



République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE DE GUELMA

*FACULTE DES SCIENCES ET DE L'INGENIERIE*

*Département de Génie des Procédés*

## THESE DE DOCTORAT

*Discipline : Chimie Industrielle*

**Sujet :**

**ETUDES DE QSAR SUR DES ACTIVITES BIOLOGIQUES UTILISANT DES  
PRODUITS D'ORIGINES NATURELS**

*Présentée par :*

**Khairedine KRAIM**

Soutenue le .../11/2009 devant les membres du jury :

Président :	M.	ABDAOUI	Professeur	Université 8 Mai 1945 (Guelma)
Rapporteur :	Dj.	KHATMI	Professeur	Université 8 Mai 1945 (Guelma)
Examineur :	F.	FERKOUS	Professeur	Université B. M. (ANNABA)
Examineur :	A.	BOUCEKKINE	Professeur	Université Renne (France)
Examineur :	Z.	DJEGHABA	Professeur	Université B. M. (ANNABA)
Examineur :	R.	NEMAMCHA	M.C	Université 8 Mai 1945 (Guelma)

**REMERCIEMENTS**

Cette thèse a été réalisée à l'université 08 Mai 1945 GUELMA, sous la direction du Professeur *KHATMI*. Je tiens à lui exprimer mes plus sincères remerciements pour son accueil chaleureux et pour la confiance qu'il m'a accordée au cours de ces quatre années. Je lui suis aussi reconnaissant de l'attention qu'il a portée à mon travail pendant ces années. Ainsi que des nombreux conseils avisés qu'il m'a prodigué au cours de longues discussions.

Je voudrais ensuite remercier Monsieur *F. FERKOUS* professeur à l'université d'Annaba de m'avoir fait l'honneur de collaborer à ce travail duquel j'ai pu bénéficier de ses conseils et de son expérience, et pour sa présence à ce jury de thèse.

Je remercie sincèrement Monsieur *M. ABDAOUI* professeur à l'université de Guelma d'avoir bien voulu présider le jury de cette thèse.

Mes plus sincères remerciements sont aussi adressés à Monsieur *A. BOUCEKKINE* professeur à l'université de Renne (France), ainsi qu'à Monsieur *R. NEMAMCHA* professeur à l'université de Guelma, et à Monsieur *Z. DJEGHABA* professeur à l'université d'Annaba pour avoir accepté et pris le temps de juger ce travail.

Ce travail n'aurait jamais abouti, sans la participation de certains jeunes chercheurs, je citerai en particulier Monsieur Youcef *SAIHI* ; Mouloud *BOUCHOUKA* ; Fateh *BOUCHAMA* ; Mohamed *BRAHIMI* ; A- Halim *ZEGHAD* ; et Lotfi *TAZIR* dont la gentillesse et la disponibilité n'ont jamais fait défaut.

Je voudrais remercier tout le personnel du département de génie des procédés de l'université de Guelma, ainsi qu'au personnel du département de chimie de l'université d'Annaba, qui m'ont encouragé beaucoup de travailler.

## Résumé

Les modèles quantitatifs de QSAR présentent une solution statistique du problème de la difficulté du calcul direct des propriétés physiques et biologiques à partir de la structure.

L'intérêt d'un modèle de QSAR est de tirer des informations à partir de l'ensemble des descripteurs numériques caractérisant la structure moléculaire et prédire ainsi les activités biologiques de nouvelles structures.

Cette thèse décrit, en deux parties, la méthodologie employée pour obtenir des relations quantitatives, structure-activité, et pour le développement des modèles QSAR à partir de différents ensembles de molécules.

La première partie est divisée en trois chapitres : le premier chapitre rapporte des généralités sur les composés naturels ( xanthones, flavanoides, curcuminoïdes), le second chapitre décrit la méthodologie de QSAR et le troisième chapitre rapporte une introduction générale sur les descripteurs moléculaires calculés à partir du serveur E-DRAGON1.

La seconde partie est consacrée pour l'application de la méthodologie de QSAR, pour modéliser l'activité inhibitrice de l' $\alpha$ -glucosidase, exprimée par la grandeur  $IC_{50}$ , à partir d'un ensemble de 57 molécules, dérivés des xanthones et curcuminoïdes en utilisant la méthode de régression linéaire multiple et les algorithmes génétiques sont utilisés dans le développement des modèles en tant que méthode d'apprentissage et de sélection respectivement.

Deux modèles de QSAR sont développés, le premier modèle est obtenu par l'utilisation de la totalité des descripteurs issus de E-DRAGON1, alors que le deuxième modèle est obtenu en utilisant seulement les descripteurs de la famille 3D-MORSE.

Les résultats obtenus de la validation et l'analyse des valeurs résiduelles normalisées prouvent la validité, la stabilité et la robustesse des deux modèles obtenus. Ils peuvent expliquer la variance des valeurs de l'activité biologique observées avec des pourcentages de 85,7% et 80,5 % respectivement.

Une deuxième étude de QSAR est réalisée sur un ensemble de 24 composés dérivés de flavonoïdes afin de modéliser leur activité inhibitrice contre le VIH1 (exprimée par la grandeur  $pIC_{50}$ ). Une étude comparative a été réalisée entre les deux méthodes de sélection : les algorithmes génétiques (GA) et l'algorithme ascendant pas à pas (Forward Stepwise).

Le modèle obtenu par GA a donné les meilleurs paramètres statistiques, pour modéliser l'activité  $pIC_{50}$ , par rapport à l'algorithme ascendant. Le modèle proposé a prouvé sa robustesse, sa bonne précision ainsi qu'une bonne stabilité après vérification par la validation interne et externe. Il peut expliquer plus de 88% de la variance des valeurs de l'activité biologique observées.

**Mots clés:** *QSAR, Régression Linéaire Multiple; Algorithmes Génétique; Algorithme ascendant pas à pas;  $\alpha$ -glucosidase; VIH-1; xanthones; curcuminoïdes; flavonoïdes; domaine d'applicabilité ; techniques de validation; test de randomisation.*

---

## Abstract

Quantitative structure- activity relationship (QSAR) models are a statistical solution to the problem of directly calculating physical and biological properties of molecules from their structure. The goal of a QSAR model is to extract information from a set of numerical descriptors characterizing molecular structure and use this information to inductively develop a relationship between structure and activity. The focus of the work reported in this dissertation is on the validation and interpretation of QSAR models and presents both applications of interpretation techniques as well as the development of validation and interpretation methodologies.

This thesis describes, by two parts, the methodology used to obtain quantitative structure – activity relationships and the development of QSAR models for several different sets of compounds. The first part was divided into three chapters cover the methodology involved in QSAR, an introduction to descriptors calculated by E-DRAGON1 server, and the last was attributed to the three natural compound derivatives: flavonoids, xanthones, and curcuminoids.

The second part of the dissertation is composed of tow application studies. The first study involves the prediction of  $\alpha$ - glucosidase inhibition. Thus, 57 xanthone and curcuminoid derivatives were evaluated as  $\alpha$ - glucosidase inhibitors, expressed by the inhibitory of these compounds ( $IC_{50}$ ). Based on these data, different molecular descriptors were used to solve this problem. A linear QSAR model was developed using Multiple Linear Regression technique, while Genetic Algorithm was adopted for selecting the most appropriate descriptors. The predictive activity of the model was evaluated by means of external validation set and the Y- randomization technique, and its structural chemical domain has

been verified by the leverage approach. It was able to describe more than 85.7 % of the variance in the experimental activity.

Also, the 3D Molecule Representation of Structures Based on Electron Diffraction (3D-MoRSE) approach has been applied to the study of the  $\alpha$ -glucosidase inhibitory activity of xanthone and curcuminoid analogues. A model capable of describing around 80.50% of the variance in the experimental activity of 45 analogues of these compounds was developed with the use of the mentioned approach and dataset. The predictive activity of the model was evaluated by means of external validation set and the Y-randomization technique, and its structural chemical domain has been verified by the leverage approach. In comparison with other descriptor classes, the model relative to the 3D-MoRSE descriptors was considered as the best.

In the second study, a structure activity relationship (QSAR) analysis was applied to a series of 24 flavonoid derivatives to predict their anti HIV-1 (IC<sub>50</sub>) activity. The best two models with four variables in each one to predict IC<sub>50</sub> have been drawn up with the help of Forward Stepwise (FS) and Genetic Algorithms (GA) as variable selection methods. The best results were achieved by the use of the GA, where the best obtained model was able to explain **88.20%** for pIC<sub>50</sub> of the experimental variance. The predictive ability of the model was evaluated by means of cross validation (leave one out, leave group out) and the Y-randomization techniques, and its structural chemical domain has been verified by the leverage approach.

**Keywords:** *QSAR, Multiple Linear Regression; Genetic Algorithms; Forward stepwise;  $\alpha$ -glucosidase; HIV-1; xanthone; curcuminoids; flavonoids; applicability domain; validation techniques; y-randomization test.*

## Table des matières

	page
<b>INTRODUCTION GENERALE</b>	<b>1</b>
<b>CHAPITRE I : FLAVONOÏDES, XANTHONES ET CURCUMINOIDES</b>	
<b>I. LES FLAVONOÏDES</b>	7
1. Structures chimiques des flavonoïdes	8
2. Les flavonoïdes et la coloration des fleurs	12
3. Activités biologiques des flavonoïdes dans les règnes végétal et animal	13
4. Activité antimicrobienne des flavonoïdes	14
5. Propriétés médicinales des flavonoïdes	16
<b>II. LES XANTHONES</b>	18
1. Classification des Xanthones	19
2. Utilisation thérapeutique des xanthones	21
<b>III. LES CURCUMINOIDES</b>	24
1. Introduction	24
2. Les dérivés des Curcuminoides	25
3. Utilisation thérapeutique des Curcuminoides :	27
<b>IV. REFERENCES</b>	29
<b>CHAPITRE II : METHODOLOGIE DE QSAR</b>	
<b>I. INTRODUCTION</b>	32
<b>II. PRESENTATION DE LA METHODE DE QSAR</b>	33
1. Dessin et optimisation de la Structure	35
2. Génération des descripteurs	35
3. Sélection des variables (les descripteurs)	36
A. La sélection objective	36
B. La sélection subjective	37
4. Développement du modèle	39
A. Régression Linéaire Multiple	40
B. Réseau de neurones artificiels	41
5. Validation du modèle	42
A. validation interne	43
B. Test de randomisation	44
C. Validation Externe	44
6. Domaine d'applicabilité	45
<b>III. CONCLUSION</b>	47
<b>IV. REFERENCES</b>	48
<b>CHAPITRE III : INTRODUCTION AUX DESCRIPTEURS GENERES PAR E- DRAGON1</b>	
<b>I. INTRODUCTION</b>	50
<b>II. DEFINITION ET CLASSIFICATION DES DESCRIPTEURS MOLECULAIRES</b>	50

<b>III. NOTIONS DE BASE SUR LES MATRICES DE CODAGE STRUCTURALE</b>	51
<b>IV. DESCRIPTEURS ISSUS DU SERVEUR E-DRAGON1</b>	54
1. Les descripteurs Constitutionnels	54
2. Les descripteurs topologiques	55
A. Descripteurs dérivés de la matrice d'adjacence	55
B. Descripteurs dérivés de la matrice de distance	56
C. Descripteurs dérivés de la matrice de distance pondérée	57
D. Descripteurs dérivés de la matrice Laplacienne	58
3. L'indice de connectivité	59
4. Les descripteurs géométriques	60
5. Représentation 3d des structures moléculaires basée sur la diffraction électronique	62
6. Fonction De Distribution Radiale	63
7. Descripteurs issus de la valeur propre de Burden	64
8. Les descripteurs WHIM	65
9. Les descripteurs GETAWAY	66
10. Les descripteurs « Walk and Path Counts »	68
11. Descripteurs topologiques de charge	68
12. Les Descripteurs 2D- Autocorrelations	69
A. Autocorrelation de la structure Topologique (ATS)	69
B. Autocorrelation de la structure Topologique de Moran (MATS)	70
C. Autocorrelation de la structure Topologique de Geary (GATSkw)	70
<b>V- CONCLUSION</b>	71
<b>VI- REFERENCES</b>	72
<b>CHAPITRE IV: MODELISATION DE L'INHIBITION DE L'<math>\alpha</math>-GLUCOSIDASE PAR LES DERIVES DES XANTHONES ET CURCUMINOIDES</b>	
<b>I. INTRODUCTION</b>	73
<b>II. METHODES EXPERIMENTALES</b>	74
1. Ensemble des données	74
2. Dessin et optimisation des structures	78
3. Génération des Descripteurs	78
4. Sélection des variables et formation du modèle	78
<b>III. RESULTATS ET DISCUSSION</b>	80
<b>1. Sélection du modèle optimale utilisant l'ensemble de descripteurs (1664)</b>	80
A. Analyse de la justesse du modèle optimal	83
B. Tests de colinéarité et Multicolinéarité	86
C. Analyse des valeurs résiduelles normalisées	88
D. Validation interne et externe du modèle de QSAR obtenu	90
E. Analyse des points aberrants sur l'axe des Y et X	93
F. Domaine d'applicabilité	95
<b>2. Le modèle des descripteurs de la famille 3D-MORSE</b>	96
A. Analyse de la justesse du modèle optimal	100
B. Tests de colinéarité et de Multicolinéarité	103

---

C. Analyse des valeurs résiduelles normalisées	103
D. Validation interne et externe	105
E. Analyse des points aberrants sur l'axe des Y et X	107
F. Domaine d'applicabilité	109
<b>3. Interprétation chimique de l'influence des descripteurs sur l'activité inhibitrice contre <math>\alpha</math>-glucosidase</b>	<b>111</b>
<b>IV. CONCLUSION</b>	<b>115</b>
<b>V. REFERENCES</b>	<b>116</b>
<b>Chapitre V: MODELISATION DE L'INHIBITION DU VIH-1 PAR LES DERIVES DES FLAVONOÏDES</b>	
<b>I. INTRODUCTION</b>	<b>118</b>
<b>II. METHODES EXPERIMENTALES</b>	<b>119</b>
1. Ensemble des données	119
2. Dessin et optimisation des structures	121
3. Génération des Descripteurs	121
4. Sélection des variables et formation du modèle	122
<b>III. RESULTATS ET DISCUSSION</b>	<b>122</b>
1. Sélection des variables	122
2. Analyse de la justesse du modèle	123
3. Tests de colinéarité et de Multicolinéarité	125
4. Analyse des valeurs résiduelles normalisées	126
5. Validation du modèle	127
6. Domaine d'applicabilité	128
7. Interprétation chimique de l'influence des descripteurs sur l'activité inhibitrice contre le VIH1	130
<b>IV. CONCLUSION</b>	<b>131</b>
<b>V. REFERENCES</b>	<b>132</b>
<b>CONCLUSION GENERALE</b>	<b>133</b>
<b>Annexe</b>	

**Liste des tableaux**

	page
<b>Chapitre I</b>	
Tableau 1. Quelques exemples de flavones	10
Tableau 2. Quelques exemples de flavonols	11
Tableau 3. Divers flavonoïdes	11
Tableau 4. Quelques couleurs associées aux flavonoïdes correspondants	13
Tableau 5. Quelques structures flavoniques ayant des propriétés antifongiques	14
Tableau 6. Flavonoïdes ayant des propriétés antibactériennes	15
Tableau 7. Quelques structures flavoniques ayant des activités antitumorale et cytotoxique	18
Tableau 8. Utilisation des dérivés du xanthone en tant que agents thérapeutiques	22
Tableau 9. Utilisation des dérivés de la curcumine comme agents thérapeutiques	27
<b>Chapitre III</b>	
Tableau 1. Descripteurs dérivés de la matrice d'adjacence	56
Tableau 2. Descripteurs dérivés de la matrice de distance	57
Tableau 3. Descripteurs dérivés de la matrice Laplacienne	58
<b>Chapitre IV</b>	
Tableau 1. Différentes classes de descripteurs	78
Tableau 2. Série des modèles optimums obtenus à différentes dimensions	81
Tableau 3. Valeurs pIC50 expérimentales et prédites pour TSET et PSET avec les descripteurs sélectionnés	82
Tableau 4. Résultats de l'analyse de la variance (ANOVA)	84
Tableau 5. Tableau des coefficients	85
Tableau 6. Matrice de corrélation de l'équation 1	87
Tableau 7. Valeurs des critères VIF et TF pour les descripteurs significatifs	88
Tableau 8. $R^2$ et $Q^2_{cv-100}$ issus du test de randomisation	92
Tableau 9 : Valeurs résiduelles normalisées et valeurs de levier	94
Tableau 10: comparaison entre six familles de descripteurs	97
Tableau 11. Les valeurs de pIC50 observées et prédites de TSET et PSET	99
Tableau 12. Résultats de l'analyse de la variance (ANOVA)	101
Tableau 13. Tableau des coefficients	102
Tableau 14. Matrice de corrélation de l'équation 2	103
Tableau 15. Valeurs des critères VIF et TF	103
Tableau 16. $R^2$ et $Q^2_{cv-100}$ issus du test de randomisation	105
Table 17. Valeurs résiduelles normalisées et valeurs de levier	108

Tableau 18. Influence des deux descripteurs (MATS7v et R4e+) sur l' $\alpha$ -glucosidase	113
---	-----

### Chapitre V

Tableau 1. Propriétés statistiques des deux modèles obtenus par les GA et FS	121
Tableau 2. Les valeurs pIC50 expérimentales et prévues avec les descripteurs sélectionnés	123
Tableau 3. Résultats de l'analyse de la variance (ANOVA)	123
Tableau 4. Tableau des coefficients	124
Tableau 4. Matrice de corrélation de l'équation 1	125
Tableau 5. Validation croisée par les deux procédures (leave one out et leave group out)	126
Table 6: $R^2$ et $Q^2_{cv-loo}$ issus du test de randomisation	127
Table 7. Valeurs résiduelles normalisées et valeurs de levier	128

**Liste des figures**

	page
<b>Chapitre I</b>	
Figure 1. Structure des flavonoïdes : A sont des isoflavones; et B sont des flavones et flavonols	8
Figure 2 : Les principaux aglycones des flavonoïdes	9
Figure 3. Deux inhibiteurs de l'enzyme VIH- transcriptase	15
Figure 4. Squelette de base des Xanthones	18
Figure 5. Dérives des xanthones biologiquement actifs	23
Figure 6. Les sources naturelles des dérivés des curcuminoïdes	24
Figure 7. Dérivés des curcuminoïdes	26
Figure 8. Dérivés des curcuminoïdes biologiquement actifs	28
<b>Chapitre II</b>	
Figure 1. Présentation des étapes aboutissant à QSAR	33
Figure 2. Les 5 étapes de base de la construction du modèle de QSAR	34
Figure 3 : Schéma représentatif d'un réseau de neurones artificiels	42
<b>Chapitre IV</b>	
Figure 1. Structures développées des dérivés du xanthone et du curcuminoïde	76
Figure 2. Les deux motifs inhibiteurs de l' $\alpha$ -glucosidase	77
Figure 3. Le nombre optimal de descripteurs du modèle optimal obtenu par GA	81
Figure 4. Evaluation de la distribution normale des valeurs résiduelles normalisées	89
Figure 5. Valeurs résiduelles normalisées en fonction des activités prédites (A) et observées (B)	90
Figure 6. Valeurs de $Q^2$ cv-loo en fonction de $R^2$ issues du test de randomisation	92
Figure 7. Valeurs prédites en fonction des valeurs observées pour TSET et PSET	93
Figure 8. Les structures des composés à grand valeurs de levier	95
Figure 9. Graphe de Williams obtenu avec les cinq descripteurs	96
Figure 10. Evaluation de la distribution normale des valeurs résiduelles normalisées	104
Figure 11. Valeurs résiduelles normalisées en fonction des activités prédites (A) et observées (B)	104
Figure 12. Valeurs de $Q^2$ cv-loo en fonction de $R^2$ issues du test de randomisation	106
Figure 13. Les valeurs prédites en fonction des valeurs observées pour les deux sous ensembles TSET et PSET	107

---

Figure 14. Les structures des composés à grand valeurs de levier	109
Figure15. Graphe de Williams obtenu avec les quatre descripteurs	110
Figure16 : Valeurs de levier en fonction des valeurs des 4 descripteurs (a, b, c, d)	110
Schéma 1. Représentation de l'influence de MATS7v et R4e+	112

## Chapitre V

Figure 1. Structures développées des dérivés du flavonoïde	119
Figure 2. Valeurs de score en fonction des valeurs résiduelles normalisées	125
Figure 3. Valeurs résiduelles normalisées en fonction de pIC50 prédites (A) et observées (B)	126
Figure 4. Valeurs de Q2 cv-loo en fonction de R <sup>2</sup> issues du test de randomisation	127
Figure 5. Graphe de Williams	128

**Liste des abréviations**

**QSAR** : Quantitative Structure Activity Relationship / Relation quantitative de structure-activité.

**MLR** : *Multiple Linear Regression/ Régression Linéaire Multiple.*

**GA** : Genetic Algorithms / Algorithmes génétiques.

**FS** : Forward Stepwise / Ascendant pas à pas.

**3D-MoRSE** : 3D Molecule Representation of Structures Based on Electron Diffraction/  
Représentation 3d des structures moléculaires basée sur la diffraction électronique.

**IC<sub>50</sub>**: concentration micro molaire d'une drogue, nécessaire pour inhiber 50% (la moitié) de l'activité enzymatique.

**VIH 1** : Virus Immunodéficientaire Humain.

## INTRODUCTION GENERALE

Le développement des médicaments est un processus lent et laborieux, exigeant un grand investissement. Une firme pharmaceutique aura besoin de huit à douze pour produire un médicament. Toute seule, la phase liée à la découverte « discovery phase » consomme une partie significative des ressources de la compagnie (pour effectuer la synthèse et les tests biochimiques) et celle-ci dure trois à cinq ans. Les dépenses pour la recherche et développement (Research and development) étaient estimées à 500 –600 millions de Dollars dans les années quatre-vingt pour dépasser les 900 millions de Dollars de nos jours, dont deux- tiers du coût sont dépensées pour la découverte des précurseurs potentielles (leads compound) qui ne surpassent malheureusement pas l'étape des tests précliniques<sup>1-2</sup>.

L'industrie pharmaceutique est emmené à réduire le temps et le coût et de produire de nouveaux produits ayant les propriétés thérapeutiques optimums et sans, si possible, d'effets secondaires indésirables, de la manière la plus efficace et la moins couteuse. Pour cette raison, l'industrie pharmaceutique continue sans cesse à chercher de nouveaux outils et des technologies modernes afin d'y parvenir<sup>3</sup>.

Aujourd'hui, après plusieurs années de développement et d'amélioration, l'ordinateur est devenu un outil indispensable dans les différents modes de vie. Sa technologie de pointe et ses utilisations répandues ont accéléré considérablement son développement. De ce fait, plusieurs nouvelles disciplines sont apparues dans presque tous les domaines scientifiques.

La chimie informatique constitue le grand exemple issu de ce développement où elle est maintenant reconnue comme subdivision de la chimie et soutien principal à la chimie

moderne. Spécialement dans la chimie des médicaments, plusieurs méthodes de recherche utilisant l'ordinateur ont été créées. Ces méthodes ont été développées pour aider les médecins chimistes (medicinal chemist) pour identifier ou concevoir des ligands et des précurseurs qui pourraient agir plus favorablement avec un récepteur donné augmentant ainsi le nombre de composés présentant une affinité puissante<sup>4,5</sup>.

QSAR (Quantitative structure Activity Relationship) est parmi les nouvelles techniques de modélisation, mettant en jeu des relations de structure avec l'activité. Elles prennent de plus en plus d'importance dans les études de la conception des précurseurs (lead compounds). Ce sont des techniques alternatives et complémentaires aux techniques expérimentales *in vitro* et *in vivo* dans les phases très précoces du développement des médicaments. Un outil qui permet une prédiction rapide d'une ou plusieurs propriétés biologique. Elle a pu être mise en place dans les laboratoires et utilisée dans l'industrie pharmaceutique. L'objectif d'une modélisation QSAR est de trouver des modèles précis, applicables et robustes afin de trouver une relation entre la structure et l'activité dans un but de prédiction mais également d'interprétation.

La puissance de la prévision de l'activité biologique utilisant l'outil QSAR l'a rendue indispensable dans les recherches pharmaceutiques. Il est intéressant de noter que QSAR a contribué principalement au succès de plusieurs molécules mises sur le marché en tant que médicaments. Parmi lesquelles nous citerons<sup>6</sup> : la quinolone carboxylic acid (Norfloxacin : antibactérien) ; l'acridines (Anticancéreux) ; la Cimitidine (antiulcère) ; le 3-amino-1-benzylpyrazolin-5-one (Muzolimine : diurétique) ; la Lomerizine (Antimigraine). Cette réussite, marquée par QSAR, a poussé les chercheurs de l'utiliser dans la prédiction de plusieurs activités biologiques (antitumorale ; anti-inflammatoire ; antioxydant ;...etc.) en

s'associant à d'autres domaines scientifiques tels que : l'informatique, statistique, mathématique, probabilité.

Cette thèse est consacrée au développement de modèles QSAR dans le but de prédire l'activité inhibitrice de deux ensembles de molécules contre l'enzyme  $\alpha$ -glucosidase et le VIH (virus immunodéficient humain).

Les modèles QSAR recherchés doivent présenter de bonnes caractéristiques : précision, justesse, stabilité et prévision. Soustraire les informations des descripteurs inclus dans chaque modèle est une tâche difficile mais également très pertinente, car elle nous permet de connaître les paramètres physicochimiques les plus influents sur l'activité et par conséquent, nous autorise à concevoir de nouvelles structures capables d'améliorer l'effet biologique.

Pour atteindre cet objectif, nous avons envisagé de produire de modèles QSAR pour un ensemble de molécules, les flavonoïdes, les xanthones et les curcuminoïdes. Ces modèles doivent s'articuler sur trois points essentiels (l'ensemble de molécules ; les descripteurs ; la sélection des descripteurs et l'analyse statistique).

1. Choix de l'ensemble des molécules : nous avons concentré notre étude sur les composés à base de flavonoïdes, xanthones et curcuminoïdes, qui sont des produits naturels. Ce choix est justifié par le succès marqué par ces produits naturels comme drogues<sup>7, 8</sup>.
2. Choix des descripteurs : nous avons utilisé les descripteurs générés par le serveur E-DRAGON1. Ce logiciel disponible sur net et ses descripteurs sont faciles à calculés. Leurs efficacités a été démontré lors de leurs utilisations dans plusieurs domaines de recherches<sup>9</sup> : environnement, médecine, industrie,...etc.

3. Choix de la méthode de sélection : les algorithmes génétiques ont été choisis comme méthode de sélection. Ce choix est justifié par le nombre élevé d'intervention de ces algorithmes dans notre vie quotidienne<sup>10</sup>. Egalement, nous avons utilisé l'analyse statistique par l'application « Régression Linéaire Multiple ». Cette technique présente l'avantage de pouvoir interpréter quantitativement l'influence de chacun des descripteurs du modèle sur l'activité .

Cette thèse est divisée en cinq chapitres précédés par une introduction générale:

Une partie Bibliographique contient trois chapitres :

- \* Chapitre1 intitulé : Flavonoïdes; les xanthones et les curcuminoïdes
- \* Chapitre2 intitulé : Méthodologie de QSAR.
- \* Chapitre3 intitulé : Introduction aux descripteurs générés par E-DRAGON1.

Et la partie du travail contient deux chapitres :

- \* Chapitre4 intitulé : Modélisation de l'inhibition de l' $\alpha$ -glucosidase par les dérivés du xanthone et curcuminoïde.
- \* Chapitre5 intitulé : Modélisation de l'inhibition du HIV-1 par les dérivés du Flavonoïde.

Et à la fin une conclusion générale.

**REFERENCES**

- [1] T. I. Oprea "*Chemoinformatics in Drug Discovery*" Ed. Wiley-vch Verlag. Allemagne, 2005.
- [2] E. A. Rezza ; P. N. Kourounakis "Chemistry and Molecular Aspects of Drug Design and Action" Ed. Taylor & Francis Group, LLC. Etats Unies, 2008.
- [3] S. Ekins "Computer Applications in Pharmaceutical Research and Development" Edt. John Wiley & Sons, Inc. Etats Unies, 2006.
- [4] F. Deanda "Development and Application of Software Tools for Computer-Assisted Drug Design" Thèse de doctorat; Août 1999. Université de Texas. Etats Unies.
- [5] K. Gubernator; H.J. Bohm "*Structure-Based Ligand Design*" Ed. Wiley-vch Verlag. Allemagne, 1998.
- [6] A. K. Ghose; V. N. Viswanadhan "Combinatorial Library Design and Evaluation: Principles, Software Tools, and Applications in Drug Discovery" Ed. Marcel Dekker, Inc. Etats Unies, 2001.
- [7] X. T. Liang; W. S. Fang "Medicinal Chemistry of Bioactive Natural Products" Ed. John Wiley & Sons, Inc. Etats Unies, 2006.
- [8] L. Zhang "Natural Products: *Drug Discovery and Therapeutic Medicine*" Ed. Humana Press Inc. Etats Unies, 2005.
- [9] J. Gasteiger "Handbook of Chemoinformatics: *From Data to Knowledge (4 Volumes)*" Ed. Wiley-vch Verlag, Weinheim. Allemagne, 2003.

- [10] B. Sanghamitra; K. P. Ankar “Classification and Learning Using Genetic Algorithms: *Applications in Bioinformatics and Web Intelligence*” Ed. Springer- Verlag. Allemagne, 2007.

**CHAPITRE I :****FLAVONOÏDES, XANTHONES ET CURCUMINOÏDES****I. LES FLAVONOÏDES****1. Introduction**

Les flavonoïdes constituent un groupe important de substances très répandues dans la nature. Ils représentent un groupe principal des antioxydants et ils forment également une importante famille de colorants naturels où dominent le jaune (*flavones*), le rouge ou le bleu (*anthocyanes*). Ces trois couleurs de base peuvent être modifiées par d'autres pigments (chlorophylles, caroténoïdes...), par chélation avec certains métaux ou par des variations de pH<sup>1</sup>. Les plus connus sont les citroflavonoïdes. Ils se trouvent dans les écorces d'agrumes: oranges, citrons ou pamplemousses. La peau de l'orange contient de minuscules vésicules baignant dans un tissu de soutien, appelé flavedo, qui doit sa couleur jaune orangée aux flavanones. En dessous de cette fine couche colorée se trouve une seconde couche blanche appelée albedo qui ne contient aucun flavanone soluble. C'est la couche externe des écorces d'orange, le flavedo, qui a prêté son nom aux flavonoïdes<sup>2</sup>.

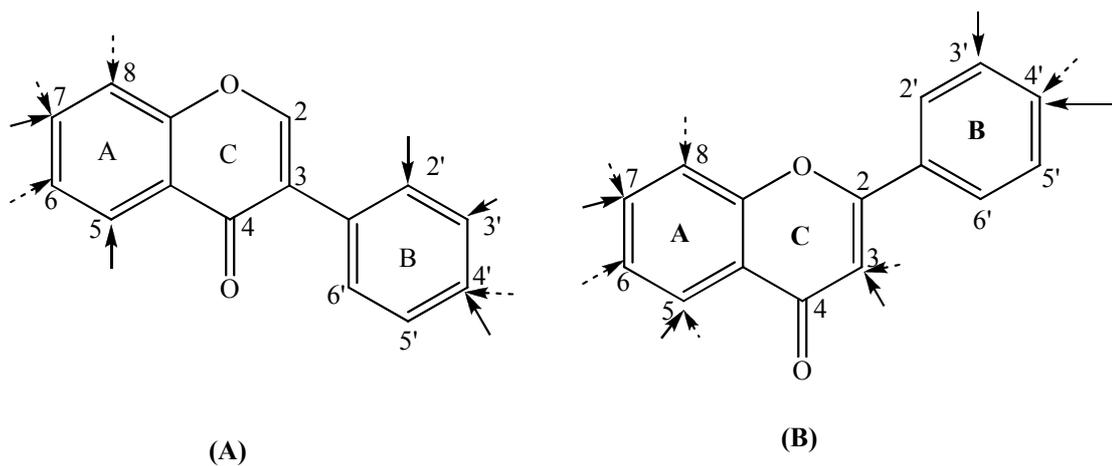
Pendant que les botanistes proposaient une classification de ce groupe des flavonoïdes, la vitamine C était découverte en 1936 par A.Szent-Gyorgi. Il a pu démontrer que les agrumes renferment, outre l'acide ascorbique (vitamine C), un autre facteur responsable de la résistance capillaire. Ce second facteur fut isolé de l'écorce de citron sous le nom de citrine en 1937 et dénommé vitamine P ou vitamine de perméabilité. D'autres auteurs parleront de facteurs C1 (acide ascorbique) et de facteurs C2 (noyau carboné de la flavone commun à tous les flavonoïdes).

Enfin, quelques années plus tard, les progrès de la biochimie permettaient de décrire leur structure moléculaire. C'est ainsi que l'on a découvert que les flavonoïdes appartenaient biochimiquement à la famille des benzopyrones, celle-ci étant scindée en deux sous-classes<sup>3</sup> :

- les  $\alpha$ -benzopyrones : les principaux dérivés utilisés en thérapeutique sont la coumarine ou les antivitamines K coumariniques;
- les  $\gamma$ -benzopyrones : les 2- $\gamma$ -phénylbenzopyrones constituent à proprement parler les flavonoïdes.

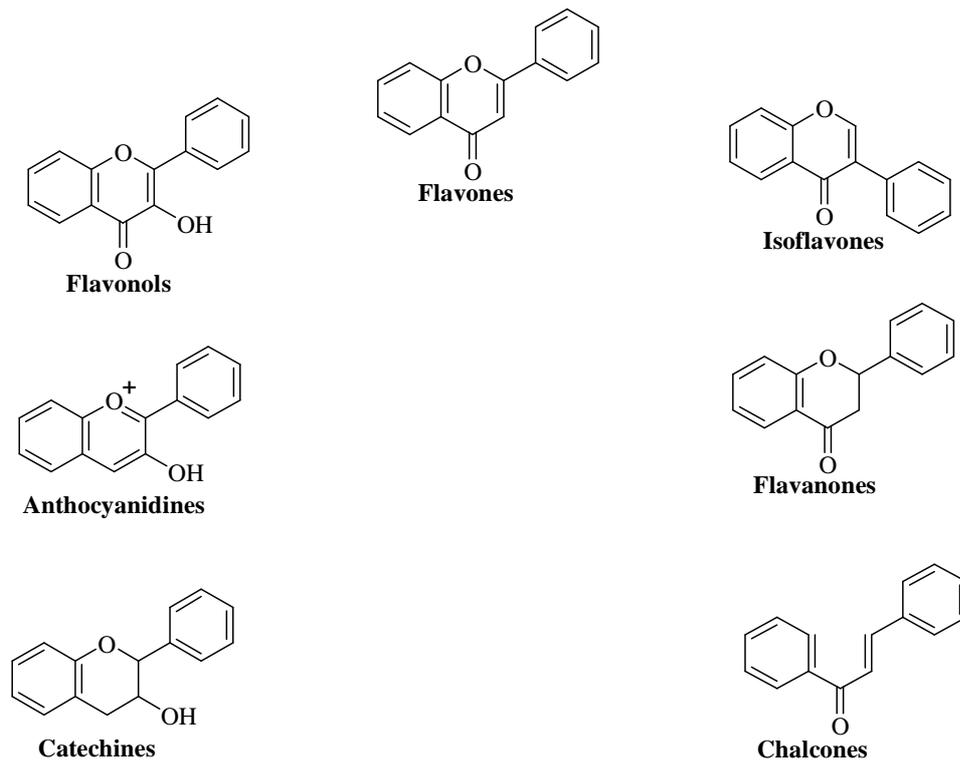
## 2. Structures chimiques des flavonoïdes

Les composés flavonoïdes sont formés d'un squelette de base à 15 carbones (C6-C3-C6) (Figure 1).



**Figure 1.** Structure des flavonoïdes : **A** sont des isoflavones; et **B** sont des flavones et flavonols. Les flèches pleines indiquent les emplacements fréquents d'hydroxylation et les flèches à tiret indiquent les emplacements fréquents d'o- et/ou de c- glycosylation.

Ces composés existent sous forme d'aglycones (génines) ou sous forme de glycosides et plus de 4000 structures sont connues à ce jour<sup>4</sup>. Les principaux aglycones sont représentés dans la figure 2.



