SOMMAIRE

Introduction générale

A l'heure actuelle, le cancer est la principale cause des maladies qui provoquent la mort de la population humaine dans certaines régions du monde, et il est prévu de continuer à devenir la principale cause de décès dans le cours des prochaines années. La chimiothérapie, ou l'utilisation d'agents chimiques pour détruire les cellules cancéreuses, est un pilier dans le traitement des tumeurs malignes. L'un des principaux avantages de la chimiothérapie est sa capacité à traiter les cancers généralisées ou métastatiques, considérant que la chirurgie et les traitements d'irradiation sont limitées au traitement des cancers pour des domaines spécifiques.

La chimiothérapie a suscité de nombreux intérêts des chercheurs et beaucoup d'efforts en cours ont été axés sur la conception et le développement de médicaments anticancéreux variés.

La molécule isatine (1H-indole-2,3-dione) est une fraction polyvalente qui affiche diverses activités biologiques, comme l'activité anticancéreuse. Les indoles N-alkylés ont également été signalés comme présentant une activité anticancéreuse.

Dans le chapitre I de ce mémoire, on définit l'activité anticancéreuse de la molécule isatine, nous parlons aussi des autres activités biologiques et de la synthèse de l'isatine.

Dans le chapitre II, on regroupe les généralités sur les descripteurs et les méthodes d'analyses permettant de déterminer une relation quantitative entre la structure et l'activité (QSAR).

Finalement dans le chapitre III, nous appliquons ces méthodes d'analyses à une série de dérivés d'isatine. L'étude de cette série a permis de sélectionner les descripteurs pertinents et de proposer un modèle quantitatif en utilisant la régression linéaire multiple (RLM) et le réseau de neurones (RN). La performance du modèle proposée a été testée avec la méthode de validation croisée (VC).

Université Sidi Mohammed Ben Abdellah

Faculté des Sciences et Techniques

www.fst-usmba.ac.ma

Chapitre I

Généralités sur les molécules bioactives de l'isatine

I. Introduction

Le cancer, qui est la deuxième cause de mortalité en France, est aujourd'hui un problème de santé publique majeur et fait l'objet de nombreuses recherches. Il existe un véritable enjeu commercial, qui pousse les laboratoires pharmaceutiques dans la course à la recherche et au développement de nouveaux traitements.

Actuellement, plusieurs traitements du cancer sont disponibles, comme par exemple la chirurgie, la radiothérapie ou la chimiothérapie. La combinaison de ces trois modes de traitement est utilisée pour guérir un certain nombre de cancers ou diminuer la prolifération des cellules cancéreuses, ce qui permet d'augmenter fortement l'espérance de vie des personnes atteintes par cette pathologie.

II. Généralités sur le cancer

II.1. Définition du cancer

De nos jours, de nombreuses pathologies tuent de nombreuses vies humaines dans tous les coins de la planète dont les maladies cardio-vasculaires, l'hypertension artérielle, cancers et autres.

Les cancers apparaissent comme un désordre de la prolifération des cellules qui se reproduisent de façon excessive et provoquent l'apparition d'une tumeur. La plupart des cellules ont des durées de vie courtes et elles doivent se diviser pour être renouvelées. Il existe une « mémoire cellulaire » suivant laquelle une cellule a, tout au long de sa vie, des caractéristiques qu'elle transmet à la cellule qui prend sa place lorsqu'elle meurt. La prolifération est donc nécessaire à la survie, elle se fait à des vitesses et selon des mécanismes propres à chaque type cellulaire. Si la croissance cellulaire n'est plus contrôlée, les cellules sont capables de se diviser à un rythme anormalement élevé. Ces désordres ont une origine cellulaire, qui peut être liée à une mutation : les cellules cessent d'obéir aux messages de freination ou bien elles se mettent à produire elles-mêmes en excès les facteurs de leur croissance **[1]**.

II.2. Causes connues du cancer

L'inhibition cancéreuse résulte d'une interaction irréversible entre un agent cancérogène (facteurs tératogènes, virus oncogènes, rayons ionisants…) et le génome d'une cellule, induisant une anomalie moléculaire transmissible aux cellules filles. L'anomalie consiste en une mutation majeure de l'ADN. Les cellules initiées engendrent par divisions successives, un ou plusieurs clones.

II.3. Données statistiques

Selon l'OMS, le taux annuel de nouveau cas de cancer dans le monde passerait de 10 millions en 2000 à 15 millions en 2020. Environ 60% de tous ces cas se produiront dans les régions les moins développées du monde. Il est la 3ème cause de décès dans le monde, et la 2ème derrière les maladies cardiovasculaires dans les pays développés. En 2000, les tumeurs malignes ont été à l'origine de 12% des quelques 56 millions de décès dans le monde, toutes causes confondues. Le cancer pulmonaire est le plus fréquent dans le monde avec 1,2 million de nouveau cas par an. On retrouve ensuite le cancer du sein avec un peu plus d'un million de cas et le cancer du côlon avec 940000 cas (World Cancer Report, OMS 2003).

