl attaché à C1(sp2) |
| **Comptes de promenade et de chemin** | SRW01 | compte de renvoi de promenade d'ordre 01 (nombre d'atomes, de nSK de non-H) |
| | SRW04 | compte de renvoi de promenade d'ordre 04 |
| | SRW06 | compte de renvoi de promenade d'ordre 06 |
| | SRW08 | compte de renvoi de promenade d'ordre 08 |
| | SRW09 | compte de renvoi de promenade d'ordre 09 |
| | SRW10 | compte de renvoi de promenade d'ordre 10 |
| | MPC02 | compte moléculaire de chemin d'ordre 02 (index de Gordon-Scantlebury) |
| | MPC03 | compte moléculaire de chemin d'ordre 03 |
| | MPC04 | compte moléculaire de chemin d'ordre 04 |
| | MPC05 | compte moléculaire de chemin d'ordre 05 |
| | MPC06 | compte moléculaire de chemin d'ordre 06 |
| | MPC07 | compte moléculaire de chemin d'ordre 07 |
| | MPC08 | compte moléculaire de chemin d'ordre 08 |
| | MPC10 | compte moléculaire de chemin d'ordre 10 |
| **Propriétés moléculaires** | Inflammat-80 | Ghose-Viswanadhan-Wendoloski anti-inflammatoire-comme l'indice à 80% |
| | Depressant-80 | Ghose-Viswanadhan-Wendoloski antidépresseur-comme l'indice à 80% |
| | Psychotic-80 | Ghose-Viswanadhan-Wendoloski antipsychotique-comme l'indice à 80% |
| | Hypnotic-80 | Ghose-Viswanadhan-Wendoloski hypnotique-comme l'indice à 80% |
| | Neoplastic-80 | Ghose-Viswanadhan-Wendoloski antinéoplastique-comme l'indice à 80 |
| | Infective-80 | Ghose-Viswanadhan-Wendoloski anti-infectant-comme l'index à 80 |
| | BLTF96 | Modèle de Verhaar de la toxicité de ligne de base de daphnie de MLOGP (mmol/l) |
| | BLTD48 | Modèle de Verhaar de la toxicité de ligne de base de daphnie de MLOGP (mmol/l |
| | BLTA96 | Verhaar model of Algae base-line toxicity from MLOGP (mmol/l) |

## IV.1.7 Codage des sorties

On exprime les différentes sorties par des valeurs logiques (-1 ou bien +1), cette méthode d'expression des sorties est la plus convenable dans le cas des prédictions qualitatives, le tableau IV.1.3 représente la forme des différentes sorties.

**Tableau IV.1.3 :** Codage des sorties

| Activité thérapeutique | Molécule possède l'activité | Molécule ne possède pas l'activité |
|---|---|---|
| Activité anti-inflammatoire. | -1 | +1 |
| Activité antibactérienne | -1 | +1 |
| Activité antibiotique | -1 | +1 |
| Activité analgésique | -1 | +1 |
| Activité antivirale | -1 | +1 |
| Activité diurétique | -1 | +1 |
| Activité glucocorticoïde | -1 | +1 |

## IV.1.8. Propriétés de la base finale

Toutes les étapes précédantes ont permis d'obtenir une base de données finale , dont les caractéristiques de cette dernière sont représentées dans le tableau IV.1.4.

**Tableau IV.1.4 :** Caractéristiques de la base finale

| Base | Dimension des entrées | Dimension des sorties |
|---|---|---|
| Base totale | 532×175 | 543× 7 |
| anti-inflammatoires | 109×175 | 109×1 |
| Antibiotiques | 160×175 | 160×1 |
| Analgésiques | 84×175 | 84×1 |
| Antiviraux | 81×175 | 81×1 |
| Diurétiques | 54×175 | 54×1 |
| glucocorticoides | 27×175 | 27×1 |
| Autres | 17×175 | 17×1 |

# Partie 2

# Modélisation neuronale

## IV.2.1. Introduction

La première partie de notre étude nous permet d'obtenir une base de données de dimension importante mais cette dernière pose un problème majeur : le model risque d'être sur-ajusté et incapable de prédire les grandeurs à modéliser, et dans ce cas on doit sélectionner le nombre de descripteur qui assure la bonne prédiction de l'activité thérapeutique, tout en tenant compte de la bonne architecture RNA qui assure une erreur minimale. Pour atteindre cet objectif, on doit passer par une étape de réduction préliminaire qui va être suivie d'une autre réduction basée sur des approches statistiques.

## IV.2.2. Réduction préliminaire

Au premier stade, on doit faire un prétraitement de la base de données, tel que la sélection des données (entrées) consiste à déterminer les variables pertinentes par rapport à la grandeur à modéliser ; elle vise deux points essentiels :

- Réduction de la dimension de l'espace de la représentation des variables du modèle.
- Élimination des entrées qui n'ont pas ou possèdent peu d'influence sur la sortie.

Donc les descripteurs éliminés dans cette première étape de réduction sont des descripteurs, soit qui portent une valeur constante ou bien qui varie dans un intervalle trop réduit.

## IV.2.3. Réduction par méthode statistique

Afin de réduire la dimensionnalité des variables d'entrées, plusieurs approches sont possibles :

- Sélectionner les variables les plus pertinentes.
- Remplacer les variables corrélées par de nouvelles variables synthétiques obtenues à partir de leurs combinaisons.

### IV.2.3.1. Méthode sélectionnée

La méthode choisie dans notre étude, c'est la méthode de la variable pertinente afin d'appliquer cette dernière, on doit procéder soit à la sélection progressive, soit à l'élimination progressive.

### IV.2.3.2. Procédure générale

Donc dans notre étude on va utiliser la méthode d'élimination progressive car c'est la méthode qui converge plus rapidement. La figure IV.2.1 présente un organigramme qui montre la procédure d'élimination progressive utilisée dans notre étude.

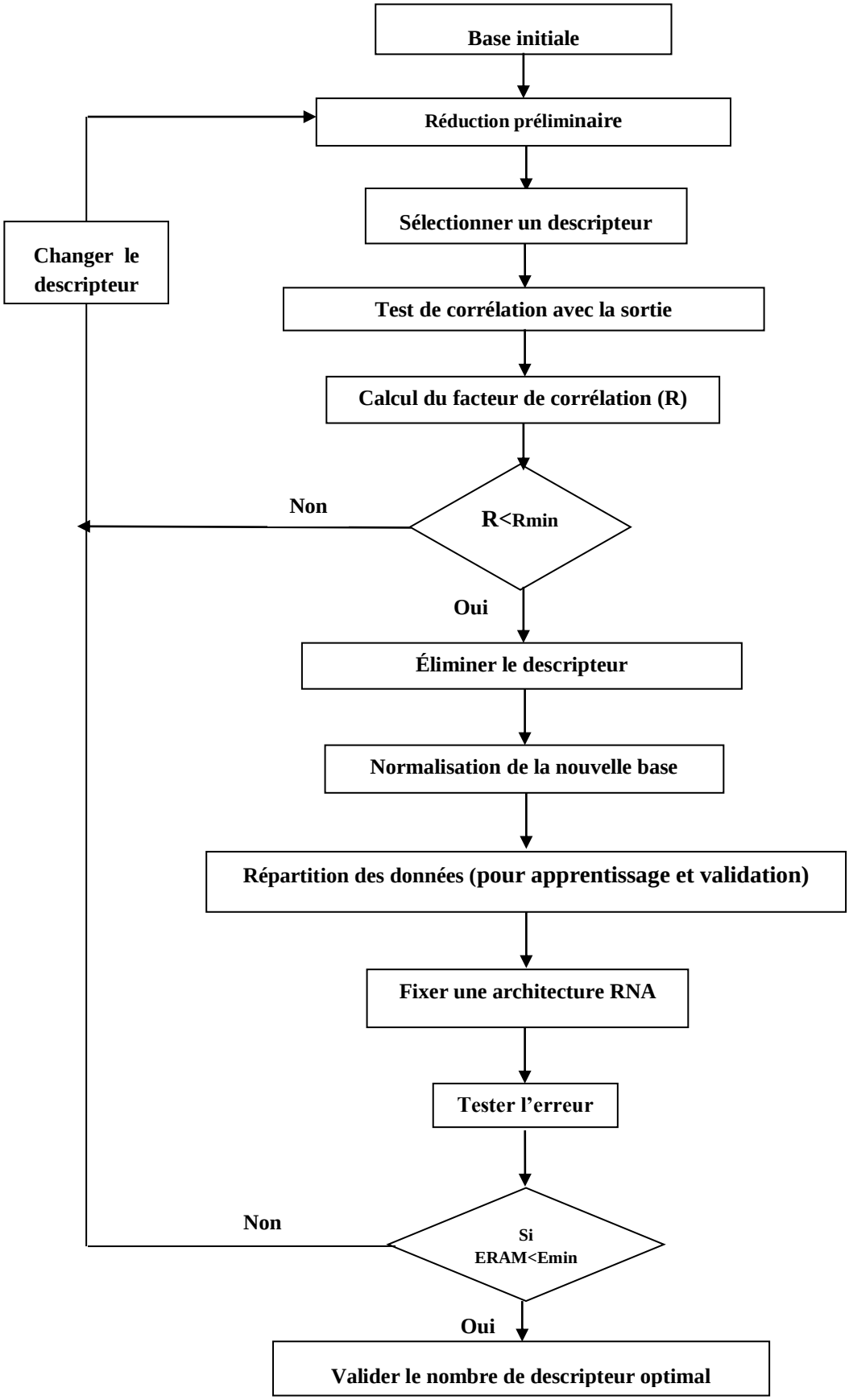



**Figure IV.2.1 :** Organigramme représentant la procédure de réduction des descripteurs par la méthode élimination progressive

## IV.2.4. Optimisation de la prédiction de l'activité anti-inflammatoire

### IV.2.4.1. Définition d'un anti-inflammatoire

Il s'agit d'un groupe de médicaments destinés à traiter une réaction inflammatoire et les maladies qui en résultent telles que les manifestations rhumatismales, les fractures, les stomatites et les lésions génitales et urinaires.

Ces médicaments ont des structures chimiques variées. Toutefois, ils présentent des caractères communs :

- Sur le plan pharmacologique : Tous les anti-inflammatoires, y compris les non stéroïdiens, inhibent la synthèse des prostaglandines.
- Sur le plan activité : Ils possèdent, en plus de leurs activités anti-inflammatoires, des effets analgésiques et antipyrétiques.
- Sur le plan chimique : Ils présentent des fonctions acides ou des caractères acides.
- Sur le plan effets secondaires : Les anti-inflammatoires provoquent une lésion de la muqueuse gastroduodénale.

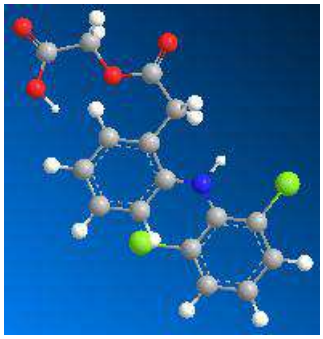
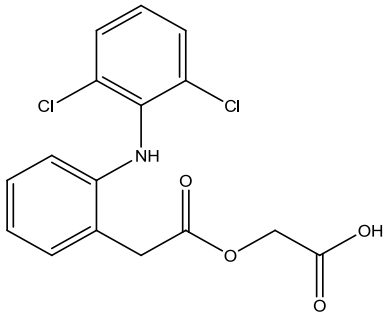
### IV.2.4.2. Liste des anti-inflammatoires utilisés

Du fait que les anti-inflammatoires possèdent des structures moléculaires très variées, on a essayé d'utiliser surtout les anti-inflammatoires non-stéroïdiens, ces médicaments agissent en inhibant l'action de la cyclo-oxygénase (COX), une enzyme médiant la formation de prostaglandines à partir de l'acide arachidonique. En plus d'être des médiateurs inflammatoires, les prostaglandines ont aussi un rôle dans l'agrégation plaquettaire, la protection de l'estomac et la fièvre. Cela explique les effets secondaires des anti-inflammatoires:

- Ils gênent l'agrégation plaquettaire et donc la coagulation du sang (on dit à tort qu'ils « fluidifient le sang »), ce qui peut prévenir la formation de caillots (thrombose), mais aggrave les hémorragies;
- Ils peuvent provoquer des problèmes gastriques (ulcères) et rénaux;
- Ils ont pour effet de combattre la fièvre (antipyrétiques).

Le tableau IV.2.1 résume toutes les molécules qui possèdent une propriété anti-inflammatoire et qui sont utilisées dans la base de données initiale.

**Tableau IV.2.1 :** Molécules anti-inflammatoire utilisées dans la base de données initiale [(Kleemann et Engel (2001)].

| Anti-inflammatoires | | | | |
|---|---|---|---|---|
| Aceclofenac non-steroidal anti-inflammatory | | | | |
|  | |  | | |
| 2-[(2.6-dichloropheny) amino]benzeneacetic acid carboxymethyl ester<br>O=C(OCC(O)=O)CC1=CC=CC=C1NC2=C(Cl)C=CC=C2Cl | | | | |
| Aceclofenac | Acemetacin | Flunoxaprofen | Fluocinonide | Piroxicam |
| acétaminophène | acide méfénamique | Fluperolone acetate | Flurbiprofen | Pirprofen |
| Alclofenac | Alclometasone dipropionate | Formebolone | Glucametacin | Pranoprofen |
| Alminoprofen | Aminophenazone | Gentisic acid | Hydroxyethyl salicylate | Prednicarbate |
| Ampiroxicam | Apigénine | Ibuprofen | Imipenem | Proglumetacin |
| Azapropazone | balsalazide | Indobufen | Indometacin | Proquazone |
| Bendazac | Benoxaprofen | Indometacin farnesil | Isoxicam | Protizinic acid |
| Betamethasone propionate | Bucloxic acid | Kebuzone | Ketorolac | Proxazole |
| Bufexamac | Bumadizone | Ketoprofen | Lornoxicam | Rofecoxib |
| Butibufen | Carprofen | Loteprednol etabonate | Loxoprofen | Rofecoxib skeletal |
| Celecoxib | Chlorthenoxazine | Lumiracoxib | Mabuprofen | Salicylate de méthyle |
| Cinmetacin | Clidanac | Meclofenamic acid | Mefenamic acid | Sulindac |
| Clobenoside | Clometacin | Meloxicam | Mesalazine | Suprofen |
| Dexelvucitabine | Diclofenac | Methadone | Metiazinic acid | Tenoxicam |
| Diacerein | Droxicam | Mofebutazone | Mofezolac | Tiaprofenic acid |
| Diflunisal | Emorfazone | Mycophenolate | Naproxen | Tiaramide |
| Etofenamate | Etoricoxib | Nimesulide | Nitrofural | Tiaprofenic acid |
| Feclobuzone | Felbinac | Oxametacin | Oxaprozin | Tiaramide |
| Fenbufen | Fenclofenac | Oxyphenbutazone | Phenylbutazone | Tinoridine |
| Fendosal | Fluazacort | Parecoxib | Pifoxime | Tolfenamic acid |
| Firocoxib | Fludroxycortide | Parsalmide | Piketoprofen | Tolmetin |
| Valdecoxib | Zileuton | Zomepirac | Pipebuzone | Pimeprofen |
| Flufenamic acid | Flumetasone | | | |

### IV.2.4.3. Optimisation du nombre de descripteurs

La sélection des entrées consiste à déterminer las variables pertinentes par rapport à la grandeur à modéliser. Elle vise deux points essentiels, [Hinton (1989)] :

- Réduction de la dimension de l'espace de représentation des variables du modèle ;
- Élimination des entrées qui n'ont pas ou peu d'influence sur la sortie.

Il est important de mentionner que chaque ajout d'une variable encombre le modèle et entraine une multiplication des paramètres du réseau. Raison pour laquelle, la représentation des entrées doit contenir seulement les facteurs pertinents par rapport à la grandeur à modéliser.

#### IV.2.4.3.1. Méthode sélectionnée : élimination progressive

On débute la modélisation par l'ensemble des descripteurs, en les éliminant progressivement, jusqu'à obtenir le meilleur jeu des composantes tout en se basant sur l'organigramme de la figure IV.2.1.

#### IV.2.4.3.2. Normalisation des données

Avant tout apprentissage, il est indispensable de normaliser et de centrer toutes les variables d'entrée. En effet, si des entrées ont des grandeurs très différentes, celles qui sont petites n'ont pas d'influence sur l'apprentissage. Il existe plusieurs approches de normalisation des données dont le choix dépend de la nature des données et de la fonction à approximer.

La normalisation utilisée a pour but d'avoir les données normalisées comprises entre -1 et +1 en utilisant l'équation de normalisation IV.1 :

$$\boldsymbol{Valeur\ normalisée} = \frac{\mathbf{2}(\boldsymbol{valeur\ réelle - Minumum})}{\boldsymbol{Maximum - Minimum}} - \mathbf{1} \qquad \textbf{IV.1}$$

Les valeurs des descripteurs calculés par DRAGON sont normalisées par la fonction « premnmx » de MATLAB qui réalise la normalisation selon l'équation **IV.1.** La fonction inverse de la dénormalisation est réalisée par la fonction « *postmnmx* ».

#### IV.2.4.3.3. Répartition des données

Dans cette étape, les données sont subdivisées en deux sous-ensembles :

- Un ensemble pour l'apprentissage ;
- Un ensemble pour la phase de validation.

L'échantillonnage des deux sous-ensembles est fait de façon à refléter parfaitement la base de données de départ, c'est-à-dire, on choisit toujours, à partir du tableau des données, deux valeurs pour l'apprentissage et la troisième valeur pour la validation.

#### IV.2.4.3.4. Architecture du réseau de neurone utilisé

Pour bien montrer l'influence du type de variables utilisées dans la modélisation, on doit bien fixer l'architecture du RNA, puis on fait varier le nombre de variables d'entrées qui représente le nombre de neurones de la couche d'entrée en déterminant à chaque fois l'erreur correspondante à ce nombre de variables jusqu'à l'obtention de l'erreur minimale pour un nombre optimal des variables considérées pertinentes selon l'organigramme (figure IV.2.1)

Donc la configuration du réseau sera comme suit :

- Algorithme d'apprentissage : trainlm ;
- Nombre de neurones dans la couche d'entrée : variable ;
- Nombre de neurones dans les deux couches cachées :
  - ✓ 18 neurones dans la première couche cachée ;
  - ✓ 2 neurones dans la deuxième couche cachée ;
- 1 seul neurone dans la couche de sortie ;
- On a adopté la fonction de transfert « tansig » pour les deux couches cachées et la fonction « purelin » pour la couche de sortie [MathWorks (2004) ; Vogl et all (1988)].

#### IV.2.4.3.5. Résultats

On commence l'élimination progressive à partir d'une base initiale contenant 165 descripteurs moléculaires et on élimine à chaque fois 5 descripteurs, de telle manière qu'on choisit les descripteurs qui représentent une corrélation minimale avec la sortie, et les différents tests de corrélations sont assurés par STATISTICA 9 ®. Les descripteurs éliminés dans chaque passage sont représentés dans le tableau IV.2.2.

**Tableau IV.2.2 :** Descripteurs éliminés dans chaque passage d'élimination progressive

| Passages d'élimination | Les descripteurs éliminés | | | | |
|---|---|---|---|---|---|
| 165 - 160 | AMW | D/Dr03 | D/Dr04 | D/Dr07 | G(N..F) |
| 160 - 155 | SRW09 | Hy | TIE | G(O..F) | G(S..S) |
| 155 - 150 | Cl-089 | T(N..Cl) | G(O..Cl) | G(O..S) | G(N..S) |
| 150 - 145 | PJI2 | Hypertens-80 | Neoplastic-50 | Neoplastic-80 | Hypnotic-50 |
| 145 - 140 | D/Dr05 | T(O..Cl) | T(O..S) | T(N..S) | G(N..Cl) |
| 140 - 135 | Infective-80 | GVWAI-50 | GVWAI-80 | C-001 | L/Bw |
| 135 - 130 | ARR | RBF | G(N..N) | C-002 | H-046 |
| 130 - 125 | G(N..O) | G(O..O) | H-051 | H-052 | PW3 |
| 125 - 120 | BLI | HNar | GNar | O-057 | AMR |
| 120 - 115 | PW4 | PW5 | MSD | T(N..N) | C-024 |
| 110 - 105 | Ui | Me | GMTIV | Xt | MLOGP2 |
| 105 - 100 | D/Dr06 | Dz | nCL | nX | ZM1V |
| 100 – 95 | Mp | MAXDN | nBnz | Jhete | Jhetp |
| 95 - 90 | Jhetm | Jhetv | S0K | Lop | Whetv |

Les résultats en termes d'erreur relative absolue moyenne (ERAM) sont représentés dans la figure (VI.2.2).

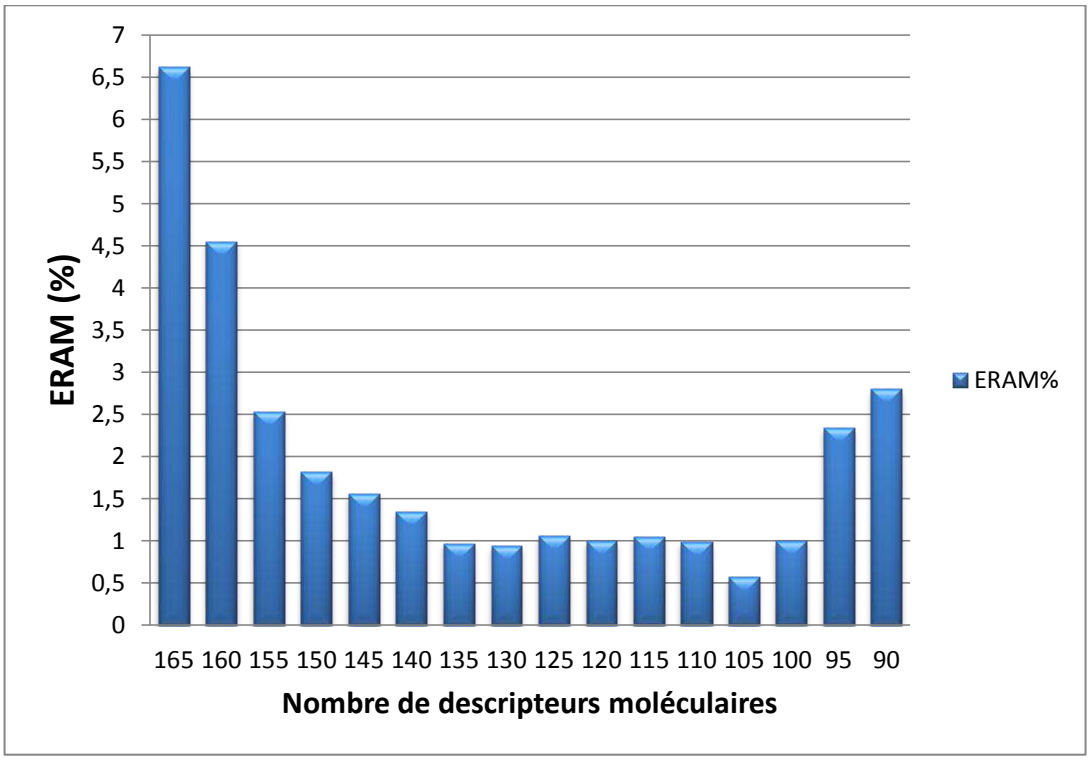



**Figure IV.2.2 :** Effet de l'élimination progressive sur l'erreur relative absolue moyenne.

Ces résultats nous permettent de dire que le nombre optimal de descripteurs qui assure un écart minimal entre les valeurs expérimentales et les valeurs calculées par le RNA est : 105 descripteurs. Une deuxième élimination a été effectuée, elle consiste à éliminer 11 molécules afin d'améliorer les performances du réseau de neurones.

### IV.2.4.4.Optimisation du modèle neuronal pour la prédiction d'activité anti-inflammatoire

L'objectif principal des travaux décrits dans cette section a été réalisé pour trouver l'architecture du réseau de neurones optimale permettant de prédire l'activité anti-inflammatoire. La conception du modèle neuronal consiste à faire une étude d'évaluation des éléments constitutifs du réseau en fonction de la performance de la modélisation souhaitée. Elle vise de fixer les éléments suivants :

1. Le type de réseau ;
2. Le nombre de couches cachées ;
3. Le nombre de neurones dans les couches ;
4. La fonction de transfert des neurones ;
5. L'algorithme d'apprentissage ;
6. Le nombre d'itérations.

L'organigramme de la méthodologie de calcul du réseau de neurones statique optimale RNSO est illustré dans la figure VI.2.3

La conception et l'optimisation du réseau ont été réalisées en trois étapes principales :

- la première étape consiste à la détermination des paramètres fondamentaux du modèle neuronal (la fonction d'activation et l'algorithme d'apprentissage),
- la deuxième étape consiste à l'étude des méthodes de prétraitement des données (la normalisation des données),
- la dernière étape, consiste à déterminer le nombre de couches cachées et le nombre de neurones dans chaque couche cachée.

Vu les recherches basées sur les réseaux de neurones artificiels, le modèle qui donne une erreur minimale sera choisi [Lee et al. (1991)]. Pendant la première étape de l'optimisation de réseau, les paramètres fondamentaux ont été examinés pour avoir une bonne performance ; Toutes les étapes ont été développées à l'aide du logiciel commerciale MATAB® et qui sont représentées dans l'organigramme de la figure IV.2.3.

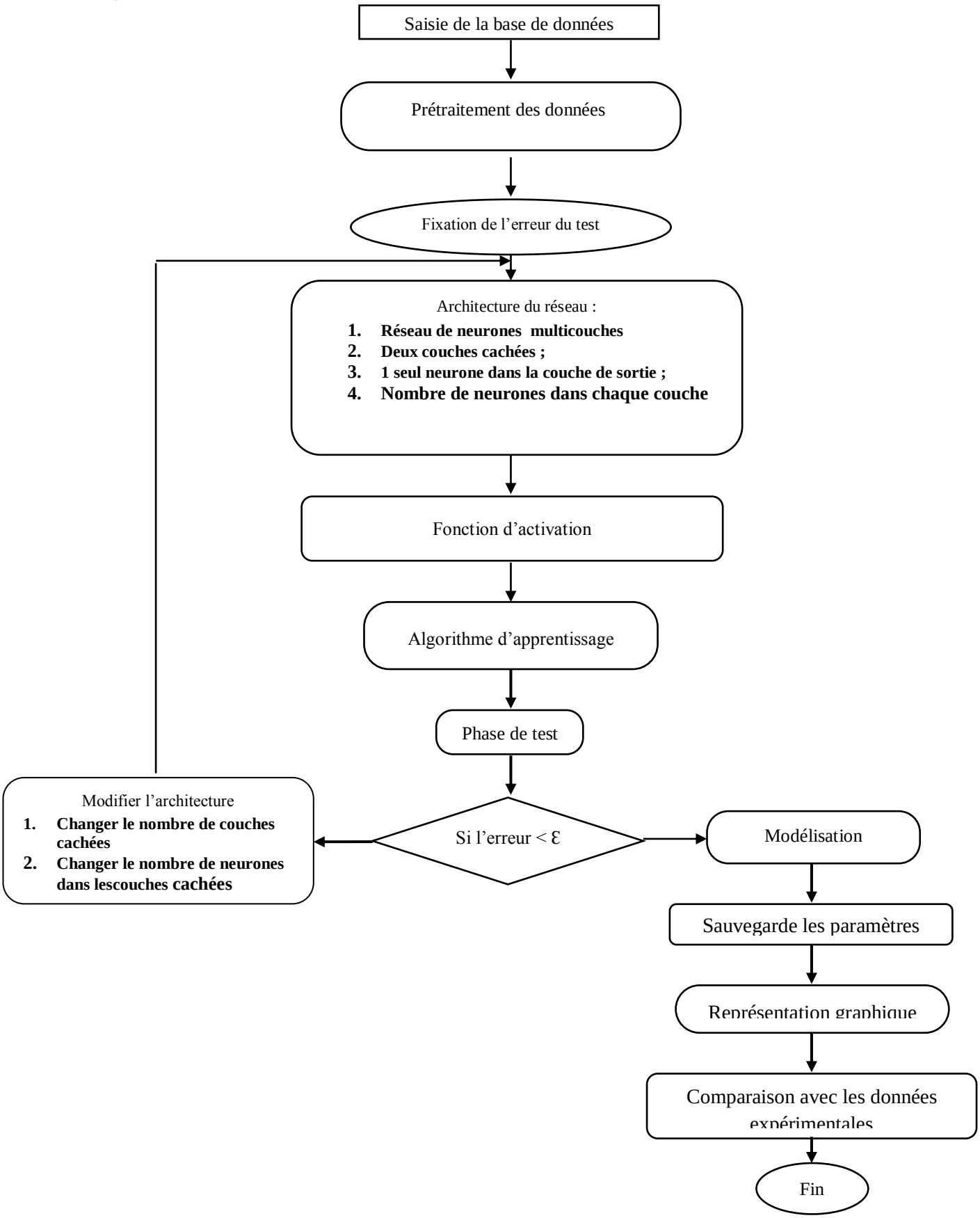



**Figure IV.2.3 :** Organigramme de la méthodologie de calcul RNSO [Omari (2009)].

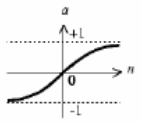
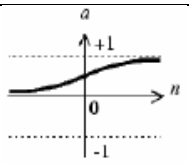
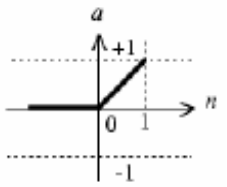
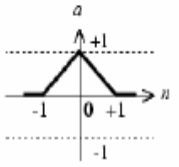
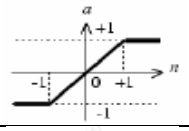
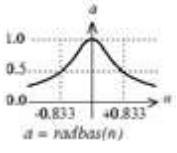
**Étape 1 : Sélection de la fonction d'activation**

Six fonctions d'activation ont été appliquées (tableau IV.2.3) pour avoir une seule fonction d'activation pour la couche cachée qui donne une erreur relative absolue minimale, la figure IV.2.4 présente la variation de l'erreur relative absolue moyenne (ERAM) pour 6 réseaux de neurones avec différentes fonctions d'activation.

**Configuration initiale du réseau**

- 105 neurones dans la couche d'entrées (nombre optimal de descripteurs) ;
- 18 neurones dans la couche cachée;
- 1 neurone dans la couche de sortie ;
- 6 fonctions d'activations ont été testées.
- On a adopté l'algorithme d'apprentissage « trainlm ».

**Tableau IV.2.3 :** Différentes fonctions d'activations utilisées dans la phase de conception du modèle neuronal.

| Fonction de transfert | Code MATLAB | Graphe |
|---|---|---|
| Tan- Sigmoïde transfer function | Tansig |  |
| Log- Sigmoïde transfer function | Logsig |  |
| Positive linear transfer function | Poslin |  |
| Triangular basis transfer function | Tribas |  |
| Symetric saturating linear transfer function | Stalins |  |
| Radial basis function | Radbas |  |

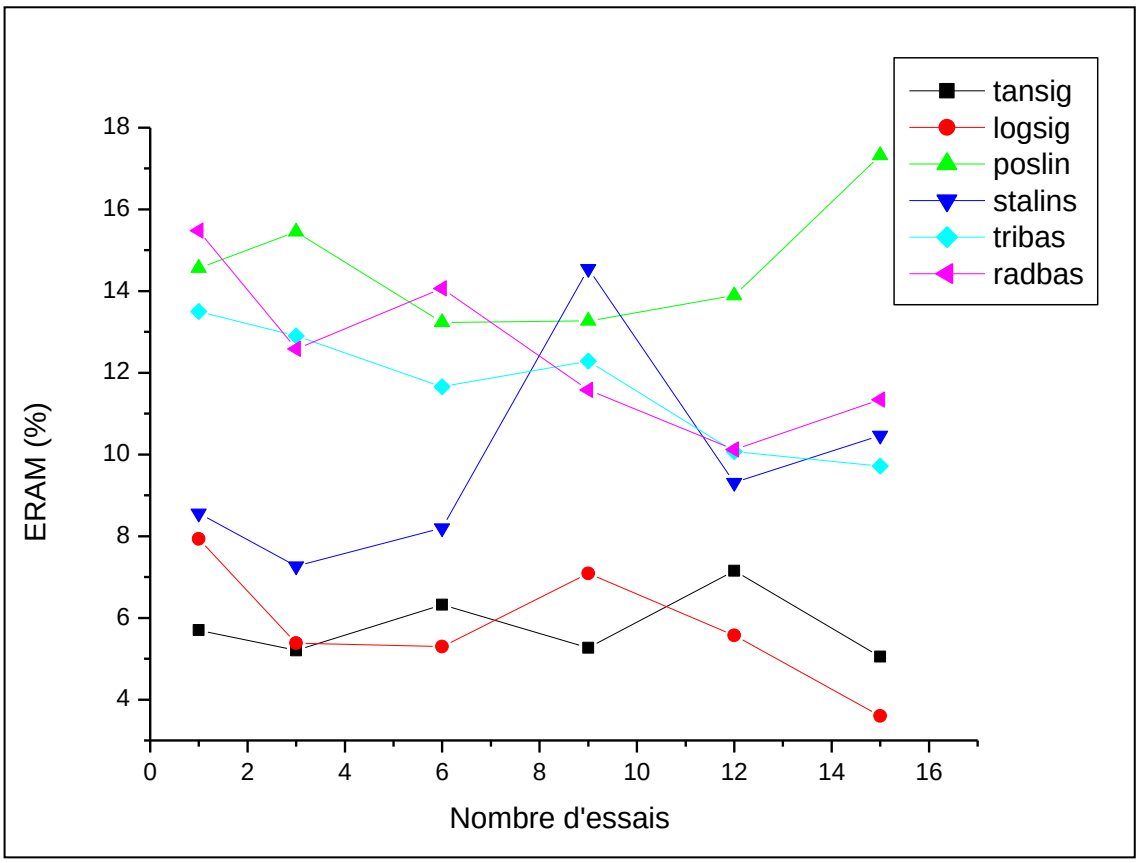



**Figure IV.2.4 :** Effet de la fonction d'activation sur les performances du réseau de neurones.

La fonction d'activation qui a été choisie pour les couches cachées est tangente Sigmoïde et la fonction identité purelin pour la couche de sortie, ces deux paramètres sont très utilisés pour les réseaux de neurones de type MLP, [Vogl et al. (1988), Hagan et al. (1996), Dreyfus et al. (2002), Mathwork. (2004)].

**Étape 2 : Sélection de l'algorithme d'apprentissage**

La sélection de l'algorithme d'apprentissage est plus difficile que celle de la fonction d'activation car il existe un nombre important des algorithmes d'apprentissage qui peuvent produire un model neuronal performant, rapide et fiable. Six algorithmes d'apprentissage ont été appliqués (tableau IV.2.4) pour avoir un seul algorithme qui donne une erreur relative absolue minimale. Initialement pour les essais de tous les réseaux, on a utilisé la même configuration initiale précédente, seulement on sélectionne la fonction « tansig» comme fonction d'activation pour les couches cachées et la fonction identité pureline pour la couche de sortie, puis on test les 6 algorithmes pour cette architecture du réseau de neurones.

Les résultats des essais sont décrits sur la figure IV.2.5 en termes d'erreur relative absolue moyenne ERAM. D'après cette figure, on constate que l'algorithme d'apprentissage qui donne la meilleure performance en terme d'ERAM minimale est bien l'algorithme Rétropropagation de l'échelle de gradient conjugué «Trainscg ».