**Figure 2 :** Les principaux aglycones des flavonoïdes

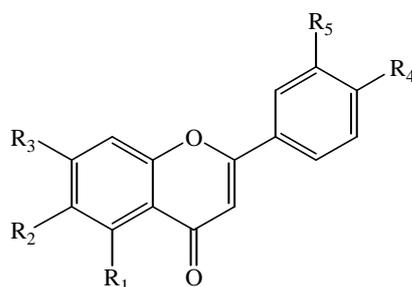
Tous les flavonoïdes peuvent être regroupés en une douzaine de classes selon le degré d'oxydation du noyau pyranique central<sup>5</sup>. Parmi ces classes on peut citer les plus répandues.

- Les *flavones* sont des 2-phénylchromones, incolores ;
- Les *isoflavones* sont des 3-phénylchromones, beaucoup moins répandues que les flavones;
- Les *flavonols* sont des 3-hydroxyflavones. Ce sont des pigments végétaux que l'on trouve souvent sous forme de glycosides ;
- Les *flavanones* sont des 2,3-dihydroflavones ;
- Les *chalcones* sont des isomères des flavanones avec ouverture du noyau pyronique entre les positions 1 et 2 ;
- Les *anthocyanidols* sont des dérivés réduits des flavonols avec formation d'un oxonium;

- Les *flavanes* ou catéchines sont également des produits de réduction, au moins formellement, des flavonols.

Les flavonoïdes les plus étudiés appartiennent aux groupes des flavones, des flavonols, en particulier, la quercétine et son hétéroside la rutine, mais aussi à ceux des flavanes, flavanones et chalcones.

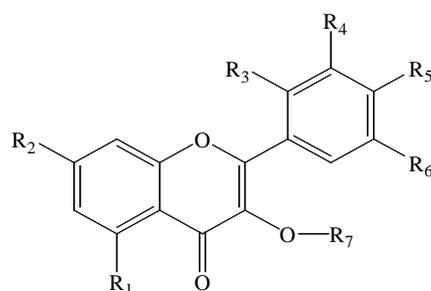
- *Groupe des flavones*



**Tableau 1.** Quelques exemples de flavones

Noms	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>
Flavones	H	H	H	H	H
Chrysin	OH	H	OH	H	H
Apigénine	OH	H	OH	OH	H
Cirsiliol	OH	O-CH <sub>3</sub>	O-CH <sub>3</sub>	OH	OH
Pédalitrine	OH	OH	O-CH <sub>3</sub>	OH	OH
Baïcaléine	OH	OH	OH	H	H

- *Groupe des flavonols*

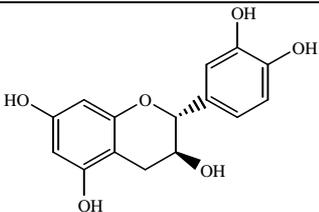
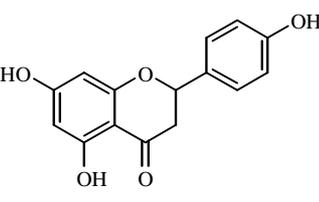
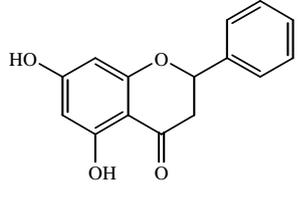


**Tableau 2.** Quelques exemples de flavonols

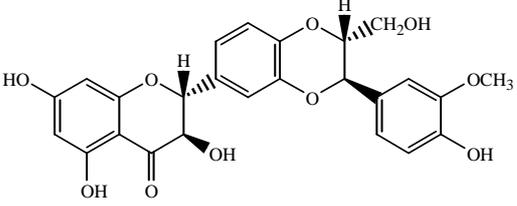
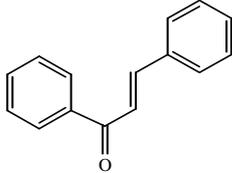
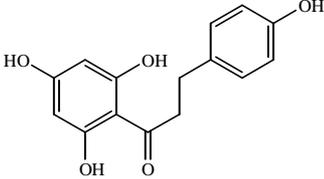
Noms	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	R <sub>7</sub>
Quercétine	OH	OH	H	OH	OH	H	H
Rutine	OH	OH	H	OH	OH	H	Rutinoose
Kaempférol	OH	OH	H	H	OH	H	H
Galangine	OH	OH	H	H	H	H	H
Morine	OH	OH	OH	H	OH	H	H
Fisétine	H	OH	H	OH	OH	H	H
Myricétine	OH	OH	H	OH	OH	OH	H

- *Divers flavonoïdes*

**Tableau 3.** Divers flavonoïdes

GROUPES	NOMS	STRUCTURES
Flavanes	Catéchine	
Flavanones	Naringénine	
	Pinocembrine	

**Tableau 3** (suite). Divers flavonoïdes

GROUPES	NOMS	STRUCTURES
Flavanonols (Silymarines (isomères))	Silybine	
Chalcones	Chalcone	
Dihydrochalcones	Phlorétine	

### 3. Les flavonoïdes et la coloration des fleurs

Une des fonctions majeures des flavonoïdes est sa contribution à la couleur des plantes et notamment à celle des fleurs. Par exemple, la couleur bleue dans les pétales est due à la présence d'une anthocyane, la delphinidine. Cependant, la plupart des glycosides de la delphinidine sont de couleur mauve et le changement vers la couleur bleue exige souvent la présence d'un co-pigment d'une flavone et parfois d'un ou plusieurs cations de métaux<sup>6</sup>. La couleur bleue exerce un effet attracteur sur les insectes tels que les abeilles ainsi que les oiseaux pollinisateurs, assurant par ce biais une étape fondamentale de la reproduction des plantes.

Le tableau suivant rassemble des couleurs de fleurs associées aux flavonoïdes correspondants<sup>7</sup>.

**Tableau 4.** Quelques couleurs associées aux flavonoïdes correspondants

Couleur	Flavonoïde
• Mauve et bleue	Delphinidine
• Magenta	Cyanidine
• Rose et orange	Pelargonidine

#### 4. Activités biologiques des flavonoïdes dans les règnes végétal et animal

Dés 1938, Szent-Gyorgyi montrait le rôle important joué par les flavonoïdes dans la respiration des végétaux qui les synthétisent. On pense maintenant qu'ils exerceraient cette action en tant que catalyseur de transport d'électrons au cours de la photosynthèse, mais pourraient aussi être impliqués comme régulateurs de la phosphorylation des canaux ioniques<sup>8</sup>.

D'autre part, dans le règne animal, l'apport en flavonoïdes qui peut avoir un rôle biologique important, compte tenu de leurs nombreuses propriétés mises en évidence chez l'animal. On a observé, par exemple, une accumulation de flavones et de leurs glucosides dans les ailes de papillons qu'ils concourent à colorer, cette coloration pourrait être impliquée dans les mécanismes de reconnaissance. En fait, ces flavones proviennent des feuilles dont se sont nourries leurs chenilles<sup>9</sup>.

On trouve aussi des flavonoïdes et en particulier de la piocembrine, de la quercétine, de la chrycine et de la galangine dans la propolis des abeilles. Ces insectes la fabriquent à partir des sécrétions des bourgeons de nombreux arbres dont le Bouleau, l'Aulne, l'Epicéa, le Sapin, le Saule, l'Orme et la modifient par leurs enzymes salivaires<sup>10</sup>. Elles mettent instinctivement à profit ses propriétés antifongiques et antibactériennes pour aseptiser leur ruche et en boucher les fentes. Les propriétés cicatrisantes et anti-infectieuses de cette substance ont été utilisées par les Egyptiens, les Grecs, les Romains et les Incas.

### 5. Activité antimicrobienne des flavonoïdes

Une des fonctions incontestées des flavonoïdes et des polyphénols est leur rôle dans la protection des plantes contre l'invasion microbienne. Leur présence dans les plantes, implique non seulement leur rôle comme agents constitutifs, mais également leur accumulation comme phytoalexines, faisant réponse à l'attaque microbienne<sup>11</sup>. Le tableau suivant<sup>12</sup> rassemble quelques structures flavoniques ayant des activités antifongiques :

**Tableau 5.** Quelques structures flavoniques ayant des propriétés antifongiques

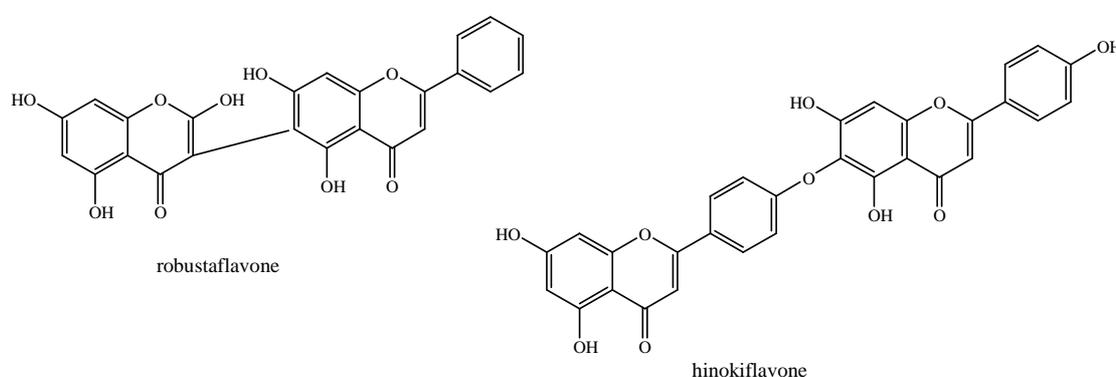
Molécule	Plante
Mackiaïne (isoflavonoïde) : 3-hydroxy-8,9-méthylène dioxypterocarpan phytoalexine	Cœur du bois des arbres légumineux Herbacées légumineuses Ex : <i>Pisum sativum</i> , <i>Trifolium</i> spp
Mucronulatol : 7,3-dihydroxy-2',4'-diméthoxyisoflavone	<i>Astragalus</i> spp
Glycoside flavone: 7-(2''-sulphatoglucoside)	<i>Thalassia testudinum</i>

L'activité antimicrobienne a été attribuée à la fonction phénolique des flavonoïdes, cette activité est sensée augmenter avec le nombre de substituants hydroxyles, méthoxy ou glucosyles. Les structures les plus efficaces étant les flavones et les flavonones<sup>13</sup>.

**Tableau 6.** Flavonoïdes ayant des propriétés antibactériennes

Composé flavonique antibactérien	Bactérie	Référence
• Retrochalcone licochalcone (4,4'-dihydroxy-2'-méthoxy-3'-prenyl)	Staphylococcus aureus	14
• 5,7,2',6'-tetrahydroxy-6-lavandulyl-4'-méthoxy flavanone	Inhibition complète de <i>S. aureus</i>	15
• 5,7-dihydroxy-3,8-diméthoxyflavone	Staphylococcus epidermis	16

Une propriété supplémentaire des flavonoïdes qui a fait l'objet de beaucoup d'études, c'est l'activité antivirale, plus particulièrement celle contre le virus du sida (VIH). Quelques flavonoïdes paraissent avoir une activité inhibitrice sur ce virus. C'est apparemment vrai pour le composé Baicaline (5, 6,7-trihydroxyflavone-7-glucoronide) séparé de la plante *Scutellaria baicalensis*<sup>17</sup>. D'autres flavonoïdes sont des inhibiteurs d'enzymes nécessaires à la reproduction virale. En effet, deux biflavonoïdes, robustaflavone et hinokiflavone, sont actifs contre l'enzyme VIH -transcriptase<sup>18</sup> (Figure 3).

**Figure 3.** Deux inhibiteurs de l'enzyme VIH- transcriptase

## 6. Propriétés médicinales des flavonoïdes

Les produits végétaux sont riches en métabolites secondaires et peuvent être absorbés et exercer des rôles biologiques divers. De nombreuses enquêtes épidémiologiques ont mis en évidence un effet préventif des fruits et des légumes vis-à-vis de l'apparition de pathologies majeures. Les polyphénols constituent une des classes des micronutriments les plus abondantes. Ces composés sont particulièrement intéressants à étudier par ce qu'ils présentent des propriétés biologiques très variées telles que :

### A. *Inhibition des enzymes par les flavonoïdes*

Des tests sur plusieurs structures de flavonoïdes étudiées, ont prouvé leur capacité d'inhiber les enzymes clés dans la respiration. Une comparaison des flavonoïdes avec des fragments variés de méthoxylations / hydroxylation, a classé les flavonoïdes selon la puissance de l'inhibition du NAHD-oxydase<sup>19</sup> :

- Robinetine avec une concentration d'inhibition  $IC_{50}$  de 19 nmol/mg de protéine ;
- Rhammetine avec une concentration d'inhibition  $IC_{50}$  de 42 nmol/mg de protéine ;
- Eupatorine avec une concentration d'inhibition  $IC_{50}$  de 43 nmol/mg de protéine ;
- Baicaleine avec une concentration d'inhibition  $IC_{50}$  de 77 nmol/mg de protéine ;
- 7,8-dihydroxyflavone avec une concentration d'inhibition  $IC_{50}$  de 277 nmol/mg de Protéine ;
- Norwogonine avec une concentration d'inhibition  $IC_{50}$  de 340 nmol/mg de protéine.

### **B. *Activité anti-inflammatoire et vasculaire des flavonoïdes***

Les flavonoïdes peuvent inhiber les voies du cyclo-oxygénase et/ou du 5-lipoxygénase du métabolisme de l'arachidonate. Plusieurs bioflavonoïdes ont montré des effets certains lors d'études expérimentales sur des malades présentant une polyarthrite rhumatoïde. Certains inhibent les décharges de l'histamine et évitent ainsi la production des leucotriènes inflammatoires. L'association thérapeutique des flavonoïdes, des enzymes protéolytiques et de la vitamine C s'est montrée plus efficace que les anti-inflammatoires non stéroïdiens<sup>20</sup>. Outre cette activité, les flavonoïdes peuvent agir sur plusieurs composantes du sang telles que les plaquettes, monocytes ainsi que sur les muscles lisses. Des travaux<sup>21</sup> ont montré des flavonoïdes qui ont une activité anti plaquettaire incluant les composés : isobavachalcone et néobavaisoflavone isolés de la plante *Psoralea corylifolia* (leguminisae). D'autres études ont signalé l'activité anti-plaquettaire et l'activité vasodilatatrice de la lutéoline<sup>22</sup>. Aussi, la quercétine, les dérivés du kaempférol et l'apigénine inhibent l'agrégation des plaquettes du lapin causée par plusieurs inducteurs<sup>23</sup>.

### **C. *Activité oestrogénique et antitumorale des flavonoïdes***

La principale famille des flavonoïdes bien connue pour ses activités oestrogéniques est celle des isoflavones, telle que la genistéine. De nouveaux phytoœstrogènes ont été caractérisés, parmi eux le 8-isopentenyl naringénine isolé de la plante thaïlandaise *anaxagoréa lutzonensis* (Annonacea)<sup>24</sup>. Des tests *in vitro* ont montré que ce flavanone a une activité oestrogénique plus grande que celle de la genistéine et que la présence du groupe 8-isopentenyl est un facteur important pour l'interaction avec le récepteur œstrogène. D'autres flavones, flavanones et flavonols avec un groupement isopentenyl en position C<sub>8</sub> ont montré une affinité considérable au récepteur de l'œstrogène, par contre, les composés 8-isopentenyl isoflavones sont inactifs.

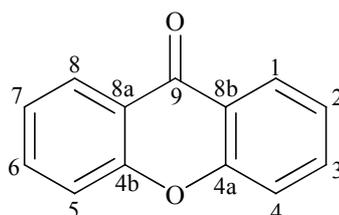
**Tableau 7.** Quelques structures flavoniques ayant des activités antitumorale et cytotoxique

Composé	Origine	Activité cytotoxique et tumorale	Réf.
8,3-dihydroxy-7,4',5'-triméthoxyflavone et son dimère	<i>Muntigia calabura</i> (Elalocarpaceae)	Cancer du sang	
6,8-diprenylriodictyol 6,8-diprenyl-3'-méthylriodictyol (Hiravanone)	<i>Monotes engleri</i> (Diptérocarpaceae)	Contre la lignée cellulaire (panel)	
6,8-di-p-hydroxybenzyltaxifoline 8-p-hydroxybenzyltaxifoline 6-p-hydroxybenzyltaxifoline	<i>Cudrania tricuspidata</i> (Moraceae)	Cancer de la peau Leucémie Cancer du colon Cancer du rein	<b>12</b>
5,2'-dihydroxy-6, 7, 8, 6'-tetraméthoxyflavone (skullcapflavone II)	<i>Scutellaria baicalensis</i> (labiatae)	Leucemie	
5,7-dihydroxy-8-méthoxyflavone 2 (S)- 5,2',5'-trihydroxy-7,8-diméthoxyflavone	<i>Scutellaria indica</i> (labiatae)	Activité cytotoxique	
4',7'',-di-O-methylamentoflavone 7''-O-méthylrobustaflavone	<i>Selaginella</i>	Cancer du sein, du poumon, du colon et de la prostate	

## II. LES XANTHONES

### 1. Introduction

Le terme xanthone est d'origine grec, dérivé du mot ζανθου (xanthos), signifie le jaune. Les dérivés de xanthone constituent une classe de composés hétérocycliques, de coloration jaune, ils possèdent un dibenzo- $\gamma$ -pyrone comme squelette de base<sup>25</sup> (Figure 4).



Dibenzo -  $\gamma$ - pyrone

**Figure 4.** Squelette de base des Xanthonnes

En 1961 Robert a pu isoler les xanthones à partir des champignons. Par la suite, un certain nombre de ses dérivés oxygénés ont été isolés à partir de plusieurs sources naturelles telles que les lichens, les champignons et d'autres plantes. Les xanthones possèdent des effets toxiques et d'autres pharmacologiques très intéressantes.

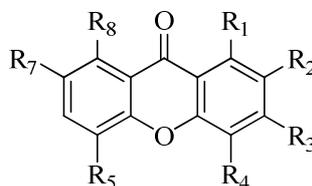
## 2. Classification des xanthones

Selon Mandal et al<sup>26</sup>, les xanthones peuvent être classés en cinq groupes :

### A. *Xanthenes oxygénés*

Qui sont subdivisés en six classes:

- a. Mono oxygéné (2) : trois composés ont été isolés dans cette classe.
- b. Di oxygéné (3) : 12 composés identifiés.
- c. Tri oxygéné (4) : le gentisine (4) a été le premier xanthone isolé.
- d. Tétra oxygéné (5) : sont nombreux et la majorité d'entre eux sont isolés à partir de la famille Gentianaceae.
- e. Penta Oxygéné (6) : seulement un petit nombre existe dans la nature.
- f. Hexa oxygéné (7).

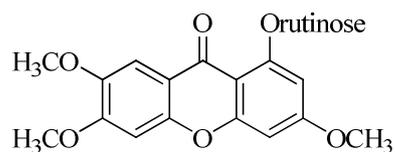


- 3-Hydroxy-2-methoxyxanthone ( $R_1=R_4=R_5=R_7=R_8=H$  ;  $R_3=OH$  ;  $R_2=OMe$ ) (3)  
 1,7-Dihydroxy-3-methoxyxanthone ( $R_2=R_4=R_5=R_8=H$  ;  $R_1=R_7=OH$  ;  $R_3=OMe$ ) (4)  
 1-Hydroxy-3, 7, 8-trimethoxyxanthone ( $R_2=R_4=R_5=H$  ;  $R_1=OH$  ;  $R_3=R_7=R_8=OMe$ ) (5)  
 1-Hydroxy-2, 3, 4, 5-tetramethoxyxanthone ( $R_7=R_8=H$  ;  $R_1=OH$  ;  $R_2=R_3=R_4=R_5=OMe$ ) (6)  
 1-Hydroxy-2, 3, 4, 5, 7-pentamethoxyxanthone ( $R_8=H$  ;  $R_1=OH$  ;  $R_2=R_3=R_4=R_5=R_7=OMe$ ) (7)

### B. *Xanthene-Glycoside*

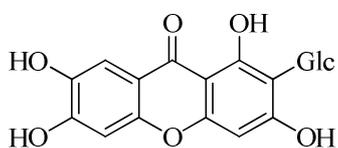
Cette classe peut être divisée en deux sous-classes O-glycosides et C-glycosides selon la nature de la liaison entre le glucose et le dérivé du xanthone:

- a. Xanthone O- glycosides (8) : ces composés ont une partie glycoside attachée à l'oxygène du xanthone positionné en 1. Ils sont facilement hydrolysables dans un milieu acide ou enzymatique.



L-Rhamnopyrannosyl (1,6) - D- glucopyrannose (8)

- b. Xanthone C-glycoside (9) : ces composés une partie glycoside attachée au carbone du xanthone positionné en 2. Ils sont moins hydrolysables en comparaison avec les O- glycoside xanthones. Mangiferin a été le premier glycoside xanthone isolé (en 1908) à partir de *Mangifera indica* (Anacardiaceae).

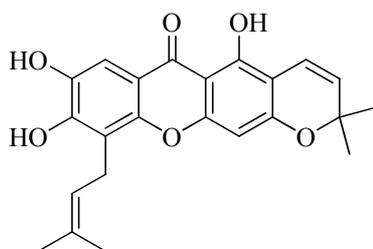


Glc: Glycoside

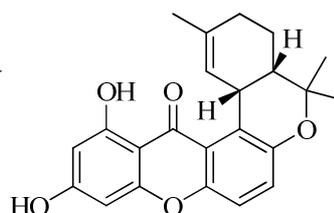
Mangiferin (9)

### C. Prenylated xanthones

La famille Guttiferae regroupe un nombre élevé de xanthones substitués par des groupements de pentenyle et geranyle.



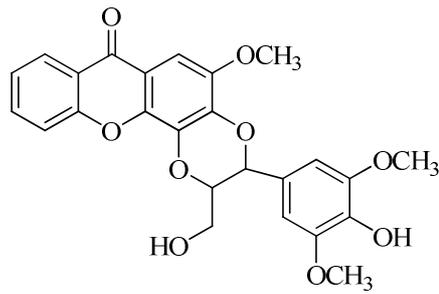
Caloxanthone (10)



Calozeyloxanthone (11)

### D. Xanthonolignoids

La structure de ces composés est le résultat d'une association du squelette de base des xanthones représenté par la **Figure 4** (page 12) avec un lignoïde. Parmi ces composés le plus connu est le cadesin D (12).



Cadensin D (12)

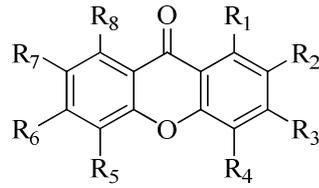
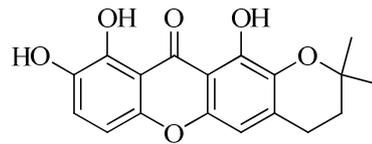
### 3. Utilisation thérapeutique des xanthones

Il a été observé sur les plantes contenant des dérivés de xanthones des activités biologiques différentes, utilisées spécialement dans la médecine indigène. Leur efficacité a été montrée dans le traitement (Tableau 8, Figure 5):

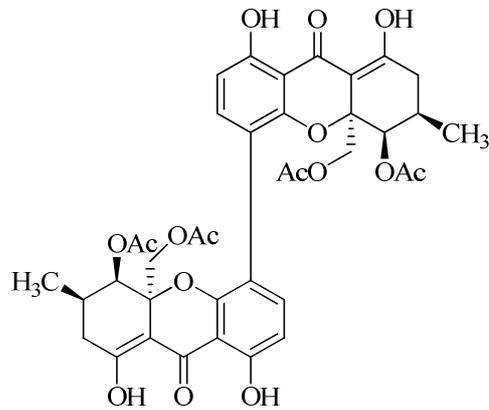
- a. Des maladies cardiovasculaires.
- b. De la malaria.
- c. Des troubles mentaux.
- d. De la fièvre.
- e. Des champignons.
- f. Des Hépatites.

**Tableau 8.** Utilisation des dérivés du xanthone en tant que agents thérapeutiques

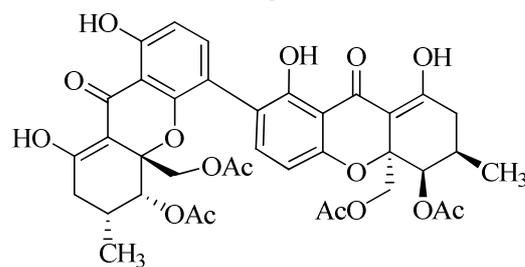
Composé	Structure	Activité Biologique	Références
1,5,8-trihydroxy-3-methoxy xanthone	<b>A:</b> R <sub>1</sub> =R <sub>5</sub> =R <sub>8</sub> =OH, R <sub>3</sub> =MeOH, R <sub>2</sub> =R <sub>4</sub> =R <sub>6</sub> =R <sub>7</sub> =H	Action sur le system nervous central	<b>27</b>
(1,3,6,7-Tetrahydroxy xanthone)	<b>A:</b> R <sub>1</sub> =R <sub>3</sub> =R <sub>6</sub> =R <sub>7</sub> =OH, R <sub>2</sub> =R <sub>4</sub> =R <sub>5</sub> =R <sub>8</sub> =H	activité Cardiovasculaire (agent vasorelaxing)	<b>28</b>
g-mangostine (1,3,6,7-tetrahydroxy-2,8-diprenylxanthone)	<b>A:</b> R <sub>1</sub> =R <sub>3</sub> =R <sub>6</sub> =R <sub>7</sub> =OH, R <sub>4</sub> =R <sub>5</sub> =H, R <sub>2</sub> =R <sub>8</sub> =Pre	Activité Anti- champignons (fongicides)	<b>29</b>
1,7,8-trihydroxy-2,2-dimethylpyrano[5',6':3,4]xanthone (globulixanthone C)	<b>B</b>	activité anti-bactérienne	<b>30</b>
Bellidifolin (1,5,8-trihydroxy-3-methoxy xanthone)	<b>A:</b> R <sub>1</sub> =R <sub>5</sub> =R <sub>8</sub> =OH, R <sub>3</sub> =MeOH, R <sub>2</sub> =R <sub>4</sub> =R <sub>6</sub> =R <sub>7</sub> =H	Anti-diabétique	<b>31, 32</b>
3,4-dihydroxy-2-methoxyxanthone & 2,3-dihydroxy-4-methoxyxanthone	<b>A:</b> R <sub>3</sub> =R <sub>4</sub> =OH, R <sub>2</sub> =MeOH, R <sub>1</sub> =R <sub>5</sub> =R <sub>6</sub> =R <sub>7</sub> =R <sub>8</sub> =H & R <sub>2</sub> =R <sub>3</sub> =OH, R <sub>4</sub> =MeOH, R <sub>1</sub> =R <sub>5</sub> =R <sub>6</sub> =R <sub>7</sub> =R <sub>8</sub> =H	activité Hépto- protectrice	<b>33</b>
Phomoxanthes <b>a et b</b>	<b>C et D</b>	activité Anti-parasite	<b>33</b>

**A**

Globulixanthone C

**B**

Phomoxanthone a

**C**

Phomoxanthone b

**D****Figure 5.** Dérives des xanthonnes biologiquement actifs

### III. LES CURCUMINOIDES

#### 1. Introduction

La curcumine a été séparée pour la première fois de la plante safran de l'inde (turmeric) en 1815 et sa structure (diferuloylmethane) a été déterminée en 1910<sup>34</sup>. Cette substance est responsable de la coloration jaune et de l'ensemble des effets thérapeutiques de la plante et elle s'y trouve avec une teneur qui varie entre 2 et 5%.

La plupart des préparations actuellement disponibles de la curcumine contiennent approximativement 77% de diferuloylmethane, 18% de demethoxycurcumin et 5% de bisdemethoxycurcumin.

La curcumine est de nature hydrophobique, elle est soluble dans le diméthylsulfoxyde, l'acétone, éthanol, et l'éther de pétrole.

Il existe plusieurs sources naturelles qui fournissent les dérivés curcuminoïdes (Figure 6) telles que les plantes suivantes : *Curcuma longa*; *Curcuma Phaeocaulis*; *Curcuma xanthorrhiza*; *Curcuma zedoaria*; ... etc.

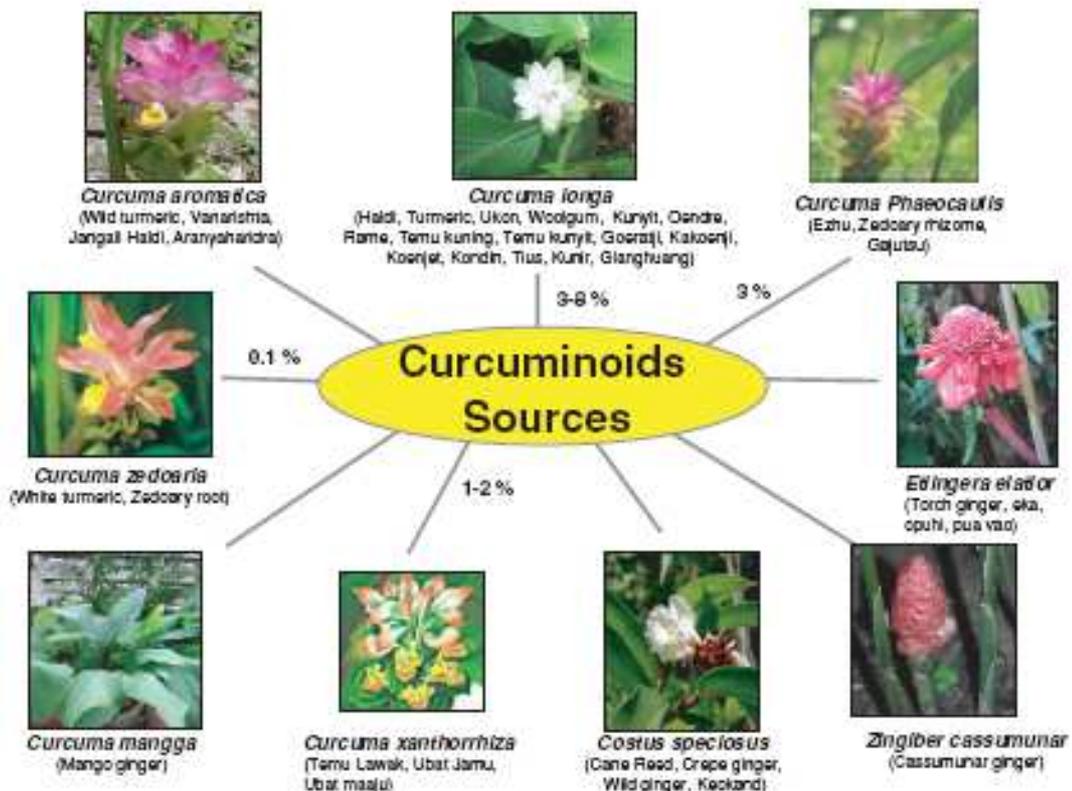
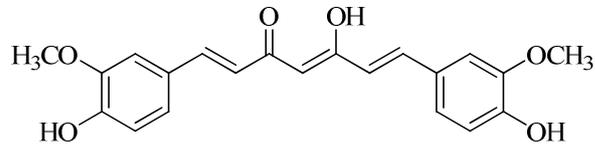


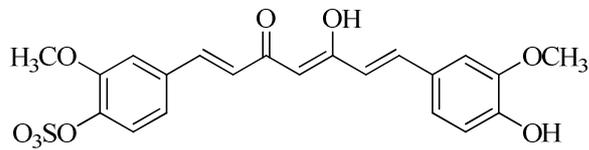
Figure 6. Les sources naturelles des dérivés des curcuminoïdes

## 2. Les dérivés des curcuminoïdes

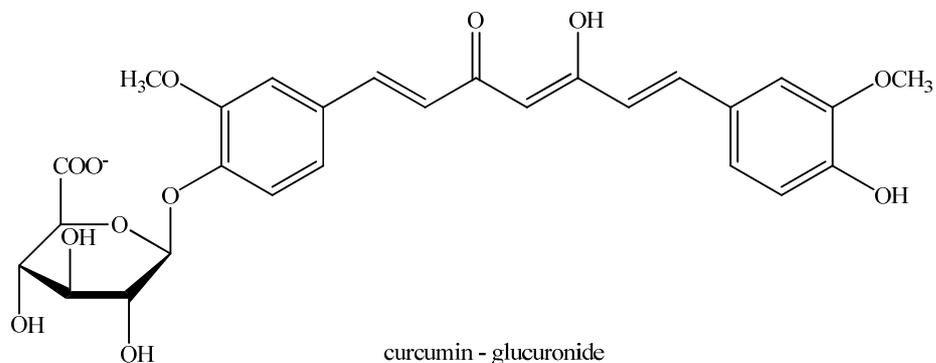
Le safran de l'Inde contient trois familles de curcuminoïdes [ex., diferuloylmethane, (nommé aussi curcumine), demethoxycurcumin, et bisdemethoxycurcumin]<sup>35</sup> (Figure 7).



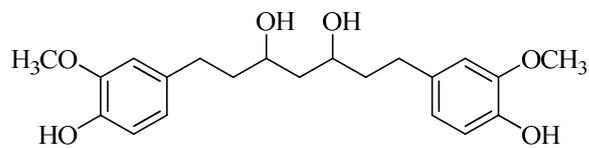
Curcumin



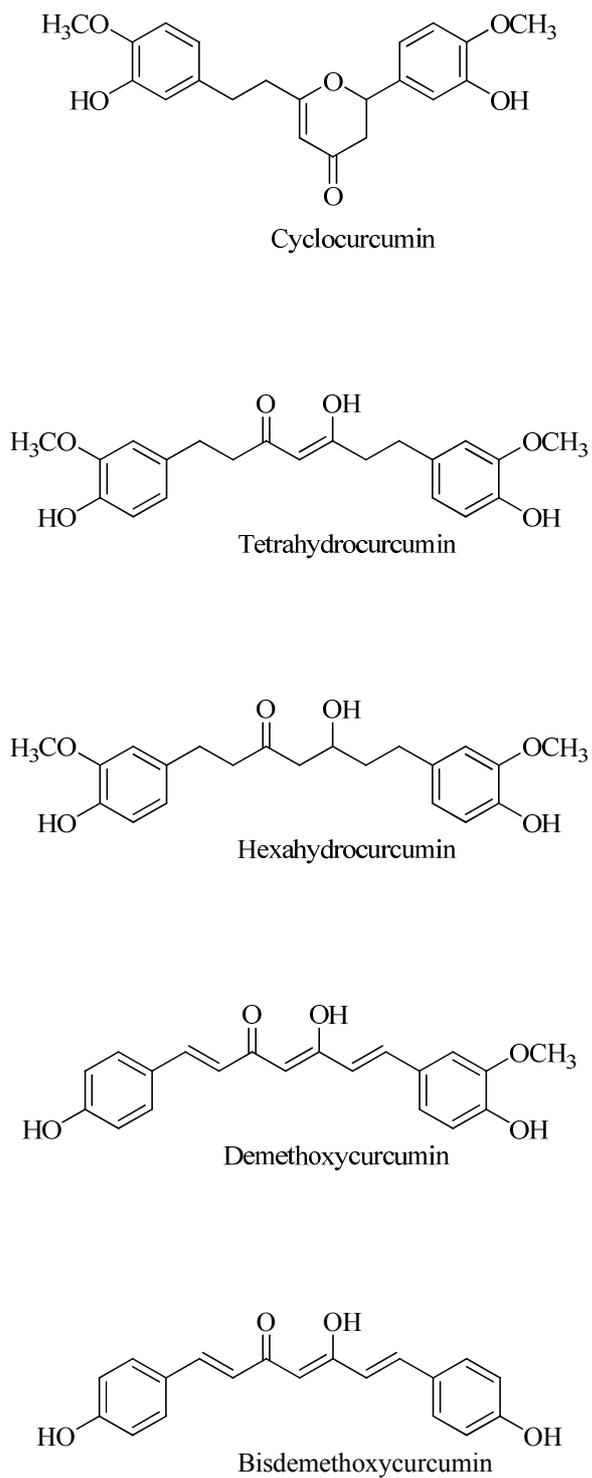
curcumin sulphate



curcumin - glucuronide



Hexahydrocurcuminol



**Figure 7.** Dérivés des curcuminoides

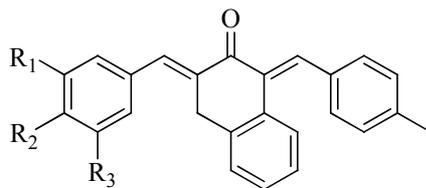
### 3. Utilisation thérapeutique des curcuminoïdes

Les dérivés curcuminoïdes ont montré une activité biologique intéressante contre une large gamme de maladies. Parmi lesquelles on trouve les propriétés suivantes (Tableau 9, Figure8) :

- antioxydantes
- anti-inflammatoires
- antitumorale
- anti-invasive
- antimétastase (antimetastatic)
- inhibiteurs d'angiogénèse (Angiogenesis)

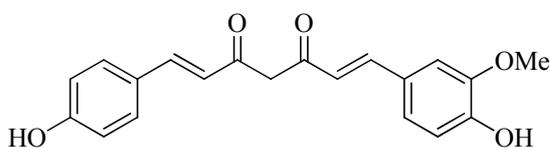
**Tableau 9.** Utilisation des dérivés de la curcumine comme agents thérapeutiques

Composé	Structure	Activité Biologique	Références
Dérivés du 3-arylidene-1-(4-nitrophenylmethylene)-3,4-dihydro-1H naphthalene-2-ones	<b>A</b>	toxicité sélective des cellules cancéreuses malignes	<b>36</b>
demethoxycurcumin et bisdemethoxycurcumin	<b>B et C</b>	inhibiteurs des angiogénèses	<b>37</b>
Curcumin	<b>D</b>	Antioxydant	<b>38</b>
Curcumin	<b>D</b>	Anti-inflammatoire	<b>39</b>
Curcumin	<b>D</b>	Des effets chimio-préventives	<b>40</b>
Curcumin	<b>D</b>	Anticancer	<b>41</b>



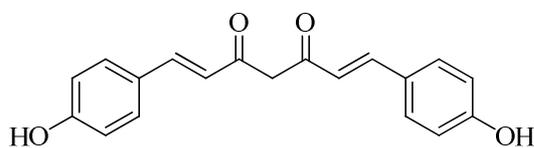
3-arylidene-1-(4-nitrophenylmethylene)-3,4-dihydro-1H naphthalene-2-ones

**A**



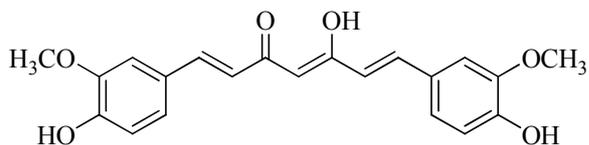
demethoxycurcumin

**B**



bis-demethoxycurcumin

**C**



Curcumin

**D**

**Figure 8.** Dérivés des curcuminoides biologiquement actifs

**IV. REFERENCES**

- [1] M.G. Hertog; E.J. Feskens; P.C. Hollman; M.B. Katan; D. Kromhout *Launcet* 342 (1993) 1007-1011.
- [2] C. Chanvallon ; P. Blanchemaison; B. Cance-Sanchez *Act Med Angiologie* 12 (1994) 3846-50.
- [3] A.A. Ramelet ; M. Monti « Phlébologie » Ed. Masson. France,1990.
- [4] F. Gert; M. Stephan *Current Opinion in Biotechnology* 12 (2001) 155-160.
- [5] J. Bruneton «Pharmacognosie et phytochimie des plantes médicinales » Ed. Tec et Docum. France, 1993.
- [6] O. Gottlieb «Micromoleculare evolution: systematics and ecology» Ed. Springer. Allemagne, 1982.
- [7] N. Saito; J.B. Harborne *Phytochemistry* 31 (1992) 3009-3015.
- [8] L.C. Cantley; G.C. Hammes *Biochemistry* 15 (1976) 1-8.
- [9] M. Barbier *Encyclopædia Universalis* 13 (1980) 66.
- [10] B. Lejeune ; B. Vernat *Parfum, cosmétiques, arômes.* 31 (1984) 2765-2774.
- [11] R.J. Grayer; J.B. Harborne; E.M. Kimmins; F.C. Stevenson; S.Wijayaguna *Acta Horticulturea* 381 (1994) 691-694.
- [12] S. Ait Moussa «Etude de la composante flavonique de la plante *Haplophyllum tuberculatum* » Thèse de magister, Université de Constantine 2003.
- [13] A.K. Picman; E.F. Schneider; J. Pieman *Biochemical Systematics and ecology* 23 (1995) 683-693.
- [14] H. Haraguchi; K. Taminoto; Y. Tamura; K. Mizutani; T. Kinoshito *Phytochemistry* 48 (1998) 125-129.
- [15] M. Inuma; H. Tsuchiya; M. Sato; J. Yokoyama; M. Ohyama; Y. Ohkama; T. Tanaka; S. Fujiwara; T. Fujii *J.Pharmacy and Pharmacology* 46 (1994) 892-895.
- [16] E. Iniesta-Sanmartin; F. Barberan; A. Guirado; F.T. Lorents *Planta Medica* 56 (1990) 643-649.