II.4. Evolution du cancer

Chaque type de cancer a probablement des facteurs déclenchant, de promotion et de progression différents. Cependant, on peut décrire un schéma général. On distingue, schématiquement, trois étapes dans la genèse d'un cancer :

- *L'initiation* correspond à une lésion rapide et irréversible de l'ADN après exposition à un carcinogène (physique, chimique, viral…etc.).
- *La promotion* correspond à une exposition prolongée, répétée ou continue, à une substance qui entretient et stabilise la lésion initiée.

- *La progression* correspond à l'acquisition des propriétés de multiplication non contrôlée, l'acquisition de l'indépendance, la perte de la différenciation, l'invasion locale et métastatique.

II.5. Traitement

Le traitement des cancers comporte l'ensemble des soins médicaux destinés à combattre la maladie pour en limiter les conséquences et entrainer la guérison. On distingue les traitements anticancéreux spécifiques dirigés contre la tumeur et les cellules néoplasiques, et les traitements non spécifiques des complications du cancer et des effets secondaires liés aux traitements spécifiques. Le choix d'un traitement dépend du type de la tumeur et de son extension locale et à distance. Les traitements spécifiques sont groupés en cinq rubriques : la chimiothérapie, l'hormonothérapie, l'immunothérapie, la radiothérapie, et la chirurgie.

II.5.1. Chimiothérapie

La chimiothérapie demeure pour la plupart des cas l'arme de la dernière chance pour les cancers en stade de dissémination métastatique. Il s'agit de l'utilisation des médicaments anticancéreux interférant dans le métabolisme et la vie cellulaire et qui, de ce fait, sont cytotoxiques, permettant d'inhiber la croissance tumorale. En association avec la chirurgie ou la radiothérapie, elle augmente la probabilité de guérison. Cependant, il faut reconnaître que la chimiothérapie est en échec devant les tumeurs solides les plus fréquentes (cancer du sein, du poumon, de la prostate, tumeurs digestives et urinaires…) **[2]**.

II.5.2. Hormonothérapie

L'hormonothérapie est un procédé médical efficace pour obtenir une régression temporaire de certains cancers disséminés. Certaines hormones sont désormais utilisées comme anticancéreux. Cette approche repose sur le fait que certaines croissances tumorales sont sous contrôle hormonal. On peut ainsi bloquer ou activer une synthèse hormonale pour arrêter la croissance d'une tumeur **[3]**.

II.5.3. Immunothérapie

L'immunothérapie consiste à stimuler les réactions de l'organisme contre les cellules malignes. On utilise ainsi des anticorps monoclonaux liés à des drogues cytotoxiques spécifiquement dirigés contre un ou plusieurs antigènes tumoraux (en général des protéines situés sur la membrane des cellules, différentes des protéines par les cellules saines) dans le but d'épargner les cellules saines. La difficulté majeure réside dans le choix de l'antigène à atteindre **[4]**.

II.5.4. Radiothérapie

La radiothérapie est basée sur l'action de rayonnements « ionisants », elle est utilisée pour les traitements locaux des tumeurs. Cependant, il se pose, le problème de la toxicité de ces rayonnements sur les tissus sains environnants. Ces rayonnements, s'ils ne tuent pas directement les cellules, ils détruisent leur appareil de reproduction de sorte que ces cellules malignes ne peuvent plus être à l'origine de nouvelles cellules filles malignes. Ils engendrent les lésions de l'ADN responsables des effets biologiques des radiations ionisants. Ces lésions peuvent être directes, liées à l'interaction d'un électron d'ionisation avec la molécule d'ADN, ou le plus souvent indirectes, par l'intermédiaire des radicaux libres créent par la radiolyse de l'eau. Les lésions « double-brin », les plus graves et les plus difficiles à réparer pour la cellule, sont classiquement considérées comme les principales lésions responsables de l'effet cellulaire létal des radiations ionisantes **[5]**.

II.5.5. Chirurgie anticancéreuse

La chirurgie est classiquement utilisée pour enlever la tumeur primaire et permet la guérison d'un grand nombre de cancers précoces. Elle est aujourd'hui la méthode la plus efficace pour les petits foyers tumoraux sans métastase. L'élimination de l'intégralité des cellules cancéreuses et la prévention de leur dissémination lors de l'opération chirurgicale peuvent être difficiles **[5]**. La chirurgie est souvent l'étape initiale du traitement des cancers, elle est fréquemment associée à d'autres traitements. L'opération peut être curative, palliative, exploratrice ou encore, plus rarement, préventive.