**Tableau IV.2.4 :** Différents algorithmes d'apprentissage utilisés lors de la phase de conception du modèle neuronal

| **Nom de l'algorithme d'apprentissage** | **Code Matlab** | **Acronyme** |
|---|---|---|
| Rétropropagation de Levenberg –Marquardt. | Trainlm | LM |
| Rétropropagation de Régularisation Bayésienne. | Trainbr | BR |
| Rétropropagation de Résilient. | Trainrp | RP |
| Rétropropagation de l'échelle de gradient conjugué. | Trainscg | SCG |
| Rétropropagation de Quasi-Newton. | Trainbfg | BFG |
| Rétropropagation de du taux d'apprentissage variable. | Traingdx | GDX |

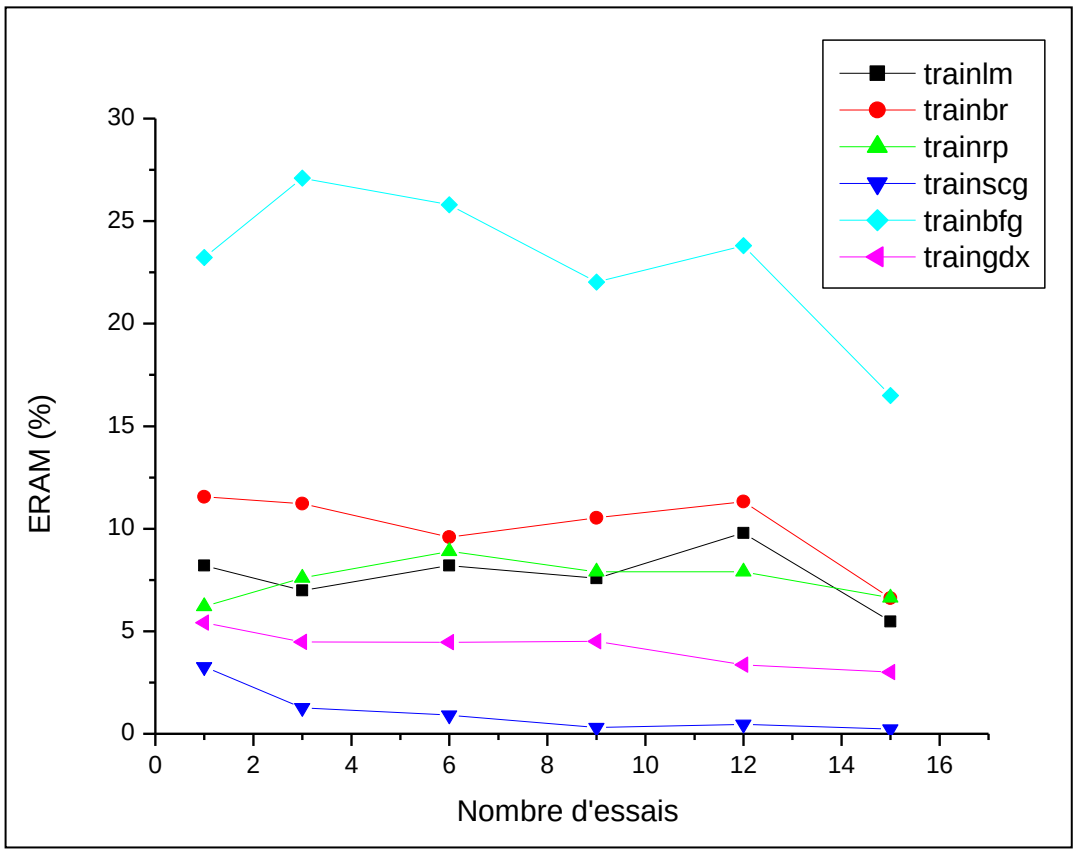



**Figure IV.2.5 :** Effet du type d'algorithme d'apprentissage sur la performance du réseau de neurones en termes d'ERAM.

**Étape 3 : Sélection de la méthode de normalisation**

La normalisation est une étape importante dans le processus de l'élaboration des données. Elle devient dans la plus part des cas, une exigence pour les matrices des inputs, du fait qu'on utilise souvent des fonctions de transfert bornées de type sigmoïde dans les modélisations statiques, en plus la normalisation accélère la convergence du réseau pendant l'apprentissage et augmente son efficacité de généralisation, [Dreyfus et al. (2002)].

Vu la grande différence qui existe entre les valeurs des entrées de notre système, on a procédé à une normalisation primaire des entrées avant de réaliser une normalisation par les fonctions disponibles dans MATLAB, et on va étudier leurs effets sur la performance du modèle neuronal.

Le tableau IV.2.5 présente les trois techniques de prétraitements de la base de données testées lors de notre étude. On a testé en premier lieu l'influence de chaque fonction de normalisation disponible sur MATLAB ensuite on a étudié l'influence de la combinaison des deux modes de normalisation. Cela a été réalisé tout en gardant la même architecture du réseau de l'étape précédente.

**Tableau IV.2.5 :** Différentes fonctions de normalisation utilisées

| Code du réseau | Méthode de prétraitement | Code en Matlab |
|---|---|---|
| Net 1 | Méthode des minimales et des maximales (MMM) ; [Hagan et al. (1996) ; MathWorks (2007)]. | Premnmx |
| Net 2 | Méthode de la moyenne et de la déviation standard (MMDS). | Prestd |
| Net 3 | Sans normalisation primaire (NP) | / |

La figure IV.2.6 montre l'influence de la méthode de normalisation sur la performance du modèle neuronal, on remarque que la normalisation primaire a une grande influence sur le modèle. L'écart en terme d'ERAM par rapport à l'apprentissage sans utilisation de la normalisation primaire, donc le bon résultat a été trouvé dans le cas de la combinaison de la normalisation primaire et la fonction « premnmx ».

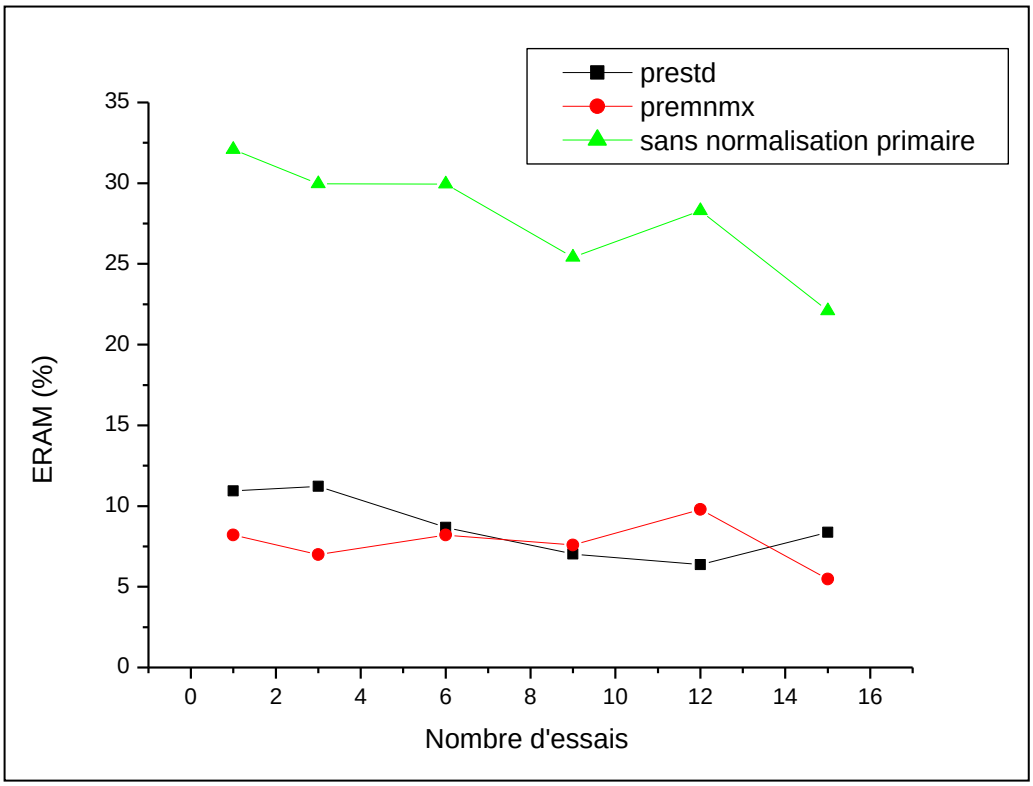



**Figure IV.2.6 :** Effet de la méthode de normalisation sur la performance du réseau de neurones en termes d'ERAM.

**Étape 4 : Optimisation du nombre de couches cachées et nombre de neurones dans chaque couche cachée**

Cette étape d'optimisation se base sur le test de différentes architectures du réseau de neurones par la modification du nombre de neurones dans chaque couche cachée. Au cours de cette étape on a utilisé deux paramètres d'évaluation de performance du réseau de neurones : l'ERAM minimale et le coefficient de régression globale.

**A) Une seule couche cachée**

En commençant par le réseau de neurones avec une seule couche cachée, le nombre de neurones a varié dans un intervalle représenté sur la figure IV.2.7.

Les figures IV.2.7 et IV.2.8 représentent respectivement la performance du réseau de neurone en termes d'ERAM et en termes du coefficient de régression globale. On a constaté que le RNA avec 30 neurones dans la couche cachée, cet architecture nous permet d'obtenir une ERAM minimale, mais le coefficient de régression n'est pas encore maximal d'où la nécessité d'ajouter une deuxième couche cachée.

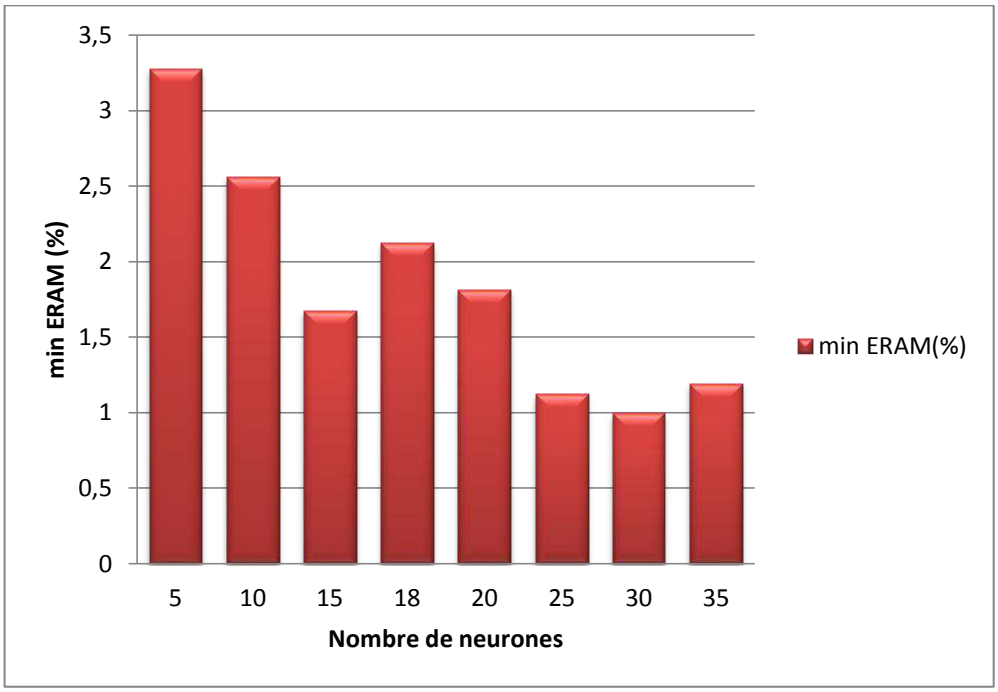



**Figure IV.2.7 :** Performance du réseau de neurones à une seule couche cachée en terme d'ERAM minimale.

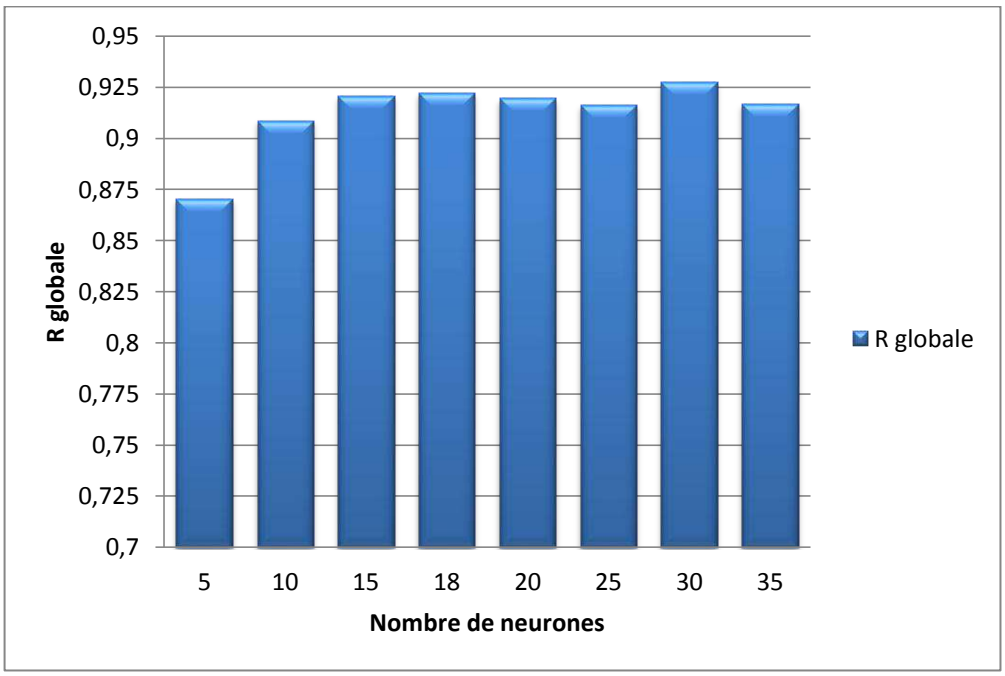



**Figure IV.2.8 :** Performance du réseau de neurones à une seule couche cachée en terme du coefficient de régression globale.

### B) Deux couches cachées

La structure du réseau de neurones a été modifiée en intégrant une deuxième couche cachée. Les résultats obtenus sont représentés dans les figures IV.2.9 et IV.2.10.

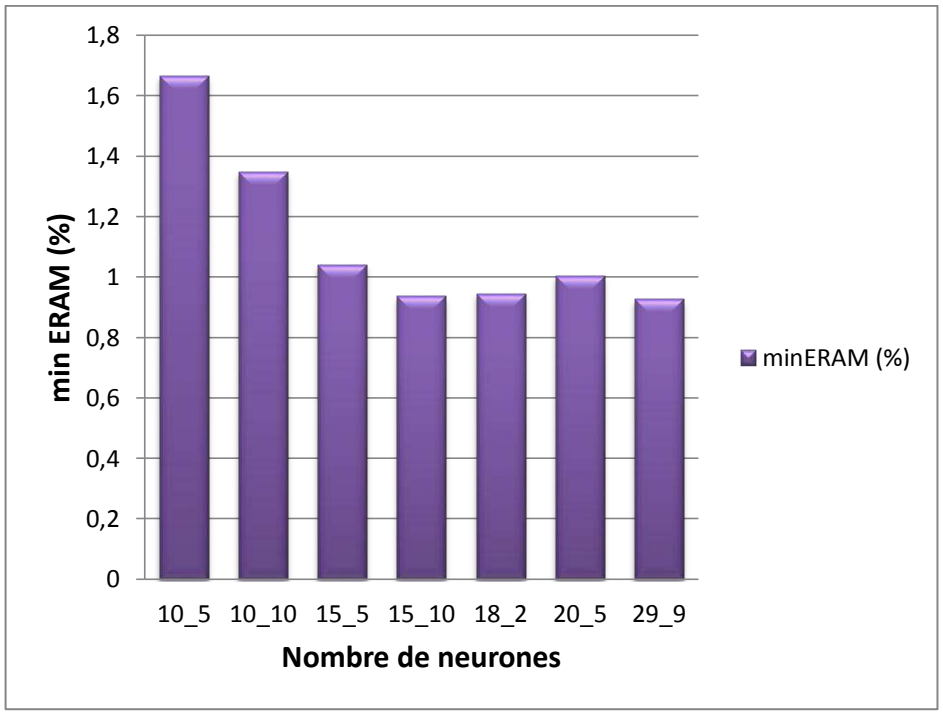



**Figure IV.2.9 :** Performance du réseau de neurones à deux couches cachées en termes d'ERAM minimale**.**

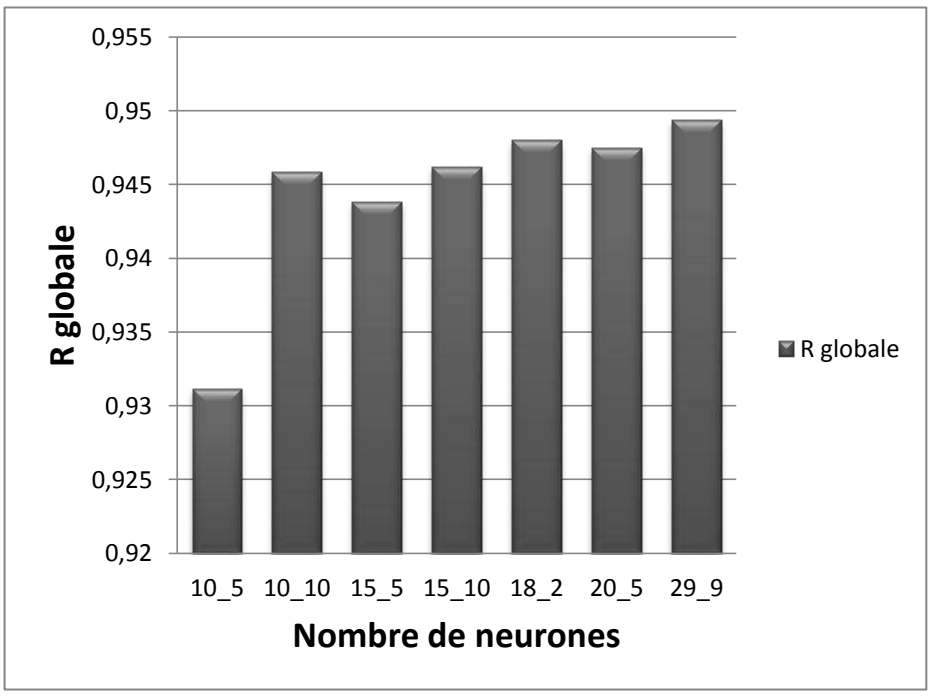



**Figure IV.2.10 :** Performance du réseau de neurones à deux couches cachées en termes du coefficient de régression globale.

### C) Trois couches cachées

La structure du réseau de neurones a été modifiée en intégrant une troisième couche cachée. Les résultats obtenus sont représentés dans les figures IV.2.11 et IV.2.12.

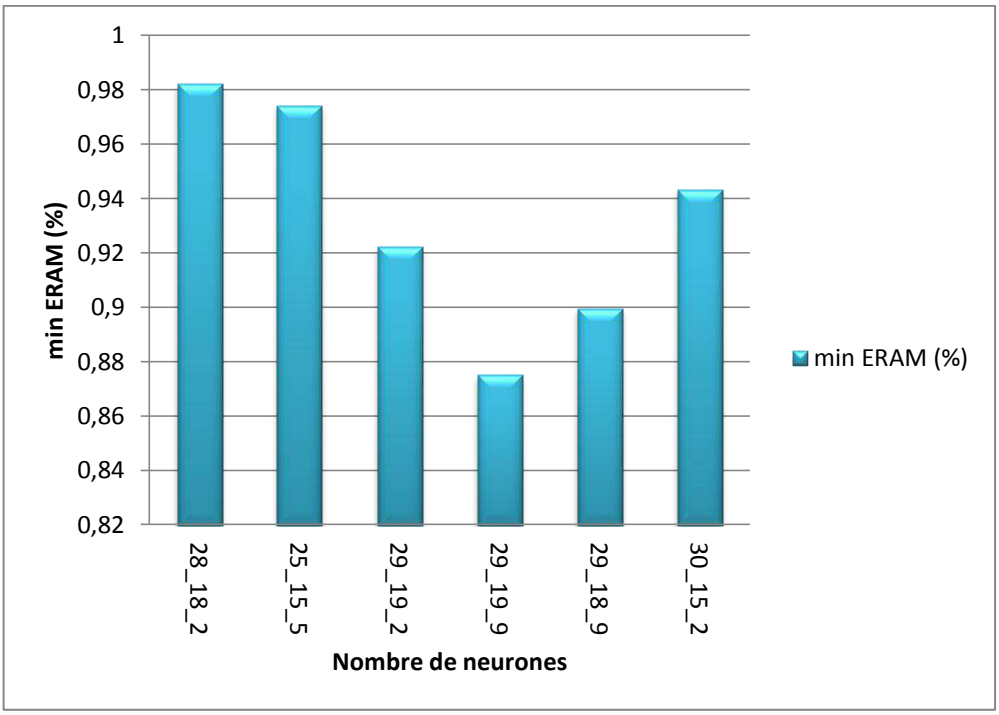



**Figure IV.2.11 :** Performance du réseau de neurones à trois couches cachées en termes d'ERAM minimale.

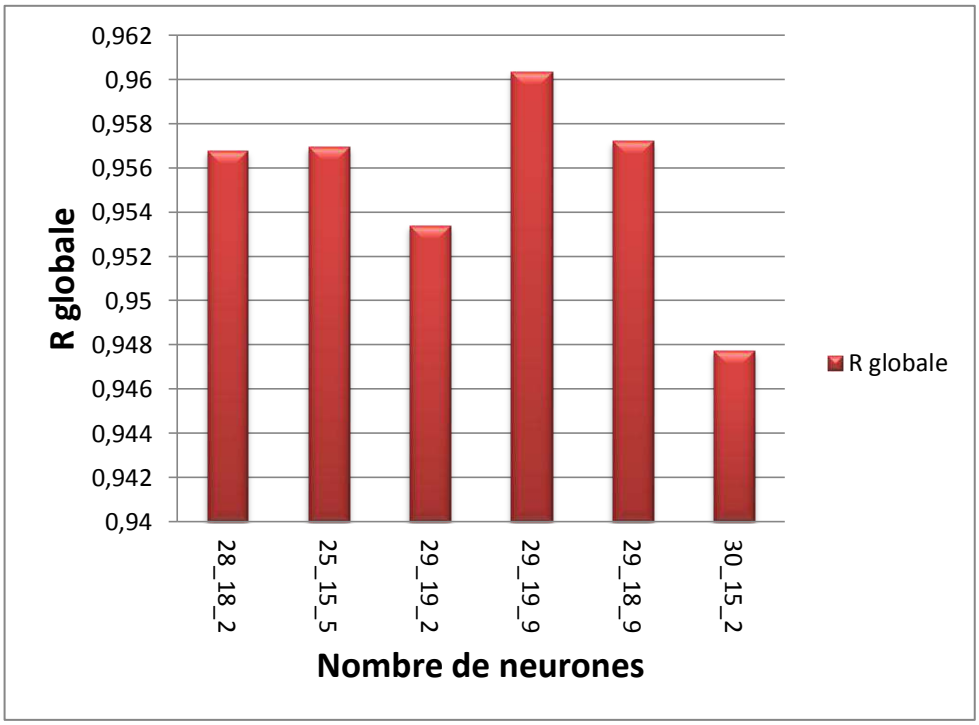



**Figure IV.2.12 :** Performance du réseau de neurones à trois couches cachées en termes du coefficient de régression globale.

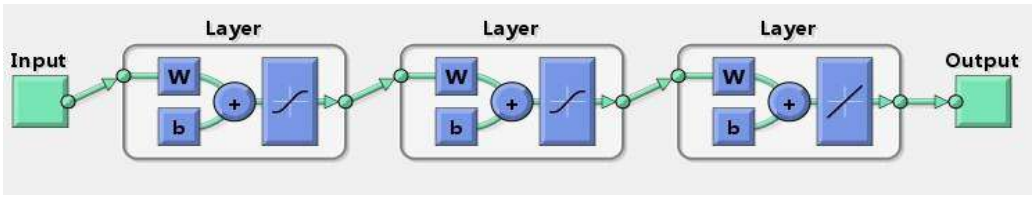
### IV.2.4.5.Structure du réseau RNSO obtenu

Le réseau adéquat vu de la monographie consultée et le test fait ne peut être en général qu'un réseau non bouclé type MLP (Multi Layer Perceptron, l'architecture de réseaux de neurones finale à trois couches cachées type MLP présenté dans la figure illustrée dans le tableau IV.2.6) avec des tangentes sigmoïdes comme fonctions de transfert des neurones cachés et une fonction linéaire (identité) pour le neurone de sortie.

L'algorithme d'apprentissage le mieux adapté d'après le test des différents algorithmes d'apprentissage est Rétropropagation de l'échelle de gradient conjugué, (Fonction trainscg de MATLAB) ®, Le tableau IV.2.6 récapitule les différents éléments structurels du réseau de neurones utilisé dans la modélisation de la prédiction de l'activité anti-inflammatoire.

**Tableau IV.2.6 :** Structure du réseau de neurones optimisé

| **Type de réseau** | **RNSO (MLP)** | |
|---|---|---|
| **Couche** | Nb de neurones | Fonction d'activation |
| **Couche d'entrée** | 105 | --- |
| **1ere Couche cachée** | 29 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **2eme Couche cachée** | 19 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **3eme couche cachée** | 9 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **Couche de sortie** | 1 | Linéaire (Fonction purelin de MATLAB®) |
| **Algorithme d'apprentissage** | Rétropropagation de l'échelle de gradient conjugué, (Fonction trainscg de MATLAB) ® | |




**105** **29** **19** **9** **1**

### IV.2.4.6 Résultats

Les figures IV.2.13, IV.2.14 présentent les courbes de régression linéaire de l'activité anti-inflammatoire obtenues durant l'élaboration du réseau de neurones, c'est-à-dire dans les deux phases de test et de généralisation. Nous remarquons que la valeur de l'erreur relative absolue moyenne par rapport aux données utilisées est égale à 0.875% et le facteur de corrélation globale est de l'ordre de 0,96.

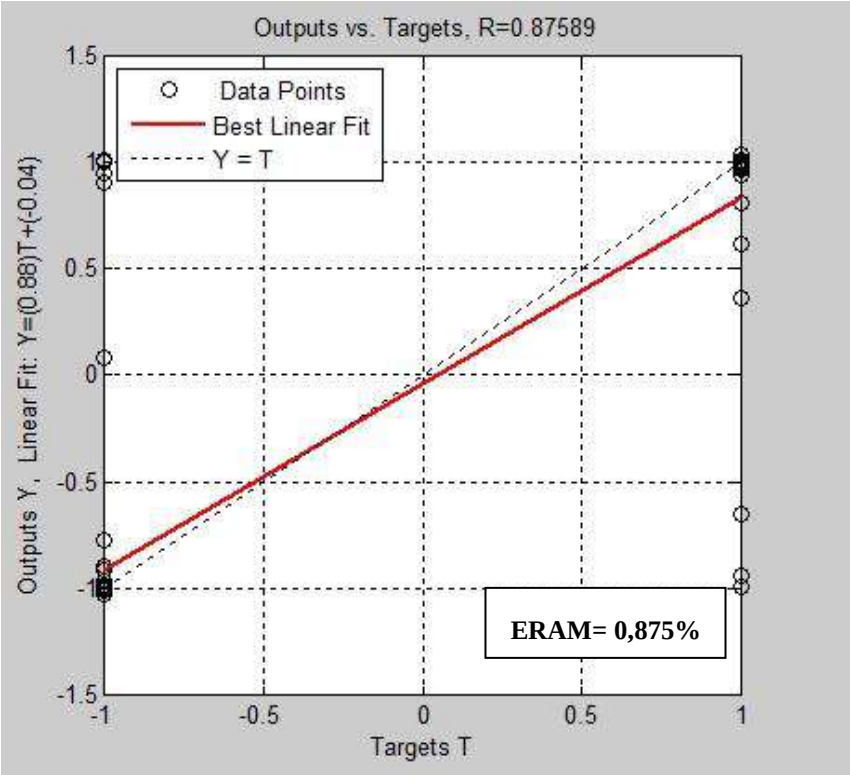



**Figure IV.2.13 :** Courbe de régression linéaire de l'activité anti-inflammatoire calculée par RNA optimisé avec l'activité anti-inflammatoire donnée expérimentalement pour le processus de test

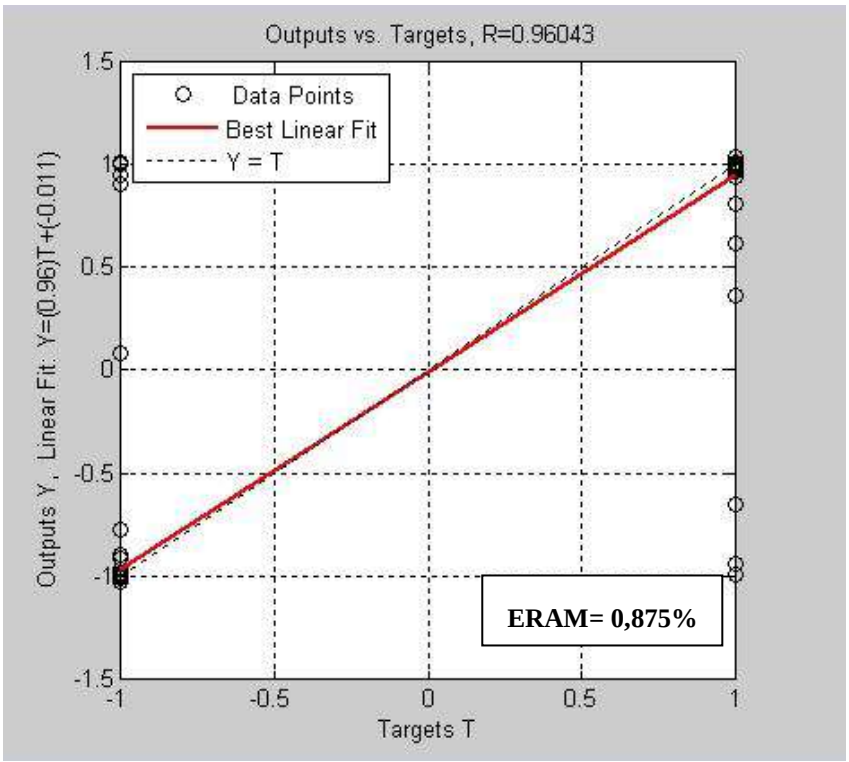



**Figure IV.2.14 :** Courbe de régression linéaire de l'activité anti-inflammatoire calculée par RNA optimisé avec l'activité anti-inflammatoire donnée expérimentalement pour le processus de généralisation

Avec

T : valeur de $Y^{\text{expérimentale}}$ ;

Y : valeur de $Y^{\text{calculée}}$.

Donc la meilleure concordance linéaire entre les valeurs calculées et les valeurs expérimentales c'est :

$Y^{\text{calculée}} = (0.96)\ (Y^{\text{expérimentale}}) + (-0.011)$ ; avec R=0,96034.

### IV.2.4.7 Analyse des résultats

L'analyse primaire des résultats obtenus, nous permet de dire que le réseau de neurone optimisé possède une capacité de prédiction de l'activité anti-inflammatoire, cette capacité est directement dépendante au facteur de corrélation obtenu, donc si on veut illustrer un intervalle de confiance pour cette capacité prédictive en terme de pourcentages, on peut dire que :

La capacité prédictive d'un RNA optimisé sur une base de données contenant 532 molécules (avec 109 molécules possèdant une activité anti-inflammatoire, et avec un nombre de descripteurs statistiquement optimisé par élimination progressive), peut générer :

- 97,5563 % des molécules sont correctement prédites (ou bien se rapprochent de la valeur désirée);
- 2,4436 % des molécules ne sont pas correctement prédites ;

Ces pourcentages obtenus sont fortement liés à plusieurs facteurs tels que :

1. La taille de la base de données
2. La nature des molécules sélectionnées dans la base de données ;
3. La méthode statistique de sélection des descripteurs les plus pertinents ;
4. Le nombre et la nature de ces descripteurs ;
5. Et finalement les propriétés du réseau de neurones optimisé.

Donc on peut améliorer les pourcentages de confiance de prédiction tout en jouant sur ces facteurs.

## IV.2.5 Optimisation de la prédiction de l'activité antibiotique

### IV.2.5.1 Définition d'un antibiotique

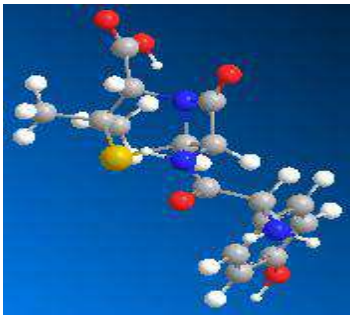
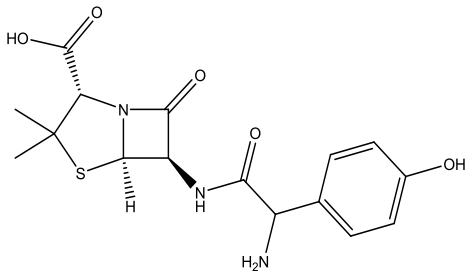
Un antibiotique est une molécule qui détruit ou bloque la croissance des bactéries. Dans le premier cas, on parle d'antibiotique bactéricide et dans le second cas d'antibiotique bactériostatique. Un même antibiotique peut être bactériostatique à faible dose et bactéricide à dose plus élevée.

Un grand nombre d'antibiotiques sont des molécules naturelles, fabriquées par des micro-organismes, des champignons ou d'autres bactéries. Ces derniers les produisent pour éliminer les bactéries concurrentes avec lesquelles ils sont en compétition dans leur biotope. Les antibiotiques agissent de manière spécifique sur les bactéries, en bloquant une étape essentielle de leur développement : synthèse de leur paroi, de l'ADN, des protéines, production d'énergie... Ce blocage se produit lorsque l'antibiotique se fixe sur sa cible, une molécule de la bactérie qui participe à l'un de ces processus métaboliques essentiels. Cette interaction entre l'antibiotique et sa cible est très sélective, spécifique des bactéries et ces composés ne sont en général pas actifs ni sur les champignons ni sur les virus. Il existe aussi d'autres molécules actives sur ces autres types d'agents infectieux que l'on 'appelle des antifongiques ou des antiviraux et qui sont distincts des antibiotiques.

### IV.2.5.2 Liste des antibiotiques utilisés

Le tableau IV.2.7 résume toutes les molécules qui possèdent une propriété antibiotique utilisé dans la base de données initiale.