- [17] B.O. Li; T. Fu; Y.D. Yan; N.W. Baylor; F.w. Ruscutti; H.F. kung Cellular Molecular Biological Research 39 (1997) 119-124.
- [18] Y.M. Lin; H. Anderson; M.T. Flavin; Y.H.S. Pai; J. Nat. Prod. 60 (1997) 884-888.
- [19] W.F. Hodnick; D.L. Duval; R.S. Pardini Biochem. Pharm. 47 (1994) 573-580.
- [20] J. Tarayre; H. Laouressrgues Arzneim-Forsch 1977, N° 27.
- [21] W. J. Tsai; W.C. Hsin; C.C. Chen J. Nat. Prod. 59 (1996) 671-672.
- [22] C.N. Lin J. Nat. Prod. 60 (1997) 851-853.
- [23] M.I. Chung; K.H. Gan; C.N. Lin; C.M. Teng J. Nat. Prod. 56 (1993) 929-934.
- [24] M. Kitaoka; H. Kadokawa; M. Sugano; A. Ichakawa; M. Taki; S. Takaishi; Y. Iijima; S. Tsutsumi; M. Boriboom; T. Akiyama Planta Medica 64 (1998) 511-515.
- [25] M.T.H. Khan ; A. Ather "Lead Molecules from Natural Products" Ed. Elsevier B.V. 2006.
- [26] S. Mandal; P. C. Das; P. C. Joshi. J. Indian Chem. Soc. 69 (1992)611–636.
- [27] D. Schaufelberger; K. Hostettmann Planta Med 54 (1988) 219–221.
- [28] F.N. Ko; C.N. Lin; S.S. Liou; T.F. Huang; C. M. Teng Eur. J. Pharmacol. 192 (1991) 133–139.
- [29] G. Gopalakrishnan; B. Banumathi; G. Suresh J. Nat. Prod. 60 (1997) 519–524.
- [30] A. E. Nkengfack; P. Mkounga; M. Meyer; Z. Fomum; B. Bodo Phytochemistry 61 (2002) 181–187.
- [31] P. Basnet; S. Kadota; M. Shimizu; T. Namba Planta. Med. 60 (1994) 507–511.
- [32] P. Basnet; S. Kadota; M. Shimizu; Y. Takata; M. Kobayashi; T. Namba Planta. Med. 61 (1995) 402–405.
- [33] E.R. Fernandez; F.L. Carvalho; F.G. Remiao; M.L. Bastos; M.M. Pinto; O.R. Gottlieb Pharm. Res. 12 (1995) 1756–1760.
- [34] M. Isaka; A. Jaturapat; K. Rukseree; K. Danwisetkanjana; M. Tanticharoen, Y. Thebtaranonth J. Nat. Prod. 64 (2001)1015–1018.

- [35] B.B. Aggarwal; Y. J. Surh; S. Shishodia" The Molecular Targets and Therapeutic Uses of Curcumin in Health and Disease" Springer Science -Business Media, 2007.
- [36] J. R. Dimmock; U. Das; H. I. Gul; M. Kawase; H. Sakagami; Z. Barath; I. Ocsovsky; J. Molnar Bioorg Med Chem Lett 15 (2005) 1633–1636.
- [37] N. M. Pandya; N. S. Dhalla; D. D. Santani Vasc Pharmacol 44(5) (2006) 265–274.
- [38] J. S. Wright J Mol. Struct (Theochem) 591 (2002) 207–217.
- [39] M. T. Huang; T. Lysz; T. Ferraro; T. F. Abidi; J. D. Laskin; A. H. Conney Cancer Res 51 (1991) 813–819.
- [40] A. Duvoix; R. Blasius; S. Delhalles; M. Schnekenburger; F. Morceau; E. Henry; M. Dicato; M. Diederich Cancer Lett. 223 (2005)181–190.
- [41] B. B. Aggarwal; A. Kumar; A. C. Bharti Anticancer Res. 23 (2003) 363–398.

**CHAPITRE II :****METHODOLOGIE DE QSAR****I. INTRODUCTION**

Le terme QSAR signifie la relation quantitative entre la structure et l'activité biologique. Ce domaine de recherche a pour objectif de corréler les structures moléculaires, codées en descripteurs numériques, avec leurs propriétés ou activité en utilisant des techniques de modélisation informatiques et mathématiques.

QSAR est devenue un outil avantageux et précieux, notamment dans la recherche pharmaceutique industrielle, plus particulièrement dans le cas où la disponibilité des échantillons est limitée ou les mesures expérimentales sont dangereuses, longues et chères.

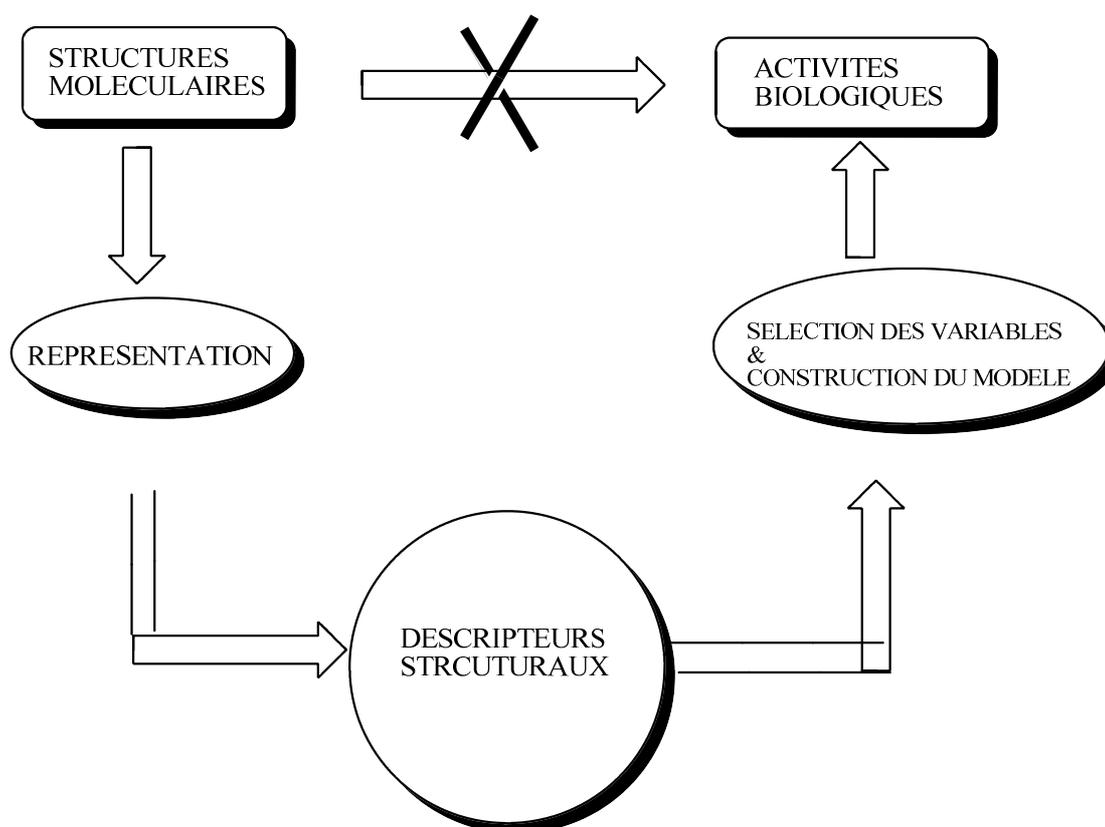
Les prévisions fournies par QSAR permettent de proposer une bonne stratégie pour le développement d'un bon pharmacophore, elle peut être ainsi un support adéquat pour valider ou rejeter une hypothèse<sup>1,2</sup>.

Sans l'utilisation de grands instruments analytiques, QSAR peut également fournir des informations utiles sur les caractéristiques structurales d'un composé responsable d'une activité biologique.

La dérivation de la relation directe avec la structure moléculaire n'est pas facile. Il est cependant possible d'identifier plusieurs facteurs structurels connus comme descripteurs moléculaires qui influent sur la propriété moléculaire choisi.

Comme il est bien illustré dans la figure 1 présentant le problème de QSAR, proposé par le groupe de Jurs<sup>3,4</sup>, l'activité ne peut pas être obtenue directement à partir de la structure moléculaire. Cette dernière doit d'abord être codée en tant que descripteurs numériques traduisant les propriétés physico-chimiques influant sur l'activité visée, tels que : la taille,

la forme, le nombre de radicaux, la polarité et la capacité d'établir des liaisons d'hydrogène.



**Figure1.** Présentation des étapes aboutissant à QSAR

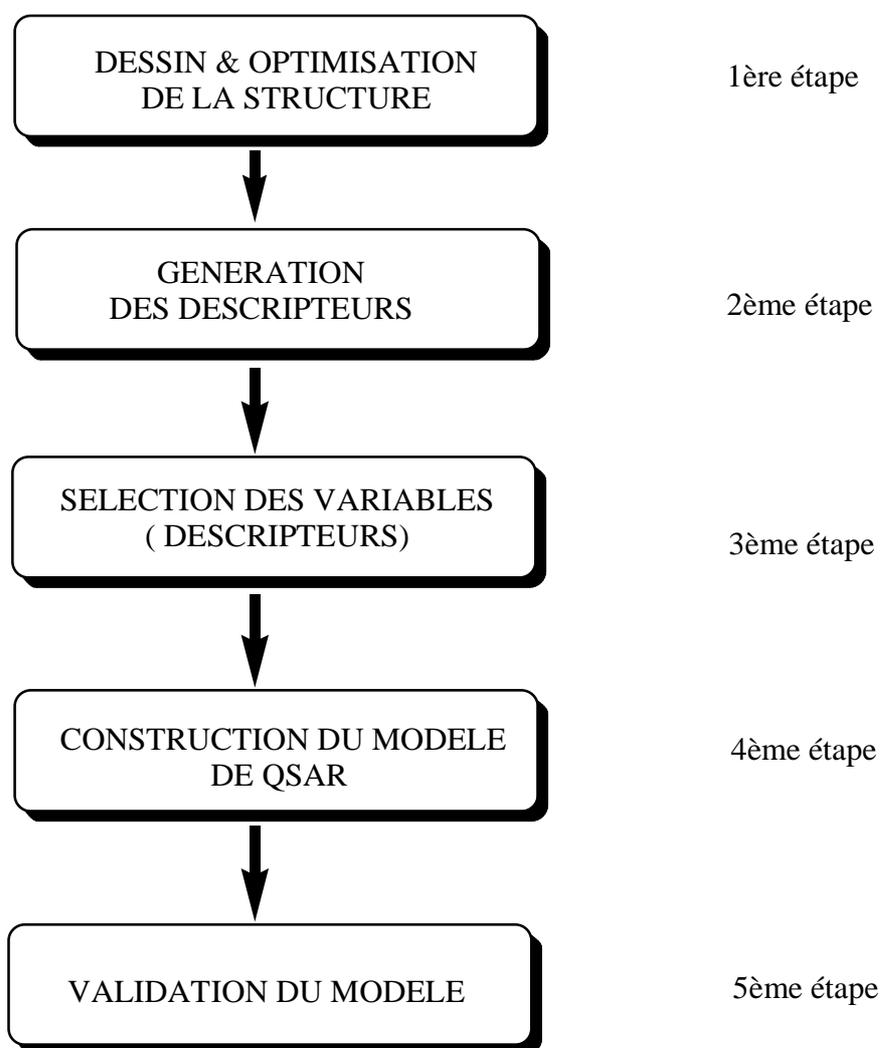
## II. PRESENTATION DE LA METHODE DE QSAR

En général, pour prédire l'activité biologique par la méthode QSAR, la construction du modèle mathématique est effectuée en 5 étapes de base (Figure 2). L'ensemble des données (l'ensemble des molécules) utilisées dans l'étude de QSAR sont collectées soit

directement à partir des mesures expérimentales, extraites de la littérature ou à partir de bases de données.

Les données collectées sont divisées en premier lieu aléatoirement en deux parties :

- ❖ Les données d'apprentissage « Training Set (TSET) » : utilisées pour la construction des modèles finaux.
- ❖ Les données de prédiction « Prédiction Set (PSET) » : utilisées pour la validation des modèles formés. Ces données ne subissent en revanche aucun traitement durant la construction des modèles et elles sont réservées uniquement pour tester la fiabilité des modèles.



**Figure 2.** Les 5 étapes de base de la construction du modèle de QSAR

## 1. Dessin et Optimisation de la Structure

La représentation des structures se fait en deux étapes successives :

La première étape consiste en la représentation structurale de tous les composés, existant dans l'ensemble des données, sous forme de croquis bidimensionnel. Plusieurs programmes spécifiques peuvent être employés pour le dessin des structures moléculaires des composés organiques on cite à titre d'exemple : Hyperchem<sup>5</sup> (Hypercube Inc., Waterloo, Ontario) ou ChemDraw<sup>6</sup> (Cambridgesoft Corp., Cambridge, MA). L'information structurale est stockée dans des tables de connections rassemblant les types d'atomes, les liaisons ainsi que leurs multiplicités. Enfin, une structure tridimensionnelle (3D) est générée par une optimisation en utilisant une méthode quantique<sup>7</sup>.

## 2. Génération des descripteurs

Les descripteurs sont des valeurs numériques qui encodent les caractéristiques physicochimiques des molécules à partir de leurs représentations structurales. La génération des descripteurs constitue l'étape cruciale lors de l'analyse de QSAR et la qualité d'un modèle dépendra essentiellement de la nature des descripteurs générés. La plupart des descripteurs, actuellement plus de 3000, peuvent être calculées théoriquement à partir de différents logiciels tels que l'OASIS<sup>8</sup>; ADAPT<sup>9</sup>; CODESSA<sup>10</sup>; et DRAGON<sup>11</sup>, etc.

Dans cette thèse nous avons utilisé le serveur **E-DRAGON1** développé par le groupe (Milano Chemometrics and QSAR Research Group). Ce dernier est divisé en 20 blocs logiques, contient une suite de programmes de génération de descripteurs capables de calculer, on line, plus de 1600 descripteurs.

### 3. Sélection des variables (les descripteurs) :

La recherche de l'ensemble des descripteurs qui forment le bon modèle, exprimant l'activité biologique avec un coût raisonnable de calcul, constitue l'étape déterminante, car le calcul de tous les modèles possibles n'est pas pratique eu égard au nombre élevé des descripteurs (plus de 1600 descripteurs).

Afin d'éviter la formation de modèles dus à la chance un contrôle rigoureux est exigé sur la taille de l'ensemble des descripteurs. Ainsi, l'approche rationnelle de la sélection des variables permet d'éviter les redondances, de diminuer le coût calculatoire et de trouver les meilleurs sous-ensembles de descripteurs.

La procédure de sélection des variables peut être divisée en deux étapes<sup>12</sup>.

- \* Sélection objective

- \* Sélection subjective

#### A. *La sélection objective :*

Elle consiste à la sélection des variables en réduisant le nombre de descripteurs sans faire participer la variable dépendante (la réponse biologique) afin de diminuer les corrélations entre les descripteurs.

La suppression d'un descripteur, ayant un pourcentage élevé de valeurs identiques pour l'ensemble des composés, aura lieu au début de la sélection. Par exemple, le nombre d'atomes d'halogène ne fournira aucune information distinctive utile si uniquement l'halogène est présent sur un seul composé sur 100 existant dans l'ensemble (TSET). Par conséquent, ce descripteur est éliminé de l'ensemble total des descripteurs. Il en est de même pour les descripteurs fournissant des informations superflues.

Ensuite, le coefficient de corrélation (R) entre les descripteurs est calculé par paires. Un des deux descripteurs est supprimé si leur combinaison possède un coefficient de

détermination supérieur au seuil requis ( $R > 0.90$ ). Cette valeur numérique de  $R$  est le seuil utilisé pour toutes les applications de sélection en réduisant le nombre de descripteurs sans pour autant perdre de l'information.

Il est parfois nécessaire, dans quelques cas, de réduire encore plus le nombre de descripteurs. A cet effet, l'analyse de l'espace vectoriel de descripteurs (VSDA) peut être utilisée. Dans cette procédure, un descripteur de l'ensemble est choisi au hasard comme descripteur (vecteur) de base initiale. Le procédé d'orthogonalisation de Gram Schmidt est appliqué pour la recherche du descripteur le plus orthogonal au descripteur de base de l'ensemble. Ensuite, ce descripteur est ajouté à l'ensemble de base (2D). Alors, le descripteur de l'ensemble le plus orthogonale au plan défini par les deux descripteurs de base est, à la fois, choisi et ajouté à l'ensemble de base. Ce procédé itératif est poursuivi jusqu'au rangement de la totalité des descripteurs selon ce principe d'orthogonalité. La formation de l'ensemble final, réduit, se fera en choisissant à partir de la base ordonnée, le sommet principal des descripteurs choisi par l'utilisateur. Le nombre de descripteurs choisis,  $x$ , sera déterminé par la taille de l'ensemble TSET.

### ***B. La sélection subjective :***

La sélection subjective des variables de l'ensemble des descripteurs déjà réduit est alors employée dans la recherche de sous-ensembles de descripteurs optimaux en utilisant l'activité visée. A cet effet, diverses méthodes d'optimisations sont utilisées pour parcourir, d'une manière efficace, la dimension du modèle (le nombre optimal de descripteurs par modèle) sans pour autant examiner toutes les combinaisons possibles.

Vu le nombre élevé de combinaisons possibles des sous-ensembles de descripteurs, des techniques statistiques et informatiques, sont employées pour explorer la dimension du modèle. Par la suite, on expose deux méthodes de sélection de descripteurs différents :

- *Sélection par les Algorithmes génétiques*

Le modèle optimal de chaque dimension est réalisé par l'exécution des algorithmes génétique en combinaison avec l'analyse de régression linéaire multiple. Les algorithmes génétiques (GA) sont des méthodes d'optimisation stochastiques inspirées des principes évolutionnaires. L'aspect distinctif des algorithmes génétiques est qu'elles découvrent plusieurs solutions simultanément, dont chacune de ces solutions explore différentes régions de l'espace.

La première étape consiste en la création d'une population choisie au hasard de N chromosomes, dont chaque chromosome est une solution candidate pour ce problème (modèle). Dans notre cas, une représentation appropriée d'un chromosome peut être une chaîne de caractère numérique codant une combinaison particulière de certains descripteurs moléculaires.

Le classement des chromosomes dans la population se fait suivant une fonction d'évaluation. Cette fonction est d'origine statistique tels que : le coefficient de détermination  $R^2$ , l'écart type  $s$ , test de Fisher F, et le coefficient de corrélation issu du cross validation  $R^2_{cv-100}$ , ce qui reflètent la qualité de chaque solution candidate.

L'étape suivante est *la reproduction*, création de nouveaux chromosomes à partir de la génération existante. Selon l'opérateur de sélection, les meilleurs chromosomes peuvent proliférer préférentiellement. Aussi, chaque chromosome a l'occasion d'échanger de l'information avec les autres via les procédures de croisement et de mutation.

Enfin, une affiliation plus convenable peut apparaître dans la prochaine génération si des mutations intéressantes ont lieu.

La recherche collective, qui a commencé au début d'une manière aléatoire, commence à gagner plus de précision et se déplace vers des régions plus optimales, car le système recueille plus de connaissance au sujet de l'espace original.

- *Sélection Ascendante pas à pas (Forward Stepwise)*

Comme son nom l'indique, il s'agit d'une technique incrémentale qui consiste à repérer dans chaque étape la variable proposant une valeur absolue de  $t$  (Student) la plus élevée et de l'ajouter dans l'ensemble (pool) courant si le coefficient est significatif, et de continuer ainsi tant que les ajouts sont possibles.

On commence par  $p$  régressions simples. Si une variable a été ajoutée, on poursuit avec  $p-1$  régressions à 2 variables, etc. L'ajout d'une variable dépend de la significativité du coefficient de la variable choisie.

#### 4. Développement du modèle

Le but d'une modélisation QSAR est de trouver une fonction permettant de relier de manière quantitative ou qualitative une propriété (ou activité) étudiée avec les descripteurs<sup>13</sup>. Plusieurs méthodes sont utilisées pour relier les descripteurs aux propriétés parmi lesquelles il y a la régression linéaire multiple (MLR), les réseaux de neurones (CNN) et la régression par les machines à vecteur de support (SVM).

Concrètement, pour un ensemble de  $n$  molécules et  $p$  descripteurs, stockés dans la matrice  $\mathbf{X}$ , un modèle QSAR détermine une fonction qui relie  $\mathbf{X}$  au vecteurs de propriétés  $\mathbf{y}$  ( $\mathbf{y} = (y_1, \dots, y_n)^T$ ).

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = f \left( \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2p} \\ x_{31} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & x_{np} \end{pmatrix} \right)$$

### A. Régression Linéaire Multiple

La régression linéaire multiple (MLR) est une méthode reliant les descripteurs à la variable à modéliser  $y$ . Elle détermine les coefficients  $a$  du vecteur  $\mathbf{A}$  et les erreurs  $b$  du vecteur  $\mathbf{B}$  dans l'équation  $\mathbf{Y} = \mathbf{X} \cdot \mathbf{A} + \mathbf{b}$  pour une série de points définis par leurs coordonnées  $(x, y)$ .  $\mathbf{X}$  représente la matrice d'attributs, c'est-à-dire l'ensemble des descripteurs pour l'ensemble de données (molécules)<sup>14</sup>. La méthode MLR établit que la propriété  $y$  d'une molécule dépend linéairement des descripteurs  $x_1, x_2, \dots, x_p$  selon la relation :

$$y = a_0 + \sum_{i=1}^p a_i x_i \quad \text{Equ. 1}$$

Pour un ensemble de données, cela revient à calculer un ensemble d'équations :

$$y_1 = a_0 + a_1 x_{11} + a_2 x_{12} + \dots + a_p x_{1p} + b_1$$

$$y_2 = a_0 + a_1 x_{21} + a_2 x_{22} + \dots + a_p x_{2p} + b_2$$

...

$$y_n = a_0 + a_1 x_{n1} + a_2 x_{n2} + \dots + a_p x_{np} + b_n$$

Au final, la méthode MLR résout l'équation  $\mathbf{Y} = \mathbf{X} \cdot \mathbf{A} + \mathbf{B}$ , où  $\mathbf{Y}$ ,  $\mathbf{X}$ ,  $\mathbf{A}$  et  $\mathbf{B}$  représentent respectivement le vecteur de propriétés, la matrice des attributs (descripteurs), la matrice des coefficients et la matrice des erreurs de régression.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2p} \\ x_{31} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & x_{np} \end{pmatrix} \quad \mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_n \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$$

La méthode des moindres carrés permet de simplifier l'équation (02) en minimisant la somme des carrés des erreurs  $b$ . L'équation devient donc :  $X.A = Y$  la matrice  $X$  n'étant pas carrée, il faut multiplier chaque membre de l'équation par la transposée  $X^T$ , pour arriver à  $X^T.X.A = X^T.Y$ . Connaissant  $X$  et la propriété  $Y$ , on en déduit que :  $A = (X^T.X)^{-1}.X^T.Y$ .

### ***B. Réseau de neurones artificiels***

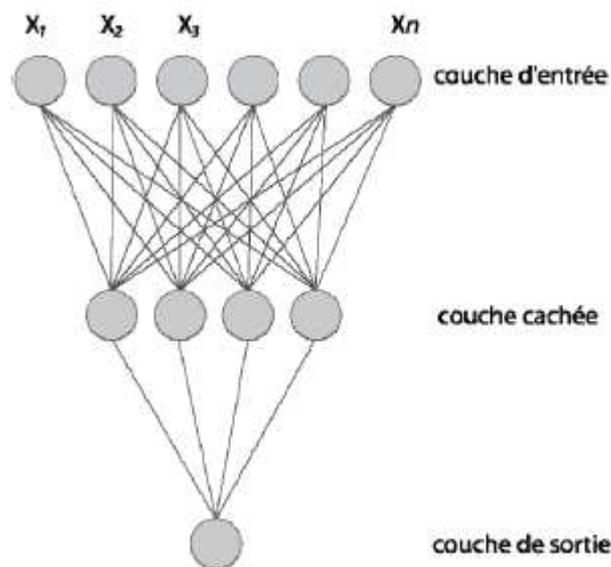
L'approche par réseau de neurones (CNN: Computational Neural Network) est analogue aux systèmes de neurones biologiques. Les neurones biologiques permettent de véhiculer et de traiter des informations en faisant circuler des messages électroniques dans un réseau constitué d'axones. Les informations sont transmises d'un neurone à l'autre, de manière unidirectionnelle, par l'intermédiaire de points de jonction appelés *synapses*. En modélisation, un neurone possède une couche d'entrée par laquelle les données arrivent. Le neurone renvoie une valeur +1 ou -1 selon que la somme pondérée dépasse un certain seuil. Les poids de chaque neurone sont ajustés au cours de l'apprentissage<sup>15</sup>.

Un réseau de neurones est constitué de multiples neurones : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Les neurones d'une couche sont interconnectés aux neurones de la couche voisine. La figure 3 représente un réseau de neurones à 3 couches, avec :

- La couche d'entrée : elle est constituée d'autant d'entrées que de descripteurs pour l'ensemble de données entier. Chaque neurone est connecté aux neurones de la couche cachée.
- La couche cachée : elle est constituée d'un nombre variable de neurones. Pour chaque neurone, le réseau effectue une opération de somme pondérée avec les différents poids de chaque neurone d'entrée. Un réseau de neurones peut avoir

plusieurs couches cachées. Un poids est associé à chaque neurone de la couche cachée.

- La couche de sortie : le nombre de neurones est égale au nombre de propriétés modélisées. Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une par une aux neurones de la couche d'entrée. Le système change itérativement les poids associés aux neurones d'entrées et de la couche cachée de façon à minimiser l'erreur entre la propriété calculée et la propriété expérimentale. A cette fin, il ajuste les poids de chaque neurone.



**Figure 3** : Schéma représentatif d'un réseau de neurones artificiels

## 5. Validation du modèle

La validation des modèles est nécessaire pour estimer leur fiabilité et détermine la reproductibilité des résultats et la pertinence d'un modèle développé pour une application donnée<sup>16</sup>.

Trois méthodes sont employées pour la validation d'un modèle de QASR, à savoir les méthodes de validation interne, externe et le test de randomisation.

### A. Validation interne

La méthode de la validation croisée, une méthode de validation interne, est la plus utilisée dans la plupart du temps dans QSAR. Elle est réalisée par différentes procédures : Leave-One-Out et Leave-n-Out où « n » représente le nombre de composés éliminés.

La procédure Leave-One-Out retire successivement une molécule de l'ensemble d'apprentissage (TSET). Un modèle QSAR est construit alors sur un ensemble de  $m-1$  de composés et la molécule retirée est prédite par le modèle formé, « m » représente le nombre de molécules utilisées dans la construction du modèle (TSET). Cette procédure est répétée  $m$  fois afin de prédire les propriétés de toutes les molécules.

La procédure Leave-n-Out correspond à un découpage en plusieurs parties de l'ensemble de données. A tour de rôle, une partie de l'ensemble de données est attribuée pour un ensemble de test interne. Les autres constituent l'ensemble d'entraînement. Les molécules de chaque groupe éliminé sont prédites par le modèle formé. Cette procédure est répétée  $p$  fois pour prédire les propriétés de toutes les molécules,  $p$  est le nombre des groupes de molécules éliminés<sup>17</sup>.

La performance des modèles de régression est estimée avec les paramètres statistiques de la validation croisée  $s_{cv-100}$  (l'écart type de la validation croisée), et  $Q^2_{cv-100}$  (coefficient de détermination de validation, il représente la capacité de la prévision du modèle QSAR).

$Q^2_{cv-loo}$  est calculé à partir de l'équation suivante :

$$Q^2_{cv-loo} = \frac{\sum_{i=1}^N (y_i - \bar{y}_i)^2 - \sum_{i=1}^N (y_i - y_{i_{cv}})^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2}$$

Avec :

$y_{i_{cv}}$  est la valeur prévue de la propriété  $y$  par la méthode de la validation croisée.  $\bar{y}_i$  est la moyenne arithmétique des  $y$ .

$Q^2_{cv-loo}$  est employé en tant qu'outil de diagnostic pour évaluer le pouvoir prédictif ou la qualité de la prévision d'un modèle.

### ***B. Test de randomisation***

Le test de randomisation permet d'affirmer que la corrélation de chance ne joue aucun rôle durant le développement du modèle. Ces tests sont exécutés quantitativement avec les modèles de QSAR et qualitativement avec les modèles issus de la classification. Les observations sont aléatoirement désorganisées dix fois. C'est-à-dire que la colonne des observations (propriétés) sera changée aléatoirement, en revanche la colonne des descripteurs reste inchangée. A la fin, on obtient dix modèles avec des caractéristiques statistiques spécifiques<sup>18</sup>.

La supposition sous-jacente de l'essai de randomisation est la suivante :

Si les capacités prédictives du modèle ne sont pas dues aux corrélations de chance, alors la désorganisation aléatoire des observations conduira à des modèles (quantitatifs ou qualitatifs) de prévisions faibles, et vice versa.

### *C. Validation Externe*

Il n'est pas suffisant de développer un modèle de QSAR avec une excellente qualité d'ajustement et de prévision, mais il est également nécessaire de généraliser ces prévisions en l'appliquant sur un échantillon externe (PSET).

Les prévisions résultantes pour PSET déterminent alors la validité externe du modèle. Il est important de noter que les composés de l'ensemble de validation externe (PSET) n'ont pas été utilisés dans le développement du modèle<sup>19</sup>.

Les mesures quantitatives de la qualité de la prévision par validation externe peuvent être identiques à celles utilisées pour la validation interne.

## **6. Domaine d'applicabilité**

L'établissement du domaine d'applicabilité pour chaque modèle constitue un des problèmes les plus importants dans l'analyse de QSAR<sup>20</sup>. L'absence de ce domaine admet que chaque modèle de QSAR peut formellement prévoir l'activité d'un produit chimique même si sa structure est complètement différente de ceux inclus dans le sous ensemble (TSET).

Ainsi, l'absence du domaine d'applicabilité comme composant obligatoire de n'importe quel modèle de QSAR mènerait à l'extrapolation injustifiée du modèle dans l'espace de chimie et, en conséquence, avec une probabilité élevée des prévisions imprécises<sup>21</sup>.

L'analyse du domaine d'applicabilité est réalisée à l'aide du graphe de Williams. Ce dernier est une représentation graphique des valeurs résiduelles normalisées  $\delta_i$ , pour chaque composé de l'ensemble (TSET), en fonction de leurs valeurs du levier  $h_{ii}$ .

Le levier  $h_{ii}$  de l'observation  $i$  est lue sur la diagonale principale de la matrice  $H$ , dite Hat Matrix, et définie de la manière suivante :

$$H = X(X^T X)^{-1} X'$$

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1 \dots n)$$

Où  $x_i$  est le rang  $i$  du vecteur des descripteurs du composé.  $X$  est la matrice du modèle contenant toutes les valeurs liées aux composés de l'ensemble TSET. L'exposant T réfère à la matrice / vecteur transposé.

La valeur critique  $h^*$  est définie comme suit :

$$h^* = 3\bar{h} = 3 \frac{\sum h_i}{n} = \frac{3p'}{n} \quad (i = 1 \dots n)$$

$n$  : nombre des composés dans l'ensemble TSET.

$p'$  : nombre des paramètres ajustés.

Les valeurs  $h_{ii}$  peuvent être calculées pour les data TSET du modèle, ainsi que pour des nouvelles structures proposées.

### III. Conclusion

Dans ce chapitre nous avons présenté la méthode de QSAR en décrivant ses différentes étapes ; la génération des descripteurs moléculaires, la sélection des variables, l'obtention et la validation des modèles de QSAR. Les notions des méthodes de sélection des variables subjective et objective, des méthodes d'apprentissage, de la validation et de domaine d'applicabilité sont détaillées. Les modèles établissent des relations entre une propriété (ou activité biologique) et des descripteurs représentant les structures chimiques par une fonction d'apprentissage. Les modèles sont alors validés par des procédés de validation interne, validation externe et test de randomisation.