**Tableau IV.2.7 :** Molécules antibiotiques utilisées dans la base de données initiale [(Kleemann et Engel (2001)].

<table>
<tr><td colspan="5">Liste des Antibiotiques</td></tr>
<tr><td colspan="2"></td><td colspan="3">Amoxicillin<br><br>O=C([C@@H](C(C)(C)S[C@]1([H])[C@@H]2NC(C(N)C3=CC=C(O)C=C3)=O)N1C2=O)O</td></tr>
<tr><td>Acide nalidixique</td><td>Acide oxolinique</td><td>Dalfopristine</td><td>Daunorubicin</td><td>Pirazinamide</td></tr>
<tr><td>Acide pipémidique</td><td>Amikacine</td><td>Demeclocycline</td><td>Dibekacin</td><td>Pivampicillin</td></tr>
<tr><td>Amoxicilline</td><td>Amphotericin</td><td>Dicloxacillin</td><td>Enoxacin</td><td>Pivmecillinam</td></tr>
<tr><td>Ampicilline</td><td>Apalcillin</td><td>Enviomycin</td><td>Epicillin</td><td>Propicillin</td></tr>
<tr><td>Aspoxicilline</td><td>Astromicin</td><td>Erythromycin</td><td>Ethambutol</td><td>Pyrrolnitrin</td></tr>
<tr><td>Azidamfenicole</td><td>Azosemide</td><td>Fleroxacin</td><td>Flomoxef</td><td>Quinupristine</td></tr>
<tr><td>Bacampicilline</td><td>Bekanamycin</td><td>Flucloxacillin</td><td>Fluperolone acetate</td><td>Ribostamycin</td></tr>
<tr><td>Bendroflumethiazide</td><td>Benzyl mustard oil</td><td>Flurbiprofen</td><td>Fluvoxamine</td><td>Ribostamycin</td></tr>
<tr><td>Benzylpenicilline</td><td>Carbenicillin</td><td>Fosfomycine</td><td>Framycetine</td><td>Rifampicin</td></tr>
<tr><td>Carfecilline</td><td>Carindacillin</td><td>Gatifloxacin</td><td>Gentamicine</td><td>Rolitetracycline</td></tr>
<tr><td>Cefacetrile</td><td>Cefaclor</td><td>Hetacillin</td><td>Isepamycin</td><td>Rosoxacin</td></tr>
<tr><td>Cefadroxil</td><td>Cefalexin</td><td>Isoconazole</td><td>Isoniazid</td><td>Secnidazol</td></tr>
<tr><td>Cefaloglycine</td><td>Cefaloridine</td><td>Kanamycin</td><td>lincomycin</td><td>Sisomicin</td></tr>
<tr><td>Cefalotine</td><td>Cefamandole</td><td>Latamoxef</td><td>Linezolid</td><td>Spectinomycin</td></tr>
<tr><td>Cefapirine</td><td>Cefatrizine</td><td>Loracarbef</td><td>Lymecycline</td><td>Stallimycin</td></tr>
<tr><td>Cefazedone</td><td>Cefazolin</td><td>Mecillinam</td><td>Meropeneme</td><td>Streptomycin</td></tr>
<tr><td>Céfépime</td><td>Céfixime</td><td>Metacycline</td><td>Metampicillin</td><td>Sulbenicillin</td></tr>
<tr><td>Cefmenoxime</td><td>Cefoperazone</td><td>Methacyclin</td><td>Meticillin</td><td>Sulfadiazine</td></tr>
<tr><td>Cefotaxime</td><td>Cefotetan</td><td>metronidazol GEN</td><td>Mezlocillin</td><td>Sulfaguanidine</td></tr>
<tr><td>Cefotiame</td><td>Céfotiam héxétil</td><td>Micronomicin</td><td>Minocyclin</td><td>Talampicillin</td></tr>
<tr><td>Cefoxitine</td><td>Cefpiramide</td><td>Midecamycin acetate</td><td>Mometasone furoate</td><td>Télithromycin</td></tr>
<tr><td>Cefpirome</td><td>Cefpodoxime proxetil</td><td>Moxifloxacin</td><td>Mupirocin</td><td>Temocillin</td></tr>
<tr><td>Cefradine</td><td>Cefsulodin</td><td>Mycophenolic acid</td><td>Nafcillin</td><td>Tetracycline</td></tr>
<tr><td>Ceftazidime</td><td>Ceftezole</td><td>Nalidixic acid</td><td>Neomycin</td><td>Thiamphenicol</td></tr>
<tr><td>Ceftizoxime</td><td>Ceftriaxone</td><td>Nifenazone</td><td>Netilmecin</td><td>Ticarcillin</td></tr>
<tr><td>Cefuroxime</td><td>Cefuroxime axetil</td><td>Nimorazol</td><td>Nitrofurantoin</td><td>Ticarcillin</td></tr>
<tr><td>Chloramphenicol</td><td>Chlortetracycline</td><td>Norfloxacin</td><td>Ofloxacin</td><td>Tinidazol</td></tr>
<tr><td>Ciclacillin</td><td>Clindamycin</td><td>Oxacillin</td><td>Pecilocin</td><td>Tobramycin</td></tr>
<tr><td>Clometocillin</td><td>Clomocycline</td><td>Oxytétracyclin</td><td>Pefloxacin</td><td>Trovafloxacin</td></tr>
<tr><td>Cloxacillin</td><td>Cycloserine</td><td>Oxytetracydine</td><td>Pheneticillin</td><td>Viomycin</td></tr>
<tr><td>Cytarabine</td><td>Griseofulvin</td><td>Paromomycin</td><td>Piperacillin</td><td>Virginiamycin</td></tr>
</table>

### IV.2.5.3 Optimisation du nombre de descripteurs

Afin de trouver le nombre optimal des descripteurs moléculaires qui assure une bonne prédiction de l'activité antibiotique, on doit passer par les mêmes étapes de la prédiction de l'activité anti-inflammatoire tout en utilisant la méthode statistique d'élimination progressive tels les tests de corrélation sont assurés par STATISTICA 9 ®.

La figure IV.2.15 représente les résultats trouvés en terme d'erreur relative absolue moyenne (ERAM%), après la réalisation de la méthode d'élimination progressive suivie d'une modélisation neuronale.

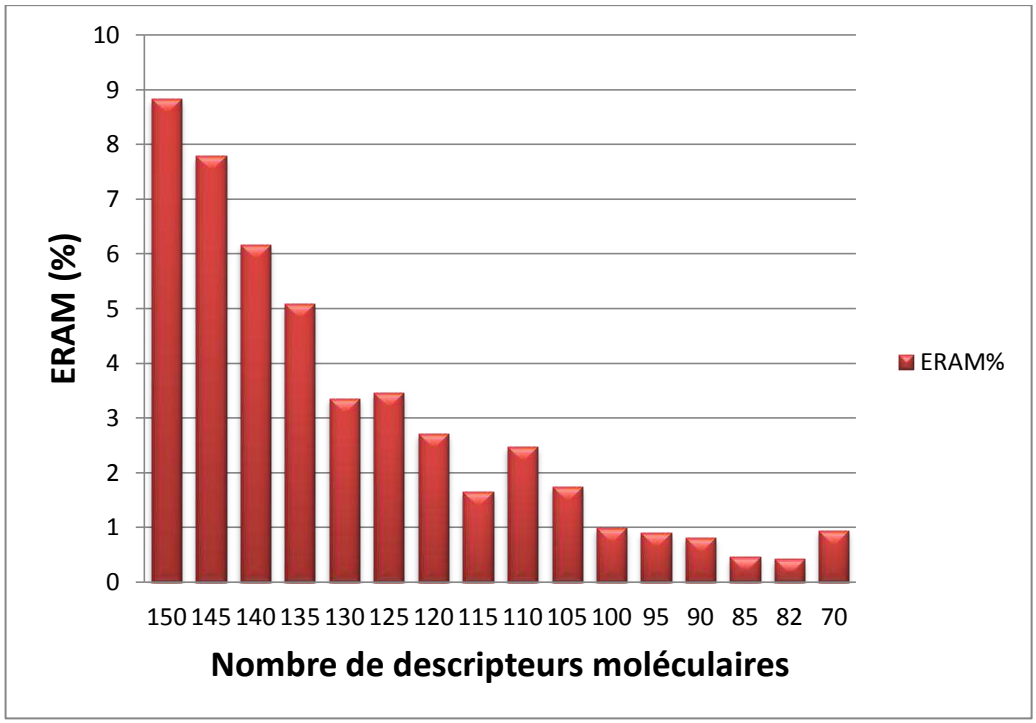



**Figure IV.2.15:** Effet de l'élimination progressive sur l'erreur relative absolue moyenne

### IV.2.5.4 Modélisation neuronale

#### IV.2.5.4.1 Type de réseau

Dans cette partie d'optimisation et par la même procédure précédente, on a sélectionné le réseau qui donne les bons résultats en terme d'erreur minimale et par conséquent, le réseau de type feedforward, fonction newff de MATLAB®, a été notre choix.

### IV.2.5.4.2. Nombre de couches cachées

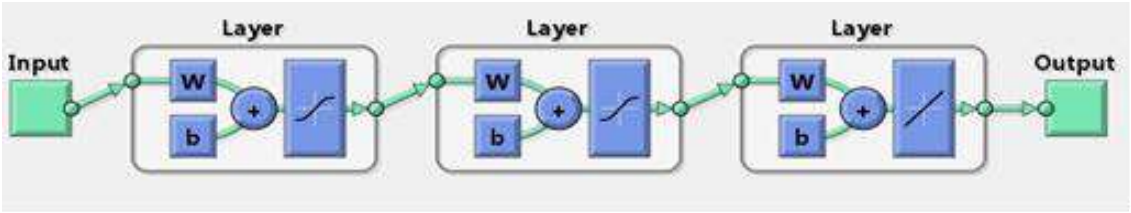
Le nombre de couches final obtenues par la stratégie d'essai et erreur, le critère d'arrêt est la capacité de généralisation. Cette dernière est évaluée en calculant un vecteur du meilleur ajustement linéaire [α, β, R] de la sortie calculée par le RNA en fonction de la sortie cible pour un ensemble de données non utilisée dans la page d'apprentissage. Le nombre final de couches cachées obtenu dans ce travail par cette stratégie est deux.

### IV.2.5.4.3. Algorithme d'apprentissage et la fonction d'activation

L'algorithme d'apprentissage utilisé est de Levenberg-Marquardt (la fonction trainlm de MATLAB®) avec l'option d'apprentissage atténué pour éviter le sur-apprentissage. La procédure commence par deux fonctions de transfert sigmoïde tangent hyperbolique (tansig) pour les couches cachées et la fonction linéaire (fonction purelin de MATLAB®) pour la couche de sortie. La structure du RNA optimisé est résumée dans le tableau IV.2.8.

**Tableau IV.2.8 :** Topologie du RNA optimisé pour l'activité antibiotique

| **Type de réseau** | **FFNN-BP (Fonction newff de MATLAB)** | |
|---|---|---|
| Couche | Nb. de neurones | Fonction d'activation |
| Couche d'entrée | 82 | - |
| 1ère Couche cachée | 21 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| 2eme Couche cachée | 2 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| Couche de sortie | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| Algorithme d'apprentissage | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |




**82** **21** **2** **1**

## IV.2.5.5. Résultats

Les figures IV.2.16, IV.2.17 présentent les courbes de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibiotique donnée expérimentalement pour les deux processus de test et généralisation.

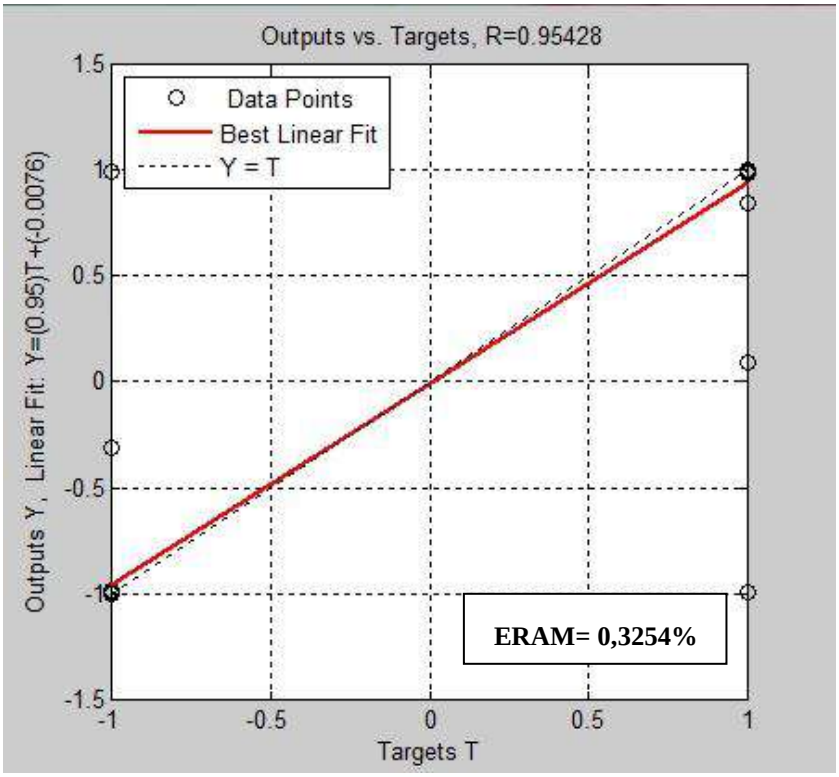



**Figure IV.2.16:** Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibiotique donnée expérimentalement pour le processus de test.

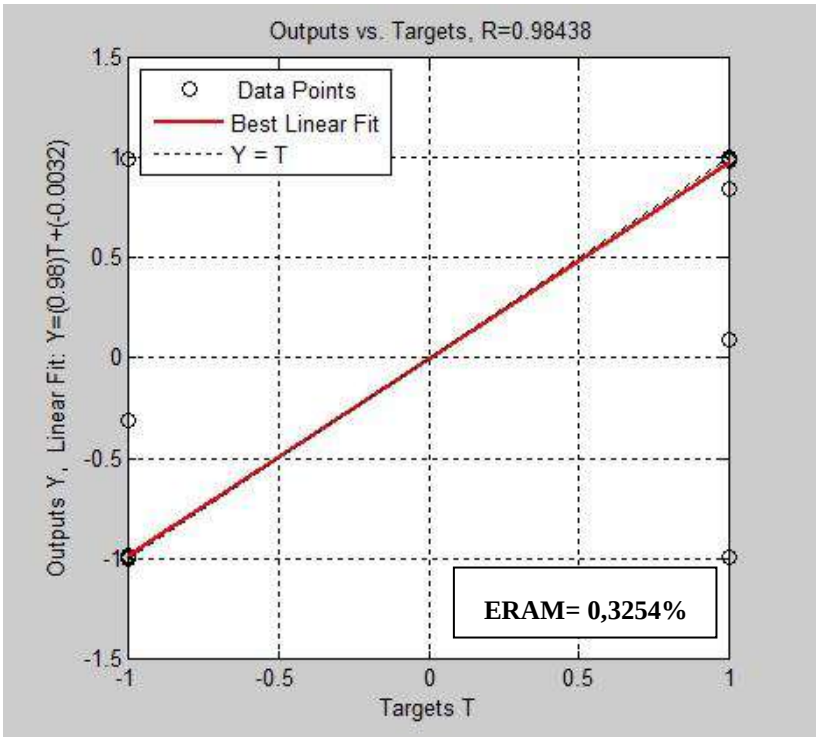



**Figure IV.2.17:** Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibiotique donnée expérimentalement pour le processus de généralisation.

Avec

T : valeur de $Y^{expérimentale}$ ;

Y : valeur de $Y^{calculée}$.

La meilleure concordance linéaire entre les valeurs calculées et les valeurs expérimentales c'est :

$$Y^{calculée} = (0.98)\ (Y^{expérimentale}) + (-0.0032)\ ;\ \text{Avec}\ \ R=0{,}98438.$$

### IV.2.5.6. Analyse des résultats

Les différentes étapes d'optimisations réalisées nous permet d'obtenir de très bons résultats en terme d'erreur relative moyenne absolue (ERAM= 0,3254%) et en terme de facteur de corrélation (98 ,43%), tel que l'analyse statistique des résultats nous permet de dire que l'architecture du réseau optimal peut prédire correctement un nombre de molécules égal à 497 parmi un ensemble total contenant 502 molécules, cependant 5 molécules sont mal prédites.

Donc en terme de pourcentages, on peut dire que le réseau de neurone optimal avec une base de données de 502 molécules et avec un nombre de descripteurs optimisé statistiquement par élimination progressive, permet de donner :

- 99,0039 % des molécules sont correctement prédites ;
- 0,9960 ~ 1% des molécules ne sont pas correctement prédites.

En terme de comparaison, on peut dire que la méthode élaborée peut générer un intervalle de prédiction pour l'activité antibiotique plus important que celui présenté par Miguel et al. (2003), tel qu'ils ont trouvé que 95.31 % des molécules testées sont correctement prédites et 4.69% sont mal prédites pour un ensemble de 64 molécules.

### IV.2.5.7. Méthode d'amélioration des capacités prédictives de l'activité antibiotique

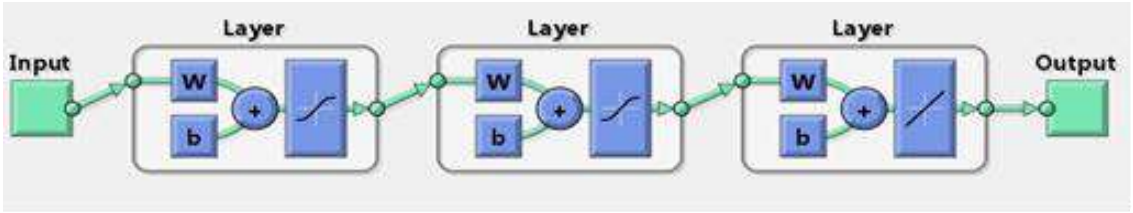
Afin d'améliorer le pourcentage de prédiction obtenu précédemment, et seulement en cas des molécules antibiotiques, on peut améliorer les capacités prédictives du système jusqu'à un pourcentage maximal, car les molécules antibiotiques possèdent un avantage qui réside dans les grandes ressemblances existants entre les structures de ces molécules.

Donc l'amélioration peut être assurée avec une base initiale contenant un nombre réduit de molécules (inférieure à 502 molécules), bien sélectionnées et avec un nombre de descripteurs différents.

Cette deuxième base de données élaborée contient des molécules antibiotiques et non-antibiotiques avec nombre totale égal à 180 molécules. Le nombre de descripteurs moléculaires utilisé dans ce cas est 112 descripteurs, tout en utilisant le même codage des sorties et la même procédure d'optimisation représentée dans l'organigramme de la méthodologie de calcul RNSO (Figure VI.2.3). La structure du RNA optimisé est résumée dans le tableau IV.2.9 :

**Tableau IV.2.9 :** Topologie du RNA optimisé pour l'activité antibiotique après réduction.

| Type de réseau | FFNN-BP (Fonction newff de MATLAB) | |
|---|---|---|
| Couche | Nb. de neurones | Fonction d'activation |
| Couche d'entrée | 112 | - |
| 1ère Couche cachée | 9 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| 2me Couche cachée | 2 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| Couche de sortie | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| Algorithme d'apprentissage | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |




112 9 2 1

Les courbes de performance et de régression linéaire de l'activité antibiotique après réduction de la taille de la base de données sont représentées dans les figures IV.2.18 et figure IV.2.19.

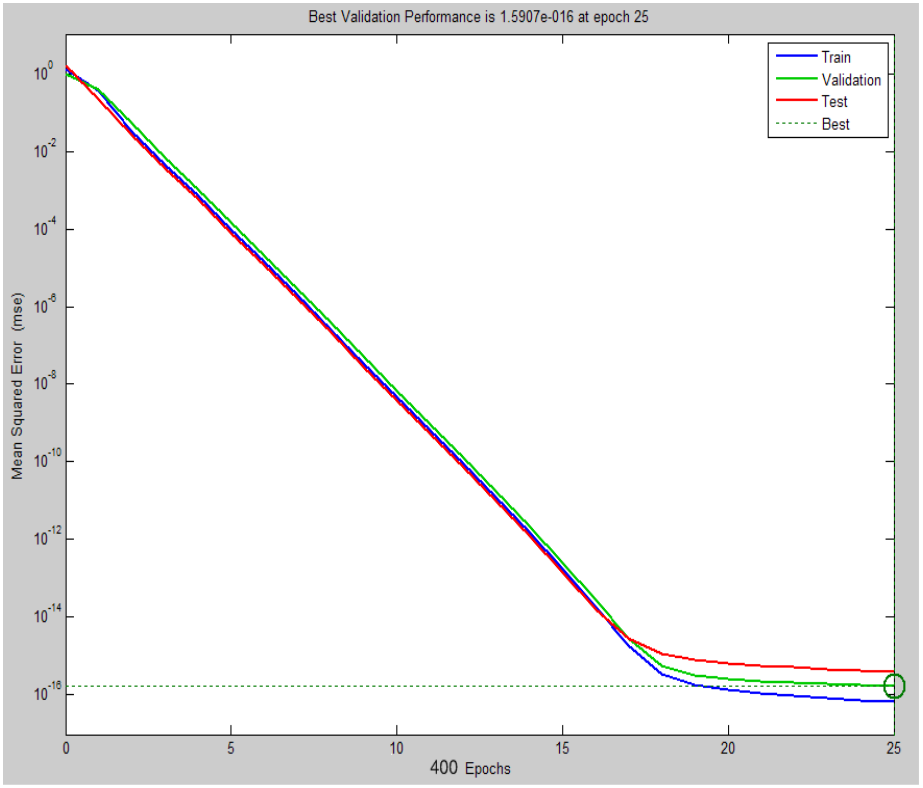



**Figure IV.2.18:** Performance du RNA optimisé.

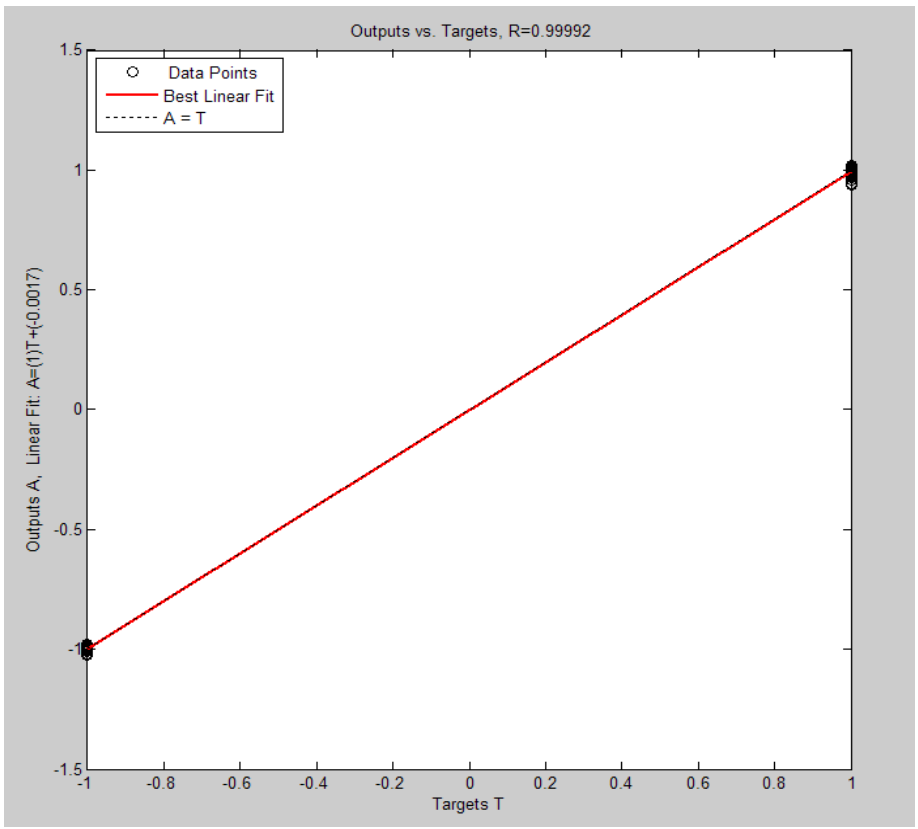



**Figure IV.2.19:** Courbe de régression linéaire de l'activité antibiotique calculée par RNA optimisé avec l'activité antibactérienne donnée expérimentalement après réduction de la base de données.

#### IV.2.5.7.1. Analyse des résultats

L'analyse statistique sous forme de comparaison entre les valeurs expérimentales données et les valeurs calculées par RNA est résumée dans les tableaux (B.1 et B.2 de l'annexe B). Donc on peut dire que la capacité prédictive du deuxième réseau optimisé pour l'activité antibiotique peut être améliorer jusqu' à un intervalle de confiance de l'ordre de 99,99 %, cela peut être assurer dans des nouvelles conditions concernant le type des molécules sélectionnées et la taille de la base de données, ainsi que l'architecture finale du réseau RNA.

## IV.2.6. Optimisation de la prédiction de l'activité analgésique

### IV.2.6.1. Définition d'un analgésique

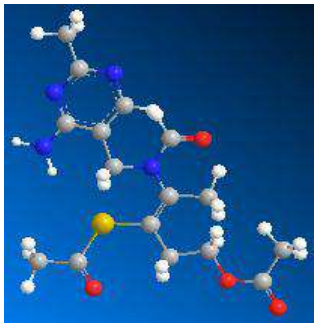
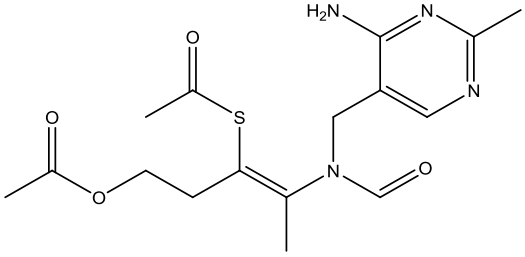
Les analgésiques sont des médicaments utilisés en médecine dans le but d'éliminer la douleur d'un patient (les antalgiques ne faisant qu'atténuer la douleur).

Les besoins en analgésiques varient selon les sujets et, pour un même sujet, selon les circonstances. Un exemple extrême en est celui observé par les médecins opérant en zone de conflit : dans le stress intense du champ de bataille, la quantité d'endorphines générée par le cerveau entraîne une autoanalgésie qui peut être suffisante pour des opérations relativement importantes.

### IV.2.6.2. Liste de molécules analgésiques utilisées

Le tableau IV.2.10 résume toutes les molécules qui possèdent une propriété analgésique utilisées dans la base de données initiale.

**Tableau IV.2.10 :** Molécules analgésiques utilisées dans la base de données initiale [(Kleemann et Engel (2001)].

| **Listes des analgésiques** | | |
|---|---|---|
| **Acetiamine**<br><br><br>ethanethioic acid S-[1-[2-(acetyloxy)ethyl]-2-[[(4-amino-2-methyl-5-pyrimidinyl)methyl]formylamino]-1-propenyl] ester<br>**CC(S/C(CCOC(C)=O)=C(N(CC1=CN=C(C)N=C1N)C=O)/C)=O** | | |
| Acecarbromal | Aceclofenac | Procaine |
| Acétaminophène | Acetiamine | Propacetamol |
| Acetylsalicylic acid | Actarit | Propiram |
| Alclofenac | Alfentanil | Propgphenazone |
| Alminoprofen | Alphaprodine | Proquazone |
| Aminophenazone | Anileridine | Proxazole |
| Antrafenine | Benorilate | Remifentanil |
| Clometacin | codéine | Salicylamide |
| Dextromoramide | Dextropropoxyphene | Salsalate |
| Diflunisal | Dihydrocodeine | Sufentanil |
| Epirizole | Ethenzamide | Sulindac |
| Ethoheptazine | Ethylmorphine | Suprofen |
| Etodolac | Feclobuzone | Taramadol |
| Felbinac | Fenbufen | Tenoxicam |
| Fenclofenac | Fendosal | Thalidomide |
| Fentanyl | Fenyramidol | Tiaramide |
| Floctafenine | Hydrochlorothiazide | Tilidine |
| Hydroxyethyl salicylate | Indometacin | Tinoridine |
| Ketamine | Ketobemidone | Viminol |
| Ketoprofen | Ketorolac | Zomepirac |
| Lactylphenetidin | Levorphanol | Harmine |
| Lobenzarit | Lonazolac | Pholcodine |
| Meptazinol | Mofebutazone | Pipebuzone |
| Mofezolac | Morphine | Piritramide |
| Naproxen | Nefopam | Phenoperidine |
| Normethadone | Oxametacin | Piketoprofen |
| Paracetamol | Pentazocine | Piperacillin |
| Pethidine | Pentetrazol | Phenazocine |

### IV.2.6.3. Optimisation du nombre de descripteurs

La figure IV.2.20 représente les résultats trouvés en terme d'erreur relative absolue moyenne (ERAM%), après réalisation d'élimination progressive, donc nous remarquons que le nombre optimal de descripteurs qui présente une erreur relative minimale (0,467%) égale à 72 descripteurs moléculaires, ces derniers sont sélectionnés en fonction de leur valeur de corrélation avec la sortie (activité analgésique).

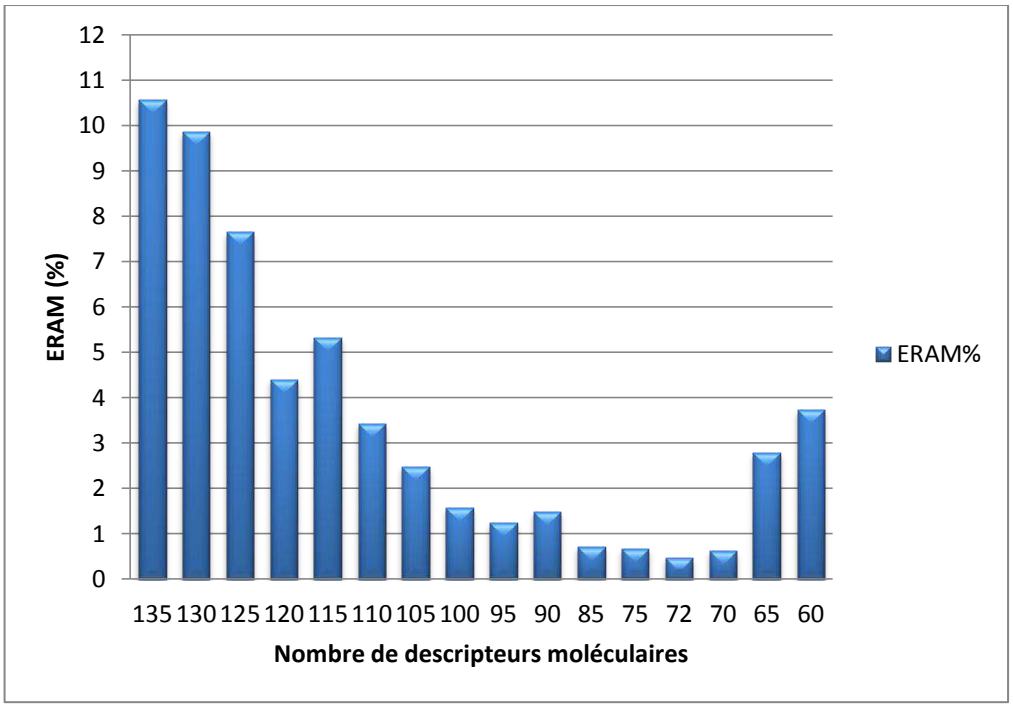


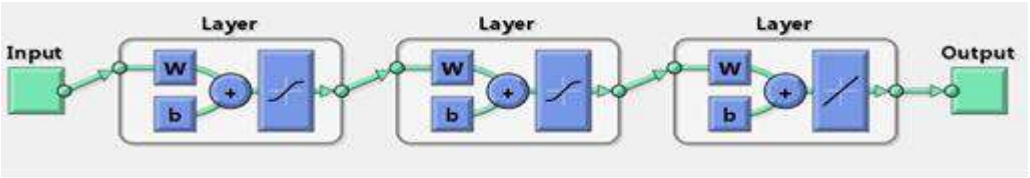

**Figure IV.2.20 :** Effet de l'élimination progressive sur l'erreur relative absolue moyenne

### IV.2.6.4. Modélisation neuronale

De la même manière précédente, on doit déterminer l'architecture du réseau de neurones qui présente les meilleures performances, l'algorithme d'apprentissage utilisé est de Levenberg-Marquardt (la fonction trainlm de MATLAB®). La procédure commence par trois fonctions de transfert sigmoïde tangent hyperbolique (fonction *tansig* de MATLAB®) pour les trois couches cachées et la fonction linéaire (fonction *purelin* de MATLAB®) pour la couche de sortie. La structure du RNA optimisé est résumée dans le tableau IV.2.11

**Tableau IV.2.11 :** Topologie du RNA optimisé pour l'activité analgésique.

| **Type de réseau** | **FFNN-BP (Fonction newff de MATLAB)** | |
|---|---|---|
| **Couche** | Nb. de neurones | Fonction d'activation |
| **Couche d'entrée** | 72 | - |
| **1ère Couche cachée** | 22 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **2me Couche cachée** | 9 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **3me Couche cachée** | 2 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **Couche de sortie** | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| **Algorithme d'apprentissage** | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |




**72** **22** **9** **2** **1**

### IV.2.6.5. Résultats

Les figures IV.2.21, IV.2.22 présentent les courbes de régression linéaire de l'activité analgésique obtenues durant l'élaboration du réseau de neurones, dans les deux phases de test et de généralisation. Nous remarquons que la valeur de l'erreur relative absolue moyenne par rapport aux données utilisées est égale 0.4333% et le facteur de corrélation globale est de l'ordre de 0,95.

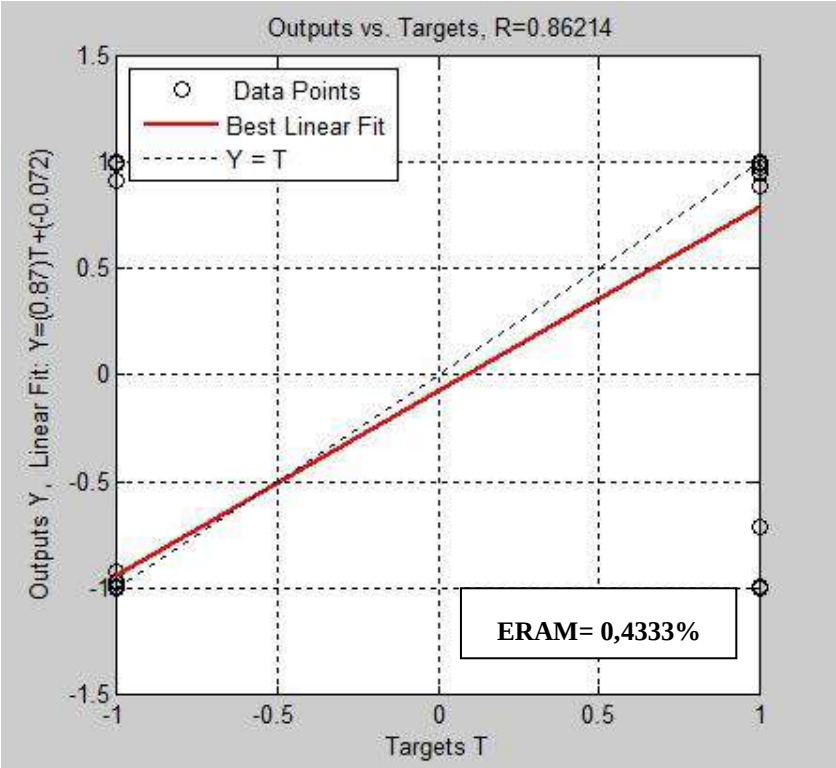



**Figure IV.2.21 :** Courbe de régression linéaire de l'activité analgésique calculée par RNA optimisé avec l'activité analgésique donnée expérimentalement pour le processus de test

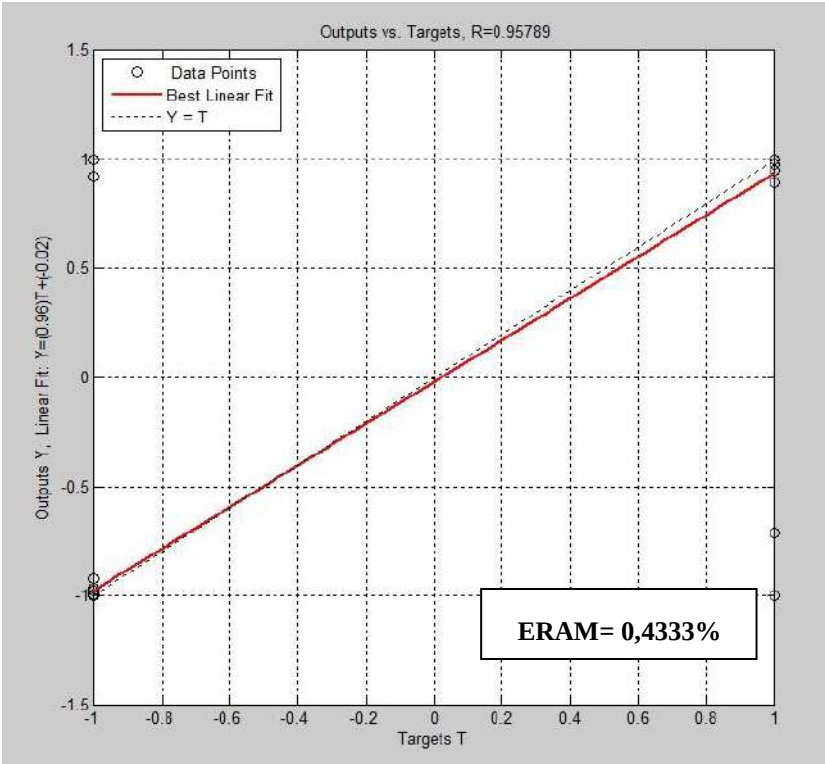



**Figure IV.2.22 :** Courbe de régression linéaire de l'activité analgésique calculée par RNA optimisé avec l'activité analgésique donnée expérimentalement pour le processus de généralisation

La meilleure concordance linéaire entre les valeurs calculées et les valeurs expérimentales c'est :

$$Y^{\text{calculée}} = (0.96)\,(Y^{\text{expérimentale}}) + (0.02)\ ;\ \text{avec } R=0{,}95789.$$

#### IV.2.6.6. Analyse des résultats

Le couplage d'élimination progressive et la modélisation neuronal pour la prédiction de l'activité analgésique permet d'obtenir de trés bons résultats en termes d'erreur relative moyenne absolue (ERAM= 0,4333%) et en terme de facteur de corrélation (95,78%). Les intervalles de confiance pour la prédiction sont données comme suit :

- 98,0079 % des molécules sont correctement prédites ;
- 1,9920 % des molécules ne sont pas correctement prédites.

### IV.2.7. Optimisation de la prédiction de l'activité antivirale

#### IV.2.7.1. Définition d'un antiviral

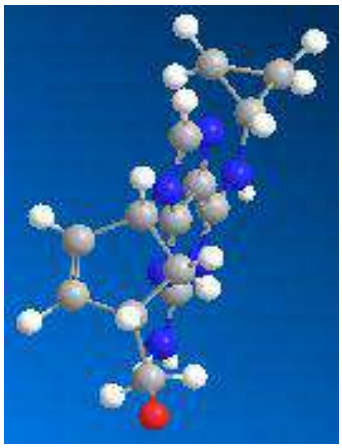
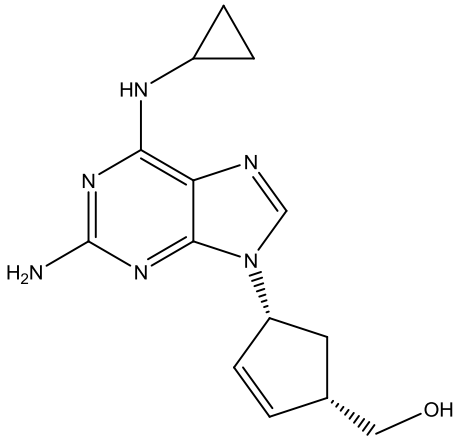
Un antiviral désigne une molécule perturbant le cycle de réplication d'un ou de plusieurs virus, permettant ainsi de ralentir mais rarement d'arrêter une infection virale. C'est avec les vaccins et la prévention, la seule méthode connue permettant de lutter contre les infections d'origines virales.

Les antiviraux présentent une méthode efficace de lutte contre les virus en attendant la mise au point d'un vaccin qui est la seule manière connue d'éradiquer un virus sur le long terme, comme l'ont montré les différentes campagnes d'éradications de la poliomyélite et surtout de la variole. Un antirétroviral désigne une classe particulière d'antiviral destinée à lutter contre les rétrovirus.

#### IV.2.7.2. Liste de molécules antivirales utilisées

Le tableau IV.2.12 résume toutes les molécules qui possèdent une propriété antivirale utilisées dans la base de données initiale

**Tableau IV.2.12 :** Molécules antivirales utilisées dans la base de données initiale [Kleemann et Engel (2001)].

| Liste des antiviraux | | |
|---|---|---|
| Abacavir  |  (1S,4R)-4-[2-amimo-6-(cyclopropylamino)-9H-purin-9-yl]-2-cyclopentene-1-methanol OC[C@@H]1C=C[C@H](N2C=NC3=C(NC4CC4)N=C(N)N=C23)C1 | |
| Abacavir | Aciclovir | Resiquimod |
| Adefovir | Alovudine | Rimantadine |
| Amantadine | Amprenavir | Saquinavir |
| Arbidol | Atazanavir | Sorivudine |
| Azlocillin | Boceprevir | Stavudine |
| Brivudine | Capravirine | Tanomastat |
| Clevudine | Cidofovir | Telbivudine |
| Coumarin | Darunavir | Theaflavin |
| Deflazacort | Delavirdine | Tipranavir |
| Delavirdine mysilate | Dextromethorphane | Tenofovir |
| Didanosine | Docosanol | Valganciclovir |
| Edoxudine | Efavirenz | Vidarabine |
| Elvitégravir | Emtricitabine | Zalcitabine |
| Entecavir | Étravirine | Zidovudine |
| Famciclovir | Fosamprenavir | Ribavirin |
| Fialuridine | Foscarnet | Rilpivirine |
| Ganciclovir | Gemcitabine | Ritonavir |
| Genistein | Guaiazulene | Stallimycin |
| Ibacitabine | Idoxuridine | Talampicillin |
| Indinavir | Inosine | Taribavirin |
| Lamivudine | Lopinavir | Terpenoid |
| Loviride | Maraviroc | Tilorone |
| Maribavir | Methisazone | Trifluridine |
| Moroxydine | Nelfinavir | Valaciclovir |
| Nevirapine | Oseltamivir | Vicriviroc |
| Penciclovir | Podophyllotoxin | Warfarin |
| Pleconaril | Raltegravir | Zanamivir |

### IV.2.7.3. Optimisation du nombre de descripteurs

La figure IV.2.23 représente les résultats trouvés en terme d'erreur relative absolue moyenne (ERAM%), après la réalisation de l'élimination progressive suivie d'une modélisation neuronale pour l'activité antivirale. Ces résultats montrent que le nombre de descripteurs qui assure une ERAM minimale (à 0,112%) est 76 descripteurs moléculaires.

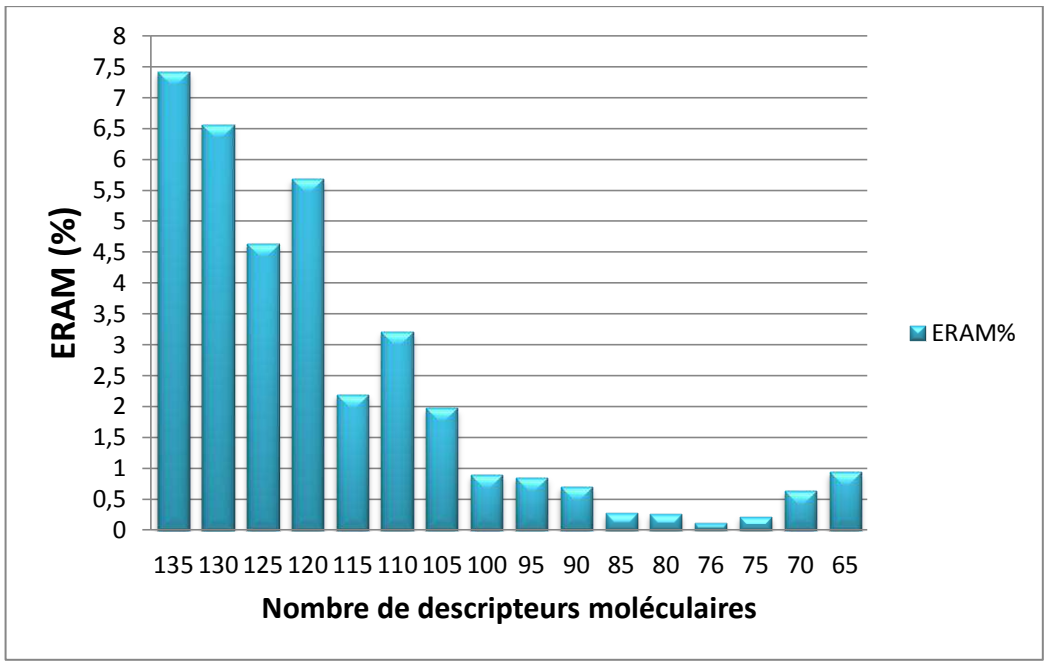


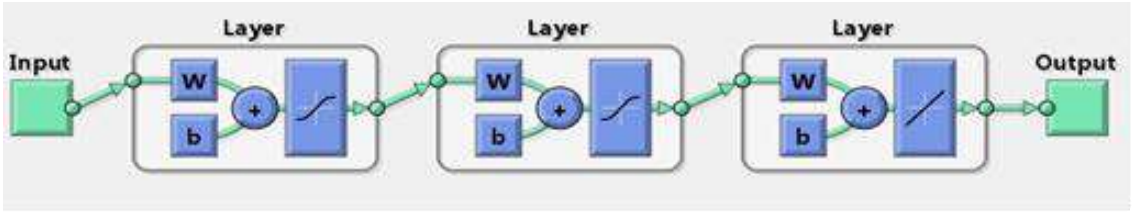

**Figure IV.2.23 :** Effet de l'élimination progressive sur l'erreur relative absolue moyenne

### IV.2.7.4. Modélisation neuronale

L'architecture du réseau de neurones qui présente les meilleures performances, l'algorithme d'apprentissage utilisé est de Levenberg-Marquardt (la fonction trainlm de MATLAB®). La procédure commence par deux fonctions de transfert sigmoïde tangent hyperbolique (fonction *tansig* de MATLAB®) pour les trois couches cachées et la fonction linéaire (fonction *purelin* de MATLAB®) pour la couche de sortie. La structure du RNA optimisé est résumée dans le tableau IV.2.13.

**Tableau IV.2.13 :** Topologie du RNA optimisé pour l'activité antivirale

| **Type de réseau** | **FFNN-BP (Fonction newff de MATLAB)** | |
|---|---|---|
| **Couche** | Nb. de neurones | Fonction d'activation |
| **Couche d'entrée** | 76 | - |
| **1ere Couche cachée** | 21 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **2eme Couche cachée** | 18 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **3eme Couche cachée** | 2 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| **Couche de sortie** | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| **Algorithme d'apprentissage** | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |




**72** **21** **18** **2** **1**

### IV.2.7.5. Résultats

Les figures IV.2.24, IV.2.25 présentent les courbes de régression linéaire de l'activité antivirale obtenues durant l'élaboration du réseau de neurones, dans les deux phases de test et de généralisation.

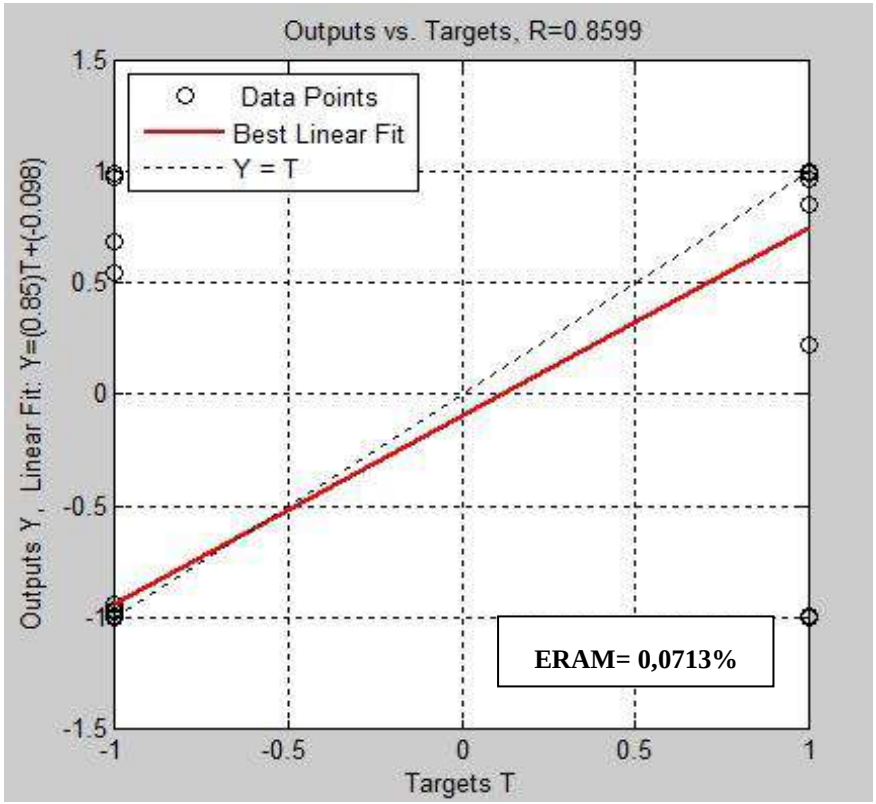



**Figure IV.2.24 :** Courbe de régression linéaire de l'activité antivirale calculée par RNA optimisé avec l'activité antivirale donnée expérimentalement pour le processus de test

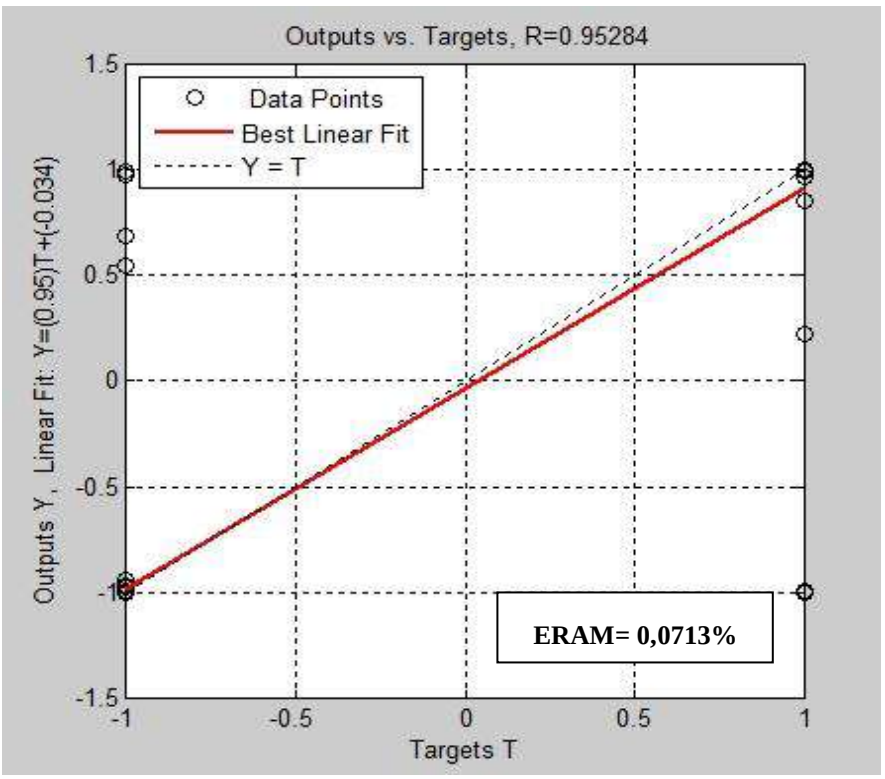



**Figure IV.2.25 :** Courbes de régression linéaire de l'activité antivirale calculée par RNA optimisé avec l'activité antivirale donnée expérimentalement pour le processus de généralisation

La meilleure concordance linéaire entre les valeurs calculées et les valeurs expérimentales : $Y^{\text{calculée}} = (0.95)\ (Y^{\text{expérimentale}}) + (-0.034)$ ; avec R=0,95284.

#### IV.2.7.6. Analyse des résultats

Les différentes étapes d'optimisations réalisées nous permettent d'obtenir de très bons résultats en terme d'erreur relative moyenne absolue (ERAM= 0,0713%) et en terme de facteur de corrélation, et en termes de pourcentage on peut dire que le réseau de neurone optimal avec une base de données de 502 molécules et avec un nombre de descripteurs optimisé statistiquement par élimination progressive, permet de donner :

- 97,8087 % des molécules sont correctement prédites ;
- 2,1912 % des molécules ne sont pas correctement prédites.

Donc on peut considérer que le réseau de neurones optimisé peut générer des forts pourcentages de prédiction de l'activité antivirale.

## IV.2.8. Optimisation de la prédiction de l'activité diurétique

### IV.2.8.1. Définition d'un diurétique

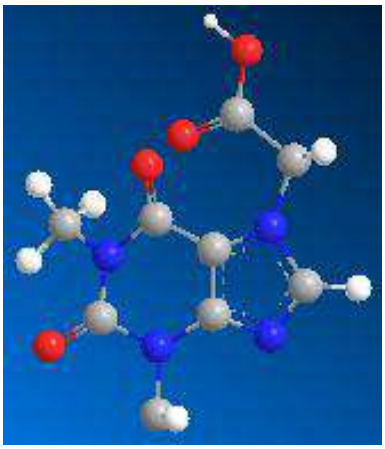
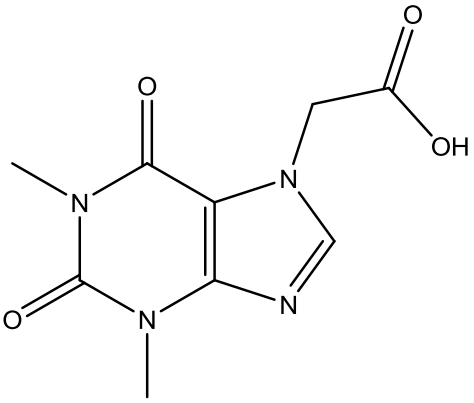
Un diurétique est une substance qui entraine une augmentation de la sécrétion urinaire et qui peut être utilisé notamment pour traiter l'hypertension artérielle, l'insuffisance cardiaque, certains œdèmes, l'hypertension portale ou l'hypokaliémie.

- Certaines sont utilisées comme médicaments : elles inhibent la réabsorption rénale des ions sodium et entrainent une augmentation de l'élimination urinaire en eau et en sodium.
- Les autres sont qualifiées de substances à effet diurétique :
  - ✓ Théobromine
  - ✓ Théophylline
  - ✓ Caféine
  - ✓ Alcool

### IV.2.8.2. Liste de molécules diurétiques utilisées

Le tableau IV.2.14 résume toutes les molécules qui possèdent une propriété diurétique utilisées dans la base de données initiale.

**Tableau IV.2.14 :** Molécules diurétiques utilisées dans la base de données initiale [Kleemann et Engel (2001)].

<table>
<tr><th colspan="3">Liste des molécules diurétiques</th></tr>
<tr><td>Acefylline<br></td><td colspan="2"><br>1,2,3,6-tetrahydro-1,3-dimethyl-2,6-dioxo-7H-purine-7-acetic acid<br>O=C(O)CN1C=NC(N(C)C(N2C)=O)=C1C2=O</td></tr>
<tr><td></td><td>Acetazolamide</td><td>Piretanide</td></tr>
<tr><td>Altizide</td><td>Ambuside</td><td>Polythiazide</td></tr>
<tr><td>Azosemide</td><td>Bemetizide</td><td>Protheobromine</td></tr>
<tr><td>Bumetanide</td><td>Butizide</td><td>Quinethazone</td></tr>
<tr><td>Caffeine</td><td>Carzenide</td><td>Spironolactone</td></tr>
<tr><td>Chlorazanil</td><td>Chlormerodrin</td><td>Triamterene</td></tr>
<tr><td>Chlorothiazide</td><td>Chlortalidone</td><td>Theobromine</td></tr>
<tr><td>Cicletanine</td><td>Clopamide</td><td>Theodrenaline</td></tr>
<tr><td>Clorexolone</td><td>Cyclopenthiazide</td><td>Theophylline</td></tr>
<tr><td>Cyclothiazide</td><td>Dopamine</td><td>Tienilic acid</td></tr>
<tr><td>Dorzolamide</td><td>Epitizide</td><td>Torasemide</td></tr>
<tr><td>Etacrynic acid</td><td>Etamiphylline</td><td>Triamterene</td></tr>
<tr><td>Ethiazide</td><td>Ethoxzolamide</td><td>Tkichlormethiazide</td></tr>
<tr><td>Etozolin</td><td>Fenquizone</td><td>Tripamide</td></tr>
<tr><td>Furosémide</td><td>Hydrochlorothiazide</td><td>Trometamol</td></tr>
<tr><td>Indapamide</td><td>D-Mannitol</td><td>Xipamide</td></tr>
<tr><td>Mefruside</td><td>Methazolamide</td><td>Muzolimine</td></tr>
<tr><td>Methyclothiazide</td><td>Meticrane</td><td>Paraflutizide</td></tr>
</table>

### IV.2.8.3. Optimisation du nombre de descripteurs

La figure IV.2.26 représente les résultats trouvés en terme d'erreur relative absolue moyenne (ERAM%). L'histogramme montre clairement que le nombre 69 peut être considéré comme un nombre optimal car il correspond à l'erreur minimale (0,3145%).

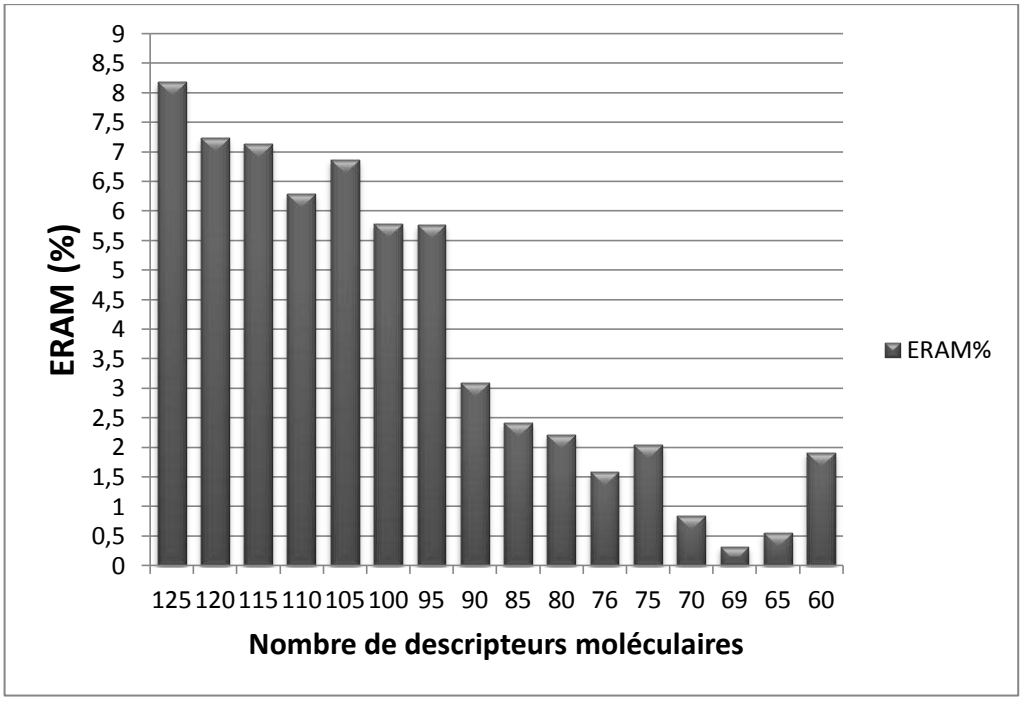
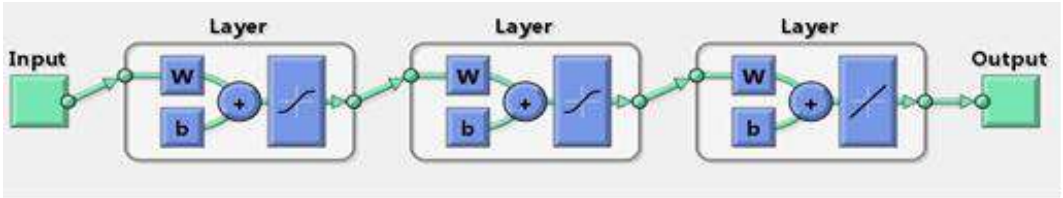



**Figure IV.2.26 :** Effet de l'élimination progressive sur l'erreur relative absolue moyenne

### IV.2.8.4. Modélisation neuronale

Le tableau IV.2.15 représente la structure du réseau optimisé pour la prédiction de l'activité diurétique.

**Tableau IV.2.15 :** Topologie du RNA optimisé pour l'activité diurétique.

| Type de réseau | FFNN-BP (Fonction newff de MATLAB) | |
|---|---|---|
| Couche | Nb. de neurones | Fonction d'activation |
| Couche d'entrée | 69 | - |
| 1ere Couche cachée | 29 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| 2me Couche cachée | 17 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| 3me Couche cachée | 2 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| Couche de sortie | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| Algorithme d'apprentissage | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |
|  | | |

Les résultats obtenus durant l'élaboration du réseau de neurones pour les phases de test et de généralisation sont illustrés dans les figures IV.2.27 et IV.2.28.

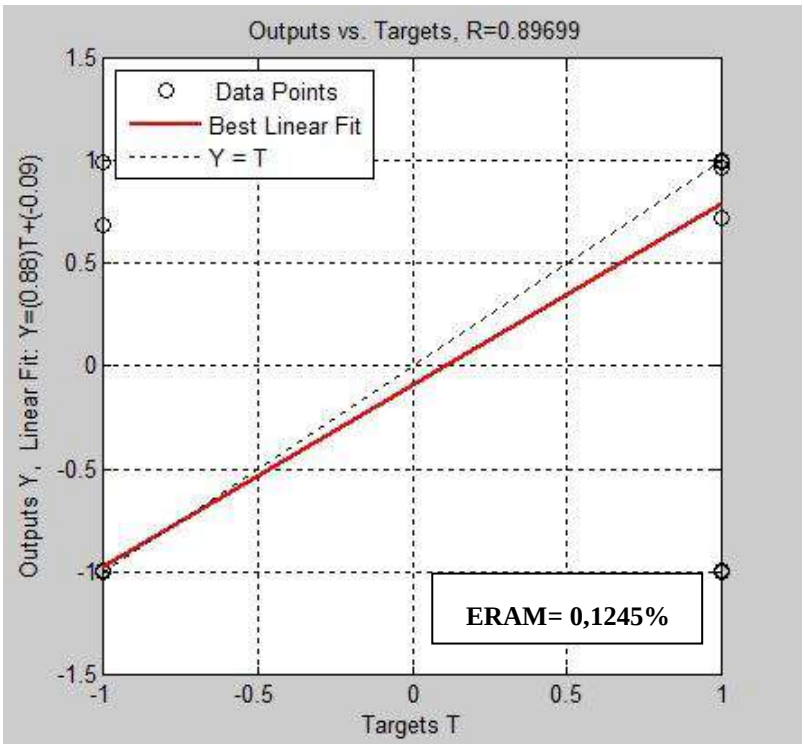



**Figure IV.2.27 :** Courbe de régression linéaire de l'activité diurétique calculée par RNA optimisé avec l'activité diurétique donnée expérimentalement pour le processus de test

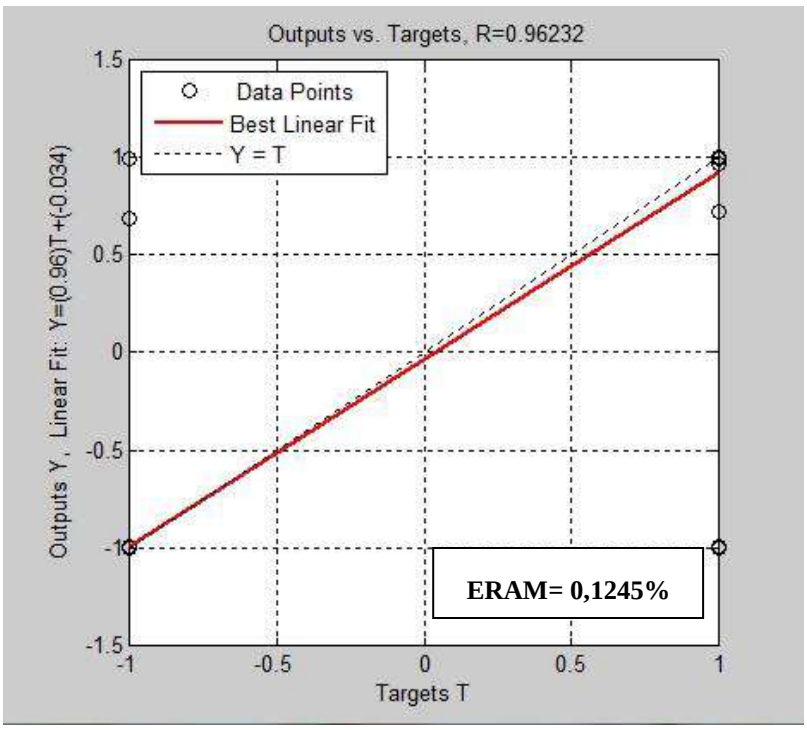



**Figure IV.2.28 :** Courbe de régression linéaire de l'activité diurétique calculée par RNA optimisé avec l'activité diurétique donnée expérimentalement pour le processus de généralisation

Avec une concordance linéaire entre les valeurs calculées et les valeurs expérimentales exprimée comme suit :

$Y^{\text{calculée}} = (0.96)\ (Y^{\text{expérimentale}}) + (-0.034)$ ; avec R=0,96232.

### IV.2.8.5. Analyse des résultats

Les différentes étapes d'optimisations réalisées nous permettent d'obtenir de bons résultats en termes d'erreur relative moyenne absolue (ERAM= 0,1245%) et en terme de facteur de corrélation, tel que l'analyse statistique des résultats nous permet de dire que l'architecture du réseau optimal peut prédire correctement un nombre de molécules égale 488 parmi un ensemble total contenant 502 molécules, cependant 14 molécules sont mal prédites.

Donc en termes de pourcentage on peut dire que le réseau de neurones optimal avec une base de données de 502 molécules et avec un nombre de descripteurs optimisé statistiquement par élimination progressive, permet de donner :

- 98,8804 % des molécules sont correctement prédites ;
- 1,1952% des molécules ne sont pas correctement prédites.

# IV.2.9. Optimisation de la prédiction de l'activité glucocorticoïde

## IV.2.9.1. Définition d'un glucocorticoïde

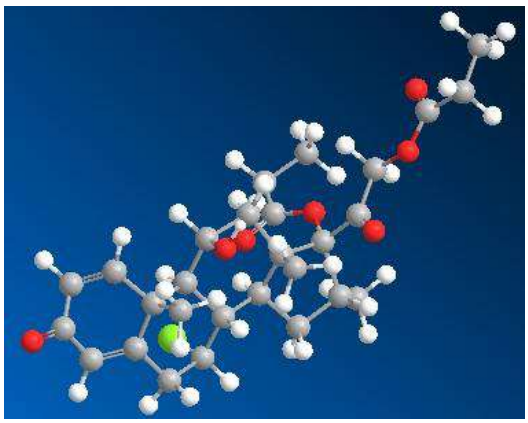
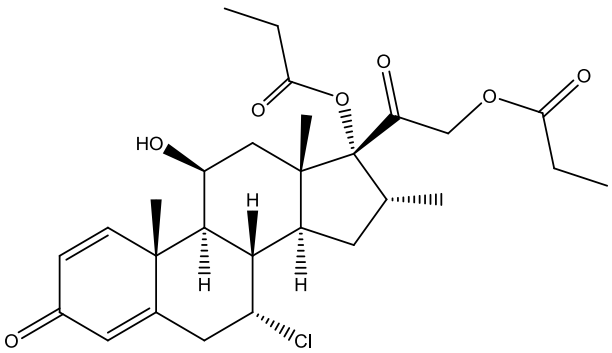
Les glucocorticoïdes sont des corticostéroïdes qui ont une action sur le métabolisme protidique et glucidique ;et qui sont des hormones stéroïdiennes naturelles sécrétées chez les êtres humains par le cortex de la glande surrénale

- glucocorticoïdes naturels (cortisone)
- glucocorticoïdes de synthèse : à effets courts (prednisone), à effets intermédiaires (paraméthasone), à effets prolongés (bétaméthasone)

## IV.2.9.2. Liste de molécules glucocorticoïdes utilisées

Le tableau IV.2.16 résume toutes les molécules qui possèdent une propriété glucocorticoïde utilisées dans la base de données initiale.

**Tableau IV.2.16 :** Molécules glucocorticoïdes utilisées dans la base de données initiale [Kleemann et Engel (2001)].

| **Liste des molécules glucocorticoïdes** | | |
|---|---|---|
| **Alclometasone dipropionate**  |  | |
| Alclometasone dipropionate | Chloroprednisone acetate | Dexamethasone 21-linolate |
| Amcinonide | Ciclometasone | Dexamethasone phosphate |
| Beclometasone | Clobetasol propionate | Dexamethasone pivalate |
| Bendacort | Clobetasone butyrate | Dexamethasone valerate |
| Betamethasone benzoate | Clocortolone | Dichlorisone |
| Betamethasone dipropionate | Cloprednol | Diorasone diacetate |
| Betamethasone divalerate | Cortisone | Difiucortolone valerate |
| Betamethasone phosphate | Desonide | Difluprednate |
| Betamethasone valerate | Desoximetasone | Dexamethasone tert-butylacetate |
| | Budesonide | Dexamethasone 21sonicotinate |

### IV.2.9.3. Optimisation du nombre de descripteurs

La figure IV.2.29 représente les résultats trouvés en terme d'erreur relative absolue moyenne (ERAM%), après la réalisation de l'élimination progressive, ces résultats montrent que le nombre de descripteurs optimal pour la prédiction de l'activité glucocorticoïde qui présente une ERAM minimale égal à 110 descripteurs moléculaires.

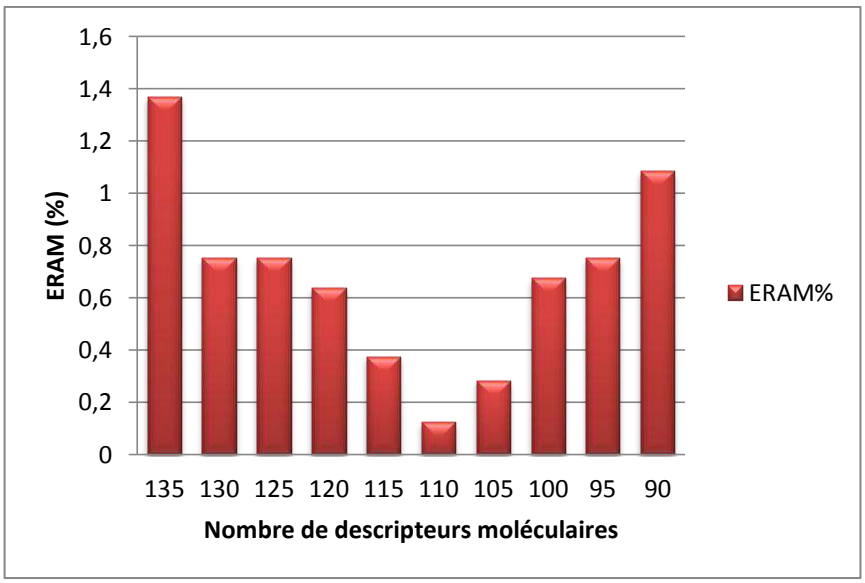


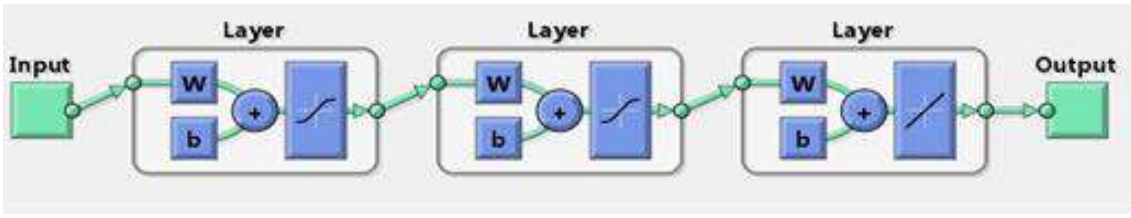

**Figure IV.2.29 :** Effet de l'élimination progressive sur l'erreur relative absolue moyenne

### IV.2.9.4. Modélisation neuronale

Le tableau IV.2.17 représente la structure du réseau optimisé pour la prédiction de l'activité glucocorticoïde.