#### IV. Références

- [1] J. Gasteiger "Handbook of Chemoinformatics: *From Data to Knowledge (4 Volumes)*" Ed. Wiley-VCH Verlag, Weinheim. Allemagne, 2003.
- [2] R. Mannhold; P. Krosgaard-Larsen; H. Timmerman "QSAR: Hansch Analysis and Related Approaches" Ed. Wiley-VCH Verlag, Weinheim. Allemagne, 1993.
- [3] L. He "QSAR and Classification modeling: Prediction of biological activity of organic compounds from molecular structure" Thèse de doctorat; Août 2005, Université de l'état de Pennsylvania. Etats Unies.
- [4] P. D. Mosier "Prediction of chemical properties and biological activities of organic compounds from molecular structure and use of probabilistic and generalizes regression neural networks" Thèse de doctorat; Décembre 2003, Université de l'état de Pennsylvania. Etats Unies.
- [5] HyperChem Release 7, HyperCube, Inc., <http://www.hyper.com>.
- [6] ChemDraw (Cambridgesoft Corp., Cambridge), [www.cambridgesoft.com](http://www.cambridgesoft.com).
- [7] M. J. S. Dewar; E. G. Zoebisch; E. F. Healy; J.J. P. Stewart; J. Am. Chem. Soc. 107 (1985) 3902- 3909.
- [8] O. Mekenyan; D. Bonchev *Acta Pharm. Jugosl.* 36 (1986) 225-237.
- [9] R. Guha; J. R. Serra; P. C. Jurs *J. Mol. Graph. Mod.* 23 (2004) 1–14.
- [10] A. R. Katritzky; V. S. Lobanov; M. Karelson. Manuel du logiciel CODESSA. Gainesville, FL. University of Florida. 1994. ([www.semichem.com/codessraefs.html](http://www.semichem.com/codessraefs.html)).
- [11] R. Todeschini, Milano Chemometrics and QSPR Group, (<http://www.disat.unimib.it/vhtml>).
- [12] J. H. Kalivas "Adaption of simulated annealing to chemical optimization problems" Ed. Elsevier Science B.V. Pays Bas, 1995.

- [13] R. Benigni "Quantitative Structure-Activity Relationship (QSAR): Models of Mutagens and Carcinogens" Ed. CRC Press LLC. Etats Unies, 2003.
- [14] S. Weisbergn "Applied Linear Regression" Ed. John Wiley and Sons. Etats Unies, 2005.
- [15] J. Devillers "*Neural Networks in QSAR and Drug Design*" Ed. Academic Press Limited. Royaume- Uni, 1996.
- [16] A. Varnek; A. Tropsha "*Chemoinformatics Approaches to Virtual Screening*" Ed. Royal Society of Chemistry. Royaume- Uni, 2008.
- [17] P. Bultinck; H. D. Winter; W. Langenaeker; J. P. Tollenaere: "Computational Medicinal Chemistry for Drug Discovery" Ed. Marcel Dekker. Etats Unies, 2004.
- [18] L. Zhang; H. Zhu; T. I. Oprea; A. Golbraikh; A. Tropsha *Pharm. Res.* 25 (8) (2008) 1902 -1914.
- [19] S. J. Patankar "Prediction of enzyme inhibition and receptor antagonist properties from molecular structure and development of radial basis function neural networks for the analysis of inhibitor binding" Thèse de doctorat; Mai 2003, Université de l'état de Pennsylvania. Etats Unies.
- [20] E. Papa; J.C. Dearden; P. Gramatica *Chemospher* 67 (2007) 351-358.
- [21] P. Gramatica; E. Giani; E. Papa *J. Mol. Graph. Mod.* 25 (2007) 755-766.

**CHAPITRE III :****INTRODUCTION AUX DESCRIPTEURS GENERES PAR  
E-DRAGON1****I. INTRODUCTION**

La prévision des propriétés physiques et biologiques des composés organiques à partir de la structure moléculaire est un problème important et encore non élucidé en chimie théorique et informatique. Ces propriétés, mesurées expérimentalement, sont quasiment exprimées d'une manière invariable en termes quantitatifs, a titre d'exemple : le point d'ébullition, l'indice de réfraction, l'énergie de l'état de transition, l'activité biologique (antivirale, antitumorale)... etc. L'étude de la relation existante entre ces propriétés et la structure moléculaire des composés chimiques constitue le parcours idéal dans la recherche d'un modèle capable de prédire ces propriétés<sup>1</sup>.

Du fait que les propriétés sont exprimées en nombres, alors que la structure moléculaire ne peut l'être, le codage de la structure moléculaire présente le premier défi pour modéliser la relation structure-propriétés. En conséquence, l'approche la plus pratique pour résoudre cette difficulté est l'utilisation des descripteurs moléculaires.

**II. DEFINITION ET CLASSIFICATION DES DESCRIPTEURS MOLECULAIRES**

Un descripteur est une représentation mathématique d'une molécule, obtenu par le biais d'un procédé qui transforme la structure moléculaire en une information structurale codée. Cette représentation mathématique doit être invariable à la fois, à la taille et au nombre d'atomes, de la

molécule, pour permettre la construction d'un modèle via des méthodes statistiques et des réseaux de neurones artificiels<sup>2</sup>.

Il existe à ce jour plus de 3000 descripteurs répertoriés, qui sont classés en 4 types :

- Les descripteurs **0-D** : Ils sont dérivés de la liste d'atomes de la molécule, comme le poids moléculaire.
- les descripteurs **1-D** : Ils sont calculés à partir de listes de sous structures de la molécule, comme les groupements fonctionnels, les cycles présents, etc.
- Les descripteurs **2-D** : Ils sont obtenus à partir des représentations des molécules sous forme de graphes bidimensionnelles.
- Les descripteurs **3-D** : Ils sont calculés à partir de la représentation tridimensionnelle de la molécule, on cite parmi eux, les descripteurs électroniques (moment dipolaire, l'électronégativité,...), le volume moléculaire, la surface moléculaire superficielle,...etc.
- les descripteurs **4-D** : Obtenus par le calcul des champs d'interactions moléculaire entre deux molécules.

### III. NOTIONS DE BASE SUR LES MATRICES DE CODAGE STRUCTURALE

Le contenu de l'information issu d'un descripteur de structure dépend de deux facteurs majeurs : la représentation moléculaire du composé, et l'algorithme employé pour le calcul du descripteur.

Avant de citer les descripteurs les plus utilisés dans ce domaine, et eu égard au rôle fondamental de la théorie des graphes<sup>3,4</sup> pour la compréhension de ces paramètres physico-chimiques, des notions de base, sur plusieurs types de matrices qui participent dans le codage des structures chimiques, sont définies ci-après :

- La matrice d'adjacence

Soit  $(G)$  un graphe donné avec  $n$  sommets, la matrice d'adjacence  $A$  est une matrice symétrique carrée de dimension  $(n \times n)$  présentée par l'équation suivante (Equ. 1) :

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdot & \cdot & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdot & \cdot & \cdot & \cdot \\ a_{3,1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & & & & & a_{n,n} \end{bmatrix} \text{ Avec } a_{i,j} = \begin{cases} 1 \forall i \neq j \wedge e_{i,j} \in E(G) \\ 0 \forall i = j \vee e_{i,j} \notin E(G) \end{cases} \quad (\text{Equ. 1})$$

$e_{ij}$ : arête.

- La matrice Laplacienne

La matrice des degrés  $DEG(G)$  est une matrice diagonale où les éléments  $D_{ii}$  correspondent au nombre de connexions au sommet  $v_i$ , c'est-à-dire à son degré. A partir de cette matrice et la précédente  $A(G)$ , on peut définir la matrice laplacienne  $L(G)$  comme suit (Equ.2).

$$L(G) = DEG(G) - A(G) \quad (\text{Equ.2})$$

Avec

$$l_{i,j} = \begin{cases} \text{deg}(v_i) & \forall i = j \\ -1 & \forall i \neq j \wedge e_{i,j} \in E(G) \\ 0 & \forall i \neq j \wedge e_{i,j} \notin E(G) \end{cases}$$

Comme les deux matrices, adjacence  $A$  et diagonale ( $DEG$ ) sont symétriques, il en résulte que la matrice Laplacienne l'est également.

- La matrice de distance

La matrice de distance  $D$  d'un graphe  $G$  de  $n$  sommets est une matrice carrée symétrique de dimension  $(n \times n)$  (Equ.3), où  $d_{i,j}$  est la distance entre les deux sommets  $v_i$  et  $v_j$  dans le graphe ( $G$ ) c'est-à-dire la longueur du plus court chemin entre ces deux sommets.

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,n} \\ d_{2,1} & d_{2,2} & \dots & d_{2,n} \\ \dots & \dots & \dots & \dots \\ d_{n,1} & \dots & \dots & d_{n,n} \end{bmatrix} \quad \text{Avec } d_{i,j} = \begin{cases} d_{i,j} & \forall i \neq j \\ 0 & \forall i = j \end{cases} \quad (\text{Equ.3})$$

- La matrice de distance pondérée

La matrice de distance pondérée dérivée de la matrice de distance pour compter simultanément la présence des hétéroatomes et la multiplicité des liaisons dans la molécule, est définie comme suit (Equ.4) :

$$[D^w]_{ij} = \begin{cases} 1 - \frac{w_c}{w_j} & \text{if } i = j \\ \sum_{b=1}^{d_{ij}} \left( \frac{1}{\pi_b^*} - \frac{w_c^2}{w_{b(1)} - w_{b(2)}} \right) & \text{if } i \neq j \end{cases} \quad (\text{Equ.4})$$

Avec :

$w_c$  : La propriété de l'atome de carbone ;

$w_i$  : La propriété de l' $i$ ème atome ;

$\pi^*$  : L'ordre de liaison conventionnel (égale 1, 2, 3 et 1,5 pour une liaison simple, double, triple, et liaison aromatique respectivement).

$d_{ij}$  : La distance topologique entre les deux sommets  $i$  et  $j$ .

$b(1)etb(2)$  : Représentent les deux sommets incidents à la liaison  $b$ .

#### IV. DESCRIPTEURS ISSUS DU SERVEUR E-DRAGON1

Les descripteurs moléculaires utilisés dans cette thèse sont ceux proposés par le serveur **E-DRAGON1**. Ce dernier est un logiciel en ligne développé par le groupe (Milano Chemometrics and QSAR Research Group) pour calculer un ensemble de plus de 1600 descripteurs moléculaires répartis en 20 blocs logiques, regroupant un ensemble de descripteurs à base de 0D, 1D, 2D et 3D.

Dans la suite de ce chapitre nous décrirons ces différents blocs de descripteurs, et nous nous limiterons surtout aux descripteurs les plus utilisés.

##### 1. Les descripteurs constitutionnels

Les descripteurs constitutionnels reflètent seulement la composition chimique du composé, sans aucune recommandation de la géométrie ou la structure électronique de la molécule. Parmi ces descripteurs, on cite<sup>4</sup> :

- ✓ Nombre total des atomes dans la molécule.
- ✓ Nombres absolu et relatif d'atomes de certaines identités chimiques (C, H, O, S, N, F,... etc.) dans la molécule.
- ✓ Nombres absolu et relatif de certains groupes et fonctionnalités chimiques dans la molécule.
- ✓ Nombre total de liaisons dans la molécule.
- ✓ Nombres absolu et relatif de liaisons simples, doubles, triples, aromatiques ou autres existantes dans la molécule.
- ✓ Nombre total des noyaux, nombre de noyaux divisé par le nombre d'atomes.
- ✓ Nombre total et relatif de phényle ou d'autres noyaux aromatiques.

- ✓ Poids moléculaire et poids atomiques moyens.

Ces descripteurs sont plus utilisés en raison de leur simplicité extrême, et du point de vue du coût calculatoire. Cependant, ils sont peu sensibles aux changements conformationnels, donc n'arrivent pas à distinguer entre deux isomères.

## 2. Les descripteurs topologiques

Les descripteurs topologiques sont basés sur les principes de la théorie des graphes. Ils codent les types d'atomes, de liaisons, des sous graphes, et aussi la nature des interconnexions entre les atomes de la molécule.

Ces descripteurs sont apparus en tant que composants importants dans plusieurs études de QSAR et de QSPR, indiquant toutefois que la topologie est un paramètre lié à beaucoup de propriétés biologiques et physiques.

En général, les descripteurs topologiques peuvent être calculés rapidement et à partir des structures sans optimisation de la géométrie. Ceci en fait des descripteurs de choix notamment dans les modèles qui sont conçus pour examiner de grandes bibliothèques ou bases de données des composés à intérêt biologique. Les différents types de descripteurs distingués dans ce bloc sont brièvement expliqués ci-dessous<sup>5</sup>:

### A. Descripteurs dérivés de la matrice d'adjacence

Le degré de sommets d'un atome ( $i$ ) est la somme correspondante de la ligne ( $i$ ) de la matrice d'adjacence qui rassemble l'information sur les paires d'atomes connectés dans un graphe moléculaire à hydrogène réduits (sans compter les atomes d'hydrogène). Ces descripteurs moléculaires sont principalement liés à la ramification moléculaire (Tableau 1).

**Tableau 1.** Descripteurs dérivés de la matrice d'adjacence

Formule	Nom du descripteur
$M_1 = \sum_a \delta_a^2 = \sum_g g^2 \cdot {}^g F$	1 <sup>er</sup> indice de Zagreb
$M_2 = \sum_b (\delta_i \cdot \delta_j)_b$	2 <sup>ème</sup> indice de Zagreb
$Q = \frac{\sum_g (g^2 - 2g) \cdot {}^g F + 2}{2}$	Indice quadratique Normalisé
$r = \sum_{\delta_i > 2} (\delta_i - 2)$	Indice de ramification

Avec :

a et b parcours tous les atomes de la molécule.

b: parcours toutes les liaisons de la molécule.

$\delta_i$  et  $\delta_j$  : les deux degrés de sommets i et j incidents à la liaison B.

g: un ordre correspond à différentes valeurs pour le degré de sommets.

${}^g F$  : compte du degré de sommets de l'ordre g.

### ***B. Descripteurs dérivés de la matrice de distance :***

Plusieurs descripteurs topologiques sont dérivés à partir de l'application de différentes opérations algébriques sur la matrice de distance qui rassemble les distances topologiques entre les paires d'atomes. Le degré de distance d'un atome (*i*) est la somme correspondante de la ligne (*i*) de la matrice de distance. Les indices topologiques basés sur les distances topologiques sont décrits ci-dessous (Tableau 2).

**Tableau 2.** Descripteurs dérivés de la matrice de distance

Formule	Nom du descripteur
$W(G) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}$	Indice de Wiener
$J = \frac{B}{C+1} \sum_b (\sigma_i \cdot \sigma_j)_b^{-\frac{1}{2}}$	Indice de Connectivité de distance de Balaban
$H = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^{-1}$	Indice de Harary

Avec :

$d_{ij}$  : Élément de la matrice de distance  $D$ .

$d_{ij}^{-1}$  : Élément de la matrice de distance réciproque  $D^{-1}$ .

$N$  : le nombre total des atomes de la molécule.

$\sigma_i$  et  $\sigma_j$  : degrés de distances de sommet de deux atomes adjacents.

$b$  : liaison moléculaire.

$B$  : le nombre total des liaisons présentées par le graphe moléculaire.

$C$  : le nombre de noyaux (cyclomatic number).

### ***C. Descripteurs dérivés de la matrice de distance pondérée :***

Le serveur E-DRAGON1 calcule cinq matrices de distance pondérées utilisant les cinq propriétés atomiques suivantes :

- Le nombre atomique ( $z$ ) ;
- La masse atomique ( $m$ ) ;
- Le volume atomique de van der Waals ( $v$ ) ;
- L'électronégativité atomique de Sanderson ( $e$ ) ;

- et la polarisabilité atomique ( $\rho$ ).

Parmi lesquels on trouve :

*Les indices dérivés de la matrice de distance pondérée - type de Wiener :* sont calculés en employant la même formule que l'indice de Wiener (W) appliquée à chaque matrice de distance pondérée.

*Les indices dérivés de la matrice de distance pondérée - type de Balaban :* sont calculés en employant la même formule que l'indice de connectivité de distance de Balaban (J) appliquée à chaque matrice de distance pondérée.

#### **D. Descripteurs dérivés de la matrice Laplacienne :**

Plusieurs types de descripteurs ont été proposés en se basant sur les éléments de la matrice Laplacienne, cette dernière est une matrice symétrique carrée représentant un graphique moléculaire à hydrogènes réduits, dont les éléments diagonaux correspondant aux degrés de sommets des atomes de la molécule, les éléments non – diagonaux des atomes adjacents (liés directement) prennent la valeur -1, les autres prennent la valeur 0.

**Tableau3.** Descripteurs dérivés de la matrice Laplacienne

<b>Formule</b>	<b>Nom du Descripteur</b>
$W^* = N \cdot \sum_{i=1}^{N-1} \frac{1}{\lambda_i}$	Indice de Quasi-Wiener
$(TI)_1 = 2 \cdot N \cdot \log\left(\frac{B}{N}\right) \sum_{i=1}^{N-1} \frac{1}{\lambda_i}$	1 <sup>er</sup> indice de Mohar
$(TI)_2 = \frac{4}{N \cdot \lambda_{N-1}}$	2 <sup>ème</sup> indice de Mohar
$T^* = \frac{1}{N} \prod_{i=1}^{N-1} \lambda_i$	Spanning Tree

Avec :

$N$  : Nombre d'atomes.

$\lambda_i$  : Valeurs propres issus de la diagonalisation de la matrice Laplacienne.

$\lambda_{N-1}$  : La première valeur propre différente de zéro.

$B$  : Nombre de liaisons

### 3. L'indice de connectivité

L'indice de connectivité est parmi les indices topologiques les plus populaires et est calculé à partir du graphe moléculaire à hydrogène réduit, dont chaque sommet (les éléments de la matrice d'adjacence) est exprimé par son degré de sommet ( $\delta$ ).

L'indice de connectivité de Randić<sup>5</sup> ou indice de ramification était le premier indice proposé, reliant seulement deux sommets. Cet indice a été étendu ensuite par Kier et Hall pour relier les sous graphes d'une molécule. En général, les indices de connectivité d'ordre  $m$  se calculent selon l'équation (Equ.5):

$${}^m\chi_t^v = \sum_{j=1}^s \prod_{i=1}^n (\delta_i^v)^{-1} \quad (\text{Equ.5})$$

Avec :

$s$ : le nombre totale de l'ordre  $m$  des sous graphes présents dans la molécule;

$n$  : le nombre des sommets existant dans le sous graphe ( $n=m+1$ ) pour les sous graphes acycliques ;

Et  $\delta_i^v$ : est la connectivité atomique de valence calculée par la formule suivante (Equ. 6) :

$\delta_i^v = Z_i^v - \eta_i$  Pour les atomes du premier rang (C, N, O,...).

$$\delta_i^v = \frac{Z_i^v - \eta_i}{Z_i - Z_i^v - 1} \text{ Pour les autres atomes.} \quad (\text{Equ.6})$$

Où :

$Z_i^v$  : Le nombre total d'électrons de valence pour l'atome  $i$  ;

$Z_i$  : Le numéro atomique de l'atome  $i$  ;

$\eta_i$  : Le nombre d'atomes d'hydrogène attachés à cet atome.

#### 4. Les descripteurs géométriques

Les descripteurs géométriques<sup>4</sup> sont définis de différentes manières, mais sont toujours dérivés de la structure tridimensionnelle de la molécule. D'une façon générale, les descripteurs géométriques sont calculés à partir de la géométrie moléculaire optimisée obtenue par des méthodes de chimie informatique ou à partir des données cristallographiques.

La connaissance des positions relatives des atomes dans l'espace tridimensionnel dans la représentation géométrique de la molécule implique que les descripteurs géométriques fournissent habituellement plus d'information et de puissance de discrimination que les descripteurs topologiques, également pour les structures et les conformations moléculaires semblables de la molécule.

En dépit de leur contenu élevé d'informations sur la molécule, les descripteurs géométriques présentent habituellement quelques inconvénients, dont celui de l'optimisation de la géométrie ce qui augmente leur coût de calcul.

Beaucoup de descripteurs moléculaires, appartenant au bloc de descripteurs géométriques, générés par le serveur E-DRAGON1 (74 descripteurs), sont généralement connus en tant

qu'indices topographiques. Ils sont calculés à partir de la représentation graphique des molécules en utilisant les distances géométriques entre les atomes au lieu des distances topologiques. Parmi lesquels en cite :

- **L'indice 3D-Wiener**, qui est un descripteur de forme, représente la somme de toutes les distances géométriques inter atomiques dans une molécule, en y intégrant la totalité des atomes y compris les atomes d'hydrogène. Cet indice montre différentes valeurs pour différentes conformations moléculaires, les plus grandes valeurs correspondent aux conformations les plus prolongées, les plus petites valeurs correspondent aux conformations les plus compactes.
- **L'indice 3D-Balaban** est calculé de la même manière que l'indice de connectivité de distance de Balaban en utilisant les degrés des distances géométriques au lieu des degrés des distances topologiques.

Il existe d'autres types de descripteurs géométriques tels que :

**Les indices gravitationnels** qui sont des descripteurs moléculaires reflétant la distribution de masse dans une molécule, sont définis comme suit (*Equ.7* et *Equ.8*):

$$G1 = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} \frac{m_i m_j}{r_{ij}^2} \quad (\text{Equ.7})$$

$$G2 = \sum_{b=1}^{nBT} \left( \frac{m_i m_j}{r_{ij}^2} \right)_b \quad (\text{Equ.8})$$

Avec :

$m_i$  et  $m_j$  : sont les masses atomiques des deux atomes considérés  $i$  et  $j$ .

$r_{ij}$  : la distance interatomique correspondante ;

$n_{AT}$  et  $n_{BT}$  : le nombre d'atomes et de liaisons respectivement.

**Le rayon de rotation**, qui est un descripteur de taille, calcule la distribution des masses atomiques dans une molécule, donné par l'expression suivante (Equ.9):

$$R_{gyr} = \sqrt{\frac{\sum_{i=1}^{n_{AT}} m_i r_i^2}{MW}} \quad (Equ.9)$$

Avec :

$r_i$  : est la distance de la  $i^{\text{ème}}$  atome du centre de masse de la molécule ;

$m_i$  : est la masse ;

$n_{AT}$  : le nombre d'atomes ;

et  $MW$  : le poids moléculaire.

**L'excentricité moléculaire** est un descripteur de forme calculé à partir des valeurs propres  $\lambda$  de la matrice d'inertie moléculaire (Equ.10) :

$$MEcc = \frac{(\lambda_1^2 - \lambda_3^2)^{1/2}}{\lambda_1} \quad (Equ.10)$$

Il s'étend de 0 à 1, dont les valeurs près de 0 correspondent aux molécules de forme sphérique et ceux près de 1 correspondent aux molécules de forme linéaire.

## 5. Représentation 3D des structures moléculaires basée sur la diffraction électronique

(En anglais : 3D Molecule Representation of Structures Based on Electron Diffraction : **3D MoRSE**)

Pour l'obtention de ces descripteurs qui sont liés directement à l'exploitation de l'information 3D des coordonnées atomiques on utilise un transformé moléculaire. Ce dernier est généralement utilisé dans les études de diffraction électronique afin de préparer des courbes de dispersion théoriques<sup>2,5</sup>.

Le transformé moléculaire est donné par l'Equ.11, où  $i(s)$  est l'intensité d'un rayonnement dispersé provoqué par une série de N atomes situés sur des points à des distances  $r_i$ .

$$i(s) = \sum_{i=1}^N f_i \cdot e^{2\pi i \cdot r_i \cdot s} \quad (\text{Equ.11})$$

$f_i$  facteur de forme, (s) représente la dispersion dans diverses directions, (distance réciproque) et dépend à la fois de l'angle de dispersion  $\nu$  et de la longueur d'onde du faisceau électronique  $\lambda$  comme c'est indiqué par l'Equ. 12.

$$s = 4\pi \sin(\vartheta/2) / \lambda \quad (\text{Equ.12})$$

Ce fondement théorique de diffraction électronique donnée par Wierl, est suivi par des modifications partielles proposées par Soltzberg et Wilkins qui ont introduit la propriété atomique pour les deux atomes  $i$  et  $j$  situés à une distance interatomique  $r_{ij}$  (Equ.13).

$$i(s) = \sum_{i=2}^N \sum_{j=1}^N p_i p_j \frac{\sin sr_{ij}}{sr_{ij}} \quad (\text{Equ.13})$$

Cette méthode de codage structurale a été largement appliquée dans le domaine de QSAR. Ainsi, les descripteurs 3D-MoRSE ont été utilisés avec succès dans la modélisation de différentes propriétés physico-chimiques et sont aussi employés dans la simulation des spectres infrarouges.

## 6. Fonction de distribution radiale

Steinhauer et Gasteiger ont développé une nouvelle famille de descripteurs tridimensionnels basés sur la fonction de distribution radiale (RDF). En général, les descripteurs RDF, dérivés de

cette fonction la plus reconnue dans la physique et en particulier dans la diffraction des rayons X, sont étroitement liés aux descripteurs 3D-MoRSE.

Formellement, l'application de la fonction de distribution radiale sur un ensemble d'atomes (molécule de  $N$  atomes) peut être interprétée comme la distribution de probabilité de trouver un atome dans un volume sphérique de rayon  $R$ .

Les descripteurs RDF sont générés à partir de l'Equ.14, où  $f$  est un facteur de graduation,  $N$  est le nombre total d'atomes dans la molécule,  $p_i$  et  $p_j$  sont les propriétés des atomes  $i$  et  $j$  respectivement.

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j=i+1}^N p_i p_j e^{-B(R-r_{ij})^2} \quad (\text{Equ.14})$$

Le terme exponentiel contient aussi le paramètre doux  $B$  qui peut être considéré comme un facteur de température qui décrit le mouvement des atomes.  $r_{ij}$  est la distance entre les deux atomes  $i$  et  $j$ .  $g(r)$  est habituellement calculé à un certain nombre de points discrets dans des intervalles bien définis<sup>1</sup>.

### 7. Descripteurs issus de la valeur propre de Burden

Les descripteurs de BCUT (Burden – CAS – University of Texas eigenvalues) sont proposés à l'origine pour rechercher la ressemblance ou la diversité existante lors du traitement de grandes bases de données, et sont basés sur une prolongation significative de l'approche de Burden<sup>5</sup>.

Les descripteurs BCUT sont calculés à partir d'un graphe moléculaire ne tenant pas compte des atomes d'hydrogène en utilisant la matrice de Burden dérivée de la matrice d'adjacence et définie comme suit : les éléments diagonaux sont les propriétés atomiques [la masse atomique ( $m$ ), volume de van der Waals atomique ( $v$ ), l'électronégativité atomique de Sanderson ( $e$ ), et la polarisabilité atomique ( $p$ )]. Les éléments non diagonaux pour deux atomes liés par une liaison

prennent la valeur de la racine carrée de l'ordre de leur liaison. Pour tous les autres (atomes non liés) la valeur attribuée est 0.001.

## 8. Les descripteurs WHIM

Les descripteurs de WHIM (Weighted Holistic Invariant Molecular descriptors) sont basés sur des indices statistiques calculés à partir des projections des atomes sur les axes principaux. Ils sont établis de façon à saisir les informations moléculaires 3D appropriées notamment celles de la taille, la forme, la symétrie et la distribution atomique en respectant les structures de référence invariants. Ils sont divisés en deux classes principales : descripteurs de WHIM directionnels et descripteurs de WHIM globaux<sup>4,6</sup>.

Dans l'approche de WHIM, la molécule est vue comme une configuration de points (les atomes) dans un espace tridimensionnel définie par les axes cartésiens (X, Y, Z). Afin d'obtenir une structure de référence unique, les axes principaux sont calculés. Ensuite, les projections des atomes sont effectuées sur chacun des trois axes principaux avec l'évaluation de la distribution des atomes autour du centre géométrique.

En effet, l'algorithme consiste à calculer les valeurs propres et les vecteurs propres d'une matrice de covariance pondérée par des propriétés atomiques. Les éléments de cette matrice sont donnés par l'Equation 15 :

$$s_{jk} = \frac{\sum_{i=1}^N w_i (q_{ij} - \bar{q}_j)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^N w_i} \quad (\text{Equ.15})$$

Avec  $n$  : le nombre des atomes dans la molécule.

$w_i$ : la propriété de l'atome  $i$ .

$q_{ij}$  et  $q_{ik}$  : la  $j^{\text{ème}}$  et  $k^{\text{ème}}$  coordonnées cartésiennes ( $j, k = x, y, z$ ) de l'atome  $i$ .

$\bar{q}_j$  et  $\bar{q}_k$  : la valeur moyenne pour la  $j^{\text{ème}}$  et la  $k^{\text{ème}}$  coordonnée ( $q_{ij}$  et  $q_{ik}$ ).

### 9. Les descripteurs GETAWAY

Les descripteurs GETAWAY (GEometry, Topology, and Atom-Weights Assembly) ont été récemment proposés en tant que descripteurs de structure chimique issus d'une nouvelle représentation de la structure moléculaire, dont l'information est stockée dans une matrice nommée la matrice d'influence moléculaire (MIM : Molecular Influence Matrix), désignée par  $H$  et définis comme suit (Equ. 16) :

$$H = M.(M^T.M)^{-1}.M^T \quad (\text{Equ.16})$$

Avec :

$M$  : la matrice moléculaire comprenant les coordonnées cartésiennes centrées ( $x, y, z$ ) des atomes de la molécule à hydrogènes réduits ;

$T$  : se rapporte à la matrice transposée.

Les éléments diagonaux  $h_{ii}$  de la matrice d'influence moléculaire  $H$ , appelés les forces de levier, rangées entre 0 et 1, codent l'information atomique liée à l'*influence* de chaque atome en déterminant la conformation de la molécule. En fait, les atomes qui se trouvent éloignés du centre géométrique moléculaire possèdent toujours une valeur de  $h_{ii}$  plus élevée par rapport à ceux situés près du centre de la molécule. D'autre part, la magnitude de levier  $h_{ii}$  maximal dans une molécule dépend de la taille et de la forme de la molécule. Au regard de la géométrie de la molécule, les valeurs  $h_{ii}$  sont efficacement sensibles aux changements conformationnels significatifs et aux longueurs de liaisons (qui expliquent le type d'atomes et la multiplicité en liaisons).

Chaque élément  $h_{ij}$  non-diagonal représente le degré d'accessibilité de l'atome  $j$  aux interactions avec l'atome  $i$ , ou en d'autres termes, la tendance des deux atomes considérés à agir entre eux. Un

signe négatif pour les éléments non-diagonaux signifie que les deux atomes occupent des régions relativement opposées par rapport au centre, par conséquent le degré de leur accessibilité mutuelle devrait être faible<sup>5</sup>.

Parmi les descripteurs dérivés de cette famille, nous citons :

Le moyen géométrique basé sur la valeur de levier (HGM) : il a été proposé pour avoir des informations liées à la forme moléculaire. En effet, il a été constaté que dans une série des isomères hydrocarbonés, le HGM croît avec le changement d'une forme moléculaire linéaire à une autre plus ramifiée, il est lié également inversement à la taille moléculaire, c'est à dire qu'il diminue lorsque le nombre d'atomes dans la molécule augmente. Ce descripteur est donné par l'expression suivante :

$$HGM = 100. \left( \prod_{i=1}^n h_{ii} \right)^{\frac{1}{2}} \quad (\text{Equ.17})$$

Descripteurs dérivés de la somme moyenne des lignes de la matrice R (RARS): ils sont basés sur les sommes des lignes de la matrice R résultante de la combinaison entre les deux matrices (valeur de levier et de distance), dont leurs éléments sont définis comme suit :

$$[R]_{ij} = \left[ \frac{\sqrt{h_{ii} - h_{jj}}}{r_{ij}} \right]_{ij} \quad \text{Avec } i \neq j \quad (\text{Equ.18})$$

$$RARS = \frac{1}{nAT} - \sum_{i=1}^{nAT} \sum_{j=1}^{nAT} \frac{\sqrt{h_{ii} - h_{jj}}}{r_{ij}} \quad (\text{Equ.19})$$

Avec :

$h_{ii}$  et  $h_{jj}$  : représentant les valeurs de levier pour les deux atomes  $i$  et  $j$  respectivement, séparés par une distance interatomique  $r_{ij}$ ;

$nAT$  : représente le nombre total des atomes dans la molécule cible.

Ces descripteurs codent l'information utile qui pourrait être liée à la présence des substituants moléculaires significatifs.

## 10. Les descripteurs « Walk and Path Counts »

### A. *Molecular Walk Counts (Compte du Parcours Moléculaire)*

Le compte du parcours (walk) moléculaire est un descripteur obtenu à partir d'un graphe moléculaire à hydrogène réduit et exprime le nombre des parcours. Le compte du parcours moléculaire de l'ordre  $k$  correspond au nombre total des parcours de longueur  $k$  dans le graphe moléculaire, dont un parcours est une suite consécutive d'arêtes (liaisons) dans une molécule et la longueur d'un parcours est son nombre d'arêtes<sup>5</sup>.

### B. *Molecular Path Counts*

Le compte de chemin moléculaire est un descripteur obtenu à partir d'un graphe moléculaire à hydrogène réduit, basé sur le chemin du graphe, qui est un parcours sans aucun sommet ou liaison répétés. Le compte de chemin moléculaire d'ordre  $k$  est le nombre total des chemins de longueur  $k$  dans le graphe moléculaire.

## 11. Descripteurs topologiques de charge

Les descripteurs topologiques de charge sont dérivés d'une matrice non symétrique (CT), dont les éléments sont définis comme suit<sup>5</sup> (Equ.20):

$$CT_{ij} = \begin{cases} \delta_i & \text{si } i = j \\ m_{ij} - m_{ji} & \text{si } i \neq j \end{cases} \quad (\text{Equ.20})$$

Avec :

$\delta_i$ : le degré de sommet de l'atome  $i$ .

$m_{ij}$ : les éléments de la matrice  $\mathbf{M}$  obtenus en multipliant la matrice d'adjacence  $\mathbf{A}$  par la matrice carrée réciproque de distance  $\mathbf{D}^{-2}$ .

$$M = A.D^{-2}$$

Les entrées diagonales de la matrice **CT** représentent la valence topologique des atomes ; alors que les entrées non diagonales  $CT_{ij}$  représentent une mesure de la charge nette transférée à partir de l'atome  $j$  à l'atome  $i$ .

## 12. Les descripteurs 2D- autocorrelation

Les descripteurs 2D- autocorrélation calculés par le serveur E-Dragon1 sont des autocorrélations spatiales calculées à partir d'un graphe moléculaire (à hydrogènes réduits) pondéré par des propriétés atomiques ( $w_i$ ). Ils décrivent comment une propriété considérée  $w_i$  soit distribuée sur une structure moléculaire topologique, et se divisent en trois types<sup>2,5</sup> :

**A. Autocorrelation de la structure topologique (ATS)** : proposé par Moreau et Broto et sont donnés par l'expression suivante :

$$ATSkw = \sum_{j=1}^{nSK-1} \sum_{j>1} w_i \cdot w_j \cdot \delta_{ij} \quad (\text{Equ.21})$$

$w_i$ : la propriété atomique ;

$nSK$  : le nombre d'atomes non hydrogène ;

$k$  : longueur d'un chemin et nommé aussi « lag » ;

$\delta_{ij}$  : delta de Kronecker ( $\delta_{ij}=1$  si la distance topologique entre les deux atomes considérés  $d_{ij}=k$  ; ou  $\delta_{ij}=0$  pour les autres cas).

**B. Autocorrelation de la structure topologique de Moran (MATS) :**

$$MATS_{KW} = \frac{\frac{1}{\Delta} \cdot \sum_{i=1}^{nSK} \sum_{j=1}^{nSK} \delta_{ij} \cdot (W_i - \bar{W}) \cdot (W_j - \bar{W})}{\frac{1}{nSK} \cdot \sum_{i=1}^{nSK} (W_i - \bar{W})^2} \quad (\text{Equ.22})$$

$W_i$  : propriété atomique ;

$\bar{W}$  : la valeur moyenne pour les atomes de la molécule ;

$nSK$  : le nombre d'atomes dans la structure non hydrogène ;

$\delta_{ij}$  : delta de Kronecker ;

$d_{ij}$  : la distance topologique entre les deux atomes i et j ;

$\Delta$  : la somme des  $\delta_{ij}$ .

**C. Autocorrelation de la structure topologique de Geary (GATSkw) :**

$$GATSkw = \frac{\frac{1}{2\Delta} \cdot \sum_{i=1}^{nSK} \sum_{j=1}^{nSK} \delta_{ij} \cdot (W_i - W_j)^2}{\frac{1}{(nSK - 1)} \cdot \sum_{i=1}^{nSK} (W_i - \bar{W})^2} \quad (\text{Equ.23})$$

$W_i$  : propriété atomique ;

$\bar{W}$  : la valeur moyenne pour les atomes de la molécule ;

$nSK$  : le nombre d'atomes dans la structure non hydrogène ;

$\delta_{ij}$  : delta de Kronecker ;

$d_{ij}$  : la distance topologique entre les deux atomes i et j ;

$\Delta$  : la somme des  $\delta_{ij}$ .