**Tableau IV.2.17 :** Topologie du RNA optimisé pour l'activité glucocorticoïde.

| **Type de réseau** | **FFNN-BP (Fonction newff de MATLAB)** | |
|---|---|---|
| Couche | Nb. de neurones | Fonction d'activation |
| Couche d'entrée | 110 | - |
| 1ère Couche cachée | 18 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| 2eme Couche cachée | 9 | Sigmoïde Tangente Hyperbolique (Fonction tansig de MATLAB®) |
| Couche de sortie | 1 | Linéaire (Fonction *purelin* de MATLAB®) |
| Algorithme d'apprentissage | Rétropropagation de Levenberg-Marquardt. (Fonction *trainlm* de MATLAB®) | |



### IV.2.9.5. Résultats

Les résultats obtenus durant l'élaboration du réseau de neurones pour les phases de test et de généralisation sont illustrés dans les figures IV.2.30 et IV.2.31.

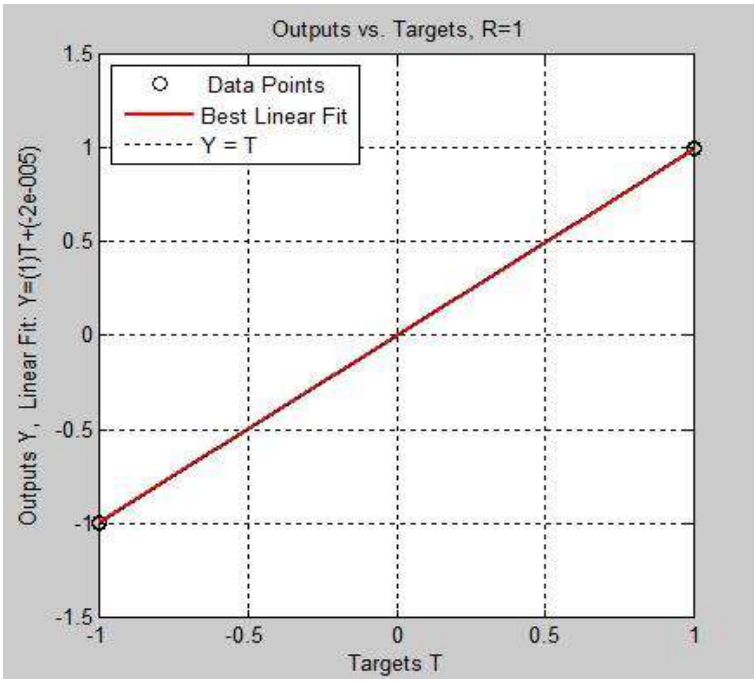



**Figure IV.2.30 :** Courbe de régression linéaire de l'activité glucocorticoïde calculée par RNA optimisé avec l'activité glucocorticoïde donnée expérimentalement pour le processus de test

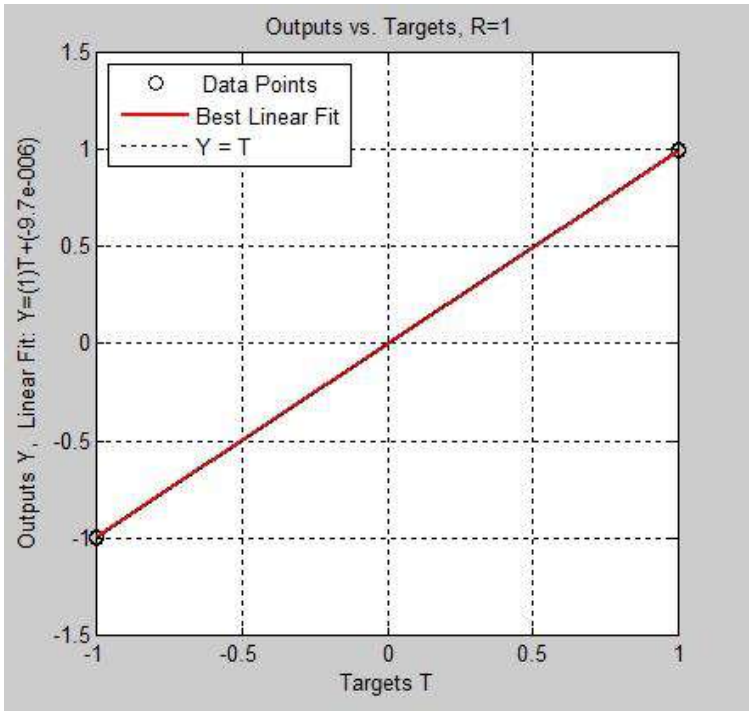



**Figure IV.2.31:** Courbe de régression linéaire de l'activité glucocorticoïde calculée par RNA optimisé avec l'activité glucocorticoïde donnée expérimentalement pour le processus de généralisation

La meilleure concordance linéaire entre les valeurs calculées et les valeurs expérimentales c'est :

$Y^{\text{calculée}} = (1)\ (Y^{\text{expérimentale}}) + (-9{,}7\ \text{e-}06)$ ; avec R=1.

Avec ERAM= (2,329 e -4)%.

### IV.2.9.5. Analyse des résultats

Le comportement du modèle neuronale vis-à-vis de la dernière base construite pour la prédiction de l'activité glucocorticoïde, illustré dans les figures IV.2.30 et IV.2.31 montre que les valeurs prédites par le modèle neuronal optimisé sont en excellent accord avec les valeurs expérimentales.

Donc en terme de pourcentage, on peut dire que le réseau de neurones optimal avec une base de données de 518 molécules et avec un nombre de descripteurs optimisé statistiquement par élimination progressive, permet de donner :

- 100 % des molécules sont correctement prédites ;

La représentation schématique du perceptron du réseau élaboré est donnée par la figure IV.2.32 et les équations IV.2.1, IV.2.2, IV.2.3.

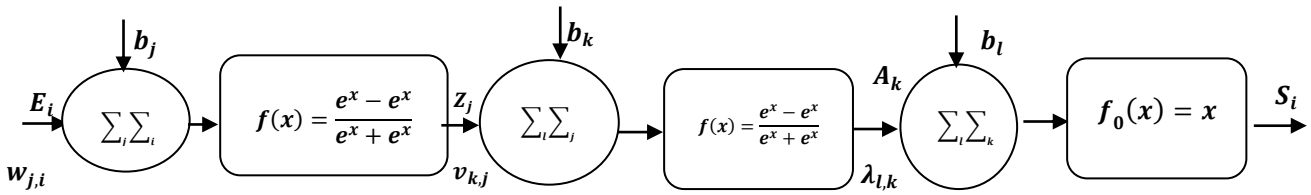



**Figure IV.2.32:** Représentation schématique du réseau de neurones multicouches pour la prédiction de l'activité des glucocorticoïdes

$$S_i = f_0\left[\sum_{k=1}^{9} A_k \lambda_{lk} + b_l\right] = \sum_{k=1}^{9} A_k \lambda_{lk} + b_l \qquad \text{IV.2.1}$$

$$A_k = f\left[\sum_{j=1}^{18} v_{kj} Z_j + b_k\right] = \frac{exp\left(\sum_{j=1}^{18} v_{kj} Z_j + b_k\right) - exp\left(-\sum_{j=1}^{18} v_{kj} Z_j + b_k\right)}{exp\left(\sum_{j=1}^{18} v_{kj} Z_j + b_k\right) + exp\left(-\sum_{j=1}^{18} v_{kj} Z_j + b_k\right)} \qquad \text{IV.2.2}$$

$$Z_j = f\left[\sum_{j=1}^{110} w_{ij} E_i + b_j\right] = \frac{exp\left(\sum_{j=1}^{110} w_{ij} E_i + b_j\right) - exp\left(-\sum_{j=1}^{110} w_{ij} E_i + b_j\right)}{exp\left(\sum_{j=1}^{110} w_{ij} E_i + b_j\right) + exp\left(-\sum_{j=1}^{110} w_{ij} E_i + b_j\right)} \qquad \text{IV.2.3}$$

### IV.2.9.6. Test de performance de prédiction du modèle construit

Dans le but de mesurer les performances du modèle neuronal conçu pour la prédiction de l'activité des glucocorticoïdes, nous avons généré une base de données constituée par des molécules possédant des valeurs de poids moléculaire supérieures à celles de la base testée initialement. Les valeurs des poids moléculaires des molécules additionnées sont comprises dans un intervalle de [627,74 ; 1007,39], tel que la base initiale est caractérisée par une valeur maximale égale à 615,73 (UI). Le tableau II.2.18 représente les valeurs prédites et les valeurs expérimentales pour les molécules additionnées pour tester les performances du modèle neuronal élaboré. Si on compare les valeurs prédites avec les valeurs expérimentales on peut remarquer clairement que la majorité des activités des molécules testées sont correctement prédites, donc on peut conclure que le modèle neuronal conçu est performant.

**Tableau IV.2.18 :** Résultats du calcul par le modèle RNA conçu pour la prédiction de l'activité des glucocorticoïdes.

| | Céfépime | Cefoperazone | Céfotiam héxétil | Dalfopristine | Enviomycin |
|---|---|---|---|---|---|
| Valeurs prévues | -0,99982156 | -1,000000301 | -1,0000006 | -1,0000038 | -1,00032 |
| valeurs expérimentales | -1,00000 | -1,00000 | -1,00000 | -1,00000 | -1,00000 |
| | Erythromycin | Midecamycin acetate | Proglumetacin | quinupristine | Rifampicin |
| Valeurs prévues | -0,99999 | -1,00012342 | -1,0039662 | -1,00010023 | -1,00001298 |
| valeurs expérimentales | -1,00000 | -1,00000 | -1,00000 | -1,00000 | -1,00000 |
| | Rimantadine | Télithromycin | Viomycin | Virginiamycin | |
| Valeurs prévues | -1,0000162 | -1,000007484 | -0,615025897 | 0,99261 | |
| valeurs expérimentales | -1,00000 | -1,00000 | -1,00000 | -1,00000 | |

Toujours pour cette activité, on a essayé de refaire les mêmes étapes de modélisation neuronale précédente et avec la même architecture du réseau mais avec une base de données de taille plus importante tout en gardant le nombre de descripteurs optimal, (les figure B.1 et B.2 de l'annexe B représentent les courbes de régression linéaires). On a remarqué que les valeurs prédites par le modèle neuronale sont en excellent accord avec les valeurs expérimentales avec un coefficient de régression égale à 99,9%, et une erreur ERAM de l'ordre de 0,036%, donc selon [Lee et al. (1991)] les résultats trouvés nous permettent de conclure que le modèle neuronale garde toujours ses capacités de prédiction pour l'activité glucocorticoïde.

## IV.2.10. Conclusion

Dans cette partie de notre étude, on est arrivé à établir une architecture neuronale convenable permettant de prédire l'activité thérapeutique d'une molécule donnée ; cette capacité de prédiction peut être évaluer en réalisant un traitement préliminaire sur la base de données initiale à l'aide d'une méthode statistique « élimination progressive », cette dernière nous permet de déterminer le nombre optimale des variables d'entrée pour la modélisation neuronale qui assure une erreur ( REAM) minimale. Les différentes études d'optimisations réalisées sur les six activités thérapeutiques nous permettent d'évaluer des intervalles de confiance pour la capacité de prédiction des modèles neuronaux conçus pour chaque activité thérapeutique donc on a obtenu :

- Pour l'activité anti-inflammatoire, on a trouvé les pourcentages suivants :
  - ✓ 97,5563 % des molécules sont correctement prédites (ou bien se rapprochent de la valeur désirée);
  - ✓ 2,4436 % des molécules ne sont pas correctement prédites.
- Pour l'activité analgésique :
  - ✓ 98,0079 % des molécules sont correctement prédites ;
  - ✓ 1,9920 % des molécules ne sont pas correctement prédites.
- Pour l'activité antivirale :
  - ✓ 97,8087 % des molécules sont correctement prédites ;
  - ✓ 2,1912 % des molécules ne sont pas correctement prédites.

- Et pour l'activité diurétique on a obtenu :
    - ✓ 98,8804 % des molécules sont correctement prédites ;
    - ✓ 1,1952% des molécules ne sont pas correctement prédites.

On note ici que les meilleures performances de prédiction ont été trouvées dans le cas des antibiotiques et glucocorticoïdes dont on signale des excellentes capacités de prédiction pour ces deux familles thérapeutiques, cela peut être interprété par la forte ressemblance qui existe entre les molécules de chaque famille thérapeutique. On a trouvé les pourcentages suivants :

- 99,0039 % des molécules antibiotiques sont correctement prédites, ce pourcentage révèle d'être plus important que celui obtenu par Miguel et al. (2003), tel qu'ils ont trouvé que 95.31 % des molécules testées sont correctement prédites pour un ensemble de 64 molécules;
- Toutes les molécules glucocorticoïdes ont été prédites d'une manière correcte (le pourcentage de prédiction dans ce cas est 100%).

Ces intervalles représentés par des pourcentages sont fortement liés au plusieurs facteurs tel que :

1. La taille de la base de données
2. La nature des molécules sélectionnées dans la base de données ;
3. La méthode statistique de sélection des descripteurs les plus pertinents ;
4. Le nombre de ces descripteurs
5. Et finalement les propriétés du réseau de neurones optimisé.

# *Partie 3*

# *Analyse en Composante Principale (ACP)*

# *&*

# *Régression Linéaire Multiple (RLM)*

## IV.3.1. Introduction

Un grand nombre de descripteurs différents sont collectés pour la modélisation de l'activité thérapeutique, 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. 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 sur-ajusté à ces exemples, et incapable de prédire la grandeur modélisée sur de nouvelles observations.

Il est donc nécessaire de réduire la dimensionnalité des variables d'entrée. Dans cette partie de notre étude on a essayé de résoudre ce problème en utilisant la méthode d'analyse en composante principale (ACP). Cette méthode nous permet d'obtenir une dimension plus réduite des variables ce qui facilite l'étape de prédiction des différentes activités thérapeutiques étudiées dans la deuxième partie de notre étude, mais cette fois elle est réalisée par un autre type de modélisation et plus simple que le mode utilisé précédemment (RNA), donc on va procéder à une régression linéaire multiple (RLM).

## IV.3.2. Présentation de la méthode d'analyse en composante principale

Cette méthode consiste à transformer un ensemble de variables originelles en un ensemble de nouvelles variables orthogonales (appelées composantes principales) par l'entremise de la création de vecteurs orthogonaux. Les composantes principales (CP) présentent d'une façon différente l'information véhiculée par l'ensemble des variables originelles. L'ACP met en relief les similitudes générales entre plusieurs variables sans référer à la causalité.

Un des buts de l'ACP est la réduction de la dimension de l'ensemble de variables originelles ou, autrement dit, la représentation de presque la même information que celle contenue dans les variables originelles avec un nombre moindre de variables.

## IV.3.3. Présentation de résultats et discussion

L'analyse ACP réalisée sur la base de données initiale de dimension (532×165) est assurée par le logiciel STATISTICA® (V.9)

### IV.3.3.1. Calcul des valeurs propres

Le calcul des valeurs propres et des pourcentages d'inertie nous permet de déterminer le nombre des axes qui expriment mieux l'information (inertie) contenue dans l'ensemble des variables. Le tableau IV.3.1 présentes un exemple des valeurs propres obtenues des 25 premiers axes (le tableau complet (tableau C.1) est représenté dans l'annexe C).

**Tableau IV.3.1 :** Valeurs propres calculées

| | Val Propre | % Total - variance | Cumul - Val Propre | Cumul % |
|---|---|---|---|---|
| 1 | 77,76652 | 47,13122 | 77,7665 | 47,1312 |
| 2 | 13,60209 | 8,24369 | 91,3686 | 55,3749 |
| 3 | 10,18420 | 6,17224 | 101,5528 | 61,5472 |
| 4 | 8,70843 | 5,27784 | 110,2612 | 66,8250 |
| 5 | 6,67416 | 4,04494 | 116,9354 | 70,8699 |
| 6 | 4,93353 | 2,99002 | 121,8689 | 73,8600 |
| 7 | 4,03655 | 2,44639 | 125,9055 | 76,3063 |
| 8 | 3,80906 | 2,30852 | 129,7145 | 78,6149 |
| 9 | 3,26502 | 1,97880 | 132,9796 | 80,5937 |
| 10 | 2,45548 | 1,48817 | 135,4350 | 82,0818 |
| 11 | 2,31657 | 1,40398 | 137,7516 | 83,4858 |
| 12 | 2,01219 | 1,21951 | 139,7638 | 84,7053 |
| 13 | 1,60005 | 0,96972 | 141,3638 | 85,6751 |
| 14 | 1,54832 | 0,93838 | 142,9122 | 86,6134 |
| 15 | 1,39034 | 0,84263 | 144,3025 | 87,4561 |
| 16 | 1,36798 | 0,82908 | 145,6705 | 88,2851 |
| 17 | 1,27069 | 0,77011 | 146,9412 | 89,0553 |
| 18 | 1,13736 | 0,68931 | 148,0785 | 89,7446 |
| 19 | 1,05160 | 0,63734 | 149,1301 | 90,3819 |
| 20 | 1,01059 | 0,61248 | 150,1407 | 90,9944 |
| 21 | 0,91365 | 0,55372 | 151,0544 | 91,5481 |
| 22 | 0,87798 | 0,53211 | 151,9323 | 92,0802 |
| 23 | 0,80596 | 0,48846 | 152,7383 | 92,5687 |
| 24 | 0,78963 | 0,47856 | 153,5279 | 93,0472 |

#### IV.3.3.1.1. Discussion

Le tableau IV.3.1 montre clairement que les résultats en valeurs propres sont ordonnés par ordre décroissant. La plus grande valeur propre est $\lambda_1$, elle vaut 77,76652 ; donc le premier axe factoriel est celui qui correspond à $\lambda_1$, dont il explique 47,1312 % de la variance totale.

Le deuxième axe factoriel sélectionné, est celui qui correspond à la deuxième valeur propre $\lambda_2$= 13,60209, cette dernière explique 8,24369% de la variance totale.

Le premier plan factoriel, engendré par les deux premiers axes factoriels, explique 55,3749 % de la variance totale, donc on peut déduire que ce pourcentage est insuffisant pour avoir une bonne explication de toutes les variables.

Alors pour mieux interpréter ces variables, on doit les représenter dans plusieurs axes principaux pour avoir un pourcentage d'inertie acceptable (66 %), et pour atteindre ce pourcentage on doit considérer les 4 premiers axes (F1, F2, F3 et F4).

Le troisième axe factoriel explique 6,17224 % de la variance total, donc par conséquent le plan factoriel (1.3) engendré par le premier et le troisième axe factoriel permet d'expliquer 53,30346% de la variance totale qui est toujours insuffisant d'où la nécessité d'utiliser le quatrième axe factoriel qui représente 5,27784% de l'inertie totale, pour atteindre un pourcentage de cumul total de 66,8250% (le pourcentage choisi doit être supérieur à 60%). Des améliorations de ce pourcentage sont très possibles mais on risque de compliquer notre étude, donc notre étude sera focalisée sur les plans : (1.2), (1.3), (2.3) et finalement le plan (1.4).

#### IV.3.3.1.2. Représentation graphique des valeurs propres et du cumul%

La figure IV.3.1 représente l'histogramme représentant la variation des valeurs propres dans 23 axes de l'analyse en composantes principales.

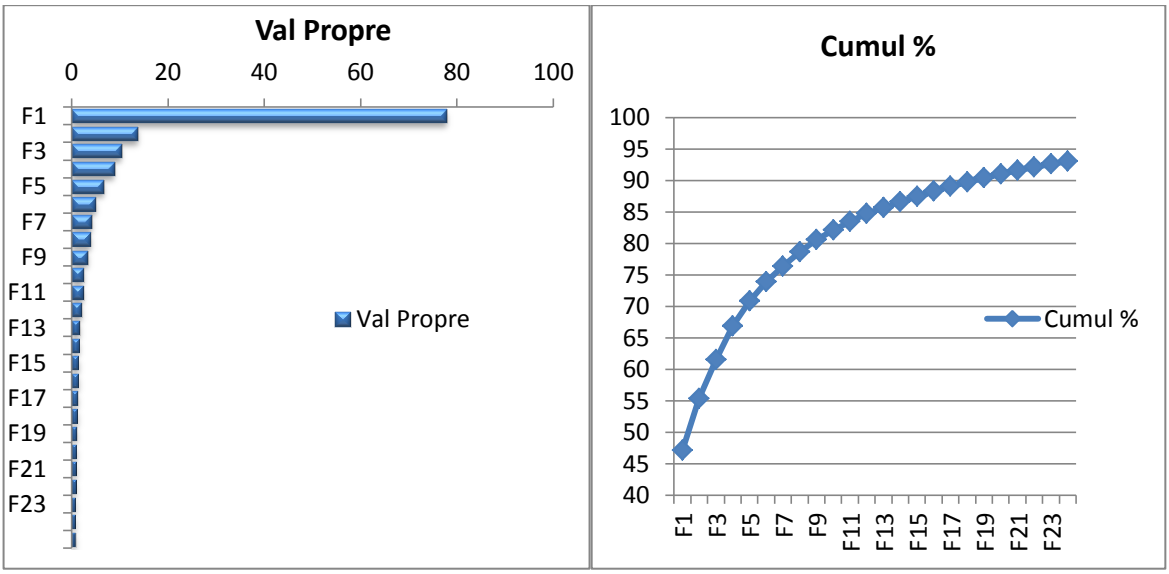



**Figure IV.3.1 :** Représentation graphique des valeurs propres et pourcentage de cumul

La représentation graphique des valeurs propres et pourcentage de cumul confirme les résultats discutés précédemment, donc il est clair que les quatre premiers axes sont ceux qui possèdent les valeurs les plus importantes en valeurs propres ce qui permet d'obtenir un cumul supérieur à 60%.

### IV.3.3.2. Interprétation des axes

Cette étape nous permet de représenter les corrélations qui existent les variables ( descripteurs moléculaires) et les différentes axes ainsi que les corrélations entre les individus (molécules médicamenteuses), le logiciel qui assure cette tâche est STATISTICA (V.9), ce dernier nous permet d'obtenir les différentes corrélations et qui sont représentées dans les tableaux IV.3.2 et IV.3.3 (les tableaux C.2 et C.3 de l'annexe C englobent toutes les corrélations pour toutes les variables et les individus).

**Tableau IV.3.2 :** Matrice des coordonnées factorielles des variables

| | **Fact. 1** | **Fact. 2** | **Fact. 3** | **Fact. 4** |
|---|---|---|---|---|
| AMW | 0,213392 | 0,220731 | 0,006933 | -0,725753 |
| MW | -0,975473 | 0,059164 | 0,025407 | -0,092216 |
| Sv | -0,975169 | -0,040102 | 0,077983 | 0,154336 |
| Se | -0,964066 | 0,019810 | -0,005424 | 0,217513 |
| Sp | -0,971081 | -0,033423 | 0,078198 | 0,160501 |
| Ss | -0,947047 | 0,113521 | -0,088398 | -0,176845 |
| Me | 0,122401 | 0,369871 | -0,319805 | -0,650660 |
| Mp | 0,205015 | -0,169997 | 0,427098 | -0,606178 |
| Ms | 0,112412 | 0,456938 | -0,411376 | -0,447212 |
| nAT | -0,955742 | 0,005765 | 0,009045 | 0,248383 |
| nSK | -0,993101 | -0,003243 | 0,050796 | -0,014161 |
| nBT | -0,961926 | -0,028038 | 0,006308 | 0,227758 |
| nBO | -0,992457 | -0,070320 | 0,041608 | -0,035414 |
| nBM | -0,267870 | -0,125692 | 0,720636 | -0,420549 |
| nDB | -0,547974 | 0,087181 | -0,273291 | -0,318184 |
| nAB | -0,070406 | -0,155581 | 0,808725 | -0,303436 |
| nH | -0,859890 | 0,012932 | -0,026582 | 0,451812 |
| nC | -0,918206 | -0,212234 | 0,187342 | 0,227840 |
| nN | -0,322899 | 0,418687 | 0,010199 | -0,435181 |
| nO | -0,740663 | 0,236903 | -0,392952 | -0,100275 |
| **nCL** | **0,064937** | **-0,043520** | **0,179497** | **-0,209094** |
| **nX** | **-0,038635** | **-0,146125** | **0,084192** | **-0,179295** |
| nCIC | -0,740489 | -0,586224 | -0,041798 | -0,195616 |
| nCIR | -0,460894 | -0,658675 | -0,285270 | 0,014882 |
| nR06 | -0,495048 | -0,558497 | 0,064526 | 0,066176 |
| ZM1 | -0,985534 | -0,127139 | -0,032967 | -0,052624 |
| ZM2 | -0,969901 | -0,194499 | -0,095310 | -0,059168 |
| BAC | -0,777765 | 0,234751 | -0,335389 | 0,115794 |
| ICR | -0,771049 | 0,002519 | 0,261774 | -0,130479 |

L'analyse des résultats représentés dans le tableau IV.3.2 nous permet de dire que la composante principale F1 possède les corrélations les plus élevées avec les variables par rapport aux autres composantes, tel qu'on remarque que certains variables sont fortement corrélées tel que MW, Se (-0,975) et ZM1 (0.98), cependant il y'a d'autres variables présentent une corrélation très faible (positive et parfois négative) tel que nCl (0,064937) et nX (-0,038), ces derniers constituent l'ensemble des descripteurs moléculaires à éliminer au cours de notre analyse des données, dont l'élimination de tel variable ( descripteur) dépende directement de la valeur de corrélation de ce dernier.

**Tableau IV.3.3 :** Coordonnées factorielles des individus

| | Fact. 1 | Fact. 2 | Fact. 3 | Fact. 4 |
|---|---|---|---|---|
| Abacavir | 3,3260 | -2,5924 | 0,92488 | -0,0521 |
| Acecarbromal | 10,5756 | 8,1780 | -2,42293 | 4,3406 |
| Aceclofenac | 3,2263 | 1,3077 | 5,08638 | -1,0900 |
| Acediasulfone | 4,5137 | 2,0279 | 3,23476 | -1,2474 |
| Acefylline | 6,6455 | 1,4529 | -3,20688 | -0,7512 |
| Acemetacin | -2,0875 | -0,8189 | 4,21857 | -1,1174 |
| acétaminophène. | 11,5169 | 3,5293 | -0,22512 | 2,2446 |
| Acetazolamide | 10,6960 | 7,6842 | -2,80409 | -1,2020 |
| Acetiamine | 1,8522 | 6,7795 | 0,42013 | 2,6608 |
| Acetylsalicylic acid | 10,7245 | 3,5216 | -0,23418 | 1,1951 |
| Acetylsulfafurazole | 4,6302 | 1,2831 | 0,06451 | -0,6879 |
| Aciclovir | 7,2331 | 2,6434 | -1,55542 | 0,0587 |
| acide méfénamique | 6,9123 | -0,9539 | 2,75058 | 1,1761 |
| Acide nalidixique | 7,2417 | -0,6521 | -1,31742 | 0,4567 |
| Acide oxolinique | 5,2782 | -2,8430 | -1,30598 | -0,9725 |
| Acide pipémidique | 2,8301 | -1,7222 | -0,73645 | -0,5410 |
| Actarit | 9,7813 | 3,5797 | 1,05910 | 1,8534 |
| Adefovir | 5,5843 | 3,6762 | -0,52683 | -1,0972 |
| Alclofenac | 9,2260 | 3,2416 | 2,34814 | 1,3459 |
| Alclometasone dipropionate | -10,0658 | -3,5898 | -3,54305 | 3,0693 |
| Alfentanil | -3,6126 | -0,5383 | 2,43565 | 1,6990 |
| Alminoprofen | 8,3669 | 2,9414 | 2,10416 | 2,9690 |
| Alovudine | 6,4666 | 0,3471 | -3,33959 | 0,8351 |
| Alphaprodine | 6,1156 | -1,7259 | 1,64210 | 3,2640 |
| Altizide | 3,4703 | 2,7174 | 0,56186 | -4,0638 |
| Amantadine | 9,9102 | -3,6843 | -3,55642 | 4,8501 |
| Ambuside | 3,1478 | 6,6870 | -0,22061 | -1,8845 |
| Amikacine | -14,3433 | 7,0188 | -4,85943 | 2,1121 |
| Aminophenazone | 7,2406 | -1,8037 | 0,64001 | 1,6380 |
| **Clindamycin** | **-0,9207** | **1,6883** | **-1,55211** | **2,2891** |

Toute en gardant la même méthodologie d'analyse précédente, l'élimination des individus (molécules médicamenteuses) se fait aussi en tenant compte de la corrélation que présente chaque individu avec les quatre composantes principales F1,F2, F3 et F4 .

### IV.3.3.3. Interprétation des plans

La représentation graphique des différents plans sélectionnés dans la première étape de cette étude nous permet d'apercevoir clairement les nuages de points qui encombre notre modèle et qui sont considérés comme variables non pertinents, donc la répartition des variable dans les différents plans confirme les résultats discutés précédemment, tel qu'on peut remarquer à partir des figures IV.3.2 et IV.3.3 que les variables qui possèdent une faible corrélation se focalise au centre du plan factoriel, cependant les variables qui sont fortement corrélés avec les axes constituant nos plans d'étude sont situés sur le cercle (corrélation proche de 1)

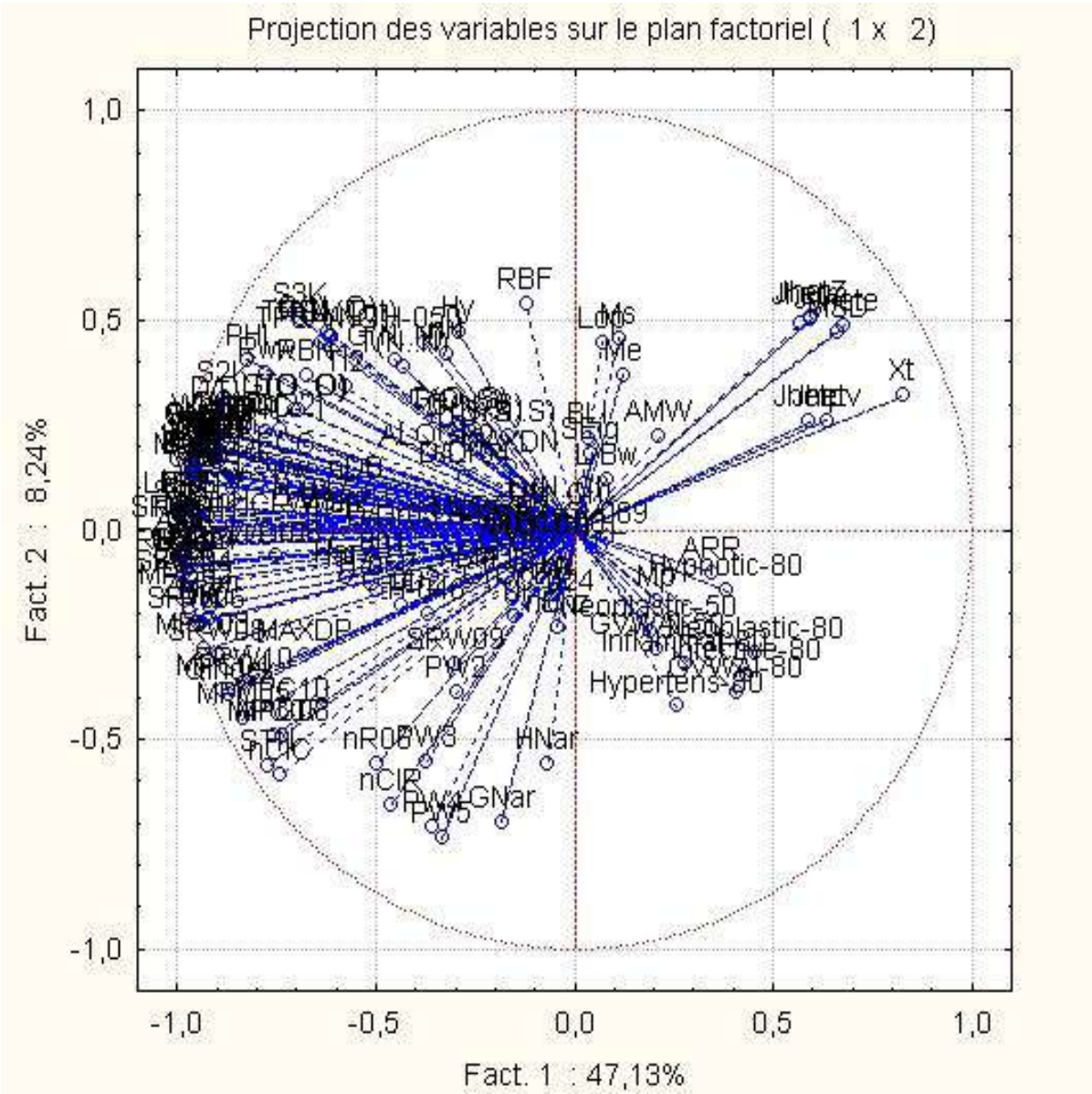



**Figure IV.3.2 :** Cercle de corrélation du plan (1.2) [STATISTICA®]

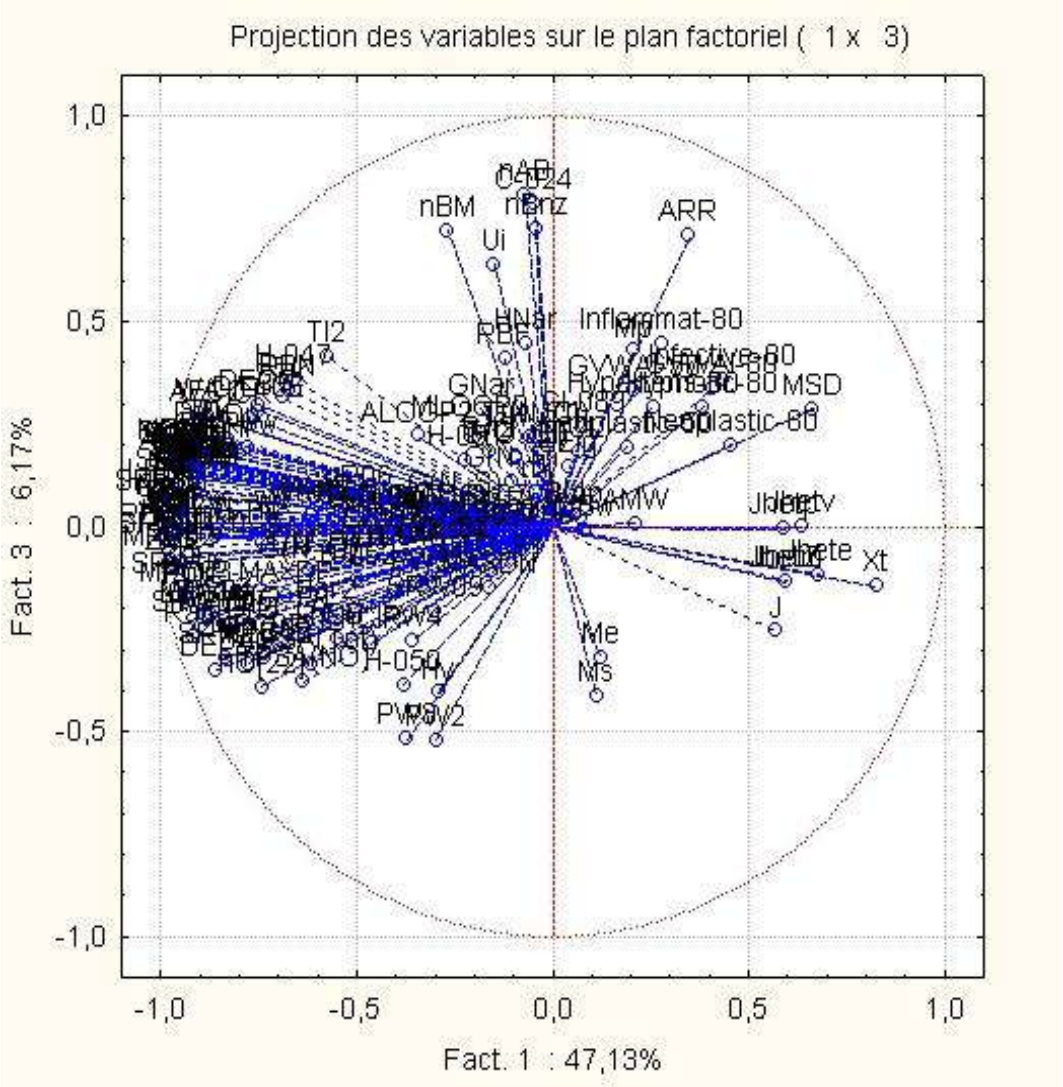



**Figure IV.3.3:** Cercle de corrélation du plan (1.3) [STATISTICA®]

Les figures IV.3.2 et IV.3.3 montrent que la majorité de l'inertie se situe sur le cercle, plus spécifiquement, ces variables présentent une corrélation négative au tour du premier axe factoriel F1, ce qui explique la forte corrélation entre les variables et cet axe.

Finalement, la bonne exploitation des résultats concernant la répartition des variables et individus sur les axes et les plans factoriels interprétés précédemment nous conduisent à la réduction de 51 descripteurs moléculaires (AMW, Me, Mp, Ms, nAB…) cette réduction sera suivie d'une réduction des individus (molécules) qui sont considérés aberrants (Abacavir , Acecarbromal, Aceclofenac , Clindamycin…) et par conséquent l'obtention d'une nouvelle base de dimension plus faible. Finalement, on a pu réduire la dimension de notre base initiale (532×165) vers une base finale de dimension (504×114) après la réalisation de l'analyse en composantes principales.

## IV.3.4. Régression linéaire multiple (RLM)

La régression linéaire multiple est la méthode la plus simple de modélisation. Elle consiste à rechercher une équation linéaire par rapport à ses paramètres, reliant la variable à modéliser Y au vecteur d'entrées X .Ces entrées peuvent être soit les variables primaires, soit des fonctions non paramétrées, ou à paramètres fixés, de ces variables (par exemple des monômes). Donc cette méthode de modélisation permet de :

- Décrire le lien linéaire existant entre des variables contrôlées (X) et une variable réponse (Y)
- Prédire la réponse Y en fonction de valeurs contrôlées X

### IV.3.4.1. Élaboration des modèles de régression multiple

#### IV.3.4.1.1. Présentation du modèle

Soit le modèle sous forme globale à k variable explicatif et n observations :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i \;\Rightarrow\; Y = Xb + e}$$

Pour i = 1…n

Tel que :

Y : vecteur des $\boldsymbol{Y_i}$ (variable à expliquer par le modèle : activité thérapeutique) ;

X : vecteur des $\boldsymbol{X_i}$ (variables explicatives : descripteurs moléculaires) ;

b : vecteur des coefficients $\mathbf{b_i^*}$ du modèle RLM ;

e : vecteur des $\boldsymbol{e_i}$ erreurs du modèle ;

n: nombre d'observations (nombre de molécules médicamenteuses).

**IV.3.4.1.2. Forme matricielle**

Afin d'en alléger l'écriture et faciliter l'expression de certains résultats, on fait habituellement recours aux notions matricielles, en écrivant le modèle observation par observation. Cette écriture rend plus aisée la manipulation du modèle linaire général, donc si on veut transformer le modèle précédent sous forme matricielle, on peut écrire :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & .. X_{1k} \\ 1 & X_{21} & X_{22} & .. X_{2k} \\ 1 & X_{31} & X_{32} & .. X_{3k} \\ 1 & . & . & .. \ . \\ 1 & . & . & .. \ . \\ 1 & . & . & .. \ . \\ 1 & . & . & .. \ . \\ 1 & X_{n1} & X_{n2} & .. X_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_3 \\ e_3 \\ \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Afin d'obtenir toujours une équation de la forme :

$$\boldsymbol{Y = Xb + e}$$

$$\boldsymbol{\widehat{Y} = Xb}$$

$$\boldsymbol{\Rightarrow e = Y - \widehat{Y} = Y - Xb}$$

Avec

$\boldsymbol{Y}$ : Valeur observée (activité moléculaire observée) ;

$\boldsymbol{\widehat{Y}}$ : Valeur prévue (activité moléculaire estimée par le modèle) ;

e : Valeur de résidu (écart entre la valeur observée et la valeur prévue).

Donc on doit estimer le vecteur des coefficients et le vecteur des résidus du modèle pour chaque activité thérapeutique, les coefficients et les résidus ont été calculés en utilisant le même logiciel précédant (STATISTICA), tel que le calcul est basé sur la méthode des moindres carrés ordinaires (MCO) qui consiste à minimiser la somme des carrés des erreurs.

## IV.3.4.2. Statistiques descriptives et résultats d'estimation des modèles

### IV.3.4.2.1. Statistiques descriptives

Les résultats d'estimation générale obtenus par l'utilisation du logiciel STATISTICA sont présentés dans le tableau IV.3.5.

**Tableau IV.3.5 :** Résultats générales de la régression linéaire multiples

| Variable dépendante | Activité anti-inflammatoire | Activité antibiotique | Activité analgésique | Activité antivirale | Activité diurétique | Activité glucocorticoïde |
|---|---|---|---|---|---|---|
| **R Multiple** | 0,66985308 | 0,81059242 | 0,65697314 | 0,64185750 | 0,40331390 | 0,90711246 |
| **$R^2$** | 0,44870315 | 0,65706007 | 0,43161371 | 0,41198106 | 0,16266210 | 0,82285301 |
| **$R^2$ ajusté** | 0,39705660 | 0, 63135794 | 0,39420399 | 0,38758193 | 0,15931275 | 0,80834528 |
| **Nombre d'observations** | 503 | 503 | 503 | 503 | 503 | 503 |
| **Test de fisher (F)** | 8,687961 | 25,56441 | 11,53748 | 16,88507 | 48,56525 | 56,71825 |
| **Degré de liberté (dl)** | 43,459 | 35,467 | 31,471 | 20,482 | 2,500 | 38,464 |
| **probabilité critique du test de Fisher** | 0,000000 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | 0,000000 |
| **Erreur-type de l'estimation** | 0,650818129 | 0,548296984 | 0,609886797 | 0,564115853 | 0,563579991 | 0,200946167 |
| **Ordonnée à l'origine** | -1,520819949 | -4,741599303 | 0,110126264 | 1,236981243 | -1,77998728 | -0,607028737 |
| **Erreur-Type** | 0,8562608 | 1,248941 | 1,459192 | 0,8093125 | 0,1036901 | 0,1479119 |
| **Test de Student (t)** | -1,776 | -3,796 | 0,07547 | 1,5284 | -17,17 | - 4,104 |
| **probabilité critique du test de Student** | 0,0764 | 0 ,0002 | 0,9399 | 0,1271 | 0,0000 | 0,0000 |

L'analyse primaire des résultats de régression montre clairement que le deuxième modèle (activité antibiotique) est le modèle le plus performant en termes des coefficients de régressions, les étapes suivantes nous permettent de bien expliquer et tester les performances de chaque modèle.

### IV.3.4.2. 2. Estimation des modèles RLM

Comme nous avons déjà présenté, tous les modèles RLM ont la même forme globale :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i}$$

Donc on doit estimer les coefficients de chaque modèle et tester la signification et la validation de ces modèles et pour atteindre cet objectif on doit passer par les étapes suivantes :

**a) Détermination des vecteurs des coefficients $b_i$* et vecteurs des résidus $e_i$ pour chaque activité**

Les tableaux (IV.3.6, IV.3.7, IV.3.8, IV.3.9 et IV.3.10) représentent l'ordonnée à l'origine ($b_0$) et les vecteurs des coefficients ($b_i^*$) des six modèles estimés par la méthode des moindres carrées (MCO) et assurée par le logiciel STATISTICA, on présente en rouge les coefficients les plus significatifs du modèle, les résidus sont illustrés dans les tableaux des annexes (C.3, C.4, C.5, C.6 C.7 et C.8).

**Tableau IV.3.6 :** Coefficients pour l'activité Anti-inflammatoire

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | -1,52082 | 0,856261 |
| nN | -0,60159 | 0,100091 |
| nBM | 0,96946 | 0,126600 |
| H-052 | -0,17866 | 0,064879 |
| G(N..N) | 0,26407 | 0,081946 |
| Har2 | -0,14488 | 0,082037 |
| ALOGP2 | 0,10909 | 0,058449 |
| Jhetp | -0,61796 | 0,341897 |
| Jhetv | 0,46189 | 0,325500 |
| SRW10 | -0,79876 | 0,358075 |
| C-001 | 0,04204 | 0,082722 |
| S3K | -0,63034 | 0,217918 |
| Se | 2,22890 | 0,483565 |
| AMR | -1,00972 | 0,521123 |
| ICR | -0,21508 | 0,202073 |
| nDB | 0,22648 | 0,074552 |
| TPSA(NO) | -0,63851 | 0,321021 |
| S0K | 1,03269 | 0,560918 |
| D/Dr06 | -0,04219 | 0,106977 |
| VAR | 0,44550 | 0,363474 |
| Whetp | 2,30897 | 0,570976 |
| SMTI | -6,86213 | 1,919098 |
| MPC08 | 1,06699 | 0,264575 |
| TPSA(Tot) | 0,47379 | 0,310057 |
| Whete | 7,81298 | 2,310924 |
| nCIR | -0,49353 | 0,132513 |
| nR06 | 0,34466 | 0,110736 |
| RHyDp | -6,64938 | 1,706901 |
| D/Dr05 | 0,15040 | 0,074644 |
| WhetZ | -0,35019 | 0,100637 |
| D/D | -0,84423 | 0,716892 |
| ww | -0,52915 | 0,221219 |

**Tableau IV.3.7 :** Coefficients pour l'activité Antibiotique

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | -4,74160 | 1,248941 |
| TPSA(Tot) | -0,19582 | 0,202249 |
| Har2 | -0,08267 | 0,044163 |
| JhetZ | -0,15102 | 0,076521 |
| nR06 | -0,29807 | 0,056066 |
| SRW10 | -3,19479 | 1,482993 |
| Xt | 0,34406 | 0,119561 |
| Qindex | -6,20659 | 0,699470 |
| SRW08 | 8,96366 | 2,040888 |
| S1K | 0,76150 | 0,691986 |
| Rww | -0,12794 | 0,094413 |
| PHI | 0,80511 | 0,550799 |
| Ram | 2,65186 | 0,578252 |
| SPI | -0,61406 | 0,623882 |
| Infective-80 | 0,11365 | 0,040687 |
| nCIC | 2,63031 | 0,414032 |
| MPC03 | -3,38017 | 0,684402 |
| TPSA(NO) | 0,45049 | 0,214715 |
| MPC10 | 0,66251 | 0,274271 |
| H-047 | 0,17987 | 0,060023 |
| T(O..O) | -0,18686 | 0,074394 |
| T(N..N) | -0,27135 | 0,084001 |
| G(N..N) | 0,22701 | 0,097928 |
| DELS | 0,87612 | 0,292442 |
| Ss | -0,64087 | 0,424775 |
| MAXDP | -0,12536 | 0,059413 |
| nDB | 0,17354 | 0,057510 |
| ALOGP2 | 0,09296 | 0,043678 |
| MPC08 | -0,66115 | 0,350687 |
| TI1 | -0,24597 | 0,105529 |
| MW | -0,54848 | 0,264441 |
| G(N..O) | -0,17283 | 0,094961 |

| | | |
|---|---|---|
| Rww | 0,65623 | 0,344152 |
| RBN | -0,35578 | 0,158758 |
| ZM2 | 3,66415 | 1,408896 |
| w | 1,07964 | 0,594056 |
| S2K | 0,90181 | 0,485173 |
| T(O..O) | -0,24797 | 0,118852 |
| SMTIV | 1,85430 | 0,962372 |
| Whetm | -2,76543 | 1,739962 |
| ECC | -2,29649 | 1,166550 |
| AECC | 0,88772 | 0,592591 |
| Jhetm | 0,12610 | 0,117415 |
| CENT | -0,54352 | 0,509602 |

**Tableau IV.3.9:** Coefficients pour l'activité antivirale

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | 1,23698 | 0,809312 |
| Infective-80 | -0,36340 | 0,047326 |
| DELS | -1,23220 | 0,208824 |
| Jhetv | 0,04293 | 0,103758 |
| SRW10 | -2,33468 | 0,286930 |
| Qindex | 1,93729 | 0,341454 |
| J | 0,77741 | 0,163703 |
| Jhete | -1,23364 | 0,188249 |
| H-047 | -0,31485 | 0,063752 |
| Xt | -0,23262 | 0,105077 |
| ZM1V | 1,17046 | 0,208900 |
| T(N..O) | -0,31120 | 0,078438 |
| TPSA(NO) | 0,64281 | 0,226361 |
| TPSA(Tot) | -0,24165 | 0,209049 |
| G(O..O) | -0,17719 | 0,068589 |
| nBM | -0,20998 | 0,081568 |
| D/Dr06 | 0,19251 | 0,067562 |
| DECC | -0,23567 | 0,072544 |
| Jhetm | 0,21733 | 0,102556 |
| nCIC | -0,68382 | 0,238691 |
| MPC05 | 0,58907 | 0,255605 |

**Tableau IV.3.10:** Coefficients pour l'activité diurétique

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | -1,7800 | 0,103690 |
| JhetZ | 17,7904 | 3,831114 |
| Jhetm | -17,4337 | 3,831114 |

| | | |
|---|---|---|
| S3K | -0,63127 | 0,316475 |
| Jhetv | 0,18458 | 0,070675 |
| nN | 0,16221 | 0,073631 |
| MSD | 0,15842 | 0,088187 |

**Tableau IV.3.8:** Coefficients pour l'activité analgésique

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | 0,11013 | 1,459192 |
| TPSA(Tot) | 0,64643 | 0,291271 |
| H-047 | 0,40154 | 0,081159 |
| MW | 0,33805 | 0,397198 |
| nC | 1,24888 | 0,437195 |
| T(N..O) | 0,00762 | 0,160459 |
| LPRS | -0,31338 | 1,809219 |
| TPSA(NO) | -0,83221 | 0,272785 |
| G(O..O) | 0,17646 | 0,097520 |
| ICR | 0,40996 | 0,114681 |
| Jhete | 0,46554 | 0,130528 |
| Jhetp | -0,29910 | 0,090370 |
| nCIR | 0,44216 | 0,109063 |
| MPC08 | -0,75275 | 0,424290 |
| Ss | 0,77015 | 0,450331 |
| MPC10 | 0,41964 | 0,364212 |
| G(N..O) | 0,25269 | 0,185632 |
| S3K | 0,12752 | 0,241217 |
| Ram | -1,34257 | 0,422650 |
| BAC | 0,61269 | 0,152948 |
| SRW04 | 3,39448 | 1,856640 |
| nDB | -0,17851 | 0,067272 |
| GMTIV | 2,82731 | 0,842699 |
| MAXDP | 0,16252 | 0,075986 |
| SMTIV | -3,19660 | 0,955513 |
| ZM1V | -0,75485 | 0,331380 |
| Wap | 0,15898 | 0,068018 |
| MSD | -0,39516 | 0,152505 |
| VDA | 0,69312 | 0,583705 |
| PHI | -1,08388 | 0,407593 |
| C-002 | 0,15839 | 0,077537 |
| ZM1 | -4,57786 | 2,623451 |

**Tableau IV.3.11 :** Coefficients pour l'activité glucocorticoïde

| | b* | Err-Type |
|---|---|---|
| OrdOrig. | -0,607029 | 0,147912 |
| Har2 | 0,40656 | 0,049768 |
| nN | -0,10227 | 0,031650 |
| H-047 | -0,09242 | 0,042012 |
| SRW10 | -0,60654 | 0,581292 |
| Ram | -0,38279 | 0,353124 |
| WhetZ | 0,49677 | 0,060040 |
| Wap | -0,43510 | 0,064099 |
| ww | 1,38913 | 0,129148 |
| w | -2,03655 | 0,292115 |
| Pol | 1,58670 | 0,423268 |
| HyDp | -1,82158 | 0,320800 |
| Whetv | 0,47908 | 0,308182 |
| G(O..O) | -0,51171 | 0,131297 |
| T(O..O) | 0,32517 | 0,127307 |
| H-052 | -0,11656 | 0,033147 |
| SNar | -1,42151 | 0,651874 |
| ZM2 | 1,70946 | 1,548077 |
| DECC | 0,19241 | 0,074722 |
| ALOGP2 | -0,08184 | 0,032822 |
| MPC07 | 0,82034 | 0,194944 |
| ZM1V | -0,62103 | 0,121264 |
| GMTIV | 1,29920 | 0,245609 |
| RBN | 0,57169 | 0,087120 |
| Rww | -1,01375 | 0,154905 |
| MPC05 | -1,22310 | 0,338292 |
| GVWAI-80 | 0,07145 | 0,027996 |
| nDB | 0,12339 | 0,039801 |
| Jhetp | -0,12808 | 0,039791 |
| UNIP | -1,82845 | 0,445185 |
| QW | 1,95497 | 0,500167 |
| SRW06 | 2,26088 | 1,185399 |
| D/Dr05 | 0,19422 | 0,052048 |
| MDDD | -0,85907 | 0,220341 |
| ECC | 6,26519 | 1,701762 |
| D/Dr06 | 0,15924 | 0,062505 |
| CSI | -4,99004 | 1,623044 |
| TIE | -0,32402 | 0,135538 |
| MPC03 | -1,42359 | 0,663830 |

**b) Présentation et tests de significations des modèles pour les six activités thérapeutiques**

**1) Activité anti-inflammatoire**

Le premier modèle a la forme suivante :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i}$$

Avec

$\mathbf{b_0}$= -1,520819949

**R Multiple**=0,66985308 $\mathbf{R^2}$ =0,44870315 $\mathbf{F_c}$=8,687961 et $\mathbf{t_c}$ =-1,776

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.6 et C.3, donc on obtient à la fin un modèle sous forme matricielle vu le grand nombre des variables explicatives du modèle, la figure IV.3.4 représente la régression linéaire du premier modèle.

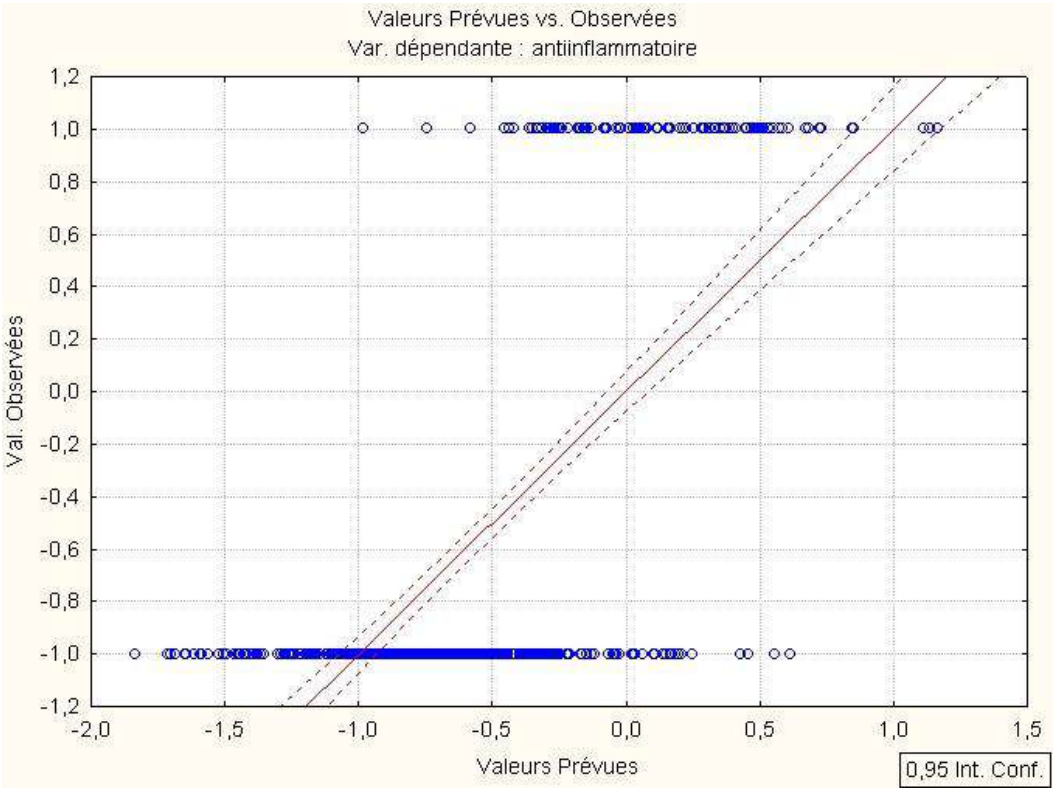



**Figure IV.3.4 :** Courbe de régression pour l'activité anti-inflammatoire

La courbe de régression linéaire de l'activité anti-inflammatoire calculée par RLM avec l'activité anti-inflammatoire donnée expérimentalement montre l'absence d'une bonne concordance entre les valeurs estimées (val. prévues) et les valeurs expérimentales qui constituent les observations du modèle RLM (val. Observées), cela peut être interprété par l'existence des variables qui n'ont pas une influence sur la sortie de notre modèle, et donc par conséquence ces variables vont élargir l'écart entre la valeur estimé et la valeur réelle, cela apparait clairement dans le tableau des résidus (tableau C.4), dont on peut remarquer une valeur importante de résidus pour quelques observations considérées aberrantes pour notre modèle, donc comme solutions pour l'amélioration des capacités prédictives du modèle : soit on procède à la réalisation d'une deuxième analyse en composante principale (ACP) avant la modélisation RLM, ou bien on élimine les individus qui présentent une valeur de résidus importante dans la première régression et on procède à une deuxième régression avec un nombre réduit d'individus.

- **Analyse statistique des résultats**
  - **Coefficient de détermination**

Le coefficient de détermination ($R^2$) est un indicateur qui permet de juger la qualité d'une régression linéaire, simple ou multiple. D'une valeur comprise entre 0 et 1, il mesure l'adéquation entre le modèle et les données observées. C'est le carré du coefficient de corrélation multiple $R_M$, et nous avons forcément $0 \leq R^2 \leq 1$.

$$\boldsymbol{R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 0,4487}$$

Avec

**SCT** : **S**omme des Carrés Totaux, traduit la variabilité totale du modèle;

**SCE** : **S**omme des Carrés Expliqués, traduit la variabilité expliquée par le modèle ;

**SCR** : **S**omme des Carrés Résiduels correspond à la variabilité non-expliquée par le modèle.

| Somme des carrés | Degrés de liberté |
|---|---|
| $\boldsymbol{SCE = \sum_i (\hat{y}_i - \bar{y})^2}$ | $\boldsymbol{k}$ |
| $\boldsymbol{SCT = \sum_i (y_i - \hat{y}_i)^2}$ | $\boldsymbol{n - k - 1}$ |
| $\boldsymbol{SCR = \sum_i (y_i - \bar{y})^2}$ | $\boldsymbol{n - 1}$ |

La valeur de ($R^2$) fournie par STASITICA pour ce premier modèle est égale à 0,448 ce qui signifie que 44,87% de la variable endogène (activité anti-inflammatoire) est expliquée par les variables exogènes (43 descripteurs moléculaires considérés pour l'explication de la sortie du modèle et qui sont figurés dans le tableau IV.5.6).

- **Tests de significativité du modèle**

1) **Test de Fisher** : Significativité globale du modèle

Le $R^2$ est un indicateur simple, on comprend aisément que plus il s'approche de la valeur 1, plus le modèle est intéressant. En revanche, il ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de y.

Nous devons nous tourner vers les tests d'hypothèses pour vérifier si la liaison mise en évidence avec la régression n'est pas un simple artefact.

La formulation du test d'hypothèse qui permet d'évaluer globalement le modèle est la suivante :

- $H_0 : b_1 = b_2 = \ldots = b_k = 0$ ;
- $H_1$ : un des coefficients au moins est non nul.

La statistique dédiée à ce test s'appuie (parmi les différentes formulations possibles) sur le $R^2$, il s'écrit :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)}}$$

et suit une loi de Fisher à (k, n-k-1) degrés de liberté.

La région critique du test est donc : rejet de H0 si et seulement si $\boldsymbol{F_c > F_t}$ pour α (k, n- k-1), où α est le facteur du risque et $F_t$ c'est le F tabulé (déterminé à partir du table du Fischer pour un degré de liberté (k, n- k-1) et pour α=5%).

Une autre manière de lire le test est de comparer la p-value (probabilité critique du test) avec α : si elle est inférieure, l'hypothèse nulle est rejetée (on prend α=0,05)

Donc pour le premier modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuse) ;
- Le nombre de variables explicatives (k=43 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve alors :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 8,687}$$

Avec

$$\boldsymbol{F_t = F^{0,05}_{(43.459)} = 1,39}$$

La comparaison entre la valeur de $\boldsymbol{F_c}$ calculé et la valeur de $\boldsymbol{F_t}$ tabulée, nous permet de dire que $\boldsymbol{F_c > F_t}$ , donc on rejette l'hypothèse $H_0$ c'est-à-dire il existe au moins un paramètre non nulle, Une deuxième comparaison entre la valeur de probabilité critique (p) et la valeur du facteur de risque α montre que $\boldsymbol{\alpha > p}$ on peut conclure alors, que le modèle globale est significatif.

**2) Test de Student:** Significations des coefficients

Pour tester la signification des coefficients du modèle, on effectue le test de student selon l'hypothèse nulle $H_0$ contre l'hypothèse alternative.

Tel que

- $H_0 : b_i = B_k = 0$ ;
- $H_1$ : un des coefficients au moins est non nul.

Avec

$$\boldsymbol{t_c = \frac{b_k - B_k}{\sqrt{var(b_k)}}}$$

Donc ce test assure mesure l'influence des variables exogènes sur la variable endogène, il est basé sur la comparaison entre la valeur ($t_c$) et la valeur tabulée ($t_t$), et la règle de décision sera

de rejeter H0 si : $|t_c| \geq t_{1-\alpha/2}(n-k)$ et généralement on prend un seuil de signification (α=0,05)**.**

Dans le cas de ce test, on peut soit tester la significativité de chaque coefficient (comme est illustré dans le tableau IV.3.6 tel que les coefficients significatifs sont représentés en rouge) ou bien on réalise un test sur la totalité des coefficients, tel que dans les deux cas la valeur de $t_c$ est fournie directement par le logiciel STATISTICA.

Donc pour ce premier modèle on obtient :

$$t_c = -1,776 \Rightarrow |t_c| = 1,776$$

$$et\ t_{1-0,05/2}(503-43) = 1,96 \quad donc\ |t_c| < t_t$$

On a aussi

$$p = 0,0764 \ \text{donc}\ p > \alpha$$

Ces deux résultats nous permet de dire que les coefficients du modèle ne sont pas tous significatifs donc on doit refaire la modélisation par un nombre réduit de coefficients, dont on utilise seulement ceux qui sont significatifs (représentés en rouge). Donc pour ce premier modèle on peut conclure qu'il est significatif globalement mais ses capacités prédictives ne sont pas vraiment fortes, cela peut être dû à l'existence des variables non significatifs (selon le test de student) ces variables encombre le modèle et minimisent ses performances.

### 2) Activité antibiotique

Le deuxième modèle a la forme suivante :

$$Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i$$

Avec

$\mathbf{b_0}$= -4,741599303

**R Multiple**=0,81059242 $\mathbf{R^2}$ =0,65706007 $\mathbf{F_c}$=25,56441 et $\mathbf{t_c}$ = -3,796

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.7 et C.4, donc on obtient à la fin un modèle sous forme matricielle vu le grand nombre

des variables explicatives du modèle, la figure IV.3.5 représente la régression linéaire du deuxième modèle.

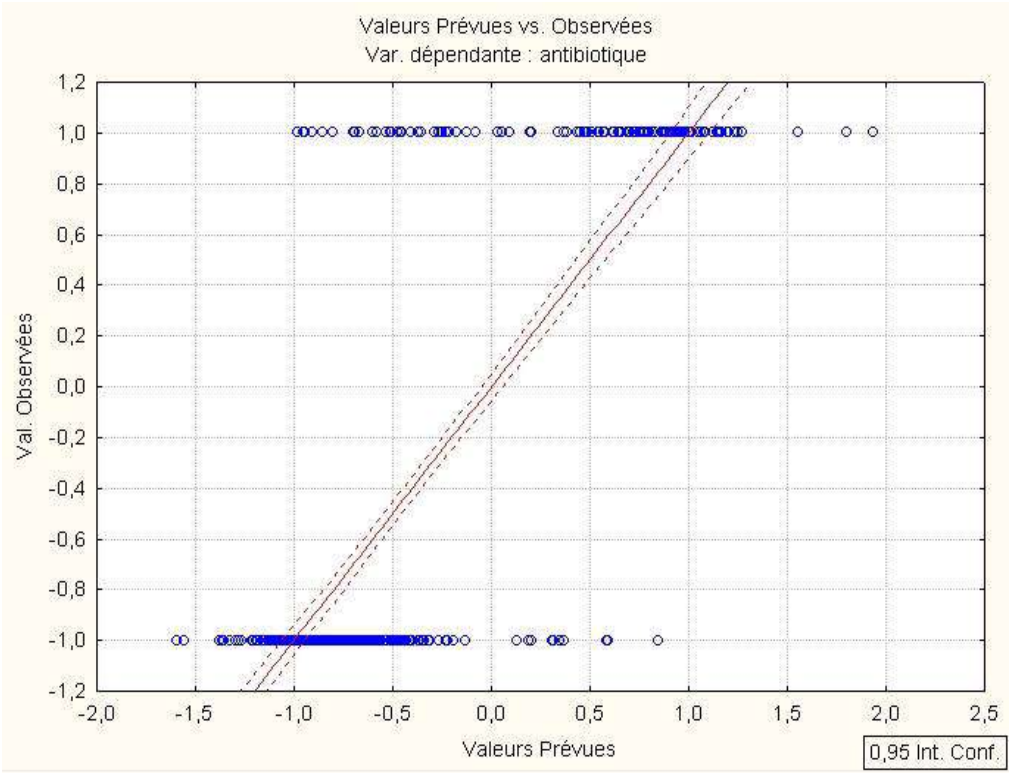



**Figure IV.3.5 :** Courbe de régression pour l'activité antibiotique

La courbe de régression linéaire de l'activité antibiotique calculée par RLM avec l'activité antibiotique donnée expérimentalement montre la présence d'une bonne concordance ($R_M$=81%) entre les valeurs estimées (val. prévues) et les valeurs expérimentales qui constituent les observations du modèle RLM (val. Observées).

- **Analyse statistique des résultats**
  - **Coefficient de détermination**

La valeur de ($R^2$) fournie par STASITICA pour ce deuxième modèle est égale à 0,657 ce qui signifie que 65,7% de la variable endogène (activité antibiotique) est expliquée par les variables exogènes (35 descripteurs moléculaires considérés pour l'explication de la sortie du modèle et qui sont figurés dans le tableau IV.5.8).

- **Tests de significativité du modèle**

**1) Test de Fisher** : Significativité globale du modèle

Pour le deuxième modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuses) ;
- Le nombre de variables explicatives (k=35 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve alors :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 25,564}$$

Avec

$$\boldsymbol{F_t = F^{0,05}_{(35.467)} = 1,46}$$

On remarque que $\boldsymbol{F_c > F_t}$ , donc on rejette l'hypothèse $H_0$ c'est-à-dire il existe au moins un paramètre non nulle, Une deuxième comparaison entre la valeur de probabilité critique (p) et la valeur du facteur de risque α montre que $\boldsymbol{\alpha > p}$ donc le modèle globale est significatif.

**2) Test de Student:** Significations des coefficients

Le test de signification des coefficients du deuxième modèle se fait en utilisant le même principe présenté dans le premier modèle, donc pour le deuxième modèle on obtient :

$$\boldsymbol{t_c = -3,796 \Rightarrow |t_c| = 3,796}$$

$$\boldsymbol{et\ t_{1-0,05/2}(503-35) = 1,96 \quad donc\ |t_c| > t_t}$$

On a aussi

$$\boldsymbol{p = 0,0002 \quad donc\ p < \alpha}$$

Ces deux résultats nous permettent de dire que les coefficients du modèle sont significatifs donc on rejette l'hypothèse $H_0$. Les différents résultats présentés pour la modélisation de l'activité antibiotique montrent que le deuxième modèle possède des bons performances pour

la prédiction de l'activité antibiotique, les différents tests de significations montrent aussi que le modèle globale est significatif, et que les coefficients sont aussi significatifs.

### 3) Activité analgésique

Le troisième modèle a la forme suivante :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i}$$

Avec

$\mathbf{b_0}$= 0,110126264

**R Multiple**=0,65697314 $\mathbf{R^2}$ =0,43161371 $\mathbf{F_c}$=11,53748 et $\mathbf{t_c}$ = 0,07547

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.8 et C.6, donc on obtient à la fin un modèle sous forme matricielle vu le grand nombre des variables explicatives du modèle, la figure IV.3.6 représente la régression linéaire du troisième modèle.

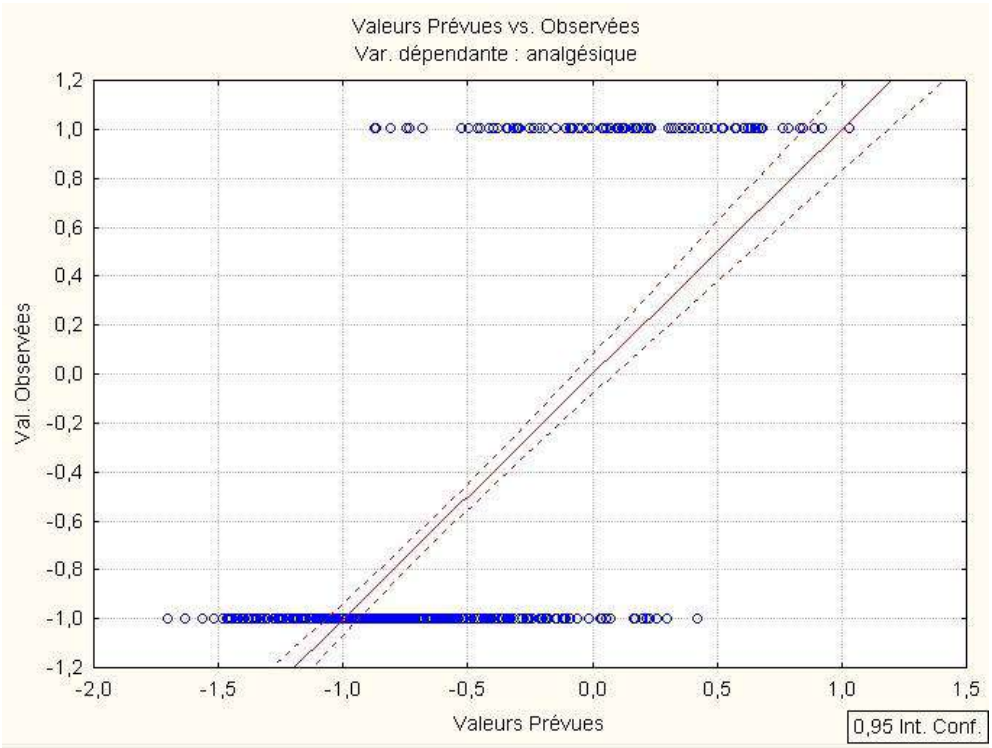



**Figure IV.3.6 :** Courbe de régression pour l'activité analgésique

La courbe de régression linéaire de l'activité analgésique montre la présence d'une concordance acceptable ($R_M$=65%) entre les valeurs estimées (val. prévues) et les valeurs expérimentales (val. Observées).