## V. CONCLUSION

Ce chapitre a été consacré aux descripteurs, nécessaires pour faire une jonction entre la structure avec une activité biologique dans les modèles QSAR. Les descripteurs sont une représentation de la molécule et contiennent une information chimique de la molécule. Une classification en 4 catégories des descripteurs, suivant leur type, a été proposée. Les types de descripteurs dépendent de la représentation de la molécule : la formule brute permet de calculer les descripteurs 1-D, la représentation en deux dimensions de la molécule permet de calculer les descripteurs 2-D, tandis que les descripteurs 3-D et 4-D nécessitent l'utilisation des conformères, donc de la représentation tridimensionnelle de la molécule. Dans cette thèse les descripteurs utilisés pour construire les modèles QSAR ceux qui sont générés à partir du serveur E-Dragon1 regroupant un ensemble de plus de 1600 descripteurs de différents types.

**VI. REFERENCES**

- [1] P. Bultinck; H. Winter; W. Langenaeker; J. Tollenaere "Computational Medicinal Chemistry for Drug Discovery" Ed. Taylor & Francis Group LLC. Etats Unies, 2004.
- [2] J. Gasteiger "Handbook of Chemoinformatics: *From Data to Knowledge (4 Volumes)*" Ed. WILEY-VCH Verlag, Weinheim. Allemagne, 2003.
- [3] J.A. Bondy; U.S.R. Murty "Graph theory" Ed. Springer, 2008.
- [4] M. Karelson "Molecular Descriptors in QSAR/QSPR" Ed. Wiley- Interscience. Etats Unies, 2000.
- [5] R. Todeschini; V. Consonni "Handbook of Molecular Descriptors: Methods and Principles in Medicinal Chemistry" Ed. Wiley – VCH, Weinheim. Allemagne, 2000.
- [6] R. Todeschini; P. Gramatica "3D QSAR in Drug Design" Ed. Kluwer/ESCOM, Dordrecht. Pays Bas, 1998.

## CHAPITRE IV:

# MODELISATION DE L'INHIBITION DE L' $\alpha$ -GLUCOSIDASE PAR LES DERIVES DES XANTHONES ET CURCUMINOIDES

## I. INTRODUCTION

Les  $\alpha$  - Glucosidases ( $\alpha$ -D-glucoside glucohydrolase EC. 3.2.1.20) sont des enzymes attachées à la membrane et situées à l'épithélium du petit intestin. Elles sont responsables de la digestion de la nourriture riche en hydrate de carbone (carbohydrate digestion). Elles hydrolysent la liaison  $\alpha$ - du glucopyranoside, libérant ainsi le  $\alpha$  - D- glucose de l'extrémité non réductrice du sucre<sup>1</sup>.

L'intérêt intense donné aux inhibiteurs de l' $\alpha$ -glucosidase (naturels et synthétiques) par les chimistes, biochimistes et pharmacologues, a contribué à leur développement en chimie. Ces derniers ont facilité l'explication du mécanisme d'action de l'  $\alpha$  -glucosidase et participé au développement des produits pharmaceutiques potentiels tels que les agents antitumoraux, antiviraux, antidiabétiques, immunorégulateurs,... etc<sup>2</sup>.

Cette enzyme est largement présente dans les tissus des micro-organismes, des plantes, et ceux des animaux, bien que la spécificité du substrat diffère considérablement selon la source de l'enzyme  $\alpha$  - glucosidase. Il y a des inhibiteurs de l'  $\alpha$  -glucosidase tels que l'acarbose et le voglibose issus des micro-organismes et le nojirimycin et 1-deoxynojirimycin issus des plantes<sup>3,4</sup>.

Le manque d'information structurale sur la nature des interactions existe entre les  $\alpha$ -glucosidases et les inhibiteurs lui a rendu difficile la découverte de bons précurseurs. Les études quantitatives structure-activité (QSAR) constituent une méthode puissante pour la conception des composés biologiquement actifs ainsi que pour la prévision de l'activité selon les propriétés physiques et chimiques résultantes<sup>5-8</sup>.

Récemment une étude de QSAR a été rapportée sur une base de 43 composés dérivés de xanthones comme inhibiteurs de l'  $\alpha$ -glucosidase<sup>9</sup>, au moyen de descripteurs électroniques, en utilisant l'analyse de régression linéaire multiple en combinaison avec l'élimination pas à pas en tant qu'algorithme de sélection. Ils ont trouvé que l'activité inhibitrice peut être modélisée par le nombre de liaisons hydrogène, le nombre de noyaux aromatiques et la valeur de mollesse.

Dans cette étude, notre objectif est de modéliser l'activité et de former un modèle de QSAR présentant de bons paramètres statistiques. Pour le faire, nous avons joué sur les trois points essentiels suivants :

- La base de données ;
- la nature des descripteurs utilisés ;
- l'algorithme employé dans la sélection des descripteurs.

Nous avons construit nos modèles de QSAR en premier lieu en utilisant tous les descripteurs du serveur E-DRAGON1 sur un ensemble de données contenant 57 molécules dérivées des motifs Xanthones et curcuminoïdes. Les descripteurs moléculaires sont sélectionnés par l'application des algorithmes génétiques.

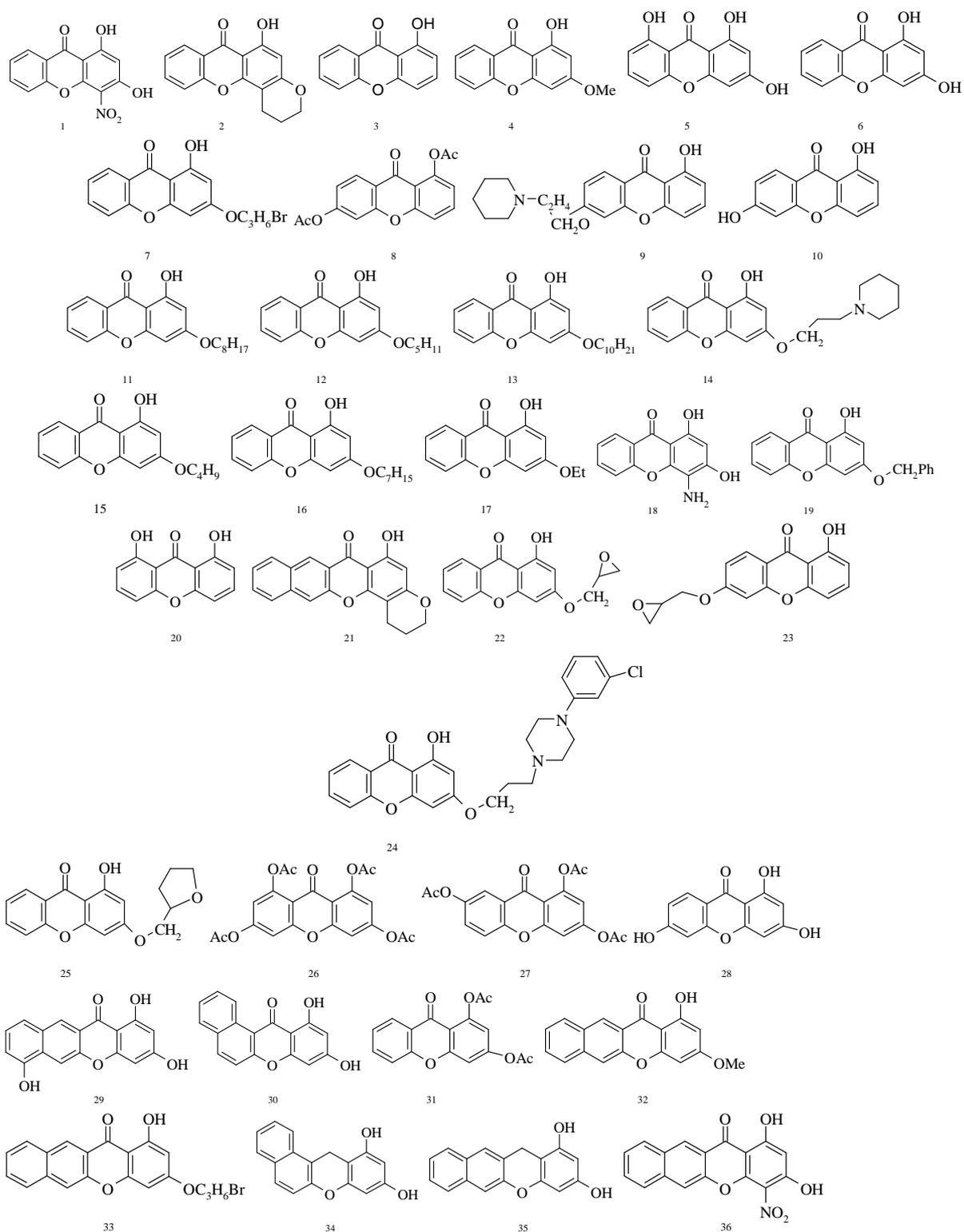
En second lieu nous avons travaillé avec les descripteurs dérivants de la famille 3D-MoRSE sur le même ensemble de données en utilisant la même procédure de sélection (GA).

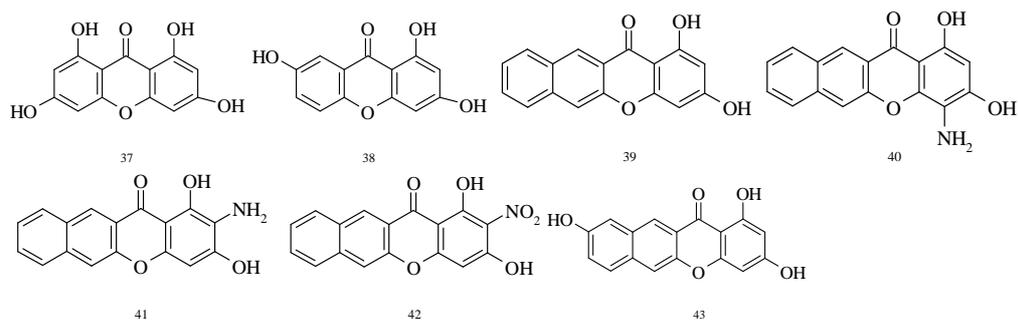
## II. METHODES EXPERIMENTALES

### 1. Ensemble des données

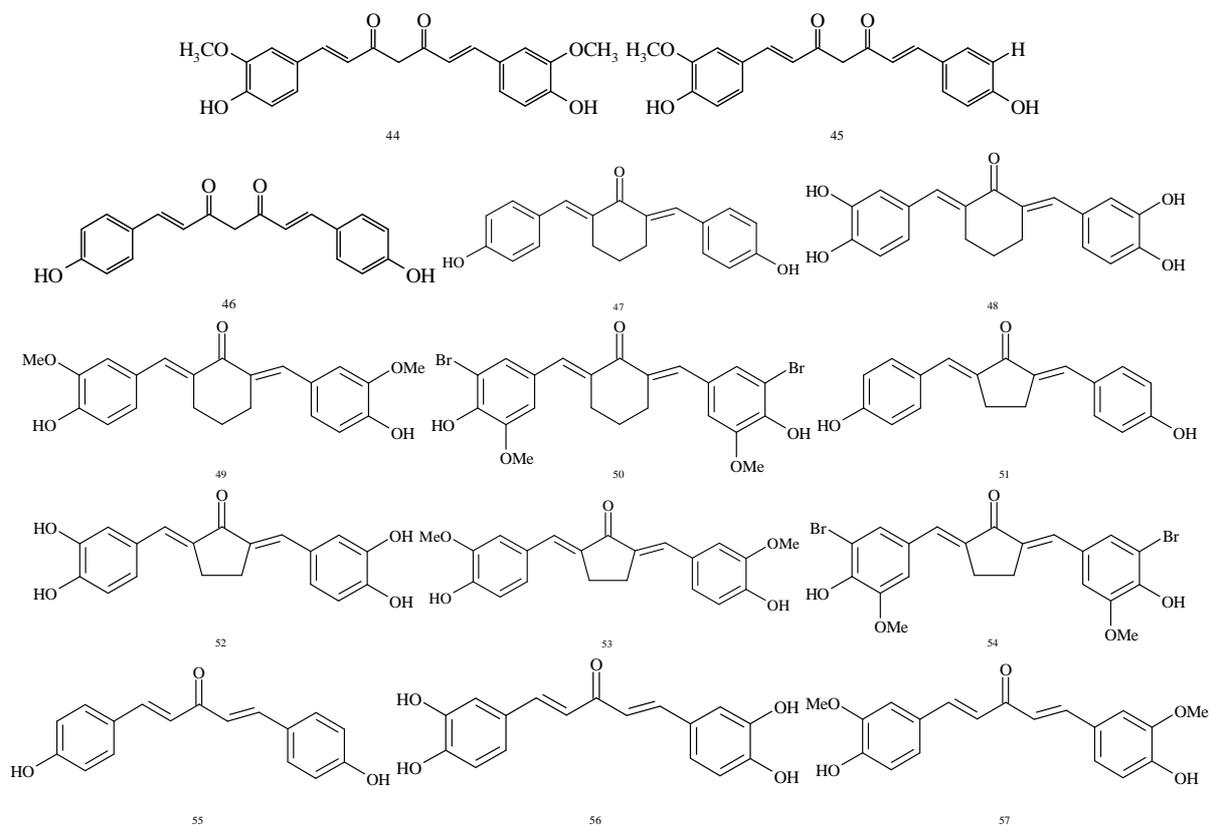
Pour cette étude, nous avons choisi de travailler sur 57 molécules dérivés des xanthones et de curcuminoïdes (Figure 1), parmi celles ayant une valeur IC50 représentant l'activité inhibitrice de l' $\alpha$ -glucosidase. Ces données ont été tirées des travaux de Liu<sup>9</sup> et Du<sup>10</sup>.

Dérivés de xanthones





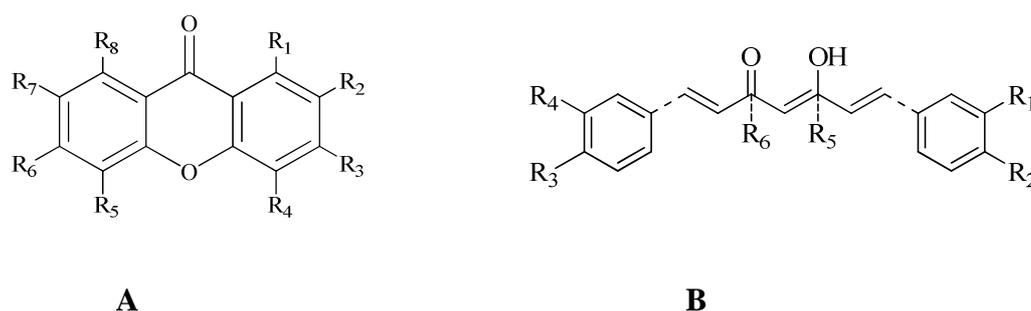
### Dérivés du Curcuminoïde



**Figure 1.** Structures développées des dérivés des xanthonés et curcuminoïdes

L'activité biologique inhibitrice est rapportée en terme  $IC_{50}$  : concentration micro molaire d'une drogue, nécessaire pour inhiber 50% (la moitié) de l'activité enzymatique. Pour notre cas nous avons exprimé l'activité inhibitrice par le rapport logarithmique  $pIC_{50}$  [ $LOG(1/IC_{50})$ ].

Les molécules de cet ensemble contiennent dans leurs structures des atomes de C, N, O, H et deux halogènes (Br, Cl) avec un nombre total d'atomes variant de 24 à 58 et un poids moléculaire variant de 212.210 à 524.220 g/mol. Les représentations structurales des molécules présentées dans la Figure 2 (A et B) donnent une idée sur la complexité structurale de cet ensemble.



**Figure 2.** Les deux motifs inhibiteurs de l' $\alpha$ -glucosidase

Celui-ci possède une diversité structurale (bicyclic, tricyclic et tétracyclic) importante et de nombreuses substitutions fonctionnelles. De plus, la présence de nombreux isomères de position présentant une différence d'activité semble être intéressante pour tester le pouvoir de diversification des descripteurs utilisés.

Cet ensemble de données présente une gamme d'activité biologique comprise entre  $pIC_{50} = -2.371$  et  $-0.204$ . Les valeurs  $pIC_{50}$  observées et prédites des structures moléculaires sont listées dans le Tableau 3 (page 82) et le Tableau 11 (page 99). Les composés contenant des ions sont exclus directement de l'ensemble, car le serveur E-DRAGON1 est incapable d'encoder les ions. Les composés retenus (57 molécules) ont tous une valeur précise de l'activité.

L'ensemble des molécules est divisé aléatoirement en deux sous-ensembles : un sous-ensemble de 13 molécules (environ 20%) choisies au hasard, en tant qu'ensemble de Prédiction (PSET), pour évaluer le pouvoir prédictif des modèles proposés. Le reste des

molécules 45 (environ 80%) constitue l'ensemble d'apprentissage (Training Set : TSET) utilisé dans la construction du modèle de QSAR.

## 2. Dessin et optimisation des structures

Les représentations 2D et 3D ont été réalisées à l'aide du programme Hyperchem. Le même logiciel est utilisé pour la recherche de la conformation la plus stable (énergétiquement) en utilisant la méthode semi empirique AM1. L'optimisation des structures est effectuée en employant l'algorithme Polack- Ribiere avec un gradient énergétique égal à 0.01 kcal/mol.Å. La structure électronique des composés à l'état fondamentale est calculée à l'aide de la méthode RHF (Restricted Hartree Fock).

## 3. Génération des Descripteurs

Les descripteurs utilisés sont ceux proposés par le serveur E-DRAGON1 qui comprend 1664 descripteurs répartis en 20 blocs comprenant les descripteurs 0D, 1D, 2D et 3D (Tableau 1) :

**Tableau 1.** Différentes classes de descripteurs

Dimension	Classes de descripteurs
0D	constitutionnels
1D	Compte des groupements fonctionnels
2D	indices d'information, indice de la charge topologique, topologiques, indices de connectivité, 2D autocorrélation, valeurs propres de Burden, indices à base de valeurs propres
3D	RDF, WHIM, géométriques, 3D-MoRSE, GETAWAY
D'autres descripteurs	de charge et propriétés moléculaires

## 4. Sélection des variables et formation du modèle

Cette procédure (sélection de variable) constitue une partie cruciale lors du développement du modèle de QSAR. Elle sert à réduire le nombre de descripteurs en deux étapes successives : sélection objective et subjective.

La sélection objective utilise seulement les variables indépendantes (les descripteurs) pour filtrer tous les descripteurs non utiles, sans employer la variable dépendante  $pIC_{50}$ . Ce procédé implique :

- Tous les descripteurs présentant les mêmes valeurs pour toutes les molécules ont été éliminés.
- Ensuite, nous ordonnons par voie décroissante les descripteurs en fonction de leurs variances. Entre temps, on calcule la matrice de corrélation de tout l'ensemble de descripteurs. A partir de cette matrice, nous éliminons un des deux descripteurs qui ont prouvé un coefficient de corrélation au-dessus de 0,9 ( $R > 0,9$ ) et présentant la petite variance.

Après élimination par la sélection objective, le nombre de descripteurs est, encore une fois, réduits par la sélection subjective en cherchant un sous ensemble optimal riche en informations. Dans cette étape on prend en considération la variable dépendante  $pIC_{50}$  pour le choix des descripteurs, et pour cela nous avons fait appel aux algorithmes génétiques (GA) comme méthode de réduction.

Les conditions de simulation par les GA appliquées dans ce contexte étaient  $10^4$  générations, le nombre de croisement est  $5 \cdot 10^3$ , le facteur de smoothness est 1, la probabilité de mutation pour ajouter un nouveau terme est 50%, et la population contient 300 modèles.

Le procédé des algorithmes génétiques est répété plusieurs fois pour confirmer que les descripteurs choisis constituent le sous-ensemble optimal pour expliquer l'activité biologique.

Le programme des algorithmes génétiques est écrit en langage C, et exécuté sur un PC personnel 3.0 GHz.

### III. RESULTATS ET DISCUSSION

#### 1. Sélection du modèle optimale en utilisant l'ensemble de descripteurs (1664)

Un bon algorithme de sélection de descripteurs et un nombre élevé d'observations (ensemble des données) sont recommandés afin d'établir un modèle de QSAR statistiquement fiable. Le modèle de QSAR est développé en utilisant des descripteurs théoriques calculés à partir d'un échantillon de 57 molécules, dérivées des xanthones et curcuminoïdes, testées en tant qu'inhibiteurs de l' $\alpha$ -glucosidase (Tableau 3).

L'application des algorithmes à l'ensemble des descripteurs nous a permis d'obtenir une série de modèles optimaux de différentes dimensions. La détermination du nombre de descripteurs par le modèle de QSAR est importante, car elle empêche dans une certaine mesure les corrélations de chance entre les descripteurs qui constituent le modèle.

Pour déterminer le nombre optimal de descripteurs dans le modèle de QSAR, nous avons utilisé une méthode simple : le point d'arrêt (breacking point). Nous avons établi consécutivement plusieurs modèles de différentes dimensions allant de 3 jusqu'à 8 descripteurs par modèle. Le modèle optimal correspondant au point d'arrêt est obtenu par l'analyse de la représentation graphique des coefficients de détermination  $R^2$  et de la validation  $Q^2_{CV-LOO}$  en fonction du nombre de descripteurs.

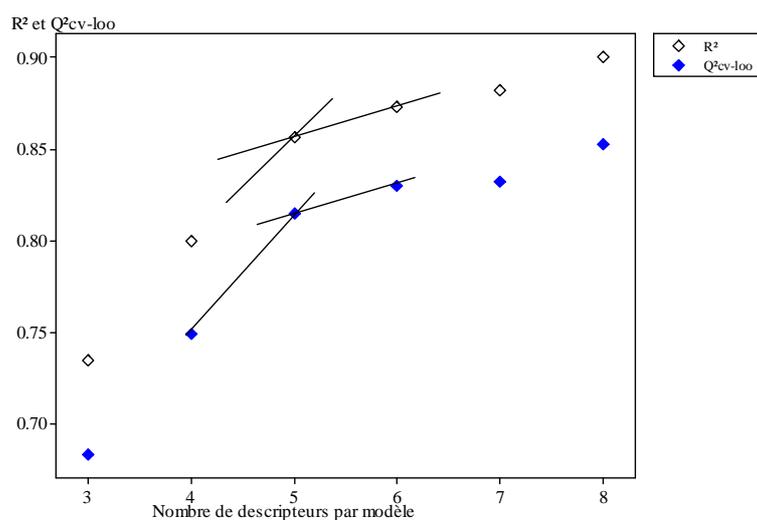
Le tableau 2 présente la série des modèles optimums obtenus à différentes dimensions allant de 3 à 8 descripteurs par modèle, ainsi que leurs coefficients de détermination  $R^2$  et de validation  $Q^2$  qui varie de 0,735 à 0,900 et de 0,684 à 0,853 respectivement.

**Tableau 2.** Série des modèles optimums obtenus à différentes dimensions

L'espace du modèle	Les descripteurs par modèle	R <sup>2</sup>	Q <sup>2</sup> <sub>cv-100</sub>
3	MATS7v; Mor15u; nArOR	0.7348	0.6835
4	MATS7v; Mor15u ; R4u; nArOR	0.7996	0.749
<b>5</b>	<b>MATS7v; Mor15u; H5u; R4e+; nArOR</b>	<b>0.8569</b>	<b>0.815</b>
6	MATS7v; Mor15u; H5u; HATS5u; R4e+; nArOR	0.8733	0.8298
7	MATS7v; Mor15u; H5u; HATS5u; H8m; R4e+; nArOR	0.8822	0.8321
8	BIC2; MATS7v; HOMA; Mor15u; Mor16u; H5u; R4e+ ; nArOR	0.9003	0.8531

Les définitions des différents descripteurs affichés dans le tableau 3 sont rapportées dans l'annexe.

La figure 3 montre la projection de R<sup>2</sup> et Q<sup>2</sup><sub>cv-100</sub> en fonction du nombre de descripteurs en utilisant les algorithmes génétiques. D'après la technique du point d'arrêt (breaking point) le nombre optimal de descripteurs est égale à cinq, avec R<sup>2</sup> = **0.8569** et Q<sup>2</sup><sub>cv-100</sub> = **0.815**.

**Figure 3.** Le nombre optimal de descripteurs du modèle optimal obtenu par GA

**Tableau 3.** Valeurs pIC50 expérimentales et prédites pour TSET et PSET avec les descripteurs sélectionnés

No.	pIC50exp.	pIC <sub>50</sub> pred.	MATS7v	Mor15u	H5u	R4e+	nArOR
<b>TSET</b>							
<b>No.</b>							
1	-2.371	-2.071	-0.681	1.055	0.484	0.074	1.000
2	-2.297	-2.086	-0.205	0.717	0.363	0.097	2.000
4	-2.238	-2.162	-0.357	1.126	0.573	0.072	2.000
5	-2.206	-2.024	-0.538	1.093	0.377	0.064	1.000
7	-2.143	-2.026	-0.111	0.811	0.468	0.078	2.000
8	-2.123	-1.94	-0.248	0.764	0.528	0.038	1.000
9	-2.119	-2.06	-0.217	0.362	1.510	0.059	2.000
11	-2.092	-2.105	-0.226	0.322	1.623	0.050	2.000
12	-2.082	-2.139	-0.209	0.558	1.104	0.058	2.000
14	-2.062	-2.182	-0.256	0.204	1.488	0.064	2.000
16	-2.045	-2.13	-0.242	0.429	1.470	0.052	2.000
17	-2.010	-2.163	-0.240	0.915	0.548	0.071	2.000
18	-1.992	-1.609	-0.501	1.456	0.481	0.086	1.000
19	-1.961	-2.042	-0.236	1.080	0.692	0.066	2.000
20	-1.913	-2.035	-0.346	0.717	0.333	0.064	1.000
22	-1.823	-1.738	0.099	0.990	0.521	0.067	2.000
23	-1.803	-1.925	-0.044	1.005	0.552	0.065	2.000
24	-1.791	-1.724	-0.073	0.270	1.989	0.060	2.000
25	-1.724	-1.495	0.163	1.130	0.778	0.063	2.000
26	-1.696	-1.564	-0.262	1.176	1.238	0.020	1.000
28	-1.618	-2.026	-0.604	1.310	0.371	0.073	1.000
29	-1.601	-1.19	-0.037	1.391	0.655	0.061	1.000
30	-1.543	-1.589	-0.206	1.300	0.542	0.048	1.000
31	-1.504	-1.758	-0.262	0.922	0.712	0.051	1.000
32	-1.496	-1.718	-0.185	1.529	0.876	0.064	2.000
33	-1.473	-1.744	-0.098	1.124	0.807	0.079	2.000
34	-1.444	-1.215	-0.244	1.370	0.631	0.102	1.000
35	-1.303	-1.009	-0.160	1.366	0.658	0.115	1.000
36	-1.233	-1.288	-0.308	1.463	0.481	0.109	1.000
38	-0.968	-1.442	-0.204	1.420	0.422	0.074	1.000
40	-0.903	-1.058	-0.213	1.879	0.752	0.079	1.000
41	-0.799	-0.873	-0.213	1.783	0.899	0.106	1.000
42	-0.771	-1.052	-0.308	1.377	0.696	0.158	1.000
43	-0.763	-0.723	0.208	1.720	0.695	0.065	1.000
45	-1.630	-1.231	-0.076	0.742	1.382	0.063	1.000
46	-1.362	-1.529	-0.274	0.328	0.931	0.048	0.000
48	-0.447	-0.694	0.071	0.985	0.974	0.066	0.000
49	-1.672	-1.764	0.105	0.161	1.701	0.052	2.000
50	-1.467	-1.434	-0.049	1.479	1.399	0.048	2.000
51	-1.512	-1.236	-0.228	0.845	0.585	0.058	0.000
52	-0.415	-0.556	0.032	1.473	0.687	0.073	0.000
53	-1.723	-1.735	0.105	0.184	1.700	0.053	2.000
54	-1.530	-1.479	-0.063	1.641	1.188	0.044	2.000
55	-1.338	-1.492	-0.160	0.452	0.680	0.039	0.000

Tableau 3. (Suite)

No.	pIC <sub>50</sub> exp.	pIC <sub>50</sub> pred.	MATS7v	Mor15u	H5u	R4e+	nArOR
<b>PSET</b>							
3	-2.249	-1.812	-0.179	0.637	0.287	0.076	1.000
6	-2.166	-1.974	-0.477	1.000	0.327	0.075	1.000
10	-2.114	-2.012	-0.477	0.943	0.331	0.073	1.000
13	-2.063	-2.131	-0.200	0.060	1.862	0.045	2.000
15	-2.056	-2.218	-0.259	0.627	0.879	0.064	2.000
21	-1.828	-1.707	-0.127	1.149	0.675	0.095	2.000
27	-1.667	-1.571	-0.047	0.859	0.978	0.025	1.000
37	-1.167	-1.991	-0.635	1.437	0.424	0.061	1.000
39	-0.919	-1.401	-0.176	1.350	0.596	0.065	1.000
44	-1.571	-1.371	0.038	1.232	1.420	0.061	2.000
47	-1.575	-1.618	-0.164	-0.243	0.873	0.056	0.000
56	-0.204	-0.53	0.208	1.131	0.797	0.059	0.000
57	-1.568	-1.384	0.238	1.131	1.320	0.035	2.000

### A. Analyse de la justesse du modèle optimal

Nous avons reconstruit le modèle optimal établi précédemment, en utilisant l'analyse de régression linéaire multiple dans le logiciel MINITAB 15 (*Equ.1*).

$$\text{pIC}_{50} = - 2.01 + 1.17 \text{ MATS7v} + 0.485 \text{ Mor15u} + 0.414 \text{ H5u} + 6.05 \text{ R4e}^+ - 0.480 \text{ nArOR} \quad (\text{Equ.1})$$

$$N = 44; \quad R = 0.925; \quad R^2 = 85.70\%; \quad R^2_{\text{aj}} = 83.80\%; \quad s = 0.197$$

Le modèle exprimé par l'équation (*Equ.1*) reliant la variable dépendante pIC<sub>50</sub> avec les cinq descripteurs est accompagné par les mesures statistiques cités ci-dessus : coefficient de corrélation multiple R, coefficient de détermination R<sup>2</sup>, coefficient de détermination ajusté R<sup>2</sup><sub>adj</sub>, et l'écart type s.

Le coefficient de corrélation multiple ( $R = 0.925$ ) nous indique qu'il existe une forte corrélation entre les valeurs observées de l'activité biologique et celles prédites par le modèle de régression. En termes de variabilité, le coefficient de détermination R<sup>2</sup> et le coefficient de détermination ajusté montrent respectivement, que 85.70% et 83.70% de la variance des valeurs observées est expliquée par les cinq descripteurs (MATS7v; Mor15u; H5u; R4e<sup>+</sup>; et nArOR).

La dispersion des données est jugée par la petite valeur de l'écart type ( $s = 0.197$ ).

Les valeurs élevées des coefficients mentionnés ci-dessus résument la corrélation des cinq descripteurs avec l'activité biologique. Cette corrélation est vérifiée en examinant le tableau d'ANOVA (Tableau 4) et le tableau des coefficients (Tableau 5).

A partir du tableau d'ANOVA, la présence ou l'absence de corrélation entre l'ensemble de descripteurs et l'activité biologique est vérifiée en examinant la statistique de Fisher  $F_{obs}$ . Pour la réaliser nous avons fait appel aux deux hypothèses : hypothèse nulle et hypothèse alternative.

**Tableau 4.** Résultats de l'analyse de la variance (ANOVA)

Source	DF <sup>a</sup>	SS <sup>b</sup>	MS <sup>c</sup>	F <sub>obs</sub>	P <sup>d</sup>
Régression	5	8.8972	1.7794	45.50	0.000
Erreur Résiduel	38	1.4853	0.0391		
Totale	43	10.3824			

a: degré de liberté ; b : somme des carrées ; c : écart moyen ; d : probabilité.

*L'hypothèse nulle ( $H_0$ ):* « aucun descripteur n'est lié à l'activité biologique :  $\beta_j = 0$  avec ( $j=0 ; 1 ; 2 ; 3 ; 4$ ),  $\beta_j$  : coefficient correspond au descripteur  $j$  »

Cette hypothèse est acceptée si la valeur de la statistique de Fischer observée est inférieure à la valeur  $F_{(0.05 ; 5 ; 38)}$ .

*L'hypothèse alternative ( $H_1$ ):* « il existe au moins un descripteur corrélé avec l'activité biologique ».

Cette hypothèse est acceptée lorsque la valeur de la statistique de Fischer observée est supérieure de la valeur  $F_{(0.05 ; 5 ; 38)}$ .

D'après le Tableau d'ANOVA, la statistique de Fischer observée ( $F_{obs}=45.50$ ) est supérieure à ( $F_{(0.05 ; 5 ; 38)}=2.53$ ), ce qui nous permet d'accepter l'hypothèse alternative et de confirmer qu'il

existe au moins un coefficient différent de zéro c'est-à-dire un descripteur corrélé avec l'activité inhibitrice expliquée par  $pIC_{50}$ .

Nous avons ensuite examiné le Tableau des coefficients pour vérifier la signification de chaque descripteur et sa contribution dans l'explication de l'activité biologique. L'utilisation des valeurs de la statistique t de Student, affichées dans le Tableau ci-dessous pour chaque descripteur, nous permet de vérifier la présence ou l'absence de corrélation entre chaque descripteur et l'activité biologique en se basant sur les deux hypothèses : hypothèse nulle et hypothèse alternative.

**Tableau 5.** Tableau des coefficients

Var. Ind. <sup>a</sup>	Coef <sup>b</sup>	E. T (Coef) <sup>c</sup>	$T_{obs}$ <sup>d</sup>	$p$ <sup>e</sup>
intercepte	-2.0090	0.1760	-11.41	0.000
MATS7v	1.1698	0.1754	6.67	0.000
Mor15u	0.48476	0.07947	6.10	0.000
H5u	0.41445	0.09274	4.47	0.000
R4e+	6.046	1.435	4.21	0.000
nArOR	-0.48043	0.04811	-9.99	0.000

a: variables indépendantes ; b : coefficients ; c : erreur type ;

d : test de student observé ; e : la valeur de probabilité.

*L'hypothèse nulle ( $H_0$ )* : « le descripteur n'est pas lié à l'activité biologique :

$\beta_j = 0$  avec ( $j=0 ; 1 ; 2 ; 3 ; 4$ ),  $\beta_j$  : coefficient correspond au descripteur  $j$  »

Cette hypothèse est acceptée si la valeur de la statistique de Student t observée est inférieure à la valeur  $t_{(0,025 ; 38)}$  à un niveau de confiance  $\alpha = 0.05$ .

*L'hypothèse alternative ( $H_1$ )* : « le descripteur est corrélé avec l'activité biologique :

$\beta_j \neq 0$  ».

Cette hypothèse est acceptée lorsque la valeur de la statistique de Student t observée est supérieure de la valeur  $t_{(0,025,38)} = 2.042$ .

D'après ce Tableau, les valeurs de la statistique observées  $t_{obs}$  sont plus élevées par rapport à celle du Tableau de la distribution  $t_{(0.025,38)}$ . Cela nous permet de rejeter l'hypothèse nulle, c'est-à-dire que les coefficients inclus dans le modèle diffèrent considérablement de zéro. Ce jugement est consolidé par les faibles valeurs de probabilité ( $p < 10^{-3}$ ) pour les six paramètres (constante + 5 coefficients) de l'équation (Equ.1).

L'intervalle de confiance pour chaque coefficient est calculé avec l'expression suivante :

$$\beta_i = \hat{\beta}_j \pm t_{(n-p-1, \alpha/2)} \cdot E.T.(\hat{\beta}_j)$$

Avec :

E.T. : l'erreur type ;  $(n-p-1 = 44-5-1=38)$  degrés de liberté du modèle ;  $\alpha$  : le niveau de confiance.

Les intervalles de confiance pour les six paramètres (avec  $\alpha = 0.05$ ) sont établis comme suit :

$-2.368 < \text{constante} < -1.650$  ;  $0.812 < \text{MATS7v} < 1.528$  ;  $0.322 < \text{Mor15u} < 0.647$  ;

$0.225 < \text{H5u} < 0.604$  ;  $3.116 < \text{R4e+} < 8.976$  ;  $-0.579 < \text{nArOR} < -0.382$ .

Il est clair que tous les coefficients ne renferment plus la valeur zéro dans leurs intervalles de confiance.

### ***B. Tests de colinéarité et Multicolinéarité***

La colinéarité et la multicolinéarité sont expliquées par la présence d'une forte corrélation d'un descripteur avec un autre (colinéarité) ou avec un ensemble de descripteurs (multicolinéarité). Les descripteurs corrélés entre eux contiennent une grande partie de la même information et provoque ainsi des problèmes calculatoires lors de la génération de la matrice inverse  $(X'X)^{-1}$ .