- **Analyse statistique des résultats**
  - **Coefficient de détermination**

La valeur de ($R^2$) du troisième modèle est égale à 0,431 ce qui signifie que 43,16% de la variable endogène (activité analgésique) est expliquée par les variables exogènes (31 descripteurs moléculaires considérés pour l'explication de la sortie du modèle et qui sont figurés dans le tableau IV.5.7).

- **Tests de significativité du modèle**

1) **Test de Fisher** : Significativité globale du modèle

Pour le troisième modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuses) ;
- Le nombre de variables explicatives (k=31 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve alors :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 11,537}$$

Avec

$$\boldsymbol{F_t = F^{0,05}_{(31.467)} = 1,46}$$

On remarque que $\boldsymbol{F_c > F_t}$ , donc on rejette l'hypothèse $H_0$ c'est-à-dire il existe au moins un paramètre non nulle, la comparaison entre la valeur de probabilité critique (p) et la valeur du facteur de risque α montre clairent que $\boldsymbol{\alpha > p}$ on peut conclure alors, que le modèle globale est significatif.

2) **Test de Student:** Significations des coefficients

Pour le troisième modèle on obtient :

$$\boldsymbol{t_c = 0,07547 \Rightarrow |t_c| = 0,07547}$$

$$\boldsymbol{et\ t_{1-0,05/2}(503-31) = 1,96 \quad donc\ |t_c| < t_t}$$

On a aussi $\boldsymbol{p = 0,9399}$ **donc** $\boldsymbol{p > \alpha}$

Ces deux résultats nous permettent de dire que les coefficients du modèle ne sont pas significatifs. Les différents résultats présentés pour la modélisation de l'activité analgésique montrent que le troisième modèle possède des performances acceptable pour la prédiction de l'activité analgésique, les différents tests de significations montrent aussi que le modèle globale est significatif, cependant il existe quelques coefficients qui ne sont pas significatifs (illustrés dans le tableau IV.5.8). Cette non significativité est due à :

- Soit à une absence de corrélation de ces variables avec la variable à expliquer (la sortie du modèle).
- Soit à une colinéarité trop élevée avec une des variables explicatives.

### 4) Activité antivirale

Le quatrième modèle a la forme suivante :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i}$$

Avec

$\mathbf{b_0}$= 1,236981243

**R Multiple**=0,64185750 $\mathbf{R^2}$ =0,41198106 $\mathbf{F_c}$=16,88507 et $\mathbf{t_c}$ = 1,5284

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.9 et C.6, , la figure IV.3.7 représente la régression linéaire du quatrième modèle.

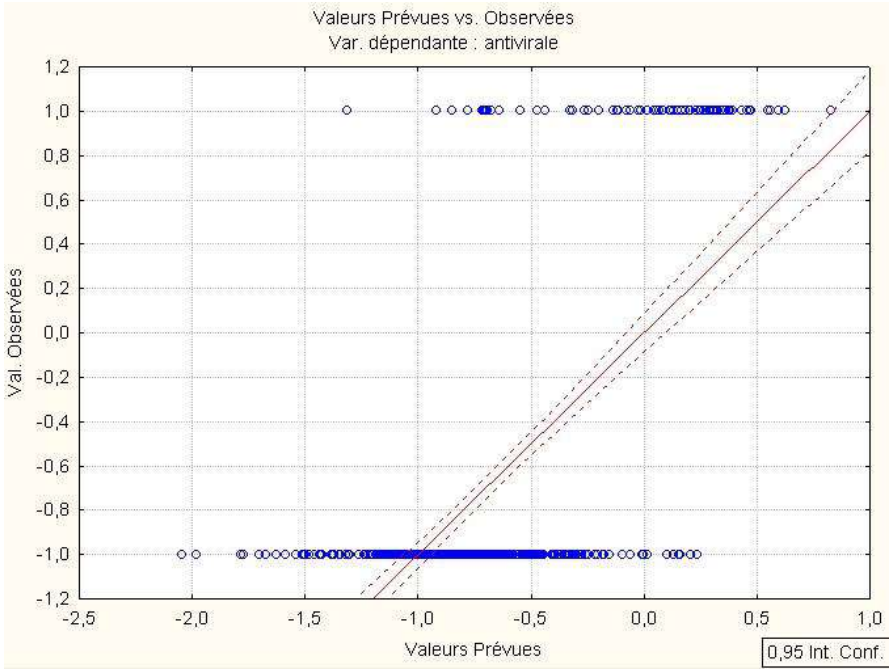



**Figure IV.3.7 :** Courbe de régression pour l'activité antivirale

La courbe de régression linéaire de l'activité antivirale montre la présence d'une concordance acceptable ($R_M$=64,1%) entre les valeurs estimées (val. prévues) et les valeurs expérimentales (val. Observées).

- **Analyse statistique des résultats**
  - **Coefficient de détermination**

La valeur de ($R^2$) ce modèle est égale à 0,411 ce qui signifie que 41,19% de la variable endogène (activité antivirale) est expliquée par les variables exogènes (20 descripteurs moléculaires considérés pour l'explication de la sortie du modèle et qui sont figurés dans le tableau IV.5.9).

- **Tests de significativité du modèle**

1) **Test de Fisher** : Significativité globale du modèle

Pour le quatrième modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuses) ;
- Le nombre de variables explicatives (k=20 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve alors :

$$F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 16,885$$

Avec

$$F_t = F^{0,05}_{(20.483)} = 1,57$$

La comparaison entre la valeur de $F_c$ calculé et la valeur de $F_t$ tabulée, nous permet de dire que $F_c > F_t$ , donc on rejette l'hypothèse $H_0$. On remarque ainsi que $\alpha > p$ on peut conclure alors, que le modèle globale est significatif.

**2) Test de Student:** Significations des coefficients

De la même manière précédente on trouve pour le quatrième modèle:

$$t_c = 1,5284 \Rightarrow |t_c| = 1,5284$$

$$et\ t_{1-0,05/2}(503-20) = 1,96 \quad donc\ |t_c| < t_t$$

On a aussi la probabilité de significativité $p = 0,1271$ **donc** $p > \alpha$ , se qui signifie que les variables sélectionnés pour l'explication de la sortie (activité antivirale) n'ont pas tous une influence sur la sortie (seulement les coefficients illustrés en rouge dans le tableau IV.5.9), donc le regroupement de toutes les résultats trouvés, nous permet de conclure que le quatrième modèle possède des performances acceptables pour la prédiction de l'activité antivirale, les différents tests de significations montrent aussi que le modèle globale est significatif, cependant il existe quelques coefficients qui ne sont pas significatifs.

### 4) Activité diurétique

Le cinquième modèle a la forme suivante :

$$Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + e_i$$

$$\Rightarrow Y_i = -1,78 + 17,79\, X_{i,1} - 17,43\, X_{i,2} + e_i$$

Avec

$X_{i,1}$: **JhetZ**

$X_{i,2}$: **Jhetm**

Donc le modèle devient :

$$Y_i = -1,78 + 17,79(\text{JhetZ}) - 17,43(\text{Jhetm}) + e_i$$

Avec

$b_0$= -1,78

**R Multiple**=0,40331390 $R^2$ =0,16266210 $F_c$=48,56525 et $t_c$ = -17,17

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.10 et C.7, la figure IV.3.8 représente la régression linéaire du dernier modèle.

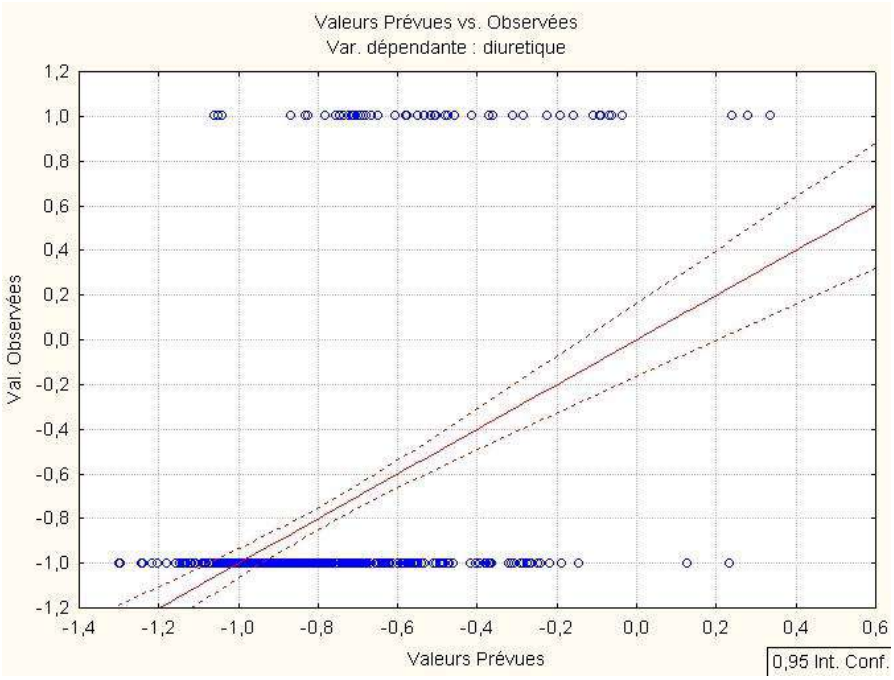



**Figure IV.3.8 :** Courbe de régression pour l'activité diurétique

La courbe de régression linéaire de l'activité diurétique montre la présence d'une faible concordance ($R_M$=40,33%) entre les valeurs estimées et les valeurs expérimentales.

- **Analyse statistique des résultats**

➢ **Coefficient de détermination**

La valeur de ($R^2$) pour ce modèle est égale à 0,1626 ce qui signifie que 16,26 % de la variable endogène (activité diurétique) est expliquée par les variables exogènes (2 descripteurs moléculaires considérés pour l'explication de la sortie du modèle et qui sont figurés dans le tableau IV.5.10).

➢ **Tests de significativité du modèle**

1) **Test de Fisher** : Significativité globale du modèle

Pour ce modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuses) ;
- Le nombre de variables explicatives (k=2 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve alors :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 48,565}$$

Avec

$$\boldsymbol{F_t = F^{0,05}_{(2.501)} = 3,00}$$

On remarque que $\boldsymbol{F_c > F_t}$ , donc on rejette l'hypothèse $H_0$ c'est-à-dire il existe au moins un paramètre non nulle, la comparaison entre la valeur de probabilité critique (p) et la valeur du facteur de risque α montre que $\boldsymbol{\alpha > p}$ donc le modèle globale est significatif.

2) **Test de Student:** Significations des coefficients

On trouve pour le cinquième modèle:

$$\boldsymbol{t_c = -17,17 \Rightarrow |t_c| = 17,17}$$

$$\boldsymbol{et\ t_{1-0,05/2}(503-2) = 1,96 \quad donc\ |t_c| > t_t}$$

On a aussi

$$\boldsymbol{p = 0,0000 \quad donc\ p < \alpha}$$

Le calcul réalisé nous permet de dire que les coefficients du modèle sont significatifs donc on rejette l'hypothèse $H_0$. Les différents résultats présentés pour la modélisation de l'activité diurétique montrent que ce modèle possède des faibles performances pour la prédiction de l'activité diurétique, cependant les différents tests de significations montrent que le modèle globale est significatif, et que les coefficients sont aussi significatifs.

### 6) Activité glucocorticoïde

Le dernier modèle a la forme suivante :

$$\boldsymbol{Y_i = b_0 + b_1^* X_{i,1} + b_2^* X_{i,2} + \cdots + b_{k-1}^* X_{i,k-1} + e_i}$$

Avec

$\mathbf{b_0}$= 1,236981243

**R Multiple**=0,90711246 $\mathbf{R^2}$ =0,82285301 $\mathbf{F_c}$=56,71825 et $\mathbf{t_c}$ = - 4,104

Les résultats en termes de coefficients ($b_i$*) et ($e_i$) sont représentés dans les tableaux IV.3.11 et C.8, la figure IV.3.9 représente la régression linéaire du sixième modèle.

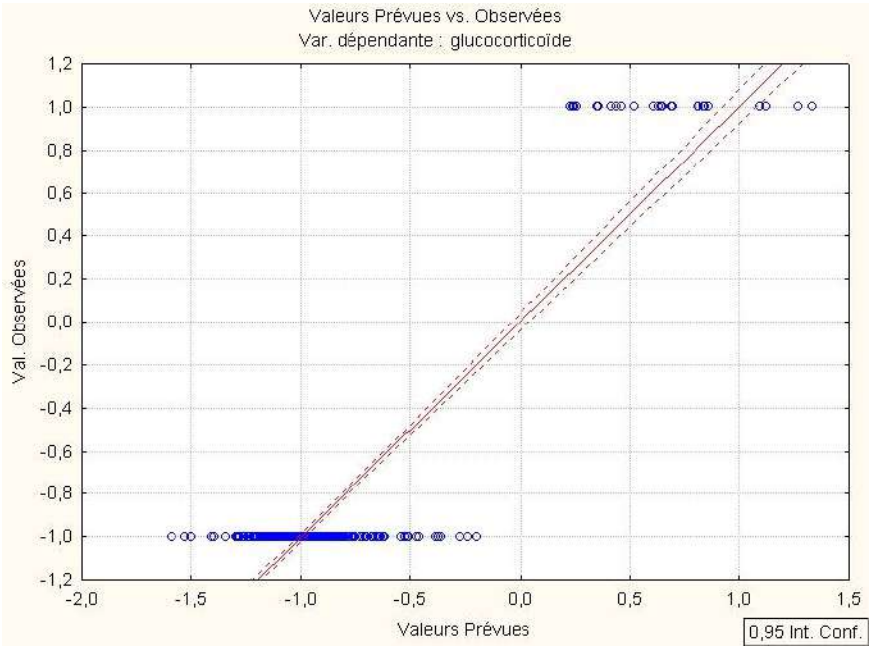



**Figure IV.3.9 :** Courbe de régression pour l'activité glucocorticoïde.

La courbe de régression linéaire de l'activité glucocorticoïde calculée par RLM montre la présence d'une bonne concordance ($R_M$=90,71%) entre les valeurs estimées (val. prévues) et les valeurs expérimentales (val. Observées).

- **Analyse statistique des résultats**

  - **Coefficient de détermination**

La valeur de ($R^2$) de ce modèle est égale à 0,8228 ce qui signifie que 82,28% de la variable endogène (activité glucocorticoïde) est expliquée par les variables exogènes. Donc on peut dire que dans ce cas les variables sélectionnés expliquent fortement la variable modélisée.

  - **Tests de significativité du modèle**

**1) Test de Fisher** : Significativité globale du modèle

Pour le sixième modèle on a les données suivantes :

- Le nombre d'observations (n= 503 molécules médicamenteuses) ;
- Le nombre de variables explicatives (k=33 descripteurs moléculaires) ;
- La probabilité critique du test de Fisher (p=0,000000)
- Le facteur de risque (α=0,05) ;

On trouve :

$$\boldsymbol{F_c = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 56,7182}$$

Avec

$$\boldsymbol{F_t = F^{0,05}_{(33.470)} = 1,46}$$

On remarque que $\boldsymbol{F_c > F_t}$ , donc on rejette l'hypothèse $H_0$ c'est-à-dire il existe au moins un paramètre non nulle, Il est claire aussi que $\boldsymbol{\alpha > p}$ on peut conclure alors, que le modèle globale est significatif.

**2) Test de Student:** Significations des coefficients

Pour le dernier modèle:

$$\boldsymbol{t_c = -4,104 \quad \Rightarrow |t_c| = 4,104}$$

$$\boldsymbol{et\ t_{1-0,05/2}(503-33) = 1,96 \quad donc\ |t_c| < t_t}$$

On a aussi la probabilité de significativité $\boldsymbol{p = 0,0000}$ **donc** $\boldsymbol{p < \alpha}$, se qui signifie que les variables sélectionnés pour l'explication de la sortie (activité glucocorticoïde) sont très significatifs pour le modèle élaboré.

## IV.3.5. Analyse comparative entre la modélisation par RLM et la modélisation neuronale

Le tableau IV.3.11 représente une comparaison entre les résultats obtenus par la méthode de modélisation traditionnelle représentée par le couplage ACP et RLM et les résultats obtenus par méthode de modélisation qu'on a élaboré représentée par le couplage de la méthode statistique « élimination progressive » et la modélisation neuronale. Cette comparaison sera en termes de facteurs de corrélations obtenues.

**Tableau IV.3.11 :** Facteurs de corrélation obtenus par RNA et RLM

| | Facteur de régression obtenu par RNA | Facteur de régression obtenu par RLM |
|---|---|---|
| Activité anti-inflammatoire | 0,96034 | 0,66985 |
| Activité antibiotique | 0,98438 | 0,81059 |
| Activité analgésique | 0,95784 | 0,65697 |
| Activité antivirale | 0,95284 | 0,64185 |
| Activité diurétique | 0,96320 | 0,40331 |
| Activité glucocorticoïde | 1,00000 | 0,90711 |

À partir des résultats présentés dans tableau IV.3.11, on peut remarquer que les facteurs de corrélation obtenus lors de la modélisation neuronale sont nettement supérieurs à ceux trouvés lors de la régression linéaire multiple, donc on peut conclure que la méthode neuronale pour l'optimisation de prédiction de l'activité thérapeutique est beaucoup plus avantageuse que la méthode traditionnelle utilisée fréquemment pour la modélisation moléculaire. Cette différence en capacité prédictive peut être expliquée par :

1) l'incapacité du RLM à prendre en compte les données manquantes, ce qui conduit souvent à rejeter beaucoup d'observations incomplètes et pourtant contenant de l'information utile dans les champs renseignés.

2) Grande sensibilité à la colinéarité entre les variables explicatives. La colinéarité la rend le modèle RLM numériquement instable et fait perdre l'interprétation des coefficients.

3) Indétermination lorsque le nombre d'observations est inférieur au nombre des variables explicatives.

## IV.3.6 Conclusion

Dans ce chapitre, nous avons élaboré une base de données de dimension importante, cette base comporte les structures moléculaires des molécules douées d'une activité thérapeutique appartenant à 6 différentes classes thérapeutiques. Ces molécules sont caractérisées par leurs descripteurs moléculaires calculés préalablement à l'aide de différents logiciels.

Dans l'objectif de prédire l'activité thérapeutique à partir du structure moléculaire nous avons introduit la méthode neuronale. La taille de la base réalisés poses de nombreuse problèmes pour la modélisation neuronale, donc on a proposé une méthode statistique (élimination progressive) permet de réduire la dimensionnalité de la base. L'architecture du RNA (nombre de couches cachées, nombre de neurones) est un paramètre important décidant de la qualité de l'apprentissage plus que les paramètres d'apprentissage (taux d'apprentissage, …). La taille et la qualité de la base de données ont une influence majeure sur l'apprentissage des RNA et sur la fiabilité du modèle neuronale à prédire les différentes activités thérapeutiques.

Finalement une étude comparative a été réalisé entre la méthode proposée (couplage élimination progressive et modélisation neuronale) et les méthodes traditionnelles utilisées pour la modélisation moléculaire tel que le couplage entre la méthode ACP et la modélisation par régression linéaire multiple (RLM) réalisée sur la même base de donnée. Cette étude montre la robustesse et la fiabilité de la méthode neuronale par rapport d'autres méthodes de modélisation.

# *Conclusion générale*

# CONCLUSION GÉNÉRALE

La modélisation de propriétés et d'activités de molécules constitue un champ de recherche important, qui permet par exemple de guider la synthèse de médicaments. Les méthodes traditionnelles de modélisation établissent des relations entre les propriétés étudiées et les caractéristiques structurelles des molécules, appelées descripteurs. Leurs principaux inconvénients résident dans la difficulté du choix des descripteurs.

Nous avons mis au point une nouvelle technique de modélisation qui s'affranchit de ces problèmes, en établissant une relation directe entre les descripteurs moléculaires et la propriété modélisée par l'utilisation d'une méthode statistique « élimination progressive » permettant de déterminer les descripteurs les plus pertinents et nous avons appliqué ensuite la modélisation neuronale. Cette méthode de modélisation par réseaux de neurones artificiels fait partie de nouvelles techniques sous l'égide de l'intelligence computationnelle ou des sciences cognitives et elle permet de développer des modèles assurant la corrélation et la prédiction de l'activité thérapeutique des molécules.

Pour atteindre cet objectif, nous avons élaboré une base de données de dimension {532, 227} contenant les structures moléculaires des molécules dotées d'une activité thérapeutique appartenant à 6 différentes classes thérapeutiques {anti-inflammatoire ; antibiotique ; analgésique ; antivirale ; diurétique et glucocorticoïde}, ces molécules sont caractérisées par leurs descripteurs moléculaires qui représentent les différentes propriétés chimiques, physiques et thérapeutiques. Le calcul des descripteurs est assuré par différents logiciels (Chembio office, Dragon) et est effectué après avoir établi les notations SMIIES pour chaque molécule. Un grand nombre de descripteurs différents ont été collectés pour la modélisation de l'activité thérapeutique et optimisé ensuite par la technique d'élimination progressive.

Pour la recherche de la topologie du réseau de neurones optimal nous avons adopté la stratégie d'essai et erreur (trial and error), avec deux critères d'arrêt : vérification d'une fonction de performance (d'erreur) et la capacité de généralisation du réseau de neurones. En tenant compte de ces deux critères, six architectures neuronales ont été optimisées pour avoir des topologies finales de: {105_ (29_19_9)_1} ; {82_ (21_2) _1} ; {72_ (22_9_2) _1} ; {76_(21_18_2) _1} ; {69_ (21_18_2) _1} et {110_ (18_9) _1} permettant de prédire respectivement l'activité thérapeutique des molécules : anti-inflammatoires ; antibiotiques ;analgésiques ; antiviraux ; diurétiques et glucocorticoïdes.

Les modèles conçus ont montré une très bonne corrélation des données expérimentales. Nous notons aussi que les meilleures performances de prédiction ont été trouvées dans le cas des antibiotiques et glucocorticoïdes dont on signale des excellentes capacités de prédiction pour ces deux familles thérapeutiques ; cela peut être expliqué par la forte ressemblance qui existe entre les molécules de chaque famille thérapeutique. On a trouvé que toutes les molécules glucocorticoïdes ont été prédites d'une manière correcte et 99,0039 % des molécules antibiotiques sont ainsi correctement prédites. En terme de comparaison ce pourcentage se révèle être plus important que celui obtenu par Miguel et al. (2003) pour un ensemble plus réduit de molécules.

Finalement une étude comparative a été réalisée entre la méthode neuronale combinée à l'élimination progressive et une des méthodes traditionnelles les plus utilisées pour la modélisation moléculaire tel que le couplage entre la méthode ACP et la modélisation par régression linéaire multiple (RLM) réalisée sur la même base de donnée. Cette étude nous a permis de montrer les avantages présentés par la méthode élaborée par rapport aux méthodes traditionnelles en terme de capacité de prédiction de l'activité thérapeutique.

A la lumière des résultats obtenus, nous pouvons tirer les conclusions suivantes :

- La modélisation par réseau de neurones est une méthode qui est basée seulement sur des données expérimentales, donc la qualité et la quantité des données décide de l'exactitude et la robustesse du modèle.
- Ainsi la modélisation neuronale peut servir comme une alternative fiable aux méthodes traditionnelles de modélisation de la prédiction des activités thérapeutiques des molécules.

- De nombreux facteurs vont intervenir sur la capacité prédictive de chaque modèle élaboré. Nous citons les plus importants selon notre étude :
    - La nature des molécules sélectionnées dans la base de données et la présence de ressemblances entre les structures des molécules testés peuvent être considérer comme des facteurs très importants pour l'amélioration des capacités prédictives du modèle. C'est le cas figuré pour les molécules antibiotiques et glucocorticoïdes dont on a obtenu une capacité de prédiction maximale.
    - La taille de la base de données
    - Le nombre et le type de descripteurs moléculaires.
    - La méthode statistique de sélection des descripteurs les plus pertinents.
    - Et finalement les propriétés du réseau de neurones optimisé.

En perspective, tout en estimant avoir atteint notre objectif tracé pour cette étude, nous envisageons d'appliquer la méthode neuronale pour la prédiction d'autres activités thérapeutiques. Nous envisageons aussi l'élaboration d'un modèle neuronale permettant la prédiction de plusieurs activités thérapeutiques à la fois. La modélisation neuronale peut être aussi appliquée en association avec de nouvelles techniques de représentation des structures moléculaires (graphs machines) qui ne demandent pas le calcul préalable des descripteurs moléculaires et qui peuvent assurer de bonne précision.

# *Références Bibliographiques*

# REFERENCES BIBLIOGRAPHIQUES

**Adina et al., 2005,** Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors Application to HIV-1 protease inhibitors Institute of Biochemistry¸ Romania, Journal of Molecular Graphics and Modelling 25 (2006) 37–45

**Balaban, A.T., 1982**, Highly discriminating distance-based topological index. Chemical Physics Letters, 1982, 89, p. 399-404.

**Borman, S., 1990**, New QSAR Techniques Eyed for Environnemental Assessments. Chem. News, 68: 20-23.

**André Bouchier ., 2006,** Analyse en composantes principales Analyses multivariés. INRA formation permanente. Janvier 2006.

**Bouzidi A., 2006,** Méthode des réseaux de neurones pour le calcul des coefficients pertinents des phénomènes de transport: Viscosité ; Conductivité et Diffusion moléculaire des gaz. *Mémoire de Magister U. Médéa*.

**Burbidge, R., et al., 2001**, Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computational Chemistry, 2001, 26(1), p. 5-14.

**Byvatov, E., et al., 2003,** Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of Chemical Information and Computer Sciences, 2003, 43 (6), p.

**Crum-Brown, A., et Frazer, T**. On the connection between chemical constitution and physiological action. *Transactions of the Royal Society of Edinburgh* 1868-69, 25, p.151-203

**Curry B. and Morgan P.H., 2006,** Model selection in neural networks: some difficulties, *Eur. J. Operational Research*, 170, 567-577.

**Demuth H. et Beale M., 2000,** Neural Network Toolbox, for use with MATLAB®, User's Guide version 4, The MathWorks Inc.

**Ding, C., et Dubchak, I., 2001,** Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 2001, 17, p. 349-358.

**Dreyfus G., Martinez J. M., Samuelides M., Gordon M. B., Badran F., Thiria S. & Herault L., 2002,** Réseaux de neurones : Méthodologie et applications, Editions Eyrolles, Paris.

**Dreyfus, G., et al., 2004,** Réseaux de neurones, méthodologie et applications. Paris : Eyrolles, 2ème édition

**Gärtner, T., 2003,** A survey of kernels for structured data. ACM SIGKDD Exploration Newsletter, 5(1), p. 49-58.

**Geman, S., Bienenstock, E., et Doursat, R., 1992,** Neural networks and the bias/variance dilemma. Neural Computation, 1992, 4 (1), p. 1-58.

**Goulon. Sigwalt. Abram, Aurélie., 2008,** Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments; *thèse de doctorat ;* l'université Pierre et Marie Curie; Électronique EDITE

**Goulon, A., Duprat, A., et Dreyfus, G., 2005,** From Hopfield nets to recursive networks to graph machines: numerical machine learning for structured data. Theoretical Computer Science, 2005, 344 (2-3), p. 298-334

**Guerra, N.E. Campillo., 2009,** Neural computational prediction of oral drug absorption based on CODES 2D descriptors; Instituto de Quımica Medica, CSIC, Juan de la Cierva Madrid. Spain European Journal of Medicinal Chemistry

**Hagan T, Demuth H.B and Beale M.H., 1996,** Neural Network Design, PWS Publishing, Boston, MA.

**Hansch, C., Leo, A., et Hoekmann, D., 1995,** Exploring QSAR: hydrophobic, electronic and steric constants. Washington, DC: American Chemical Society.

**Hansch, C., 1969,** A Quantitative Approach to Biochemical Structure-Activity Relationships.

**Hansch, C., Leo, A., and Taft, R.W., 1991,** A Survey of Hammett Substituent Constants and Resonance and Field Parameters. Chem. Rev. , 91: 165-195

**Haussler, D., 1999,** Convolution kernels on discrete structures. Rapport technique, UCSCCRL-99-10, 1999, UC Santa Cruz.

**Haykin S., 1999,** Neural Networks: A Comprehensive Foundation, 2nd Ed., Prentice Hall, and Upper Saddle River, NJ.

**Hebb D., 1949,** The Organization of Behavior, Editions Wiley, New York

**Heritage, T.W., et al., 1998,** EVA: A novel theoretical descriptor for QSAR studies Perspectives in Drug Discovery and Design, 1998, 9-11 (0), p. 381-398.

**Hinton G., 1989,** Connectionnist Learning Procedures, Artificial Intelligence.

**Hopfield J. J., 1982,** Neural Networks and Physical Systems with Emergent, Collective Computational Abilities, Proc. Natl. Acad. Sci., Vol. 79, pp. 2554-2558.

**Jaakkola, T., Diekhans, M., et Haussler, D. A., 2000,** discriminative framework for detecting remote protein homologies. Journal of Computational Biology, 2000, 7 (1-2), p. 95-114.

**Jolliffe, I.T., 2002,** Principal Component Analysis. New-York, NY: Springer, 2ème edition.

**Kashima, H., Tsuda, K., et Inokuchi, A., 2003,** Marginalized kernels between labeled graphs. Twentieth International Conference on Machine Learning, 2003 AAAI Press, p. 321-328.

**Kier, L.B., et Hall, L.H., 1976,** Molecular connectivity in chemistry and drug research. New- York : Academic Press, 1976.

**Kleemann A., Engel J., 2001,** Pharmaceutical Substances; (4th Ed.2001 in two volumes)

**Klopman, G., et al., 1994,** Computer automated (logP) calculations based on an extended group approach. Journal of Chemical Information and Computer Sciences, 1994, 34 (4), p. 752-781.

**Korichi M, Gerbaud V, Talou T, Raynaud C et Floquet P., 2006,** Relation Structure moléculaire – Odeur Utilisation des Réseaux de Neurones pour l'estimation de l'Odeur Balsamique ; Toulouse (France)

**Lee Y., Hoon S. Oh et Won Kim M., 1999,** The Effect of Initial Weights on Premature Saturation in Back-Propagation Learning, International Joint Conference on Neural Networks, Vol. 1, pp. 765-770

**Leo, A., et al., 1975,** Calculation of hydrophobic constant (logP) from π and f constants. Journal of Medicinal Chemistry, 1975, 18, p. 865.

**Leslie, C., Eskin, E., et Noble, W.W.S., 2002,** The spectrum kernel: a string kernel for SVM protein classification. Proceedings of the Pacific Symposium on Biocomputing, p. 564-575.

**Lhuguenot J.C., 2001,** Relation quantitative structure-activité (QSAR) : une méthode mal reconnue car trop souvent mal utilisée.

**Lind, P., et Maltseva, T., 2003,** Support vector machines for the estimation of aqueous solubility. *Journal of Chemical Information and Computer Sciences*, 2003, 43 (6), p. 1855-1859.

**Lipnick, RL. Charles Ernest Overton., 1986,** Narcosis Studies and a Contribution to General, Pharmacology. Trends Pharmacol. Sci., 7: 161-164.

**Mahé, P., et al., 2004,** Extensions of marginalized graph kernels, in Twenty-first International Conference on Machine Learning. 2004, Banff, Alberta, Canada: ACM Press.

**Mahé, P., et al., 2005,** Graph kernels for molecular structure-activity relationship analysis with support vector machines. Journal of Chemical Information and Modeling, 2005, 45 (4), p. 939-951.

**MathWorks, I., 2004,** Neural Network Toolbox User's Guide. MATLAB 7.0, Release 14, MATLAB Manual.

**McCulloch W.S. et Pitts W., 1943,** A logical Calculus for the Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, Vol. 5, pp. 115-133.

**Micheli, A., Portera, F., et Sperduti, A., 2003,** QSAR/QSPR studies by kernel machines, recursive neural networks and their integration. 14th Italian Workshop on Neural Nets,WIRN, 2003, Vietri sul Mare, Italy : Springer, p. 308-315.

**Miguel Murcia-Soler ,Facundo Pe´rez-Gime´nez, Francisco J. G,M. Salabert-Salvador, Wladimiro Dıaz-Villanueva., 2003,** Artificial Neural Networks and Linear Discriminant Analysis: A Valuable Combination in the Selection of New Antibacterial Compounds; J. Chem. Inf. Comput. Sci. 2003, 44, 1031-1041

**Monge O., 2006,** Création et utilisation de chimiothèques optimisées pour la recherche in silico de nouveaux composes bioactifs. Novembre 2006, réalisé à Orléans.

M**uller, K.R., et al., 2005,** Classifying 'drug-likeness' with kernel-based learning methods. Journal of Chemical Information and Modeling, 45 (2), p. 249-253.

**Omari T., 2009,** Modélisation numérique d'un capteur-stockeur solaire. *Mémoire de Magister U. Médéa*.

**Opera T., 2000,** Property distribution of drug-related chemical databases. J Comput Aid Mol Des; 14: 251-64.

**Plumb A. P., Rowe R. C., York P. and Brown M., 2005,** Optimisation of the predictive ability of artificial neural network (ANN) models: A comparison of three ANN programs and four classes of training algorithm, *Eur. J. Pharm. Sci.*, 25, 395-405

**Randić, M.**, On characterization of molecular branching. Journal of the American Chemical Society, 1975, 97, p. 6609-6614

**Richton G., 2003,** Nonleadlikness and leadlikness in biochemical screening Drug Discov Today; 8: 86-96.

**Rosenblatt F., 1962**, Principles of Neurodynamics, Spartan Books, New York.

**Rossi S., 2004,** Australian Medicines Handbook 2004. Adelaide: Australian Medicines Handbook. (ISBN 0-95785-214-2)

**Rumelhart D. E., Hinton G.E. et Williams R.J., 1986,** Learning Internal Representations by Error Propagation, MIT Press, Cambridge.

**Samarasinghe S., 2006,** Neural Networks for Applied Sciences and Engineering, Taylor & Francis.

**Schuur, J.H., Selzer, P., et Gasteiger, J., 1996,** The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. Journal of Chemical Information and Computer Sciences, 1996, 36 (2), p. 334-344.

**Serdler J, Mc Govern S, Danan T, Shaichet B., 2003,** Identification and prediction of promiscuous aggregating inhibitors among known drugs. J Med Chem; 46 : 4447-86

**Si-moussa C., 2009,** Modélisation des équilibres de phases à pressions élevées par les réseaux de neurones artificiels, *Thèse de Doctorat* ; Ecole Nationale Supérieure Polytechnique, Alger

**Sperandio O, Miteva MA, Villoutreix BO., 2008,** Combining ligand-and structure based methods in drug design projects. Curr Comput Aid Drug Design; 4: 250-8.

**Stoppiglia, H., 1997,** Méthodes statistiques de sélection de modèles neuronaux ; applications financières et bancaires [thèse en ligne]. Paris : Université Pierre et Marie Curie, 1997.

**Stoppiglia, H., et al., 2003,** Ranking a random feature for variable and feature selection. Journal of Machine Learning Research, 2003, 3, p. 1399-1414.

**Toshio Fujita., 1995,** « QSAR AND DROG DESIGNE NEW DEVELOPMENTS AND APPLICATIONS » IS B N 0-444-88615-X , Department of Agricultural Chemistry, Kyoto University, Kyoto, and EMIL PROJECT, Fujitsu Kansai Systems Laboratory, Osaka, Japan

**Tsuda, K., Kin, T., et Asai, K., 2002,** Marginalized kernels for biological sequences. Bioinformatics, 2002, 18, p. 268-275.