Le problème de colinéarité entre les descripteurs inclus dans le modèle final de QSAR est testé par l'examen de la matrice de corrélation, en calculant le coefficient de corrélation pour toutes les combinaisons paires possibles des cinq descripteurs. Les valeurs élevées du

coefficient de corrélation  $R \geq 0,9$  correspondent aux fortes corrélations entre les descripteurs du modèle (*Equ.1*). Les résultats obtenus sont récapitulés dans le tableau 6.

**Tableau 6.** Matrice de corrélation de l'équation 1

	MATS7v	Mor15u	H5u	R4e+
MATS7v	1			
Mor15u	-0.063	1		
H5u	0.366	-0.523	1	
R4e+	-0.186	0.395	-0.361	1
nArOR	0.178	-0.232	0.369	-0.063

L'examen de la matrice de corrélation (Tableau 6) confirme l'absence du problème de colinéarité entre les descripteurs du modèle, expliquée par les faibles valeurs des coefficients de corrélation ( $R < 0,9$ ).

Nous avons ainsi examiné la multicollinéarité par l'approche de la valeur d'inflation de la variance (VIF : Variance Inflation Factor) exprimée par l'équation 2.

$$VIF(x_k) = \frac{1}{1 - r_k^2} \quad (\text{Equ.2})$$

Avec  $VIF(x_k)$  est la valeur du facteur d'inflation de la variance pour le descripteur  $x_k$  ( $k=1 ; 2 ; \dots ; 5$ ), et  $r_k^2$  est le coefficient de corrélation carrée résultant de la régression du descripteur  $x_k$  sur le reste des descripteurs. Une valeur de  $VIF \geq 10$  signifie la présence d'une forte colinéarité entre les descripteurs, l'absence de la multicollinéarité est signifiée par une valeur de  $VIF < 10$ .

On peut aussi confirmer la présence ou l'absence de la forte multicollinéarité par une autre grandeur, similaire au VIF, appelée facteur de tolérance TF donnée par l'équation (*Equ.3*):

$$TF(x_k) = \frac{1}{VIF(x_k)} = 1 - R_{x_k}^2 \quad (\text{Equ.3})$$

Les valeurs de TF varient dans l'intervalle  $0 < TF < 1$ . Les valeurs de  $TF > 0,5$  et  $TF < 0,5$  correspondent à la présence et à l'absence de fortes multicollinéarité entre les descripteurs respectivement.

Les valeurs correspondantes au  $VIF(x_k)$  issues des cinq descripteurs de notre modèle sont réunies dans le Tableau 7.

**Tableau 7.** Valeurs des critères VIF et TF pour les descripteurs significatifs

Variable	VIF	TF
MATS7v	1.206	0.829
Mor15u	1.539	0.650
H5u	1.787	0.560
R4e+	1.262	0.792
nArOR	1.178	0.849

D'après ce Tableau, les cinq descripteurs ne présentent aucun problème de multicollinéarité avec une valeur maximale de  $VIF_{H_{3u}} = 1.787 < 10$  et les valeurs de TF toutes inférieures à 0,5.

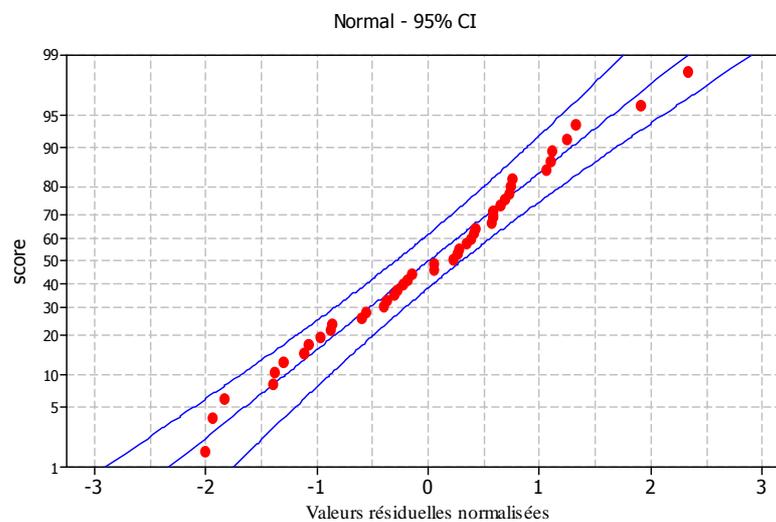
### C. Analyse des valeurs résiduelles normalisées

Cette étape consiste à examiner les valeurs résiduelles issues du modèle de QSAR obtenu. Une valeur résiduelle est la différence entre les valeurs observées et prédites de l'activité inhibitrice, présentée par les dérivés des xanthones et curcuminoïdes, contre l'enzyme  $\alpha$ -glucosidase. La justesse de notre modèle est vérifiée par l'analyse de la distribution normale et la linéarité des valeurs résiduelles.

Parmi les différentes méthodes utilisées pour l'analyse des valeurs résiduelles nous avons choisi, dans ce travail, d'examiner la linéarité par la représentation graphique des valeurs résiduelles normalisées en fonction des valeurs observées et prédites du pIC50. L'évaluation de la distribution normale des valeurs résiduelles normalisées est effectuée par la

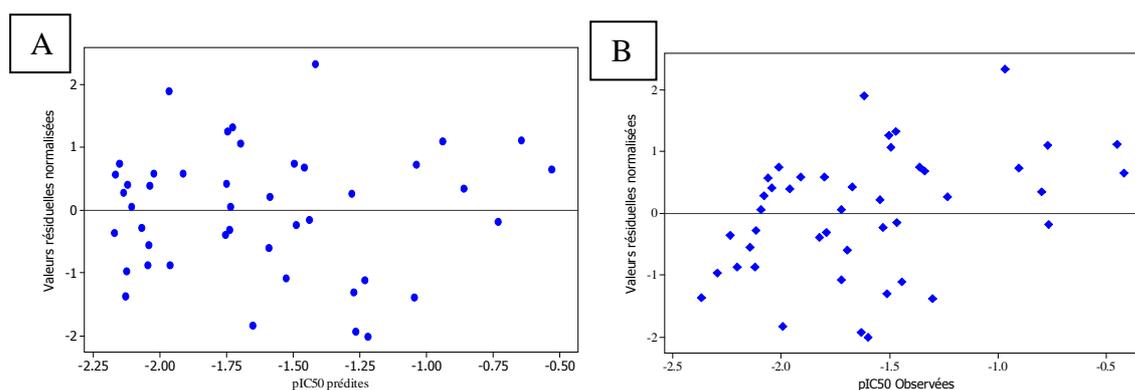
représentation graphique de chaque valeur résiduelle normalisée en fonction de son score. Les résultats de l'analyse obtenus sont illustrés dans les deux figures suivantes (Figures 4 et 5).

A partir de la figure 4, les valeurs résiduelles normalisées forme approximativement une droite linéaire dans un intervalle de confiance de 95%. Ce qui montre que les valeurs sont distribuées normalement.



**Figure 4.** Evaluation de la distribution normale des valeurs résiduelles normalisées

La figure 5 présente les deux graphes qui correspondent à la représentation graphique des valeurs résiduelles normalisées respectivement en fonction des valeurs prédites et observées du  $pIC_{50}$ . La bonne dispersion des points des deux cotés du zéro, sans tendance d'accroissement ou de diminution, indique que la variance est une constante, alors qu'il n'y a aucun problème de non linéarité de notre modèle.



**Figure 5.** Valeurs résiduelles normalisées en fonction des activités prédites (A) et observées (B)

#### ***D. Validation interne et externe du modèle de QSAR obtenu***

La validation du modèle QSAR obtenu est nécessaire pour estimer sa fiabilité. Dans cette étude nous avons utilisé trois méthodes de validation : la validation interne (validation croisée et test de randomisation) et la validation externe.

La validation croisée est effectuée par la procédure leave-one-out, on retire successivement une molécule de l'ensemble d'apprentissage. Cette procédure est répétée  $n$  fois ( $n$  est le nombre des molécules qui constituent l'ensemble d'apprentissage) afin de prédire les propriétés de toutes les molécules.

La valeur élevée du coefficient de détermination ( $Q^2_{cv-loo} = 0,815$ ) et la petite valeur de l'écart type ( $s_{cv-loo} = 0,208$ ) de la validation croisée du modèle obtenu prouvent la puissance prédictive de cette approche et la stabilité du modèle.

La bonne mesure de prévisibilité du modèle de QSAR obtenu par la validation interne (validation croisée) ne suffit pas. En effet, nous devons en plus généraliser ces prévisions en les appliquant sur un échantillon externe.

Pour cela, nous avons utilisé l'équation du modèle obtenu (*Equ.1*) pour prédire l'activité inhibitrice de l' $\alpha$  - glucosidase d'un ensemble externe de molécules (ensemble de prédiction) mentionné au départ. Les valeurs prédites de l'activité de molécules de l'ensemble externe

sont affichées dans la troisième colonne du tableau 2. La valeur du coefficient de détermination issu de la validation externe est  $Q_{ext}^2 = 0.66$ .

Selon *Tropsha et ses collaborateurs*<sup>11</sup>, un modèle de QSAR ne possède une puissance prédictive acceptable que si les conditions suivantes sont satisfaites :

$$Q_{cv-loo}^2 > 0.5;$$

$$R^2 > 0.6;$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \quad \text{Ou} \quad \frac{R^2 - R_0'^2}{R^2} < 0.1$$

$$0.85 \leq k \leq 1.15 \quad \text{Ou} \quad 0.85 \leq k' \leq 1.15 \quad Q_{cv-loo}^2 = 0.66 > 0.5.$$

- $R^2 = 0.81$ .
- $R_0^2 = 0.772 \Rightarrow (R^2 - R_0^2) / R^2 = (0.81 - 0.772) / 0.81 = 0.046 < 0.1$ .
- $k = 1 \Rightarrow 0.85 < k = 1 < 1.15$ .

Avec :

$R^2$ , des équations 3 et 4, est le coefficient de détermination entre les valeurs observées et celles prédites par le modèle (seulement pour l'ensemble d'apprentissage : 44 molécules).

$-R_0^2$  : coefficient de détermination issu de la régression des activités observées sur les activités prédites pour tout l'ensemble (57 molécules).

$R_0'^2$  : coefficient de détermination issu de la régression des activités prédites sur les activités observées pour tout l'ensemble (57 molécules).

$k$  et  $k'$  sont les pentes des lignes de régression qui passe par l'origine.

Afin de tester la robustesse du modèle obtenu, nous avons ensuite procédé au test de randomisation. Ce test consiste en la permutation aléatoire des composantes du vecteur de la réponse  $Y$  sans toucher à la matrice des descripteurs  $X$ , suivie par une reconstruction du modèle. Ce procédé est réitéré dix fois.

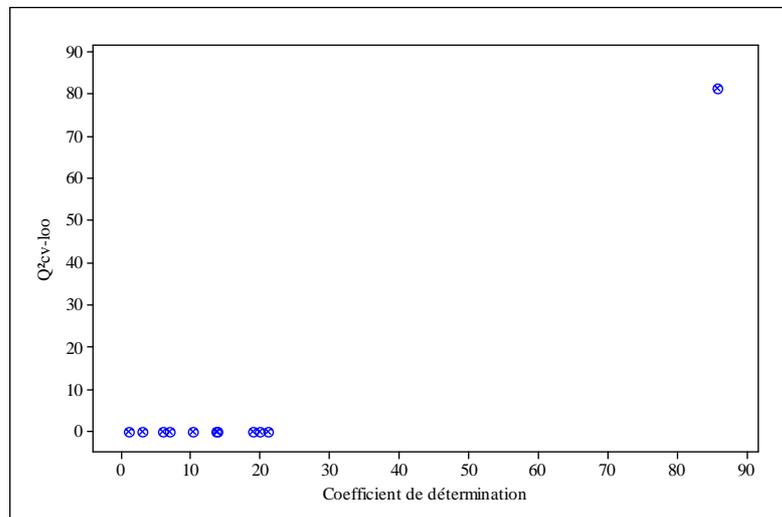
Les paramètres statistiques issus de chaque reconstruction du modèle et celui de l'Equ.1 sont cités dans le tableau suivant (Tableau 8).

**Tableau 8.**  $R^2$  et  $Q^2_{cv-100}$  issus du test de randomisation

Iteration	$R^2$	$Q^2_{cv-100}$
1	13.9	0.0
2	21.0	0.0
3	3.0	0.0
4	13.7	0.0
5	6.0	0.0
6	1.0	0.0
7	10.3	0.0
8	20.0	0.0
9	7.0	0.0
10	19.0	0.0
11*	85.7	81.5

\*: les paramètres statistiques issus de l'Equ.1

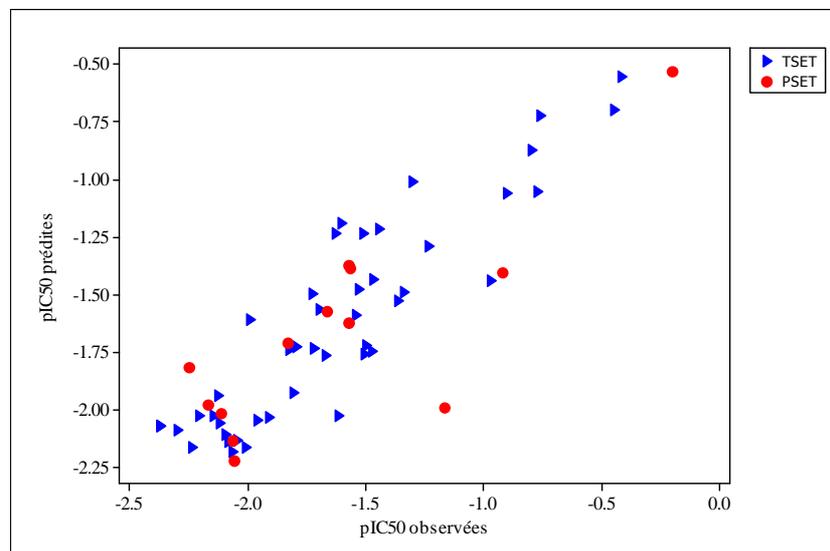
La représentation graphique des valeurs du coefficient  $Q^2_{cv-100}$  en fonction du coefficient de détermination (Figure 6) confirme que le modèle présenté par l'Equ.1 est plus robuste et n'est pas dû à la chance.



**Figure 6.** Valeurs de  $Q^2_{cv-100}$  en fonction de  $R^2$  issues du test de randomisation

La figure 7 représente la représentation linéaire des valeurs prédites en fonction des valeurs expérimentales pour les deux sous-ensembles d'apprentissage et de validation (TSET et PSET). Les résultats obtenus montrent que la technique linéaire de MLR combinée avec les

algorithmes génétiques en tant que procédure de sélection de variable est adéquate pour produire un modèle efficace de QSAR capable de modeler et de prédire l'activité inhibitrice de l'  $\alpha$  -glucosidase.



**Figure 7.** Valeurs prédites en fonction des valeurs observées pour TSET et PSET

### *E. Analyse des points aberrants sur l'axe des Y et X*

Les points aberrants sont localisés loin des valeurs de l'activité  $pIC_{50}$  prédites (points aberrants sur l'axe des Y) ou loin des valeurs des descripteurs (points aberrants sur les axes des X).

Généralement, les points possédants des valeurs résiduelles normalisées supérieures à 3 fois de l'écart type sont considérés comme points aberrants sur l'axe des Y.

On peut déterminer ainsi les points aberrant sur les axes des descripteurs en utilisant la valeur

de levier  $h_{ii}$ , dont les points possédants des valeurs supérieures à  $\frac{2(p+1)}{n} = \frac{2(5+1)}{44} = 0.272$

sont considérés comme points à grand levier, avec p et n nombres de descripteurs dans le modèle et nombres d'observations respectivement.

Les résultats obtenus lors de l'analyse des points aberrants et points à grand valeur de levier pour notre modèle sont affichés dans le tableau 9.

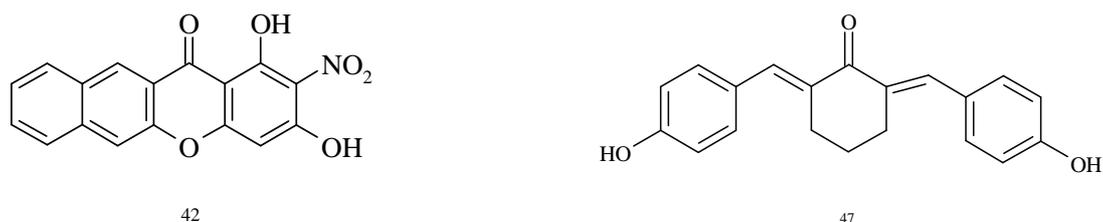
**Tableau 9 :** Valeurs résiduelles normalisées et valeurs de levier

Molécule Training set			Molécule Test set		
No.	$\delta_i$	$h_i$	No.	$\delta_i$	$h_i$
1	-1.3722	0.1855	3	-1.2976	0.1525
2	-0.9630	0.1865	6	-0.5701	0.0978
4	-0.3654	0.0883	10	-0.3029	0.1019
5	-0.8662	0.1208	13	0.2019	0.1864
7	-0.5512	0.1288	15	0.4810	0.0677
8	-0.8703	0.1104	21	-0.3593	0.0889
9	-0.2828	0.1036	27	-0.2851	0.1183
11	0.0583	0.1247	37	2.4468	0.2006
12	0.2776	0.0608	39	1.4312	0.0444
14	0.5666	0.1287	44	-0.5939	0.1197
16	0.4091	0.0989	47	0.1277	0.3889
17	0.7438	0.0835	56	0.9680	0.2573
18	-1.8293	0.1070	57	-0.5464	0.1945
19	0.3925	0.0633			
20	0.5835	0.1093			
22	-0.3911	0.1719			
23	0.5866	0.1070			
24	-0.3034	0.1973			
25	-1.0740	0.1459			
26	-0.5959	0.2123			
28	1.9063	0.1486			
29	-2.0064	0.0685			
30	0.2239	0.0767			
31	1.2545	0.0502			
32	1.0661	0.1042			
33	1.3288	0.0596			
34	-1.1134	0.0765			
35	-1.3909	0.1279			
36	0.2619	0.0994			
38	2.3321	0.0520			
40	0.7321	0.1260			
41	0.3402	0.1655			
42	1.0985	0.4043			
43	-0.1816	0.1983			
45	-1.9328	0.0852			
46	0.7528	0.2048			
48	1.1190	0.1943			
49	0.4277	0.1600			
50	-0.1493	0.1821			
51	-1.2994	0.1349			
52	0.6489	0.1761			
53	0.0561	0.1576			
54	-0.2294	0.2007			
55	0.6894	0.2124			

D'après ce Tableau, on remarque que toutes les valeurs résiduelles normalisées sont inférieures à 3, ce qui se traduit par l'absence des points aberrants sur l'axe des Y ( $pIC_{50}$ ).

Nous enregistrons ainsi, en examinant les valeurs de levier affichées dans le tableau, la présence de deux points à grand valeur de levier, qui dépassent le seuil critique ( $h = 0,272$ ).

Ces deux points correspondent aux molécules 42 et 47 (Figure 8).



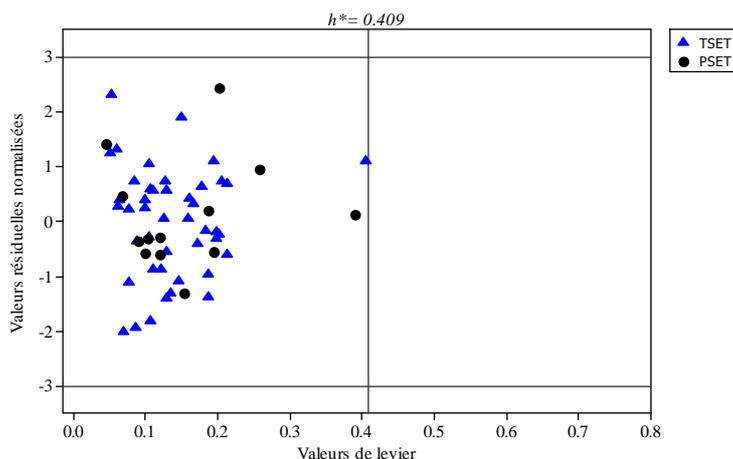
**Figure 8.** Les deux structures considérées comme des points aberrants sur l'axe des X.

#### ***F. Domaine d'applicabilité***

Le modèle représenté par l'équation (*Equ.1*) est capable de prédire quantitativement l'activité inhibitrice de l'enzyme  $\alpha$ -glucosidase. De ce fait, toute nouvelle structure appartenant au domaine d'applicabilité du modèle pourrait avoir une mesure quantitative de son activité inhibitrice en se basant sur l'équation mathématique précédente.

Le calcul des valeurs de levier  $h_{ii}$  et des valeurs résiduelles normalisées  $\delta_i$  (Tableau 8), pour tous les composés de la base de données, nous a permis de définir le graphe de Williams. Ce dernier est une représentation graphique des valeurs résiduelles normalisées en fonction des valeurs de levier de chaque molécule (Figure 9) à partir du quel on peut déterminer les molécules correspondantes aux points aberrants ainsi que les produits chimiques influents sur le modèle.

D'après la figure 9 nous remarquons l'absence des points aberrants et des points influents dans le modèle, car tous les composés appartenant à l'ensemble de données sont situés en dessous du domaine d'application du modèle (valeurs résiduelles bornées entre -3 et 3 et les valeurs de levier sont aussi toutes limitées par la valeur critique  $h^* = 0,409$ ).



**Figure 9.** Graphe de Williams obtenu avec les cinq descripteurs

## 2. Le modèle des descripteurs de la famille 3D-MoRSE

Dans l'étude de QSAR sur l' $\alpha$ -glucosidase, nous avons utilisé, en premier lieu, la totalité des descripteurs issus du serveur E-DRAGON1. Le modèle de régression obtenu (Equ. 2) a prouvé sa validité et sa robustesse dans l'explication de la relation entre les descripteurs moléculaires significatifs et l'activité inhibitrice contre l' $\alpha$ -glucosidase. Ce résultat attrayant nous a poussés de faire une étude comparative entre différents blocks du E-DRAGON1, pour voir quelle est la famille de descripteurs qui explique mieux la relation entre ces derniers et l'activité  $pIC_{50}$ . Les descripteurs utilisés dans cette étude appartiennent aux familles les plus reconnues dans la littérature : 2D-autocorrelation, Geometrical, Getway, RDF, WHIM, et 3D-MoRSE.

Le développement des modèles de QSAR sont toujours effectués avec la même base structurale : 45 molécules dérivées de xanthonnes et 12 molécules dérivées de curcuminoïdes (Figure 1). Nous essayons, par ces modèles, de quantifier l'activité biologique inhibitrice de l'enzyme  $\alpha$ -glucosidase  $pIC_{50}$ .

Nous poursuivons la même méthodologie lors du développement de chaque modèle de QSAR, en utilisant toujours les algorithmes génétiques comme méthode de sélection des descripteurs.

Avant de commencer la construction de chaque modèle, nous avons divisé l'ensemble de molécules en deux sous ensembles, ensemble d'apprentissage (TSET) et ensemble de prédiction (PSET), réservés respectivement pour la construction du modèle et la validation externe.

Après détermination des modèles optimums de différentes dimensions par les algorithmes génétiques (allant de 3 à 8 descripteurs pour chaque famille), nous avons limité la dimension optimale des modèles à 4 descripteurs en utilisant la technique de point d'arrêt (breacking point). Les résultats obtenus sont affichés dans le tableau 10.

La définition des descripteurs utilisés dans les modèles sera présentée dans le tableau 2 (Annexe : Chapitre IV).

**Tableau 10:** comparaison entre six familles de descripteurs

Familles de descripteurs	N° var.	R <sup>2</sup>	S	F	Q <sup>2</sup>
2D-autocorrélation	4	57.04	0.334	13.28	46.74
Géométrique	4	48.54	0.3655	9.43	36.22
GETAWAY	4	49.16	0.3633	9.67	37.46
<b>3D-MORSE</b>	<b>4</b>	<b>80.50</b>	<b>0.225</b>	<b>41.27</b>	<b>75.68</b>
RDF	4	56.57	0.3358	13.03	44.77
WHIM	4	43.6	0.3827	7.73	29.61

Les descripteurs inclus dans les modèles sont :

2D-autocorrélation	GATS2m; GATS7m; GATS1v; GATS3v
Géométrique	HOMA; RCI; AROM; HOMT
GETAWAY	HATS7u; R3u; R4u; R3e+
<b>3D-MORSE</b>	<b>Mor06u; Mor18u; Mor31u; Mor27v</b>
RDF	RDF025m; RDF130v; RDF010e ; RDF060p
WHIM	E2u; G2m; G2e; E1s

Ce tableau rassemble les paramètres statistiques (R<sup>2</sup> ; Q<sup>2</sup><sub>cv-100</sub> ; s ; et F) correspondant au six modèles optimaux obtenus par les approches suivantes : 2D-autocorrélation, Géométrique, Getaway, RDF, WHIM, et 3D- MoRSE.

D'après les résultats statistiques affichés dans le tableau, nous remarquons que le modèle obtenu par l'approche 3D-MoRSE présente les meilleurs paramètres statistiques ( $R^2=80.50\%$  ;  $Q^2_{cv-100}=75.68\%$  ;  $F=41.27$  ;  $s=0.225$ ) en comparaison avec les résultats modérés issus des autres approches.

Les bons résultats obtenus par les descripteurs de la famille 3D-MoRSE nous ont poussé à avancer dans leur traitement pour évaluer la robustesse et la fiabilité du modèle et prouver l'efficacité de ces descripteurs à interpréter et modéliser l'activité inhibitrice des dérivés des xanthones et curcuminoïdes contre l' $\alpha$ -glucosidase (Tableau 11).

**Tableau 11.** Les valeurs de pIC50 observées et prédites de TSET et PSET

No.	pIC50exp.	pIC <sub>50</sub> pred.	Mor06u	Mor18u	Mor31u	Mor27v
<b>TSET</b>						
<b>No.</b>						
1	-2.371	-1.892	-1.876	-0.945	0.068	-0.321
2	-2.297	-2.26	-1.911	-1.294	0.235	-0.388
4	-2.238	-2.185	-2.015	-1.154	0.155	-0.416
5	-2.206	-1.856	-2.013	-1.225	0.092	-0.361
7	-2.143	-2.308	-2.757	-1.339	0.397	-0.379
8	-2.123	-1.859	-2.153	-1.282	0.116	-0.372
9	-2.119	-1.798	-3.496	-1.572	0.39	-0.356
11	-2.092	-1.805	-2.601	-2.028	0.592	-0.177
12	-2.082	-2.054	-3.263	-1.679	0.55	-0.29
14	-2.062	-2.022	-3.464	-1.526	0.462	-0.356
16	-2.045	-2.052	-2.868	-1.916	0.614	-0.224
17	-2.01	-2.09	-2.899	-1.343	0.282	-0.429
18	-1.992	-1.704	-1.48	-1.183	-0.039	-0.338
19	-1.961	-2.045	-3.924	-1.783	0.381	-0.585
20	-1.913	-2.062	-1.92	-1.274	0.172	-0.355
22	-1.823	-1.874	-2.845	-1.39	0.204	-0.42
23	-1.803	-1.751	-2.982	-1.42	0.178	-0.424
24	-1.791	-2.072	-3.466	-2.076	0.368	-0.596
25	-1.724	-1.83	-3.629	-1.409	0.322	-0.404
26	-1.696	-1.596	-2.141	-1.111	-0.102	-0.43
28	-1.618	-1.614	-2.436	-1.078	-0.003	-0.369
29	-1.601	-1.08	-2.811	-1.377	-0.078	-0.366
30	-1.543	-1.473	-2.23	-1.461	0.028	-0.35
31	-1.504	-1.851	-2.442	-1.241	0.145	-0.361
32	-1.496	-1.474	-2.747	-1.47	0.021	-0.44
33	-1.473	-1.593	-3.355	-1.619	0.237	-0.403
34	-1.444	-1.28	-2.741	-1.608	0.058	-0.358
35	-1.303	-1.356	-3.024	-1.453	0.06	-0.385
36	-1.233	-1.207	-2.649	-1.167	-0.08	-0.327
38	-0.968	-1.642	-2.137	-1.126	-0.022	-0.35
40	-0.903	-1.012	-2.217	-1.406	-0.192	-0.348
41	-0.799	-0.75	-2.27	-1.454	-0.266	-0.34
42	-0.771	-0.866	-2.496	-1.285	-0.199	-0.302
43	-0.763	-0.866	-2.755	-1.333	-0.214	-0.371
44	-1.571	-1.484	-2.937	-1.258	0.235	-0.196
45	-1.63	-1.522	-3.121	-1.064	0.319	-0.106
46	-1.362	-1.801	-2.673	-1.078	0.318	-0.129
48	-0.447	-0.285	-5.101	-1.258	0.176	-0.092
49	-1.672	-1.539	-3.813	-1.497	0.163	-0.505
50	-1.467	-1.823	-1.975	-1.532	-0.024	-0.52
51	-1.512	-1.408	-3.781	-1.352	0.428	-0.125
52	-0.415	-0.866	-3.534	-1.408	0.132	-0.131
53	-1.723	-1.433	-3.804	-1.511	0.142	-0.49
54	-1.53	-1.613	-1.9	-1.385	-0.007	-0.371
55	-1.338	-1.548	-3.145	-1.149	0.208	-0.246

Tableau 11. (suite)

No.	pIC <sub>50</sub> exp.	pIC <sub>50</sub> pred.	Mor06u	Mor18u	Mor31u	Mor27v
<b>PSET</b>						
3	-2.249	-2.09	-2.171	-1.192	0.196	-0.364
6	-2.166	-1.87	-2.296	-1.158	0.108	-0.37
10	-2.114	-1.85	-2.286	-1.117	0.09	-0.363
13	-2.063	-2.11	-1.747	-2.121	0.662	-0.08
15	-2.056	-2.15	-3.259	-1.521	0.5	-0.34
21	-1.828	-1.59	-2.713	-1.555	0.117	-0.412
27	-1.667	-1.79	-1.453	-1.446	0.046	-0.345
37	-1.167	-1.54	-2.165	-1.213	-0.033	-0.362
39	-0.919	-1.15	-2.991	-1.397	-0.053	-0.39
47	-1.575	-1.33	-3.959	-1.672	0.382	-0.249
56	-0.204	-0.92	-2.949	-1.023	-0.079	-0.205
57	-1.568	-1.03	-2.706	-1.382	-0.073	-0.308

### A. Analyse de la justesse du modèle optimal

Le modèle 3D-MoRSE obtenu est donné par l'expression suivante :

$$pIC_{50} = -2.327 - 2.332 \text{ Mor31u} + 2.250 \text{ Mor27v} - 0.624 \text{ Mor18u} - 0.354 \text{ Mor06u} \quad (\text{Equ.2})$$

$$N= 45; \quad R=0.897 \quad R^2= 80.50\%; \quad R^2_{aj}= 78.5\%; \quad s = 0.225$$

Les paramètres statistiques mentionnés ci-dessus montre que les coefficients de détermination  $R^2$  et  $R^2$  ajusté expliquent respectivement plus de 80% et 78% de la variance des valeurs observées de l'activité biologique.

Le modèle mentionné ci-dessus, reliant la variable dépendante pIC<sub>50</sub> avec les quatre descripteurs, est accompagné par les mesures statistiques suivantes: coefficient de corrélation multiple R, coefficient de détermination  $R^2$ , coefficient de détermination ajusté  $R^2_{adj}$ , et l'écart type s.

Le coefficient de corrélation multiple ( $R = 0.897$ ) indique qu'il existe une forte corrélation entre les valeurs observées et prédites de l'activité biologique. Les valeurs des coefficients de détermination  $R^2$  et  $R^2$  ajusté prouve une bonne variabilité du modèle et expliquent

respectivement plus de 80% et de 78% de la variance des valeurs observées de l'activité biologique. La dispersion des données est jugée par la petite valeur de l'écart type ( $s = 0.225$ ).

La corrélation des quatre descripteurs sélectionnés avec l'activité biologique est testée en examinant la table d'ANOVA (Tableau 12) et le tableau des coefficients (Tableau 13).

A partir de la table d'ANOVA, la statistique de Fisher F nous permet de vérifier la présence ou l'absence de corrélation entre l'ensemble de descripteurs et l'activité biologique. Pour ce faire nous avons fait appel aux deux hypothèses : hypothèse nulle et hypothèse alternative.

**Tableau 12.** Résultats de l'analyse de la variance (ANOVA)

Source	DF <sup>a</sup>	SC <sup>b</sup>	CM <sup>c</sup>	F <sub>obs</sub>	P <sup>d</sup>
Régression	4	8.3602	2.0900	41.27	0.000
Erreur Résiduel	40	2.0255	0.0506		
Totale	44	10.3857			

a: degré de liberté ; b : somme des carrés ; c : Mean squares ; d : probabilité.

*L'hypothèse nulle ( $H_0$ ):* « aucun descripteur n'est lié à l'activité biologique :  $\beta_j = 0$  avec ( $j=0 ; 1 ; 2 ; 3 ; 4$ ),  $\beta_j$  : coefficient correspond au descripteur  $j$  »

Cette hypothèse est acceptée si la valeur de la statistique de Fischer observée est inférieure à la valeur  $F_{(0.05 ; 4 ; 40)}$ .

*L'hypothèse alternative ( $H_1$ ):* « il existe au moins un descripteur corrélé avec l'activité biologique ».

Cette hypothèse est acceptée lorsque la valeur de la statistique de Fischer observée est supérieure à la valeur  $F_{(0.05 ; 4 ; 40)}$ .

Sachant que  $F_{(0.05 ; 4 ; 40)} = 2.61$ , cela nous permet de rejeter l'hypothèse nulle et de garder l'hypothèse alternative, car la valeur de la statistique de Fischer est supérieure à 2,61 ( $F_{obs} = 41,26$ ), on peut conclure donc qu'il existe au moins un coefficient différent de zéro qui participe dans l'explication de l'activité biologique.

Nous avons ensuite examiné la table des coefficients pour vérifier la signification de chaque descripteur et sa contribution dans l'explication de l'activité biologique. L'utilisation des valeurs de la statistique t de Student, affichées dans la table ci-dessous pour chaque descripteur, nous permet de vérifier la présence ou l'absence de corrélation entre chaque descripteur et l'activité biologique en se basant sur les deux hypothèses : hypothèse nulle et hypothèse alternative.

**Tableau 13.** Tableau des coefficients

descripteur	Coef <sup>a</sup>	Er. T (Coef) <sup>b</sup>	$T_{obs}$ <sup>c</sup>	$p^d$	Interval de Conf. <sup>e</sup>	
					Limite inf.	Limite sup.
Constante	-2.3278	0.2365	-9.84	0.000	-2.806	-1.850
Mor06u	-0.35452	0.05457	-6.5	0.000	-0.465	-0.244
Mor18u	-0.6248	0.1801	-3.47	0.001	-0.989	-0.261
Mor31u	-2.3323	0.2034	-11.47	0.000	-2.743	-1.921
Mor27v	2.2502	0.3342	6.73	0.000	1.575	2.926

a : coefficients ; b : erreur type ; c : test de student observé ; d : la valeur de probabilité ; e : intervalle de confiance.

*L'hypothèse nulle ( $H_0$ )* : « le descripteur n'est pas lié à l'activité biologique :  $\beta_j = 0$  avec ( $j=0 ; 1 ; 2 ; 3 ; 4$ ),  $\beta_j$  : coefficient correspond au descripteur  $j$  »

Cette hypothèse est acceptée si la valeur de la statistique de student t observée est inférieure à la valeur  $t_{(0,025 ; 40)}$ .

*L'hypothèse alternative ( $H_1$ )* : « le descripteur est corrélé avec l'activité biologique :  $\beta_j \neq 0$  ».

Cette hypothèse est acceptée lorsque la valeur de la statistique de student t observée est supérieure de la valeur  $t_{(0,025 ; 40)}$ .

$$t_{obs} > t_{(0,025,40)} = 2.021$$

A partir des valeurs de la statistique t, obtenues à un niveau de confiance  $\alpha = 0.05$ , affichées dans le tableau 13, nous pouvons rejeter l'hypothèse nulle. Ce qui prouve la présence de

corrélation entre les descripteurs du modèle et l'activité  $pIC_{50}$ . Les valeurs faibles de  $p$  des paramètres, enregistrées dans la 5<sup>ème</sup> colonne du tableau confirment ainsi que les coefficients diffèrent de manière significative de zéro.

### ***B. Tests de colinéarité et de Multicolinéarité***

Nous avons entamé le problème de colinéarité et de multicolinéarité respectivement par l'examen de la matrice de corrélation et la vérification par les deux critères VIF et TF. Les résultats obtenus présentés dans les deux tableaux 14 et 15.

**Tableau 14.** Matrice de corrélation de l'équation 1

	MOR06U	MOR18U	MOR31U
MOR06U	1		
MOR18U	0.351	1	
MOR31U	-0.453	-0.471	1
MOR27V	-0.142	0.281	0.201

**Tableau 15.** Valeurs des critères VIF et TF

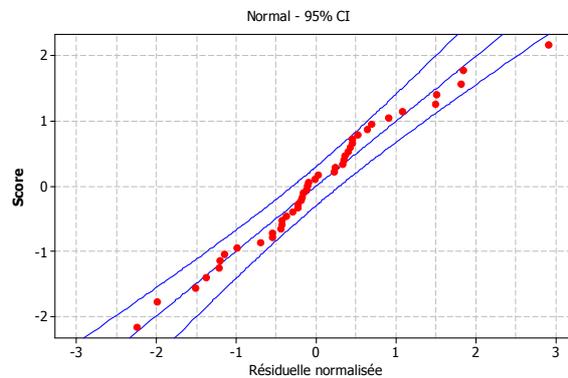
Variable	VIF	TF
MOR06U	1.4	0.71
MOR18U	1.7	0.59
MOR31U	1.6	0.63
MOR27V	1.3	0.77

D'après le tableau 14, les faibles valeurs du coefficient de corrélation, nous confirment l'absence d'une forte corrélation entre les descripteurs. En effet, la valeur maximale de ce coefficient, produite par la combinaison entre les deux descripteurs (MOR31U et MOR18U) est  $R = -0.471 < 0.9$ .

L'examen des valeurs des deux critères VIF et TF (Tableau 15) prouve aussi l'absence de la multicolinéarité entre les variables indépendantes de l'équation (*Equ.2*).

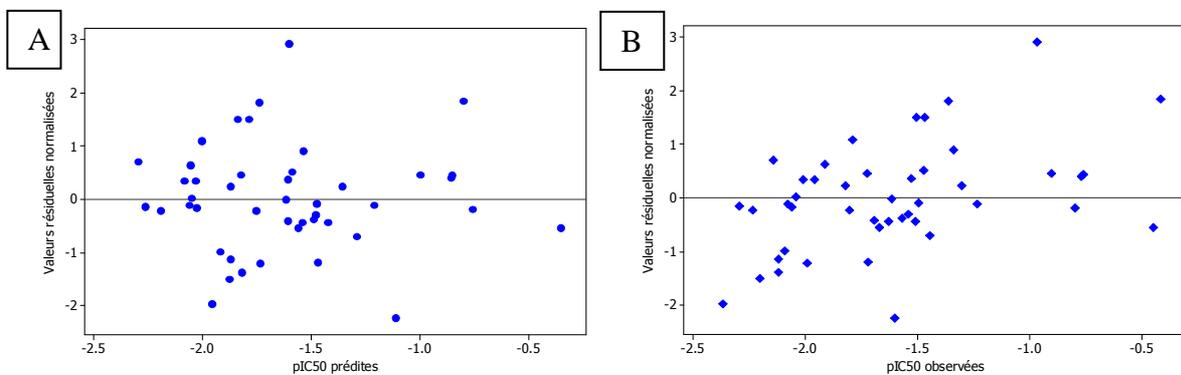
### C. Analyse des valeurs résiduelles normalisées

L'examen graphique des valeurs résiduelles normalisées peut nous aider à contrôler la justesse du modèle développé précédemment. Nous avons entamé en premier lieu la représentation graphique des valeurs de score en fonction de ces dernières pour vérifier leur distribution normale. Ensuite, les valeurs résiduelles normalisées sont projetées en fonction des valeurs de  $pIC_{50}$  observées et prédites. Les deux graphes obtenus sont présentés dans les deux figures 10 et 11 suivantes.



**Figure 10.** Evaluation de la distribution normale des valeurs résiduelles normalisées

A partir de ce graphe présenté par la figure 10, nous observons que les valeurs résiduelles normalisées projetées en fonction du score forment approximativement une droite linéaire dans un intervalle de confiance de 95%. Ce qui montre que les valeurs sont distribuées normalement.



**Figure 11.** Valeurs résiduelles normalisées en fonction des activités prédites (A) et observées (B)

Les deux graphes (Figure 11) montrent que la dispersion des valeurs résiduelles normalisées est nulle. Cette dispersion nulle est confirmée aussi par les faibles valeurs des coefficients de corrélation issus de la régression des valeurs des résiduelles normalisées absolues contre les valeurs de pIC<sub>50</sub> ajustées et observées.

#### *D. Validation interne et externe*

Nous avons contrôlé la stabilité et la validité de notre modèle issu de la famille 3D-MoRSE par la validation interne qui implique à la fois la validation croisée (Leave – One-Out) et le test de randomisation et par la validation externe.

La valeur élevée du coefficient de corrélation multiple ( $Q^2_{cv-100} = 0.757$ ) et la faible valeur de l'écart type ( $s_{cv-100} = 0.236$ ) résultant de ce procédé démontre la stabilité du modèle et le pouvoir prédictif de cette approche.

La robustesse du modèle est ensuite vérifiée par le test de randomisation. Ce test consiste en la permutation aléatoire des composantes du vecteur de l'activité pIC<sub>50</sub>, sans changer la matrice des descripteurs X, suivie par une reconstruction du modèle. Ce procédé est réitéré dix fois.

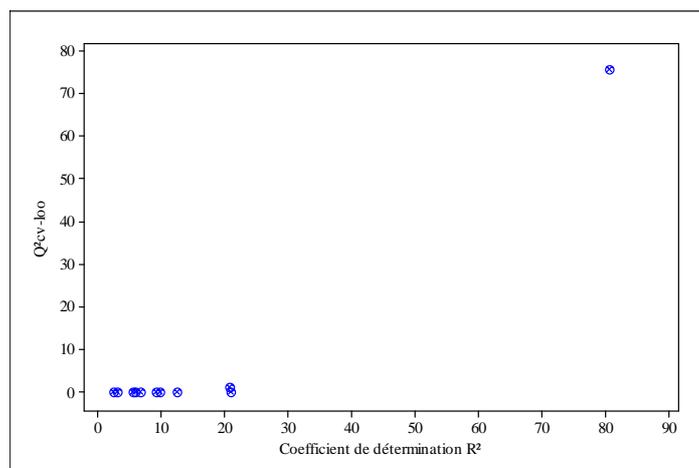
Les paramètres statistiques issus du modèle présenté par l'Equ.2 et des autres modèles sont cités dans le tableau suivant (Tableau 16).

**Tableau 16.**  $R^2$  et  $Q_{cv-100}^2$  issus du test de randomisation

Iteration	$R^2$	$Q^2$
1	05.50	00.00
2	09.10	00.00
3	02.50	00.00
4	05.90	00.00
5	09.70	00.00
6	20.70	01.22
7	21.00	00.06
8	12.50	00.00
9	06.70	00.00
10	03.00	00.00
11*	80.50	75.69

\*: les paramètres statistiques issus de l'Equ.5

Selon le graphe (Figure 12), le point présenté par les coordonnées cartésiennes issu des paramètres statistiques de l'équation (Equ2) se positionne solitairement et cela confirme la robustesse de notre modèle.

**Figure 12.** Valeurs de  $Q_{cv-100}^2$  en fonction de  $R^2$  issues du test de randomisation

Ensuite nous avons vérifié le pouvoir prédictif externe par la prédiction de l'activité biologique de l'ensemble de prédiction (PSET) (Tableau 11).

Nous avons obtenu des résultats qu'on peut estimer satisfaisants qui sont en accord avec les conditions exigées par Tropsha :

$$Q_{ext}^2 = 0.66.$$

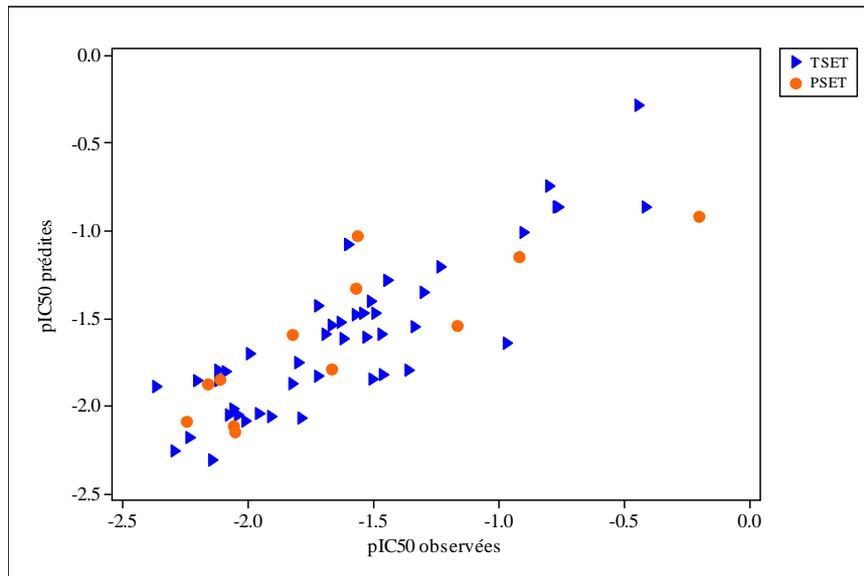
$$Q_{cv-100}^2 = 0.757 > 0.5 ;$$

$$R^2 = 0.758 ;$$

$$R_0^2 = 0.734 \Rightarrow \frac{(R^2 - R_0^2)}{R^2} = 0.031 < 0.1 ;$$

$$0.85 \leq k = 1.01 \leq 1.15 .$$

La représentation des valeurs prédites en fonction des valeurs observées est donnée par la figure suivante. La ligne droite présentée par cette figure montre une forte corrélation entre ces valeurs (prédites et observées), et confirme ainsi la précision de notre modèle.



**Figure 13.** Les valeurs prédites en fonction des valeurs observées pour les deux sous ensembles TSET et PSET

### *E. Analyse des points aberrants sur l'axe des Y et X*

Les valeurs résiduelles normalisées  $\delta_i$  et de levier  $h_{ii}$  de chaque observation des deux ensembles (TSET et PSET) sont exposés dans le tableau 17 :

D'après ce tableau, toutes les observations issues de TSET et PSET sont bornées par les deux limites 3 et -3 sur l'axe des résiduelles normalisées, confirmant ainsi l'absence de points aberrants.

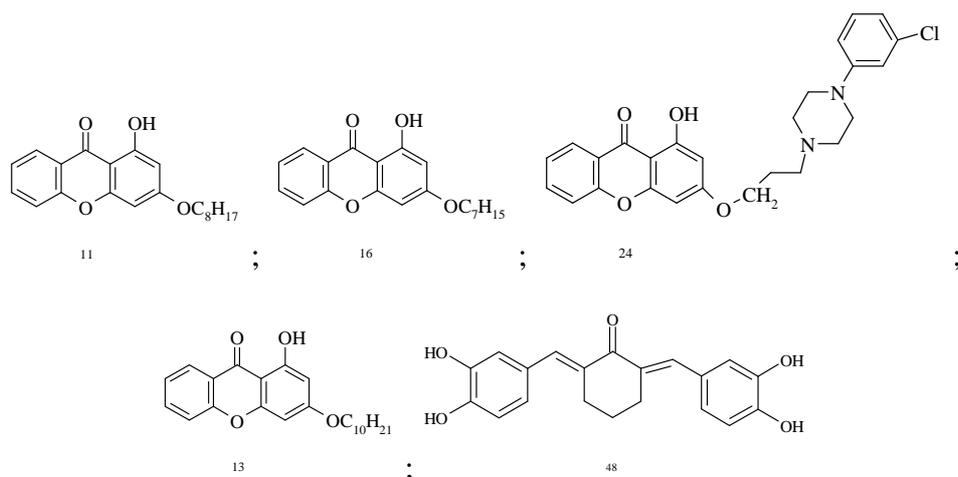
L'examen des valeurs de levier nous amène à cinq observations présentant des valeurs de

levier plus grands que le seuil  $\frac{2(p+1)}{n} = \frac{2(4+1)}{44} = 0.227$ .

**Table 17.** Valeurs résiduelles normalisées et valeurs de levier

Molécule Training set			Molécule Test set		
No.	$\delta_i$	$h_{ii}$	No.	$\delta_i$	$h_{ii}$
1	-1.9873	0.1271	3	-0.4609	0.0728
2	-0.1586	0.0922	6	-0.8642	0.0574
4	-0.2247	0.0995	10	-0.7489	0.0639
5	-1.5105	0.0556	<b>13</b>	0.1440	<b>0.7965</b>
7	0.6971	0.0913	15	0.2592	0.0864
8	-1.1452	0.0449	21	-0.6913	0.0415
9	-1.3837	0.0587	27	0.3457	0.1338
<b>11</b>	-0.9939	<b>0.3921</b>	37	1.0658	0.0481
12	-0.1161	0.1069	39	0.6625	0.0627
14	-0.1717	0.0784	47	-0.7201	0.1104
<b>16</b>	0.0258	<b>0.2534</b>	56	2.0740	0.1218
17	0.3408	0.0692	57	-1.5555	0.0732
18	-1.2118	0.1020			
19	0.3333	0.2108			
20	0.6367	0.0707			
22	0.2203	0.0393			
23	-0.2281	0.0375			
<b>24</b>	1.0801	<b>0.2509</b>			
25	0.4507	0.0854			
26	-0.4253	0.0821			
28	-0.0185	0.0672			
29	-2.2420	0.0627			
30	-0.3009	0.0579			
31	1.5110	0.0397			
32	-0.0977	0.0477			
33	0.5175	0.0513			
34	-0.7002	0.0723			
35	0.2320	0.0392			
36	-0.1110	0.0590			
38	2.9065	0.0576			
40	0.4533	0.1211			
41	-0.1956	0.1849			
42	0.3985	0.1121			
43	0.4296	0.1148			
44	-0.3731	0.0626			
45	-0.4398	0.1563			
46	1.8074	0.1426			
<b>48</b>	-0.5545	<b>0.4071</b>			
49	-0.5486	0.1352			
50	1.4936	0.1073			
51	-0.4306	0.1403			
52	1.8459	0.1505			
53	-1.2066	0.1244			
54	0.3562	0.0683			
55	0.9014	0.0699			

Les structures chimiques à grand effet de levier sont :

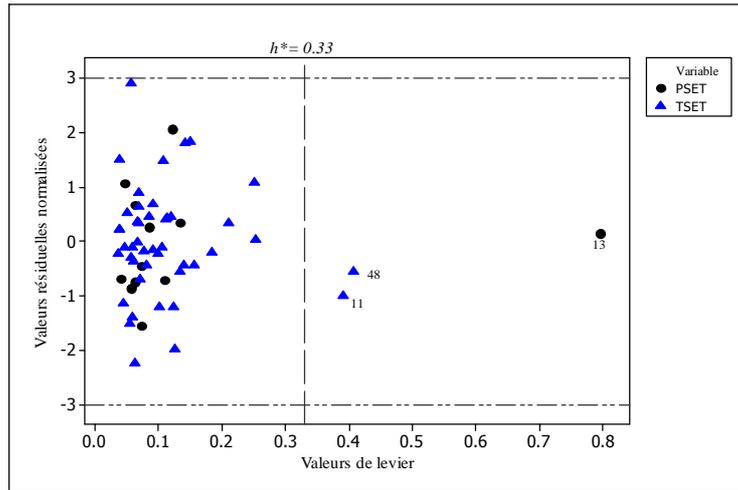


**Figure 14.** Les structures des composés à grand valeurs de levier

Après la comparaison de ces structures avec le reste des molécules, nous pouvons noter que les seules différences concernent le poids moléculaire et le volume moléculaire élevés de ces composés (11 ; 13 ; 16 et 24) par rapport aux autres. Ces différences sont dues au groupement alkoxy attaché au motif du xanthone. Cette observation est différente pour le composé n°48.

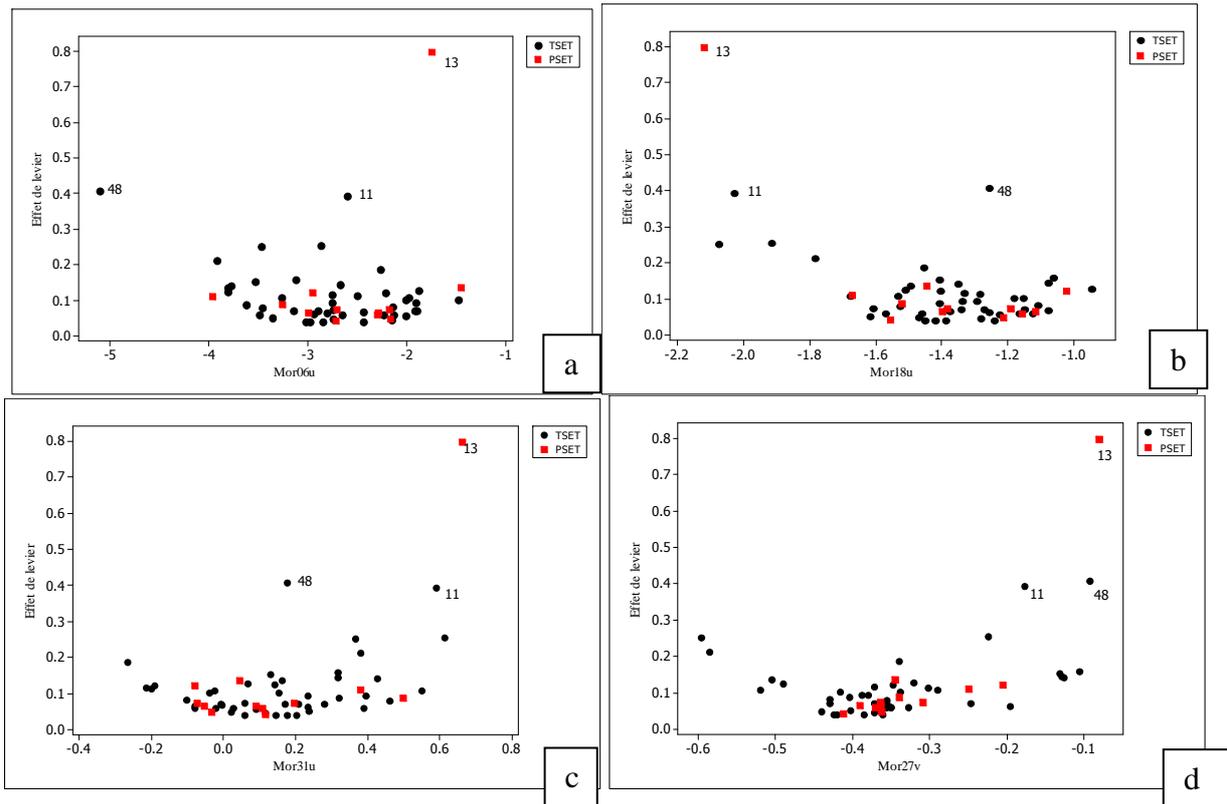
#### *F. Domaine d'applicabilité*

L'analyse du domaine d'applicabilité représenté par la figure ci-dessous (Figure. 15), montre que les trois composés (n°11, 48 et 13) sont caractérisés par des valeurs élevées de  $hi > h^* = 0.33$  (Tableau 17). Ils sont considérés comme produits chimiques influents sur les deux sous ensembles TSET et PSET.



**Figure15.** Graphe de Williams obtenu avec les quatre descripteurs

Pour tirer plus d'information concernant ces trois composés, nous avons procédé à un examen approfondi, en consultant la représentation graphique des valeurs de levier en fonction des valeurs de chaque descripteur. Les résultats obtenus sont résumés dans la figure suivante :



**Figure16 :** Valeurs de levier en fonction des valeurs des 4 descripteurs (a, b, c, d)

A partir de la figure précédente, nous pouvons apporter les remarques essentielles suivantes :

1. Les trois composés, représentant les points influents, se retrouvent toujours en dehors de l'agglomérat.
2. Les deux composés (11 et 13) qui présentent une similitude structurale démontrée lors de l'analyse des points aberrants sur l'axe des X, apparaissent dans ces graphes (Figure 16) dans la même région par rapport à l'axe des X lequel est en relation avec les descripteurs.
3. Selon le graphe (a), le composé n°48 est clairement séparé de l'ensemble, ce qui montre que le descripteur Mor06u est le responsable majoritairement de la valeur élevée de levier.
4. Le graphe (d) montre que les trois composés se retrouvent dans une même région par rapport à l'axe des X lequel est lié au descripteur Mor27v en rapport avec les valeurs de van de Waals.

### **3. Interprétation chimique de l'influence des descripteurs sur l'activité inhibitrice contre $\alpha$ -glucosidase**

Toute activité biologique étant directement liée à la forme moléculaire des produits chimiques (électronique, géométrique, constitutionnelle...etc.); il est ainsi possible de déterminer les facteurs responsables de l'inhibition observée des composés, en interprétant les descripteurs recueillis dans les deux modèles obtenus.

#### Le premier modèle

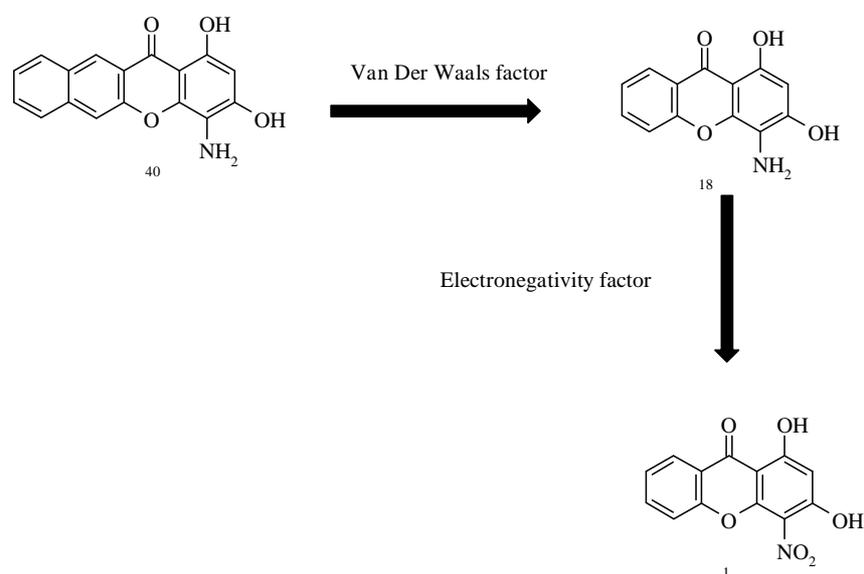
L'équation du modèle de QSAR, obtenu par l'utilisation de la totalité des descripteurs issus du E-DRAGON1, présenté par l'équation (*Equ.1*) rassemble cinq descripteurs de classe différentes :

Le descripteur bidimensionnel MATS7v (Moran autocorrelation - lag 7 / pondéré par le volume atomique de van der Waals) appartenant à la classe des descripteurs 2D-

Autocorrelation, vient de la combinaison de l'information topologique des composés avec leurs volumes atomiques de van der Waals; les deux descripteurs tridimensionnels : H5u et R4e+, dérivés de la famille GETAWAY; Mor15u un descripteur tridimensionnel dérivé de la famille 3D-MoRSE; et enfin nArOR qui comptabilise le nombre de liaisons de type éthers aromatiques.

Les deux descripteurs MATS7v et R4e<sup>+</sup> sont les plus importants, car ils présentent les coefficients les plus élevés dans l'équation (*Equ.1*). Ils influent positivement sur l'activité. En effet, d'après notre modèle, l'activité inhibitrice des analogues du xanthone et du curcuminoïde contre l'enzyme  $\alpha$ -glucosidase augmente avec les valeurs de ces deux descripteurs.

Ces derniers sont liés aux deux propriétés chimiques suivantes: l'électronégativité et le volume de van der Waals. Afin de comprendre l'influence de ces deux facteurs (électronégativité et volumes de van der Waals), nous avons fait une comparaison entre les composés N°40, 18, et 1, (Schéma1).



**Schéma 1.** Représentation de l'influence de MATS7v et R4e<sup>+</sup>

Les deux composés **40** et **18** ont des structures très proches, sauf que le composé n°40 possède un noyau aromatique de plus. Cette différence mène à augmenter la valeur de

MATS7v de -0,213 à -0,501, contribuant à l'augmentation de la valeur de l'activité pIC<sub>50</sub>. D'autre part, si on fait la comparaison entre les composés 18 et 1, on trouve que la seule différence entre les deux structures (schéma 1) est la substitution du groupe amine (-NH<sub>2</sub>) du composé 18 par un groupe nitro (-NO<sub>2</sub>) du composé 1. Malgré cette faible différence du point de vue structurale, le composé **18** présente une activité 2.3 fois supérieure par rapport au composé 1, augmentant ainsi la valeur de la valeur IC<sub>50</sub> de 102,3 µM (composé 18) à 235,2 µM (composé n°1) (Tableau 18).

Selon le schéma précédent, le passage du composé n°1 au composé n°40, qui correspond à la présence des deux facteurs simultanément, peut engendrer une réduction remarquable de la valeur IC<sub>50</sub> de 235.2 à 8.3 µM.

**Tableau18.** Influence des deux descripteurs (MATS7v et R4e<sup>+</sup>) sur l'α- glucosidase

	Composés		
	40	18	1
<b>MATS7V</b>	-0.213	-0.501	-0.681
<b>R4e+</b>	0.079	0.086	0.074
<b>pIC<sub>50</sub></b>	-0.903	-1.992	-2.371
<b>IC<sub>50</sub></b>	8.3	102.3	235.2

#### Le deuxième modèle

Le modèle généré par l'approche 3D-MoRSE, représenté par l'équation (*Equ.2*) regroupe quatre descripteurs définis comme suit : [Mor31u (3D-MoRSE - signal 31 / non pondéré); Mor18u (3D-MoRSE - signal 18 / non pondéré); Mor06u (3D-MoRSE - signal 06 / non pondéré)], et Mor27v (3D-MoRSE - signal 27 / pondéré par le volume atomique de van der Waals).

D'après l'équation (*Equ.2*), seul le descripteur Mor27v contient une propriété chimique dans son expression (volume de van der Waals).

Le signe positif de Mor27v dans l'équation de régression indique que pIC<sub>50</sub> est directement proportionnel à ce descripteur. Par conséquent, l'inhibition peut être favorisée par les molécules qui possèdent de grandes valeurs pour Mor27v. Ainsi, nous pouvons constater que le renforcement du volume de van der Waals augmente l'inhibition de l'enzyme  $\alpha$ -glucosidase. Les trois descripteurs Mor31u; Mor18u; Mor06u reçoivent des coefficients négatifs dans l'équation de régression, ce qui indique que l'inhibition est inversement proportionnelle à ces trois descripteurs.

Il est important de noter que les deux descripteurs (Mor27v et Mor31u), avec leurs coefficients élevés, sont les deux variables indépendantes importantes présentés dans cette équation.

#### IV. CONCLUSION

Ce chapitre a été consacré à l'application de la méthodologie de QSAR, pour modéliser l'activité inhibitrice contre l' $\alpha$ -glucosidase à partir d'un ensemble de 57 molécules, dérivées de xanthones et de curcuminoïdes. La méthode de régression linéaire multiple et les algorithmes génétiques sont utilisés dans le développement des modèles en tant que méthode d'apprentissage et de sélection respectivement.

Nous avons développé deux modèles de QSAR, le premier (Equ. 1) est obtenu par l'utilisation de la totalité des descripteurs issus du serveur E-DRAGON1, alors que le deuxième modèle (Equ 2) est obtenu en utilisant seulement les descripteurs de la famille 3D-MoRSE.

Les deux modèles ont été validés par la validation croisée, le test de randomisation et la validation externe (sur PSET).

Les résultats obtenus de la validation et l'analyse des valeurs résiduelles normalisées prouvent la validité, la stabilité et la robustesse des deux modèles obtenus.

Nous avons complété notre étude par l'analyse des points influents en utilisant le domaine d'applicabilité chimique et l'analyse des points aberrants. Le domaine d'applicabilité sert à filtrer les composés influents et de donner les limites de l'application des modèles obtenus.

**V. REFERENCES**

- [1] Y.P. Zhu; L. J. Yin; Y. Q. Cheng; K. Yamaki; Y. Mori; Y. C. Su; L. T. Li *Food Chem.* 109 (2008) 737–742.
- [2] W. H. Xu; F. G Dai; G. Z. Liu; J. F. Wang; H. M. Liu *Bioorg. Med. Chem.* 15 (2007) 4247–4255.
- [3] K. Kimura; J.H. Lee; I.S. Lee; H.S. Lee; K.H. Park; S. Chiba; D.M. Kim *Carbohydr. Res.* 339 (2004) 1035–1040.
- [4] K. Y. Kim; K. A. Nama; H. Kurihara; S. M. Kim *Phytochemistry.* 69 (2008) 2820–2825.
- [5] M. Karelson “Molecular Descriptors in QSAR/QSPR” Ed. Wiley- Interscience. Etats Unies, 2000.
- [6] M. Fernandez; J. Caballero; A. M. Helguera; E. A. Castro; M. P. Gonzalez *Bioorg. Med. Chem.* 13 (2005) 3269 - 3277.
- [7] S. Liane; M. P. Gonzalez; Y. Fall; G. Gomez *Eur. J. Med. Chem.* 42 (2007) 64-70.
- [8] A. Tropsha; W. Zheng *Curr. Pharm. Des.* 7 (2001) 125 – 133.
- [9] Y. Liu; Z. Ke; J. Cui; W. H. Chen; L. Ma; B. Wang *Bioorg. Med. Chem.* 16 (2008) 7185–7192.
- [10] Z. Y. Du; R. R. Liu; W. Y. Shao; X. P. Mao; L. Ma; L. Q. Gu; Z. S. Huang; A. S. C. Chan *Eur. J. Med. Chem.* 41 (2006) 213–218.
- [11] A. Golbraikh; A. Tropsha *Molecular Diversity* 5 (2002) 231-243.

## CHAPITRE V:

### MODELISATION DE L'INHIBITION DU VIH-1

#### PAR LES DERIVES DU FLAVONOIDE

##### I. INTRODUCTION

Le syndrome acquis d'immunodéficientaire (SIDA) provoqué par le virus de type (VIH-1), est devenu une pandémie mondiale majeure. Trois millions de personnes sont mortes du SIDA et 40 millions vivaient avec à la fin de l'année 2003<sup>1,2</sup>.

Malgré la disponibilité de plusieurs drogues anti-VIH-1 permettant d'améliorer la vie des personnes portant ce virus, la résistance de ce dernier aux médicaments pousse encore les scientifiques à la recherche de nouveaux agents thérapeutiques. Ainsi, beaucoup de produits d'origines naturelles, en raison de leur diversité structurale, ont été expérimentés pour déceler des effets antiviraux. Parmi les quels on peut citer: les alcaloïdes, les polysaccharides sulfatés, les polyphénols, les flavonoïdes, les coumarines, les composés phénoliques, les tannins, les triterpènes, les phloroglucinols, les lactones, les depsidones, les dérivés d'O-caffeoyl, les protéines, les saponines, les xanthones, les naphthodianthrones, les phospholipides, les quinines et les peptides<sup>3-5</sup>.

Hu<sup>6</sup> et ses collaborateurs ont testé l'activité antivirale d'une série étudié de dérivés flavonoïdes, ils cherchaient à trouver une relation entre les modifications structurales de ces dérivés flavonoïdes avec leurs activités antivirales mesurées.

Récemment, des études de QSAR ont été rapportées sur un ensemble de dérivés de flavonoides composé de 24 molécules. Ils ont cherché de trouver, au moyen de descripteurs électroniques, une relation quantitative entre ces structures et leur activité inhibitrice contre le VIH-1. Ces études ont été réalisées avec des méthodes statistiques différentes, y compris l'Analyse de Composantes Principales (ACP); l'Analyse Hierarchical Cluster (AHC); la Régression Linéaire Multiple (MLR) et les moindres carrés partiel (PLS)<sup>7,8</sup>.

Dans ce chapitre nous exposerons une étude de QSAR sur 24 dérivés de flavonoïdes substitués sur des positions différentes (Figure 1). Nous essayons à partir de cette étude de trouver un modèle de QSAR reliant l'activité inhibitrice de ces composés contre le VIH-1 avec leur structure moléculaire. Les descripteurs moléculaires utilisés dans cette étude de QSAR sont ceux existant sur le serveur E-DRAGON1. Les méthodes de sélection utilisées sont les algorithmes génétiques et la méthode ascendante pas à pas (FS).

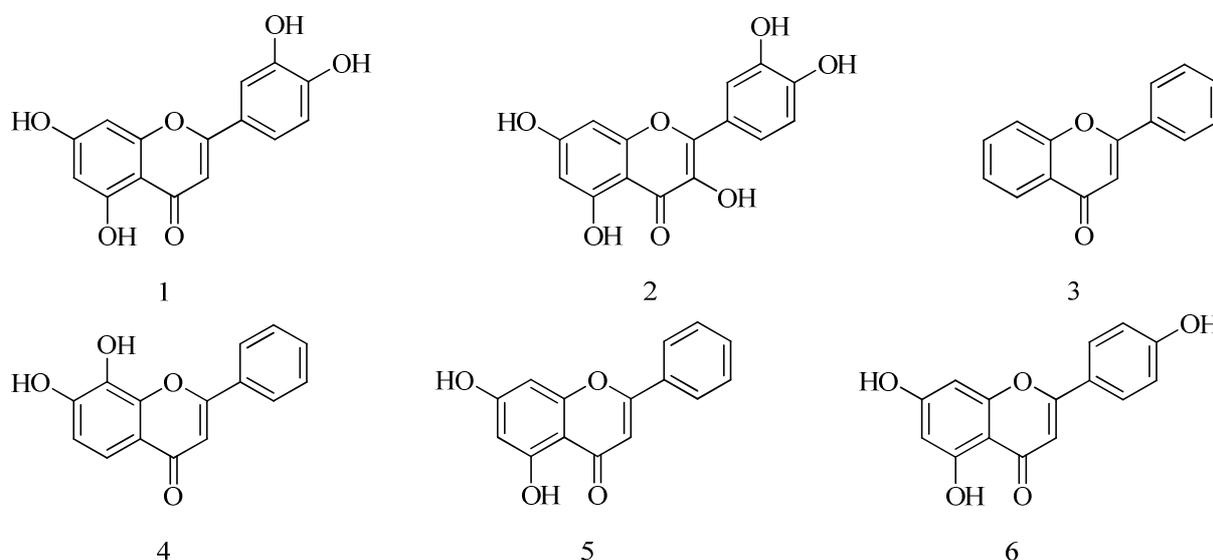
## II. METHODES EXPERIMENTALES

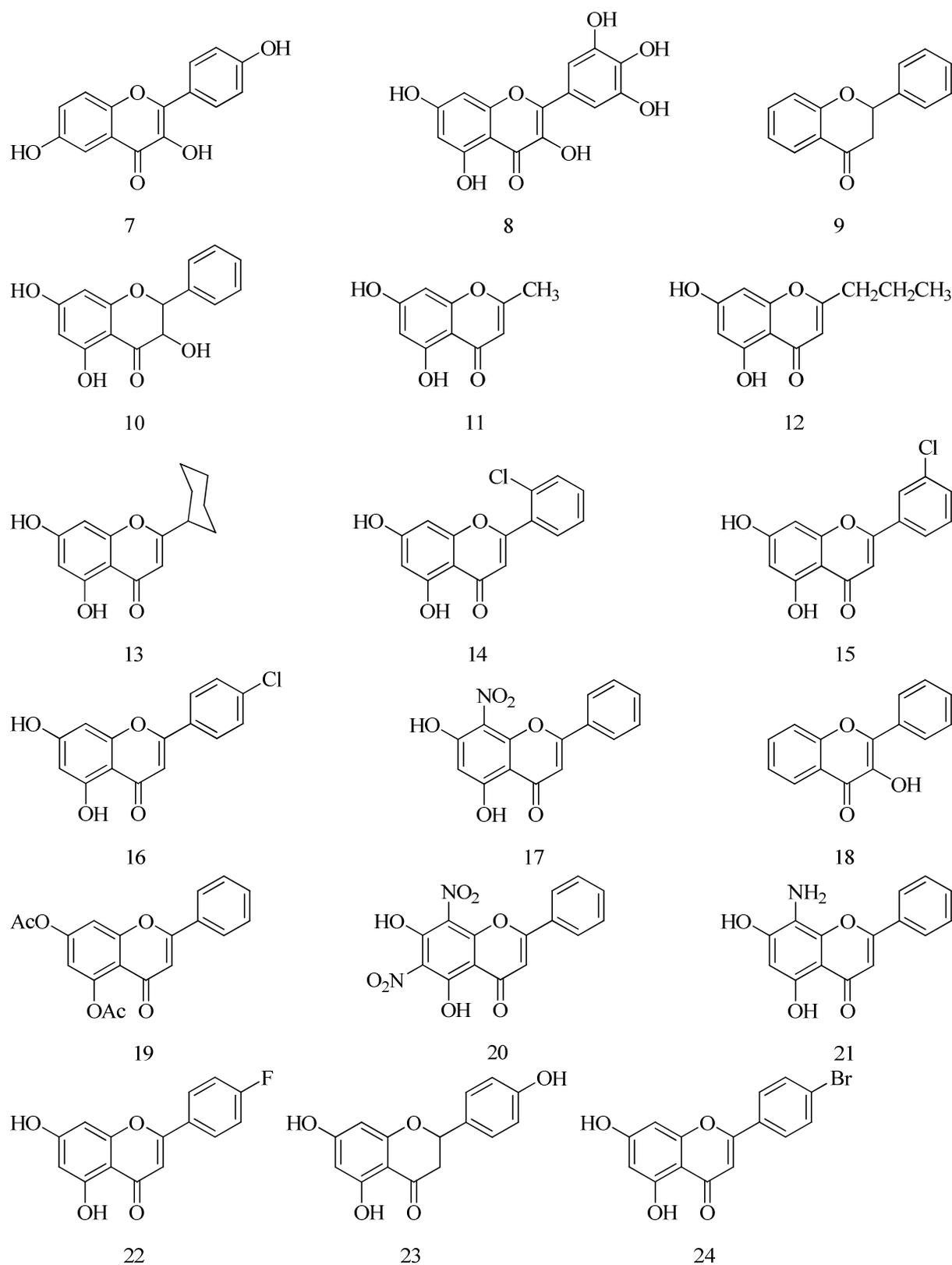
### 1. Ensemble des données

Pour cette étude, nous avons choisi de travailler avec 24 molécules dérivés du flavonoïde (Figure 1) parmi celles ayant une valeur de IC<sub>50</sub> représentant l'activité inhibitrice contre le virus VIH-1. Ces données ont été tirées d'un travail de Hu<sup>6</sup>.