**Vert, J.-P., Saigo, H., et Akutsu, T., 2004,** Local alignment kernels for biological sequences, in *Kernel Methods in Computational Biology*, Schölkopf, B., Tsuda, K., et Vert, J., Editors. 2004 MIT Press. p. 131-154.

**Vogl T. P., Mangis, J. K., Rigler, A. K., Zink, W. T., & Alkon, D. L., 1988,** Accelerating the Cnvergence of the Bckpropagation Method, Biological Cybernetics, Vol. 59.

**Weininger, D., J. Chem., 1988,** Inf. Comput. Sci., 28 (1988), 31-36.

**Wiener, H.**, **1947,** Structural determination of paraffin boiling points. Journal of Chemical Information and Computer Sciences, 1947, 69, p. 17-20.

www.opensmiles.org

**Wold, H., 1966,** Estimation of principal components and related models by iterative least squares, in Multivariate Analysis, Krishnaiaah, P.R., Editor. 1966, New York: Academic Press. p. 391-420.

**Yao, X.J., et al., 2004,** Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regressions. Journal of Chemical Information and Computer Sciences, 2004, 44(4), p. 1257-1266.

# *Annexes*

# ANNEXE A

## Échantillons des tests de corrélations réalisés

XLSTAT 2011.1.02 - Tests de corrélation - le 17/02/2011 à 09:54:29
Tableau observations/variables : Classeur = Classeur1.xlsx / Feuille = Feuil4 / Plage = Feuil4!$A:$B / 532 lignes et 2 colonnes
Type de corrélation : Pearson

Statistiques descriptives :

| Variable | Observations | Obs. avec données manquantes | Obs. sans données manquantes | Minimum | Maximum |
|---|---|---|---|---|---|
| BLTF96 | 532 | 0 | 532 | -7,170 | 3,670 |
| Anti-inflammatoire | 532 | 0 | 532 | -1,000 | 1,000 |

Matrice de corrélation (Pearson) :

| Variables | BLTF96 | Anti-inflammatoire |
|---|---|---|
| BLTF96 | **1** | **-0,362** |
| Anti-inflammatoire | **-0,362** | **1** |

*Les valeurs en gras sont différentes de 0 à un niveau de signification alpha=0,05*

p-values :

| Variables | BLTF96 | Anti-inflammatoire |
|---|---|---|
| BLTF96 | **0** | **< 0,0001** |
| antiinflam | **< 0,0001** | **0** |

*Les valeurs en gras sont différentes de 0 à un niveau de signification alpha=0,05*

Coefficients de détermination ($R^2$) :

| Variables | BLTF96 | Anti-inflammatoire |
|---|---|---|
| BLTF96[2898] | **1** | **0,131** |
| Anti-inflammatoire | **0,131** | **1** |

# ANNEXE B

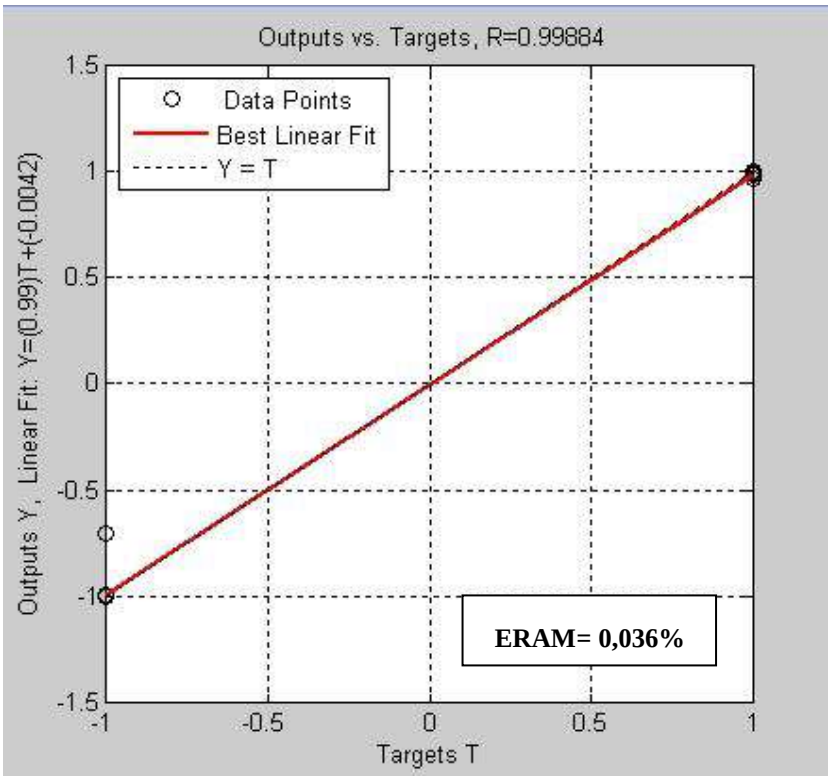



**Figure B.1:** Courbe de régression linéaire des valeurs prédites par RNA avec les valeurs données expérimentalement pour le processus de test pour la prédiction de l'activité des glucocorticoïdes.

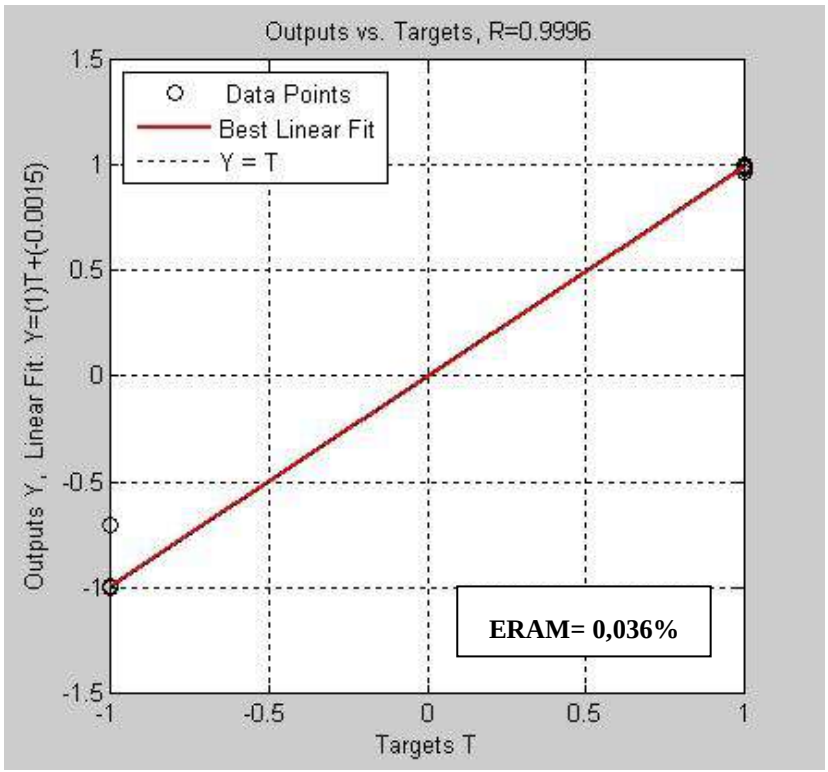



**Figure B.2 :** Courbe de régression linéaire des valeurs prédites par RNA avec les valeurs données expérimentalement pour le processus de généralisation pour la prédiction de l'activité des glucocorticoïdes

**Tableau B.1:** Résultats du calcul du réseau de neurone artificiel, processus d'apprentissage.
(Ensemble des molécules non antibiotiques et molécules antibiotique)

| Processus d'apprentissage | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Molécules | Vexp | Vcal | Molécules | Vexp | Vcal | Molécules | Vexp | Vcal |
| Bucetin | -1 | -0.9989 | Sulfadiazine | -1 | -1.0015 | Céftriaxone | 1 | 0.9994 |
| Acetanilide | -1 | -0.9986 | Sulfadimethoxine | -1 | -0.9976 | Céfuroximeaxétil | 1 | 0.9991 |
| Alanine | -1 | -0.9984 | Sulfafurazole | -1 | -0.9973 | Chloramphénicol | 1 | 0.9985 |
| Aspartic acid | -1 | -0.9985 | Sulfalene | -1 | -0.9985 | Clindamycine | 1 | 0.9988 |
| Benzydamine | -1 | -0.9991 | Sulfamethoxazole | -1 | -0.9977 | Dalfopristine | 1 | 0.9999 |
| Bufexamac | -1 | -0.9988 | Sulfametoxydiazine | -1 | -1.0014 | Déméclocycline | 1 | 0.9993 |
| Butyl hydroxytoluol | -1 | -0.999 | Sulfametrole | -1 | -0.9975 | Dicloxacilline | 1 | 0.9994 |
| Butylmethoxyphenol | -1 | -0.9985 | Sulfanilamide | -1 | -0.9992 | Doxycycline | 1 | 0.9992 |
| Chloroxylenol | -1 | -0.9986 | Sulfaperin | -1 | -1.0016 | Enoxacine | 1 | 0.9982 |
| Cinoxacin | -1 | -0.9989 | Sulfaquinoxaline | -1 | -0.9995 | Ertapénème | 1 | 0.9991 |
| Clioquinol | -1 | -0.9989 | Sulfasalazine | -1 | -0.9985 | Ethambutol | 1 | 0.9978 |
| Cysteine | -1 | -0.9985 | Sulfisomidine | -1 | -0.9974 | Flucloxacilline | 1 | 0.9993 |
| Cystine | -1 | -0.9988 | Sulindac | -1 | -1.0032 | Fluméquine | 1 | 0.9981 |
| Diazoxide | -1 | -0.9989 | Taurolidine | -1 | -1.0002 | Fosfomycine | 1 | 0.9973 |
| Diflunisal | -1 | -0.999 | Tetroxoprim | -1 | -0.9959 | Framycetine | 1 | 0.9998 |
| Ethylparaben | -1 | -0.9987 | Viminol | -1 | -0.9998 | Gatifloxacine | 1 | 0.9987 |
| Felodipine | -1 | -0.9992 | Acidenalidixique | 1 | 0.9973 | Gentamicine | 1 | 0.9992 |
| Gentisicacid | -1 | -0.9986 | Acideoxolinique | 1 | 0.9976 | Grepafloxacine | 1 | 0.9981 |
| Glutamic acid | -1 | -0.9989 | Amikacine | 1 | 0.9997 | Hetacilline | 1 | 0.9904 |
| Glycine | -1 | -0.9984 | Amoxycilline | 1 | 0.9985 | Imipénème | 1 | 0.9953 |
| Histidine | -1 | -0.9987 | Ampicilline | 1 | 0.9984 | Isoniazide | 1 | 0.9929 |
| Isoleucine | -1 | -0.9987 | Astromicine | 1 | 0.9988 | Kanamycine | 1 | 0.9997 |
| Leucine | -1 | -0.9987 | Aztréonam | 1 | 0.9987 | Latamoxef | 1 | 1.0001 |
| Levoglutamide | -1 | -0.9986 | Benzylpénicilline | 1 | 0.9985 | Lévofloxacine | 1 | 0.9987 |
| Minoxidil | -1 | -0.9989 | Céfaclor | 1 | 0.9986 | Lincomycine | 1 | 0.9991 |
| Muzolimine | -1 | -0.9991 | Céfadroxil | 1 | 0.9986 | Loracarbef | 1 | 0.9987 |
| Nalidixicacid | -1 | -0.9982 | Céfalexine | 1 | 0.9982 | Lymécycline | 1 | 0.9992 |
| Naproxen | -1 | -0.9988 | Céfalotine | 1 | 0.9985 | Méropénème | 1 | 0.9952 |
| Nifedipine | -1 | -0.9991 | Céfamandole | 1 | 0.999 | Metampicilline | 1 | 0.9995 |
| Nifuratel | -1 | -0.9989 | Céfatrizine | 1 | 0.9988 | Métronidazole | 1 | 0.9938 |
| Pentazocine | -1 | -0.9991 | Céfazoline | 1 | 0.999 | Mezlocilline | 1 | 1.0049 |
| Nilvadipine | -1 | -0.9992 | Céfixime | 1 | 0.9988 | Minocycline | 1 | 0.9975 |
| Oxolinicacid | -1 | -0.9985 | Céfoperazone | 1 | 0.9998 | Mupirocine | 1 | 1.0146 |
| Phenolphthalein | -1 | -0.999 | Céfotétan | 1 | 0.9994 | Néomycine | 1 | 0.9996 |
| Piretanide | -1 | -1.0004 | Céfotiamhéxétil | 1 | 0.9999 | Nimorazole | 1 | 0.9981 |
| Salacetamide | -1 | -0.9993 | Céfoxitine | 1 | 0.9986 | Norfloxacine | 1 | 0.9988 |
| Salicylamide | -1 | -0.9985 | Céfpirome | 1 | 0.9993 | Ofloxacine | 1 | 0.9987 |
| Salicylicacid | -1 | -0.9987 | Céfradine | 1 | 0.9983 | Paromomycine | 1 | 1.0019 |
| Sulfacarbamide | -1 | -0.9996 | Céfsulodine | 1 | 0.9993 | Phénoxyméthylp | 1 | 0.9986 |
| Sulfacetamide | -1 | -0.9994 | Céftazidime | 1 | 0.9994 | Pipéracilline | 1 | 1.0039 |

**Tableau B.2:** Résultats du calcul du réseau de neurone artificiel, processus de validation. (Ensemble des molécules non antibiotiques et molécules antibiotique)

| **Processus de validation** | | | **Processus de test** | | |
|---|---|---|---|---|---|
| **Molécules** | **Vexp** | **Vcal** | **Molécules** | **Vexp** | **Vcal** |
| Acetylaminonitro propoxybenzene | -1 | -1.0007 | Benzylisothiocyanate | -1 | -1 |
| Aspirin | -1 | -0.9985 | Chiniofon | -1 | -1.0401 |
| Bisacodyl | -1 | -0.998 | Hydroxyproline | -1 | -1.0001 |
| Chlorocresol | -1 | -0.9987 | Normethadone | -1 | -0.9987 |
| Clofenotane | -1 | -1.0152 | Phenacetin | -1 | -0.9986 |
| Dantron | -1 | -0.9981 | Phthalylsulfathiazole | -1 | -1.0003 |
| Ethenzamide | -1 | -0.998 | Piromidicacid | -1 | -0.9982 |
| Floctafenine | -1 | -0.9991 | Sorbitol | -1 | -0.9967 |
| Glafenine | -1 | -0.9971 | Sulfadimidine | -1 | -0.9974 |
| Hexachlorophene | -1 | -1.0268 | Sulfamethoxypyridazine | -1 | -0.9982 |
| Methylphenobarbital | -1 | -0.9983 | Sulfamoxole | -1 | -0.9968 |
| Nifuroxazide | -1 | -0.9994 | Acidepipémidique | 1 | 0.9987 |
| Nitrofurantoin | -1 | -0.9984 | Bacampicilline | 1 | 0.9994 |
| Paracetamol | -1 | -0.999 | Céfaloridine | 1 | 1.0002 |
| Phenazone | -1 | -0.9983 | Céfapirine | 1 | 0.9978 |
| Piperylone | -1 | -1.0027 | Céfépime | 1 | 1.0004 |
| Propyphenazone | -1 | -0.999 | Céfotaxime | 1 | 1.001 |
| Salsalate | -1 | -0.9985 | Céfotiam | 1 | 1.0009 |
| Succinylsulfathiazole | -1 | -1.002 | Céfpodoximeproxétil | 1 | 0.9983 |
| Sulfadoxine | -1 | -0.9968 | Céfuroxime | 1 | 0.9988 |
| Sulfamerazine | -1 | -0.9995 | Chlortétracycline | 1 | 0.9992 |
| Sulfathiazole | -1 | -0.9997 | Cloxacilline | 1 | 1.0009 |
| Urapidil | -1 | -0.9929 | Dibekacine | 1 | 0.998 |
| Bénéthamine benzylpénicilline | 1 | 0.9981 | Isépamycine | 1 | 1.0019 |
| Céftizoxime | 1 | 0.9977 | Linézolide | 1 | 1.001 |
| Ciprofloxacine | 1 | 0.9989 | Méthacycline | 1 | 0.9991 |
| Loméfloxacine | 1 | 0.9939 | Nitrofurantoine | 1 | 0.9969 |
| Moxifloxacine | 1 | 0.9986 | Oxacilline | 1 | 1.001 |
| Netilmicine | 1 | 1.001 | Péfloxacine | 1 | 0.9987 |
| Oxytétracycline | 1 | 0.9979 | Pirazinamide | 1 | 0.992 |

# ANNEXE C

**Tableau C.1 :** Échantillon du tableau des valeurs propres fourni par STAISTICA® pour la première base

Val. Propres (matrice de corrél.) & stat. associées
(Feuille_de_données1_(Récupéré)) Variables actives seules

| | Val Propre | % Total - variance | Cumul - Val Propre | Cumul - % |
|---|---|---|---|---|
| **1** | 77,76652 | 47,13122 | 77,7665 | 47,1312 |
| **2** | 13,60209 | 8,24369 | 91,3686 | 55,3749 |
| **3** | 10,18420 | 6,17224 | 101,5528 | 61,5472 |
| **4** | 8,70843 | 5,27784 | 110,2612 | 66,8250 |
| **5** | 6,67416 | 4,04494 | 116,9354 | 70,8699 |
| **6** | 4,93353 | 2,99002 | 121,8689 | 73,8600 |
| **7** | 4,03655 | 2,44639 | 125,9055 | 76,3063 |
| **8** | 3,80906 | 2,30852 | 129,7145 | 78,6149 |
| **9** | 3,26502 | 1,97880 | 132,9796 | 80,5937 |
| **10** | 2,45548 | 1,48817 | 135,4350 | 82,0818 |
| **11** | 2,31657 | 1,40398 | 137,7516 | 83,4858 |
| **12** | 2,01219 | 1,21951 | 139,7638 | 84,7053 |
| **13** | 1,60005 | 0,96972 | 141,3638 | 85,6751 |
| **14** | 1,54832 | 0,93838 | 142,9122 | 86,6134 |
| **15** | 1,39034 | 0,84263 | 144,3025 | 87,4561 |

**Tableau C.2** : Échantillon de la matrice des coordonnées factorielles des variables pour quelques molécules

Coord. factorielles des var., basées sur les corrélations
(Feuille_de_données1_(Récupéré))

| | Fact. 1 | Fact. 2 | Fact. 3 | Fact. 4 |
|---|---|---|---|---|
| **AMW** | 0,213392 | 0,220731 | 0,006933 | -0,725753 |
| **MW** | -0,975473 | 0,059164 | 0,025407 | -0,092216 |
| **Sv** | -0,975169 | -0,040102 | 0,077983 | 0,154336 |
| **Se** | -0,964066 | 0,019810 | -0,005424 | 0,217513 |
| **Sp** | -0,971081 | -0,033423 | 0,078198 | 0,160501 |
| **Ss** | -0,947047 | 0,113521 | -0,088398 | -0,176845 |
| **Me** | 0,122401 | 0,369871 | -0,319805 | -0,650660 |
| **Mp** | 0,205015 | -0,169997 | 0,427098 | -0,606178 |
| **Ms** | 0,112412 | 0,456938 | -0,411376 | -0,447212 |
| **nAT** | -0,955742 | 0,005765 | 0,009045 | 0,248383 |
| **nSK** | -0,993101 | -0,003243 | 0,050796 | -0,014161 |
| **nBT** | -0,961926 | -0,028038 | 0,006308 | 0,227758 |
| **nBO** | -0,992457 | -0,070320 | 0,041608 | -0,035414 |
| **nBM** | -0,267870 | -0,125692 | 0,720636 | -0,420549 |

**Tableau C.3 :** Échantillon du tableau de régression et résidus des anti-inflammatoires

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : anti-inflammatoires | | | |
|---|---|---|---|
| | Valeur observée | Valeur prévue | Résidus |
| Acediasulfone | -1,00000 | -0,31398 | -0,68602 |
| Acefylline | -1,00000 | -0,82473 | -0,17527 |
| Acemetacin | 1,00000 | 0,49868 | 0,50132 |
| acétaminophène. | -1,00000 | -0,31674 | -0,68326 |
| Acetazolamide | -1,00000 | -1,07069 | 0,07069 |
| Acetiamine | -1,00000 | -0,96230 | -0,03770 |
| Acetylsalicylic acid | -1,00000 | 0,10632 | -1,10632 |
| Acetylsulfafurazole | -1,00000 | -0,16479 | -0,83521 |
| Aciclovir | -1,00000 | -1,05011 | 0,05011 |
| acide méfénamique | 1,00000 | 0,46818 | 0,53182 |
| Acide nalidixique | -1,00000 | -0,49858 | -0,50142 |
| Acide oxolinique | -1,00000 | -0,37158 | -0,62842 |
| Actarit | -1,00000 | -0,32968 | -0,67032 |
| Adefovir | -1,00000 | -0,91221 | -0,08779 |
| Alclofenac | 1,00000 | -0,02109 | 1,02109 |
| Alclometasone diprop | 1,00000 | 0,04011 | 0,95989 |
| Alminoprofen | 1,00000 | -0,03319 | 1,03319 |
| Alovudine | -1,00000 | -0,48304 | -0,51696 |
| Alphaprodine | -1,00000 | -0,03397 | -0,96603 |
| Altizide | -1,00000 | -0,90991 | -0,09009 |
| Amantadine | -1,00000 | -1,06970 | 0,06970 |
| Ambuside | -1,00000 | -1,21856 | 0,21856 |
| Amikacine | -1,00000 | -0,90008 | -0,09992 |
| Aminophenazone | 1,00000 | -0,43488 | 1,43488 |
| Amoxicillin | -1,00000 | -0,97816 | -0,02184 |
| Amphotericin B | -1,00000 | -1,44342 | 0,44342 |
| Ampicillin | -1,00000 | -0,87750 | -0,12250 |
| Ampiroxicam | 1,00000 | -0,29302 | 1,29302 |
| Amprenavir | -1,00000 | -1,06786 | 0,06786 |
| Anileridine | -1,00000 | -0,53100 | -0,46900 |
| Antrafenine | 1,00000 | 0,85336 | 0,14664 |
| Apalcillin | -1,00000 | -1,06627 | 0,06626 |
| Apigénine | 1,00000 | 0,16904 | 0,83096 |
| Arbidol | -1,00000 | -0,24209 | -0,75791 |

**Tableau C.4:** Échantillon du régression et résidus pour l'activité antibiotique

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : antibiotique | | | |
|---|---|---|---|
| | Valeur Observée | Valeur Prévue | Résidus |
| Acediasulfone | -1,00000 | -0,77072 | -0,22929 |
| Acefylline | -1,00000 | -0,61598 | -0,38402 |
| Acemetacin | -1,00000 | -0,86347 | -0,13653 |
| acétaminophène. | -1,00000 | -0,88845 | -0,11155 |
| Acetazolamide | -1,00000 | -1,07596 | 0,07596 |
| Acetiamine | -1,00000 | -0,92432 | -0,07568 |
| Acetylsalicylic acid | -1,00000 | -0,93798 | -0,06202 |
| Acetylsulfafurazole | -1,00000 | -0,85742 | -0,14258 |
| Aciclovir | -1,00000 | -0,78099 | -0,21901 |
| acide méfénamique | -1,00000 | -0,93601 | -0,06399 |
| Acide nalidixique | 1,00000 | -0,69933 | 1,69933 |
| Acide oxolinique | 1,00000 | -0,90970 | 1,90970 |
| Actarit | -1,00000 | -0,77646 | -0,22354 |
| Adefovir | -1,00000 | -1,10423 | 0,10423 |
| Alclofenac | -1,00000 | -0,67079 | -0,32921 |
| Alclometasone diprop | -1,00000 | -0,60460 | -0,39540 |
| Alminoprofen | -1,00000 | -0,60007 | -0,39993 |
| Alovudine | -1,00000 | -0,75524 | -0,24477 |
| Alphaprodine | -1,00000 | -1,14657 | 0,14657 |
| Altizide | -1,00000 | -0,68931 | -0,31069 |
| Amantadine | -1,00000 | -1,14250 | 0,14250 |
| Ambuside | -1,00000 | -0,40802 | -0,59198 |
| Amikacine | 1,00000 | 1,12081 | -0,12081 |
| Aminophenazone | -1,00000 | -0,84927 | -0,15073 |
| Amoxicillin | 1,00000 | 0,88231 | 0,11769 |
| Amphotericin B | 1,00000 | 1,05248 | -0,05248 |
| Ampicillin | 1,00000 | 0,79897 | 0,20103 |
| Ampiroxicam | -1,00000 | -0,63681 | -0,36319 |
| Amprenavir | -1,00000 | -0,78457 | -0,21543 |
| Anileridine | -1,00000 | -0,93110 | -0,06890 |
| Antrafenine | -1,00000 | -1,05231 | 0,05231 |
| Apalcillin | 1,00000 | 0,62407 | 0,37593 |
| Apigénine | -1,00000 | -0,84727 | -0,15273 |
| Arbidol | -1,00000 | -1,06172 | 0,06172 |

**Tableau C.5 :** Échantillon du Tableau de régression et résidus des analgésiques

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : analgésique | | | |
|---|---|---|---|
| | Valeur Observée | Valeur Prévue | Résidus |
| Acediasulfone | -1,00000 | -0,39045 | -0,60955 |
| Acefylline | -1,00000 | -0,58969 | -0,41031 |
| Acemetacin | -1,00000 | -0,51252 | -0,48748 |
| acétaminophène. | 1,00000 | -0,14439 | 1,14439 |
| Acetazolamide | -1,00000 | -1,12096 | 0,12096 |
| Acetiamine | 1,00000 | -0,67787 | 1,67787 |
| Acetylsalicylic acid | 1,00000 | -0,00075 | 1,00075 |
| Acetylsulfafurazole | -1,00000 | -0,79913 | -0,20087 |
| Aciclovir | -1,00000 | -0,75065 | -0,24935 |
| acide méfénamique | -1,00000 | -0,09016 | -0,90984 |
| Acide nalidixique | -1,00000 | -0,43042 | -0,56958 |
| Acide oxolinique | -1,00000 | -0,79415 | -0,20585 |
| Actarit | 1,00000 | -0,45843 | 1,45843 |
| Adefovir | -1,00000 | -1,07731 | 0,07731 |
| Alclofenac | 1,00000 | 0,13304 | 0,86696 |
| Alclometasone diprop | -1,00000 | -1,14428 | 0,14428 |
| Alminoprofen | 1,00000 | -0,10540 | 1,10540 |
| Alovudine | -1,00000 | -0,73687 | -0,26313 |
| Alphaprodine | 1,00000 | 0,46407 | 0,53593 |
| Altizide | -1,00000 | -0,96567 | -0,03433 |
| Amantadine | -1,00000 | -0,55728 | -0,44272 |
| Ambuside | -1,00000 | -0,66426 | -0,33574 |
| Amikacine | -1,00000 | -0,82245 | -0,17755 |
| Aminophenazone | 1,00000 | 0,36991 | 0,63009 |
| Amoxicillin | -1,00000 | -0,87309 | -0,12691 |
| Amphotericin B | -1,00000 | -1,14132 | 0,14132 |
| Ampicillin | -1,00000 | -0,75193 | -0,24807 |
| Ampiroxicam | -1,00000 | -0,81924 | -0,18076 |
| Amprenavir | -1,00000 | -0,51038 | -0,48962 |
| Anileridine | 1,00000 | 0,41150 | 0,58850 |
| Antrafenine | 1,00000 | 0,62408 | 0,37592 |
| Apalcillin | -1,00000 | -1,00091 | 0,00091 |
| Apigénine | -1,00000 | -0,42123 | -0,57877 |
| Arbidol | -1,00000 | -0,11190 | -0,88810 |
| Aspoxicillin | -1,00000 | -0,77736 | -0,22264 |

**Tableau C.6:** Échantillon du tableau de régression et résidus des antiviraux

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : antivirale | | | |
|---|---|---|---|
| | Valeur Observée | Valeur Prévue | Résidus |
| Acediasulfone | -1,00000 | -1,09672 | 0,09671 |
| Acefylline | -1,00000 | -0,34890 | -0,65110 |
| Acemetacin | -1,00000 | -0,60480 | -0,39520 |
| acétaminophène. | -1,00000 | -0,79508 | -0,20492 |
| Acetazolamide | -1,00000 | -0,29996 | -0,70004 |
| Acetiamine | -1,00000 | 0,15484 | -1,15484 |
| Acetylsalicylic acid | -1,00000 | -0,74927 | -0,25073 |
| Acetylsulfafurazole | -1,00000 | -0,84444 | -0,15556 |
| Aciclovir | 1,00000 | -0,11518 | 1,11518 |
| acide méfénamique | -1,00000 | -0,83784 | -0,16216 |
| Acide nalidixique | -1,00000 | -0,82088 | -0,17912 |
| Acide oxolinique | -1,00000 | -0,45377 | -0,54623 |
| Actarit | -1,00000 | -0,93328 | -0,06672 |
| Adefovir | 1,00000 | 0,06622 | 0,93378 |
| Alclofenac | -1,00000 | -1,47946 | 0,47946 |
| Alclometasone diprop | -1,00000 | -0,44711 | -0,55289 |
| Alminoprofen | -1,00000 | -1,06831 | 0,06831 |
| Alovudine | 1,00000 | 0,24877 | 0,75123 |
| Alphaprodine | -1,00000 | -0,83954 | -0,16046 |
| Altizide | -1,00000 | -1,30905 | 0,30905 |
| Amantadine | 1,00000 | -0,11170 | 1,11170 |
| Ambuside | -1,00000 | -1,66996 | 0,66996 |
| Amikacine | -1,00000 | -1,01750 | 0,01750 |
| Aminophenazone | -1,00000 | -0,94694 | -0,05306 |
| Amoxicillin | -1,00000 | -0,92357 | -0,07643 |
| Amphotericin B | -1,00000 | -1,69921 | 0,69921 |
| Ampicillin | -1,00000 | -0,97555 | -0,02445 |
| Ampiroxicam | -1,00000 | -0,93636 | -0,06364 |
| Amprenavir | 1,00000 | -0,11297 | 1,11297 |
| Anileridine | -1,00000 | -0,70382 | -0,29618 |
| Antrafenine | -1,00000 | 0,15913 | -1,15913 |
| Apalcillin | -1,00000 | -0,53704 | -0,46296 |
| Apigénine | -1,00000 | -0,77752 | -0,22248 |

**Tableau C.7 :** Échantillon du Tableau de régression et résidus des diurétiques

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : diurétique | | | |
|---|---|---|---|
| | Valeur Observée | Valeur Prévue | Résidus |
| Acediasulfone | -1,00000 | -0,53942 | -0,46058 |
| Acefylline | 1,00000 | -0,54978 | 1,54978 |
| Acemetacin | -1,00000 | -0,81006 | -0,18994 |
| acétaminophène. | -1,00000 | -0,49368 | -0,50632 |
| Acetazolamide | 1,00000 | 0,28037 | 0,71963 |
| Acetiamine | -1,00000 | -0,26278 | -0,73722 |
| Acetylsulfafurazole | -1,00000 | -0,24086 | -0,75914 |
| Aciclovir | -1,00000 | -0,67902 | -0,32098 |
| acide méfénamique | -1,00000 | -0,66754 | -0,33246 |
| Acide nalidixique | -1,00000 | -0,55578 | -0,44423 |
| Actarit | -1,00000 | -0,56811 | -0,43189 |
| Adefovir | -1,00000 | -0,78009 | -0,21991 |
| Alclofenac | -1,00000 | -0,48581 | -0,51419 |
| Alclometasone diprop | -1,00000 | -0,97150 | -0,02850 |
| Alovudine | -1,00000 | -0,81211 | -0,18789 |
| Alphaprodine | -1,00000 | -0,69444 | -0,30556 |
| Altizide | 1,00000 | -0,28531 | 1,28531 |
| Amantadine | -1,00000 | -0,94238 | -0,05762 |
| Ambuside | 1,00000 | 0,24098 | 0,75902 |
| Amikacine | -1,00000 | -0,87086 | -0,12914 |
| Aminophenazone | -1,00000 | -0,63020 | -0,36980 |
| Amoxicillin | -1,00000 | -0,95565 | -0,04435 |
| Amphotericin B | -1,00000 | -0,94238 | -0,05762 |
| Ampicillin | -1,00000 | -0,94923 | -0,05077 |
| Ampiroxicam | -1,00000 | -0,69067 | -0,30933 |
| Amprenavir | -1,00000 | -0,86615 | -0,13385 |
| Anileridine | -1,00000 | -1,04001 | 0,04001 |
| Antrafenine | -1,00000 | -1,23991 | 0,23991 |
| Apalcillin | -1,00000 | -1,11667 | 0,11667 |
| Apigénine | -1,00000 | -0,80543 | -0,19457 |
| Arbidol | -1,00000 | -0,66909 | -0,33091 |
| Aspoxicillin | -1,00000 | -0,91711 | -0,08289 |
| Atazanavir | -1,00000 | -0,85810 | -0,14190 |

**Tableau C.8 :** Échantillon du Tableau de régression et résidus des glucocorticoïdes.

| Valeurs Prévues & Résidus (Feuille_de_données1_(Récupéré)) Var. dépendante : glucocorticoïde | | | |
|---|---|---|---|
| | Valeur Observée | Valeur Prévue | Résidus |
| Acediasulfone | -1,00000 | -0,95851 | -0,041485 |
| Acefylline | -1,00000 | -1,07331 | 0,073312 |
| Acemetacin | -1,00000 | -0,94922 | -0,050781 |
| acétaminophène. | -1,00000 | -0,96632 | -0,033685 |
| Acetazolamide | -1,00000 | -1,15396 | 0,153956 |
| Acetiamine | -1,00000 | -1,29479 | 0,294786 |
| Acetylsalicylic acid | -1,00000 | -1,00534 | 0,005342 |
| Acetylsulfafurazole | -1,00000 | -1,05676 | 0,056757 |
| Aciclovir | -1,00000 | -1,00662 | 0,006622 |
| acide méfénamique | -1,00000 | -0,97311 | -0,026887 |
| Acide nalidixique | -1,00000 | -1,03092 | 0,030920 |
| Acide oxolinique | -1,00000 | -1,01002 | 0,010017 |
| Actarit | -1,00000 | -1,02032 | 0,020316 |
| Adefovir | -1,00000 | -0,97560 | -0,024396 |
| Alclofenac | -1,00000 | -0,89132 | -0,108680 |
| Alclometasone diprop | -1,00000 | -0,45735 | -0,542646 |
| Alminoprofen | -1,00000 | -0,91109 | -0,088908 |
| Alovudine | -1,00000 | -0,74911 | -0,250893 |
| Alphaprodine | -1,00000 | -0,93945 | -0,060550 |
| Altizide | -1,00000 | -0,79019 | -0,209814 |
| Amantadine | -1,00000 | -1,01661 | 0,016607 |
| Ambuside | -1,00000 | -1,18352 | 0,183517 |
| Amikacine | -1,00000 | -1,07589 | 0,075891 |
| Aminophenazone | -1,00000 | -0,98899 | -0,011007 |
| Amoxicillin | -1,00000 | -1,06799 | 0,067987 |
| Amphotericin B | -1,00000 | -0,63563 | -0,364370 |
| Ampicillin | -1,00000 | -1,06494 | 0,064936 |
| Ampiroxicam | -1,00000 | -0,64301 | -0,356988 |
| Amprenavir | -1,00000 | -0,93554 | -0,064463 |
| Anileridine | -1,00000 | -1,08737 | 0,087373 |
| Antrafenine | -1,00000 | -1,14554 | 0,145539 |
| Apalcillin | -1,00000 | -0,90563 | -0,094369 |