L'activité biologique a été évaluée avec la grandeur (IC<sub>50</sub>), mesurant la concentration de flavonoïde nécessaire pour réduire la croissance non infectée (sans contamination) de cellules de lymphocyte à 50% de la culture de cellules non traitées.

Dans cette thèse, nous avons exprimé l'activité antivirale par le rapport logarithmique, désigné par pIC<sub>50</sub> (LOG(1/IC<sub>50</sub>)).





**Figure 1.** Structures développées des dérivés du flavonoïde

## 2. Dessin et optimisation des structures

Les structures ont été dessinées à l'aide de l'interface du logiciel HyperChem. Les structures tridimensionnelles sont ensuite optimisées avec la méthode semi empirique AM1. Les conformations optimales 3D obtenues seront utilisées pour la génération des descripteurs dépendants de la géométrie.

## 3. Génération des Descripteurs

La génération des descripteurs moléculaires a été faite à l'aide du serveur E-DRAGON1. Le nombre de descripteurs obtenu a été réduit en utilisant la procédure de la sélection objective. Elle consiste à éliminé tout descripteur qui donne soit les mêmes valeurs pour toutes les structures ou possédant un coefficient de corrélation supérieur à 0,9 ( $R > 0,9$ ) lors de sa combinaison avec un autre descripteur.

## 4. Sélection des variables et formation du modèle

Les descripteurs obtenus par la sélection objective ont subi, dans une deuxième étape, la sélection subjective afin de trouver un sous ensemble optimal riche en informations. A cet effet deux méthodes de sélection ont été utilisées: les algorithmes génétiques et la méthode ascendante pas à pas (Forward stepwise).

## III. RESULTATS ET DISCUSSION

Dans notre cas, nous avons construis deux modèles de QSAR capables de modéliser l'activité inhibitrice d'un ensemble de 24 dérivés de flavonoïde contre l'enzyme VIH-1. Le modèle obtenu est le résultat d'une longue procédure dont la sélection de variables constitue l'étape la plus importante dans ce parcours.

Pour le faire, nous avons fait appel à deux méthodes différentes: la sélection par les algorithmes génétiques et la sélection ascendante pas à pas (forward stepwise). L'analyse des paramètres statistiques (coefficient de corrélation ; coefficient de détermination; coefficient de

détermination ajusté ; et l'écart type  $s$ ) des modèles obtenus a permis la comparaison entre ces deux méthodes.

Vu le nombre réduit des molécules, nous avons décidé d'utiliser les 24 structures comme ensemble d'apprentissage (TSET), de même nous avons limité la dimension des modèles à quatre descripteurs tout en respectant la règle de Topliss.

Le meilleur modèle, ainsi obtenu, sera analysé statistiquement pour vérifier sa justesse, sa robustesse et sa prédiction. Dans une deuxième étape nous étudierons l'effet de la diversité structurale sur la grandeur biologique  $pIC_{50}$ , qui explique l'activité inhibitrice de cet ensemble par l'examen des descripteurs inclus dans le modèle sélectionné.

### 1. Sélection des variables

Après la sélection objective, la matrice réduite de descripteurs est utilisée pour générer le modèle optimal en utilisant les deux méthodes mentionnées précédemment. Les résultats obtenus sont affichés dans le tableau suivant (Tableau 1) :

**Tableau 1.** Propriétés statistiques des deux modèles obtenus par les GA et FS

	N° Modèle	Méthode de sélection	Descripteurs sélectionnés	Paramètres statistiques			
				R	R <sup>2</sup>	R <sup>2</sup> aj	s
pIC <sub>50</sub>	01	GA <sup>a</sup>	RDF110e; Mor26m; G2u; E1m	0.939	88.20	85.70	0.184
	02	Asc <sup>b</sup>	GGI10; RDF085v; nX; G2p	0.861	74.10	68.70	0.273

a: Les algorithmes génétiques ; b : ascendante.

c: La définition des descripteurs affichés par ce tableau est donnée dans le tableau 1 (Annexe : Chapitre V)

A partir de ces résultats, les paramètres utilisés pour la comparaison indiquent clairement que les deux algorithmes utilisés ont donné des résultats complètement différents. Nous remarquons que pour la même grandeur biologique, les deux méthodes de sélection de variables fournissent deux ensembles de descripteurs complètement différents présentant des mesures statistiques différentes. Mais il est important de noter la supériorité remarquable des algorithmes génétiques dans l'exploration de l'espace de recherche des meilleurs descripteurs par rapport à l'algorithme ascendant pas à pas.

Le modèle (N°1) avec  $R^2 = 88,20\%$  explique mieux la variance des données expérimentales ( $pIC_{50}$ ) que le modèle N°2 ( $R^2 = 74,10\%$ ).

La faible valeur de l'écart type  $s$  enregistrée avec les algorithmes génétiques consolide également la même conclusion.

Le modèle de QSAR obtenu avec les GA comme procédure de sélection de variables, fournit les meilleurs paramètres statistiques. Ces résultats préliminaires obtenus, nous ont poussés à traiter ce modèle statistiquement et l'analyser afin de confirmer sa robustesse, sa validité, et son pouvoir à prédire des activités inhibitrices pour de nouvelles structures.

## 2. Analyse de la justesse du modèle

Nous avons reconstruit le modèle N°1 mentionné ci-dessus par l'application de la régression linéaire multiple incluse dans le logiciel MINITAB 15.

Le modèle est affiché comme suit:

Le modèle mentionné ci-dessous relie la variable dépendante  $pIC_{50}$  avec les quatre descripteurs (Tableau 2). Dans le même tableau nous avons rapporté les mesures statistiques suivantes : coefficient de corrélation multiple  $R$ , coefficient de détermination  $R^2$ , coefficient de détermination ajusté  $R^2_{adj}$ , et l'écart type  $s$ .

---


$$pIC_{50} = 4.48 - 0.625 \text{ RDF110e} - 2.62 \text{ Mor26m} - 9.44 \text{ G2u} + 2.52 \text{ E1m} \quad (\text{Equ.1})$$

$$N = 24; \quad R = 0.939 \quad R^2 = 88.20\%; \quad R^2_{aj} = 85.70\%; \quad s = 0.184$$


---

Les résultats statistiques présentés par l'Equ.1 sont similaires à ceux déjà obtenus avec le modèle de  $pEC_{50}$ . On remarque une forte corrélation entre les variables expérimentales et celles prédites, exprimée par une valeur de  $R = 0.939$ . La variabilité a été exprimée par des valeurs plus élevées de deux paramètres  $R^2$  et  $R^2_{aj}$ , qui expliquent respectivement 88.20% et 85.70% de la variance des valeurs  $pIC_{50}$  expérimentales.

**Tableau 2.** Les valeurs pIC<sub>50</sub> expérimentales et prévues avec les descripteurs sélectionnés

No.	pIC <sub>50</sub> exp.	pIC <sub>50</sub> pred.	RDF110e	Mor26m	G2u	E1m
1	4.796	4.458	1.011	-0.388	0.182	0.519
2	3.879	3.852	1.218	-0.504	0.245	0.445
3	4.167	4.383	0.001	-0.341	0.195	0.336
4	4.854	4.821	0.001	-0.495	0.190	0.331
5	4.328	4.253	0.004	-0.368	0.228	0.381
6	4.456	4.518	0.678	-0.383	0.188	0.488
7	3.804	3.919	1.477	-0.419	0.188	0.411
8	4.161	4.287	1.291	-0.549	0.214	0.473
9	4.347	4.232	0.005	-0.252	0.190	0.351
10	4.357	4.575	0.003	-0.304	0.184	0.410
11	3.670	3.728	0.000	-0.183	0.211	0.300
12	4.137	4.198	0.000	-0.188	0.174	0.343
13	4.377	4.188	0.150	-0.175	0.193	0.461
14	4.770	4.681	0.000	-0.431	0.190	0.342
15	4.854	4.704	1.687	-0.569	0.190	0.626
16	4.796	4.870	2.029	-0.569	0.190	0.777
17	4.921	4.668	0.002	-0.439	0.186	0.314
18	4.770	4.621	0.002	-0.379	0.172	0.305
19	3.428	3.536	3.113	-0.598	0.159	0.370
20	3.429	3.229	2.764	-0.489	0.197	0.417
21	4.469	4.824	0.039	-0.459	0.186	0.364
22	4.886	4.886	0.017	-0.349	0.209	0.584
23	3.532	3.734	1.358	-0.312	0.200	0.464
24	4.678	4.701	2.679	-0.177	0.190	1.279

La corrélation entre les descripteurs et la variable dépendante est justifiée par l'examen du tableau d'ANOVA et le tableau des coefficients (Tableau 3 et Tableau 4) respectivement.

**Tableau 3.** Résultats de l'analyse de la variance (ANOVA)

Source	DF <sup>a</sup>	SC <sup>b</sup>	CM <sup>c</sup>	F <sub>obs</sub>	P <sup>d</sup>
Régression	4	4.8133	1.2033	35.42	0.000
Erreur Résiduel	19	0.6454	0.0340		
Totale	23	5.4588			

a: degré de liberté ; b : somme des carrées ; c : moyenne des carrées ; d : probabilité.

L'existence d'une corrélation entre l'ensemble des descripteurs avec la variable dépendante pIC<sub>50</sub> est testée en premier lieu en examinant le tableau d'ANOVA. Ce test est effectué en comparant la valeur de la statistique F observée à la valeur de F tabulée.

A partir de ce tableau la valeur de F observée ( $F_{obs}=35.42$ ) est plus élevée que la valeur théorique  $F_{(0.05; 4; 19)}=1.47$ . On peut conclure donc qu'il existe au moins un coefficient différent de zéro ce qui indique que le descripteur qui lui correspond est significativement corrélé avec l'activité biologique.

L'examen du tableau des coefficients permet de vérifier la signification des descripteurs dans la corrélation avec  $pIC_{50}$ , en comparant, pour chacun d'eux, avec la valeur statistique théorique  $t_{(0.025,19)} = 2.093$ .

**Tableau 4.** Tableau des coefficients

descripteur	Coef <sup>a</sup>	Er. T (Coef) <sup>b</sup>	$t_{obs}$ <sup>c</sup>	$p$ <sup>d</sup>
Constante	4.4846	0.4417	10.15	0.000
RDF110e	-0.62506	0.05579	-11.20	0.000
Mor26m	-2.6200	0.3659	-7.16	0.000
G2u	-9.440	2.1970	-4.30	0.000
E1m	2.5180	0.2495	10.09	0.000

a : coefficients ; b : erreur type ; c : test de student observé ; d : la valeur de probabilité.

D'après ce tableau, les valeurs élevées du  $t_{obs}$  ( $t_{obs} > t_{(0.025,19)} = 2.093$ ) pour les cinq paramètres (les 4 descripteurs et la constante) confirment la présence de corrélation entre chaque descripteur du modèle et l'activité  $pIC_{50}$ . Les valeurs inférieures à  $10^{-3}$  de p enregistrés dans la 5<sup>ème</sup> colonne du tableau confirment que les coefficients diffèrent de manière significative de zéro.

### 3. Tests de colinéarité et de Multicolinéarité

Nous avons abordé le problème de colinéarité et de multicolinéarité respectivement par l'examen des résultats affichés dans le tableau suivant (Tableau 4).

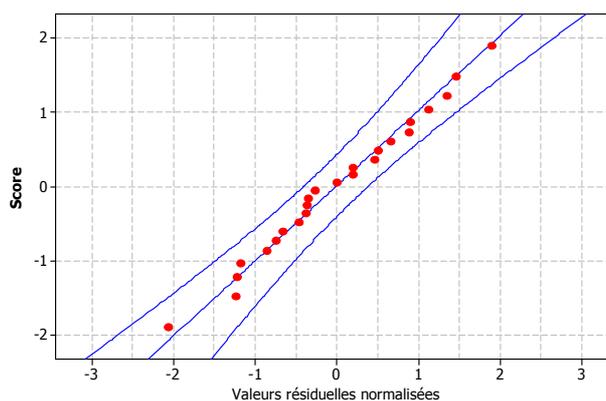
**Tableau 4.** Matrice de corrélation de l'équation 1

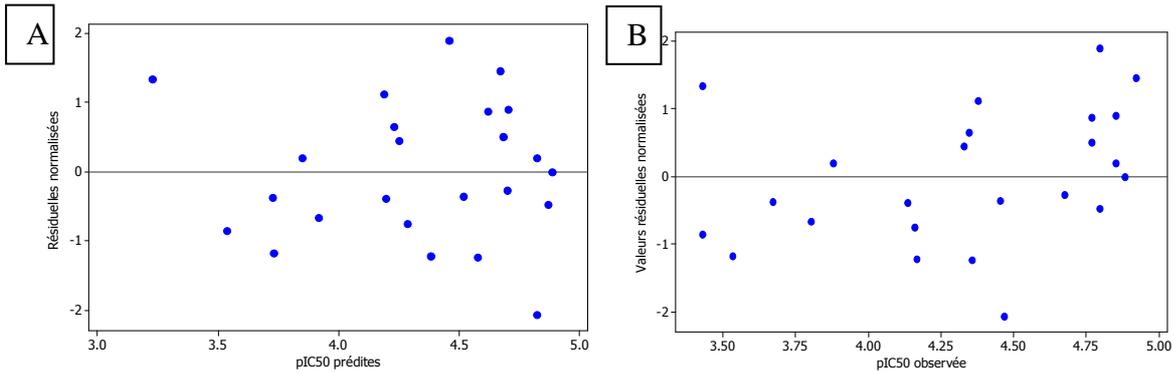
	Coefficient de corrélation			VIF
	RDF110e	Mor26m	G2u	
RDF110e	1			2.223
Mor26m	-0.421	1		1.531
G2u	-0.121	-0.006	1	1.047
E1m	0.547	0.104	0.032	1.817

D'après ce tableau, les faibles valeurs enregistrées du coefficient de corrélation et du critère VIF indiquent l'absence de problèmes de colinéarité et de multicollinéarité.

#### 4. Analyse des valeurs résiduelles normalisées

La figure 2 qui présente les valeurs scores en fonction des valeurs résiduelles normalisées nous indique que les points ainsi projetés forment approximativement une droite linéaire dans un intervalle de confiance de 95%. Ce qui montre que les valeurs résiduelles normalisées sont distribuées normalement.

**Figure 2.** Valeurs de score en fonction des valeurs résiduelles normalisées



**Figure 3.** Valeurs résiduelles normalisées en fonction de pIC<sub>50</sub> prédites (A) et observées (B)

Les représentations graphiques des valeurs résiduelles normalisées en fonction pIC<sub>50</sub> prédites et observées respectivement sont indiquées sur la figure 3. On constate d’après cette figure que la dispersion des valeurs résiduelles normalisées est presque nulle, ce qui prouve que la précision de notre modèle est bonne.

### 5. Validation du modèle

Nous avons contrôlé la stabilité et la validité de notre modèle en utilisant deux procédures : validation croisée (cv-loo : leave – one- out et cv-lgo : leave – groupe- out) et le test de Randomisation.

Les résultats issus de la validation sont rassemblés dans le tableau suivant :

**Tableau 5.** Validation croisée par les deux procédures (leave one out et leave group out)

	validation croisée	
	Leave one out	Leave groupe out
$Q^2$	83.14	84.19
s	0.1959	0.1895

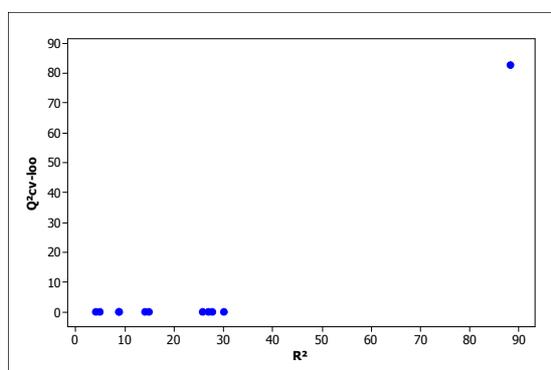
Les résultats montrent des valeurs assez élevées du coefficient de détermination issus de la validation croisée ( $Q^2_{cv-loo} = 83.14$  ;  $Q^2_{cv-lgo} = 84.19$ ) en revanche les valeurs de l’écart type sont faibles ( $s_{cv-loo} = 0.1959$  ;  $s_{cv-lgo} = 0.1895$ ), ceci prouve une bonne stabilité et une bonne prédiction du modèle par rapport aux processus d’inclusion et d’exclusion.

La robustesse du modèle est ensuite vérifiée par le test de randomisation en examinant les deux paramètres statistiques ( $R^2$  et  $Q^2_{cv-100}$ ). Les valeurs issues de chaque itération sont affichées et présentées dans le tableau 6 et la figure 4 :

**Table 6:**  $R^2$  et  $Q^2_{cv-100}$  issus du test de randomisation

Iteration	$R^2$	$Q^2_{cv-100}$
1	8.60	0
2	26.80	0
3	27.70	0
4	4.80	0
5	8.70	0
6	3.90	0
7	13.90	0
8	14.70	0
9	30.00	0
10	25.70	0
11*	88.20	83.10

\* : modèle représenté par l'équation 1.



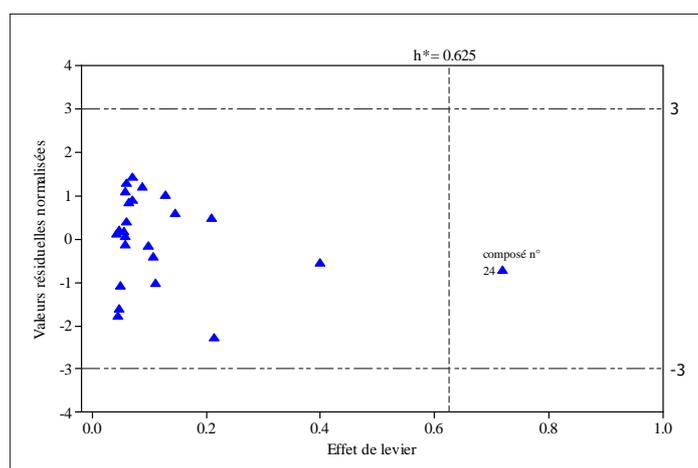
**Figure 4.** Valeurs de  $Q^2_{cv-100}$  en fonction de  $R^2$  issues du test de randomisation

## 6. Domaine d'applicabilité

Le calcul des valeurs de  $h_i$  et  $\delta_i$  (Tableau 7) pour tous les composés de la base nous a permis de définir le graphe de Williams. Ce dernier est une représentation graphique des résidus standardisés en fonction des valeurs de levier pour chaque molécule (Figure 5).

**Table 7.** Valeurs résiduelles normalisées et valeurs de levier

Molécule No.	pIC50	
	$\delta_i$	$h_i$
1	1.9620	0.0681
2	0.2965	0.4490
3	-1.5026	0.0760
4	-0.1561	0.1761
5	0.4347	0.2162
6	-0.5429	0.0568
7	-0.7471	0.0895
8	-0.9538	0.1664
9	0.6500	0.1104
10	-1.5630	0.0986
11	-0.2858	0.2675
12	-0.4938	0.2216
13	1.2481	0.1622
14	0.2736	0.1142
15	0.7623	0.1858
16	-0.7594	0.2760
17	1.3126	0.1279
18	0.6786	0.1658
19	-0.9765	0.5252
20	1.7017	0.3487
21	-2.0641	0.1311
22	-0.2016	0.1676
23	-1.1853	0.1356
24	0.1049	0.7950

**Figure 5.** Graphe de Williams définie par les quatre descripteurs

Selon cette représentation graphique, le modèle ne présente aucun point aberrant, c'est-à-dire que toutes les observations sont bornées par les deux limites 3 et -3. Concernant les points

influent, on observe la présence d'un seul produit (n°24) uniquement pour le modèle représenté par l'équation 04 avec ( $h_i > h^* = 0.625$ ).

### 7. Interprétation chimique de l'influence des descripteurs sur l'activité inhibitrice contre le VIH-1

Le modèle de QSAR présenté par l'équation 1 renferme des descripteurs de natures différentes et chacun d'eux quantifie une caractéristique structurale spécifique. A partir de cette équation, nous pouvons tirer les deux points essentiels suivants :

1. Les quatre descripteurs appartiennent aux trois familles suivantes :

- WHIM : représentée par les deux descripteurs (G2U et E1m).
- 3D-MoRSE : représentée par Mor26m.
- RDF : représentée par RDF110e.

2. Sur le plan des propriétés chimiques, nous observons que :

- E1m et Mor26m : utilisent la masse atomique.
- RDF110e : utilise l'électronégativité dans sa formule.
- G2U : un descripteur qui explique la symétrie de la molécule et n'utilise aucune propriété chimique dans sa formule.

Le G2U, indice issu de la 2<sup>nd</sup> composante directionnelle symétrique, est l'indice le plus influent sur la variable pIC<sub>50</sub>, car il présente le coefficient le plus élevé dans l'*Equ.1*. D'autre part, et à partir du tableau 2 (page 124) (qui affiche les descripteurs et les valeurs de pIC<sub>50</sub> observées), ce descripteur est corrélé inversement avec la grandeur pIC<sub>50</sub>, c'est-à-dire que les composés présentant des petites valeurs de G2U sont considérés plus actifs (des valeurs de pIC<sub>50</sub> plus élevées). De même, le descripteur RDF110e pondéré par l'électronégativité est inversement corrélé avec pIC<sub>50</sub>. Les deux autres descripteurs (en relation avec la masse atomique) sont corrélés positivement avec la variable dépendante.

#### IV. CONCLUSION

Dans ce chapitre, nous avons rapporté une étude de QSAR sur un ensemble de 24 composés dérivés de flavanoïde afin de modéliser leur activité inhibitrice contre le VIH1. La grandeur biologique utilisée est  $pIC_{50}$ .

Au début, une étude comparative a été réalisée entre les deux méthodes de sélection : les algorithmes génétiques (GA) et l'algorithme ascendant pas à pas (Forward Stepwise).

Le modèle obtenu avec GA a donné les meilleurs paramètres statistiques pour modéliser l'activité  $pIC_{50}$  par rapport à l'algorithme ascendant. Le modèle proposé a prouvé sa robustesse, sa bonne précision ainsi que sa bonne stabilité après vérification avec respectivement la méthode de validation croisée (LOO et LGO) et le test de randomisation. Le modèle présenté par l'équation 1 présente la capacité d'expliquer plus de 88% de la variance des valeurs de l'activité biologique observées. Le domaine d'applicabilité chimique a été défini pour déceler les points aberrants et les composés chimiques influents. Cette étude a montré qu'il n'y avait pas de points aberrants et que seul le composé chimique influent était le composé N° 24.

**V. REFERENCES**

- [1] R. Garg ; B. Bhatarai *Top Heterocycl Chem* 3 (2006) 181–271.
- [2] X. Li; R. Vince *Bioorg. Med. Chem.* 14 (2006) 5742–5755.
- [3] Q. Wang; Z. H. Ding; J. K. Liu; Y. T. Zheng *Antiviral Research* 64 (2004) 189-194.
- [4] K. K. Carroll; N. Guthrie; F. V. So; A. F. Chambers “Flavonoids in Health and Disease” Ed. Marcel Dekker. Etats Unies, 1998.
- [5] K. H. Park; Y. D. Park; J. M. Han; K. R. Im; B. W. Lee; I. Y. Jeong; T. S. Jeong; W. S. Lee *Bioorg. Med. Chem. Lett.* 16 (2006) 5580.
- [6] C. Hu; K. Chen; Q. Shi; R.E. Kilkuskie; Y. Cheng; K. Lee *J. Nat. Prod.* 57 (1995) 42-51.
- [7] J. J. Lameira; C.N. Alves; V. Moliner; E. Silla *Eur J. Med. Chem.* 41 (2006) 616.
- [8] J. Lameira; I. G. Medeiros; M. Reis; A. S. Santos; C. N. Alves *Bioorg. Med. Chem.* 14 (2006) 7105.

## CONCLUSION GENERALE

L'objectif de notre travail est de modéliser les activités inhibitrices de  $\alpha$ -glucosidase et du virus VIH-1 pour former des modèles de QSAR robustes, stables, et précis capables de prédire efficacement ces activités.

En premier, nous avons choisis 57 molécules d'origines naturelles dérivés des xanthones et de curcuminoïdes afin de modéliser l'activité inhibitrice de  $\alpha$ -glucosidase exprimée par la grandeur  $pIC_{50}$ . Deux modèles ont été construits en utilisant la méthode des algorithmes génétiques comme méthode de sélection.

Le premier modèle est obtenu en utilisant la totalité des descripteurs issus du serveur E-DRAGON1. En revanche, le deuxième modèle est obtenu seulement avec les descripteurs de la famille 3D-MoRSE.

Les résultats de la validation externe et interne des deux modèles et l'analyse de leurs caractéristiques statistiques ont montré la robustesse, la stabilité, la validité et la bonne prédiction de ces deux modèles.

Deuxièmement et suivant la même démarche, nous avons modélisé l'activité inhibitrice  $IC_{50}$  du virus VIH-1 à partir d'une série de 24 dérivés de flavonoïdes en utilisant tout les descripteurs du E-DRAGON1. Deux méthodes de sélection ont été utilisées : les algorithmes génétiques (GA) et la méthode ascendante (forward stepwise FS). On a pu démontrer que le modèle obtenu par les algorithmes génétiques est plus robuste, plus précis et d'une bonne stabilité par rapport au deuxième modèle développé par la méthode ascendante (FS), ce qui montre que la méthode de GA, dans notre exemple, est meilleure dans la sélection des descripteurs par rapport à la méthode FS.

Enfin, nous avons pu soutirer des informations utiles à partir des équations des modèles. Ainsi, nous avons pu interpréter la contribution des descripteurs sur la variation de l'activité à l'aide des descripteurs pondérés tels que le volume de van der Waals, l'électronégativité de Sanderson et la masse atomique. En revanche, les autres descripteurs non pondérés sont difficiles à interpréter car ils ne sont pas liés aux paramètres physicochimiques.

A travers les différents résultats obtenus au cours de ce travail, nous pouvons dire que le but que nous nous sommes fixés au départ a été largement atteint. Ainsi, les ensembles de molécules, les traitements statistiques et les techniques informatiques utilisés, lors du développement et l'analyse des modèles de QSAR, ont donné de bons résultats, ce qui nous permet d'entrevoir des perspectives assez prometteuses dans ce domaine par l'amélioration des traitements statistiques et par l'utilisation d'autres méthodes de sélection. Ce travail a un prolongement évident dans la modélisation qualitative et précisément avec les méthodes de classification comme l'analyse par composantes principales ACP, la fonction de discrimination linéaire (LDA), machine à support vecteur SVM...etc.

## Chapitre IV

**Tableau 1.** Symboles et définitions des descripteurs utilisés dans le chapitre IV.

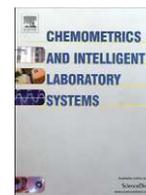
Symbole	Définition
MATS7v	Moran autocorrelation - lag 7/weighted by atomic van der Waals volumes
Mor15u	3D-MoRSE - signal 15 / unweighted
nArOR	Number of ethers (aromatic)
R4u	R autocorrelation of lag 4 / unweighted
H5u	H autocorrelation of lag 5 / unweighted
R4e+	R maximal autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities
HATS5u	leverage-weighted autocorrelation of lag 5 / unweighted
H8m	H autocorrelation of lag 8 / weighted by atomic masses
BIC2	bond information content (neighborhood symmetry of 2-order)
HOMA	Harmonic Oscillator Model of Aromaticity index
Mor16u	3D-MoRSE - signal 16 / unweighted

**Tableau 2.** Symboles et définitions des descripteurs utilisés dans le chapitre IV.

Symbole	Définition
<b>2D-autocorrélation</b>	
GATS2m	Geary autocorrelation - lag 2 / weighted by atomic masses
GATS7m	Geary autocorrelation - lag 7 / weighted by atomic masses
GATS1v	Geary autocorrelation - lag 1 / weighted by atomic van der Waals volumes
GATS3v	Geary autocorrelation - lag 3 / weighted by atomic van der Waals volumes
<b>Geometrical</b>	
HOMA	Harmonic Oscillator Model of Aromaticity index
RCI	Jug RC index
AROM	Aromaticity (trial)
HOMT	HOMA total (trial)
<b>GETAWAY</b>	
HATS7u	Leverage-weighted autocorrelation of lag 7 / unweighted
R3u	R autocorrelation of lag 3 / unweighted
R4u	R autocorrelation of lag 4 / unweighted
R3e+	R maximal autocorrelation of lag 3 / unweighted
<b>3D-MORSE</b>	
Mor06u	3D-MoRSE - signal 06 / unweighted
Mor18u	3D-MoRSE - signal 18 / unweighted
Mor31u	3D-MoRSE - signal 31 / unweighted
Mor27v	3D-MoRSE - signal 27 / weighted by atomic van der Waals volumes
<b>RDF</b>	
RDF025m	Radial Distribution Function - 2.5 / weighted by atomic masses
RDF130v	Radial Distribution Function - 13.0 / weighted by atomic van der Waals volumes
RDF010e	Radial Distribution Function - 1.0 / weighted by atomic Sanderson electronegativities
RDF060p	Radial Distribution Function - 6.0 / weighted by atomic polarizabilities
<b>WHIM</b>	
E2u	2nd component accessibility directional WHIM index / unweighted
G2m	2st component symmetry directional WHIM index / weighted by atomic masses
G2e	2st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities
E1s	1st component accessibility directional WHIM index / weighted by atomic electrotopological states

**Chapitre V****Tableau 1.** Symboles et définitions des descripteurs utilisés dans le chapitre V.

Symbol	Définition
nX	Number of halogen atoms(constitutional descriptors)
RDF110e	Radial Distribution Functio-11.0/ weighted by atomic atomic Sanderson electronegativities
RDF085v	Radial Distribution Functio-8.5/ weighted by atomic van der Waals volumes
E1m	1 <sup>st</sup> component accessibility directional WHIM index/ weighted by atomic masses
G2p	2 <sup>nd</sup> component symmetry directional WHIM index/ weighted by atomic polarizabilities
GGI10	Topological charge index of order 10
G2u	2 <sup>nd</sup> component symmetry directional WHIM index/ unweighted
Mor26m	3D -MorSE – signal 26/ weighted by atomic masses



## Quantitative structure activity relationship for the computational prediction of $\alpha$ -glucosidase inhibitory

Khairedine Kraim<sup>a,\*</sup>, Djameleddin Khatmi<sup>a</sup>, Youcef Saihi<sup>b</sup>, Fouad Ferkous<sup>b</sup>, Mohamed Brahimi<sup>c</sup>

<sup>a</sup> Department of Industrial Chemistry, Faculty of Sciences and Techniques of Engineer, University of 08 Mai 1945 Guelma, Algeria

<sup>b</sup> Department of Chemistry, Faculty of Sciences, University of BADJI Mokhtar Annaba, Algeria

<sup>c</sup> National College of Computer Sciences, Algiers, Algeria

### ARTICLE INFO

#### Article history:

Received 3 December 2008

Revised 18 January 2009

Accepted 4 March 2009

Available online 18 March 2009

#### Keywords:

$\alpha$ -glucosidase

Xanthone derivatives

Genetic Algorithm

Multiple Linear Regression

Applicability domain

### ABSTRACT

Quantitative structure–activity relationship (QSAR) models are useful in understanding how chemical structure relates to the biological activity of natural and synthetic chemicals and for design of newer and better therapeutics. In the present study, 57 xanthone and curcuminoid derivatives were evaluated as  $\alpha$ -glucosidase inhibitors, expressed by the cytotoxicity of these compounds ( $IC_{50}$ ). Based on these data, different molecular descriptors were used to solve this problem. A linear QSAR model was developed using Multiple Linear Regression technique, while Genetic Algorithm was adopted for selecting the most appropriate descriptors. The predictive activity of the model was evaluated by means of external validation set and the *Y*-randomization technique, and its structural chemical domain has been verified by the leverage approach. It was able to describe more than 85.7% of the variance in the experimental activity.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

It is well known that  $\alpha$ -glucosidase (EC 3.2.1.20, 3.2.1.10, 3.2.1.48 and 3.2.1.106) are exo-acting carbohydrases, which catalyze release of  $\alpha$ -D-glucopyranose from the non-reducing ends of various carbohydrate substrates [1–3]. These enzymes have drawn a special interest of the pharmaceutical research community because it was revealed that the inhibition of its catalytic activity led to the retardation of glucose absorption and the decrease in postprandial blood glucose level. This indicates that the effective  $\alpha$ -glucosidase inhibitors may serve as chemotherapeutic agents for clinic use in the treatment of diabetes, hypertension, dyslipidemia, obesity and cardiovascular diseases in patients with metabolic syndrome [4–6].

Since the discovery of acarbose [7] (Fig. 1) that is the first member of  $\alpha$ -glucosidase inhibitors approved for the treatment of type 2 diabetes, a many  $\alpha$ -glucosidase inhibitors [8] have been reported, such as voglibose from microorganisms, and 1-deoxynojirimycin isolated from plants, however, they are confined to glucosidic derivatives [9,10]. They can also be synthesized chemically [1].

The lack of structural information about the nature of the interactions between  $\alpha$ -glucosidases and the inhibitors has thus made it a difficult task to discover good lead compounds. Quantitative structure-activity relationship (QSAR) studies are a powerful method

for the design of bioactive compounds and the prediction of activity according to the physical and chemical properties [11–14].

Recently a QSAR study on a data set of 43 xanthone derivatives as  $\alpha$ -glucosidase inhibitors was reported by means of quantum chemical descriptors using Multiple Linear Regression in combination with the Elimination Stepwise as variable selection algorithm. They showed that the inhibitory activity can be modeled by the number of hydrogen bond forming, the number of aromatic rings and the softness value [15].

The aim of this study is to develop a QSAR model of the inhibitory of 57 xanthone and curcuminoid derivatives against  $\alpha$ -glucosidase, to better understand the structural features of these types of compounds, using 0, 1, 2 and 3D molecular descriptors calculated using the E-Dragon. This study may help us to design new analogues with better biological profile.

### 2. Materials and methods

#### 2.1. Dataset and biological data

The database consists of 57 recently discovered xanthone and curcuminoid derivatives as  $\alpha$ -glucosidase inhibitors [15,16]. Their structures and *in vitro* activity are listed in Fig. 2 and Table 1.

Activities were converted into the corresponding  $-\log_{10}IC_{50}$  values ( $pIC_{50}$ ), where  $IC_{50}$  is the effective concentration of compound required to achieve 50% of inhibition of  $\alpha$ -glucosidase.

\* Corresponding author. Tel.: +213 663 58 58 41.  
E-mail address: [kkhchem@gmail.com](mailto:kkhchem@gmail.com) (K. Kraim).



ERROR: undefinedresource  
OFFENDING COMMAND: findresource

STACK:

/0  
/CSA  
/0  
/CSA  
-mark-