III. Activités biologiques des isatines

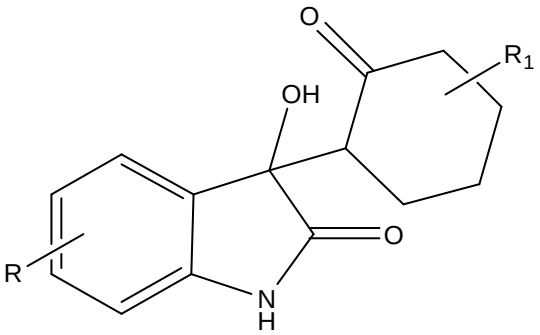
III.1. Activité sur le système nerveux central

III.1.1. Activité antidépresseur

Les dérivés de l'isatine présentent un large spectre d'activités. Ils agissent comme inhibiteurs puissants de la monoamine-oxydase (**MAO**). Ils possèdent des propriétés anticonvulsivantes et anxiolytiques. Des propriétés de l'isatine relatives à l'augmentation de la vigilance ont été également mises en évidence **[6-8]**.

III.1.2. Activité anticonvulsivante

Bhattachaya et coll. **[9]** ont montré que l'isatine exerce des effets anxiolytiques sur le système nerveux central. A des doses plus élevées des effets anticonvulsivantes significatifs contre la pentylènetétrazole (**PTZ**) ont été mises en évidence. Ils ont observé également que l'isatine intervient comme un antagoniste puissant agissant comme anticancéreuse **Figure 1**.



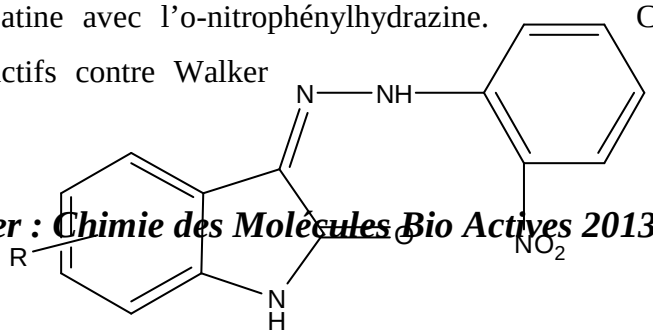
R= H, 1-CH_3, 5-Br, 5-NO_2, 4-Cl, 7-CH_3

R_1= 2-méthyl cyclohexanone, 2-cyclohexyl cyclohexanone

Figure 1

III.2. Activité anticancéreuse

Popp et pajouhesh. **[10]** ont préparé la 3-o-nitrophénylhydrazone isatine, par condensation de l'isatine avec l'o-nitrophénylhydrazine. Ces composés ont été trouvés, pour être actifs contre Walker carcinoma-256 et



inactifs contre la lignée cellulaire (leucémie) **Figure 2**.

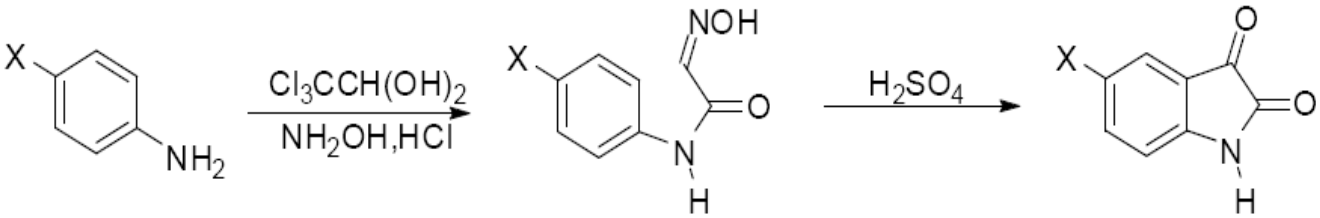
R= H, 4-CF_3, 5-Br, 5-Cl, 5-F, 5-OCH_3, 5-CH_3, 5-NO_2

Figure 2

IV. Méthodes de synthèse d'isatine

IV.1. Méthode de Sandmeyer

Elle consiste à la condensation de l'aniline avec le trichloroéthandiol et le chlorhydrate d'hydroxylamine en présence du sulfate de sodium aqueux. L'isonitroacétanilide formé est traité avec l'acide sulfurique concentré, pour donner l'isatine correspondante.



X= H, Cl, Br

Schéma 1

IV.2. Utilisation des nitroacétanilides

Kearny et coll. **[11]** ont obtenu le nitroacétanilide à partir du 1-arylamino-1-méthylsulfanylo-2-nitroéthène, qui par hydrolyse alcaline, est aisément cyclisé en isatine-3-oxime. Ce dernier traité par l'acide sulfurique concentré ou l'acide trifluorométhane sulfurique, à température ambiante, conduit à l'isatine correspondante.

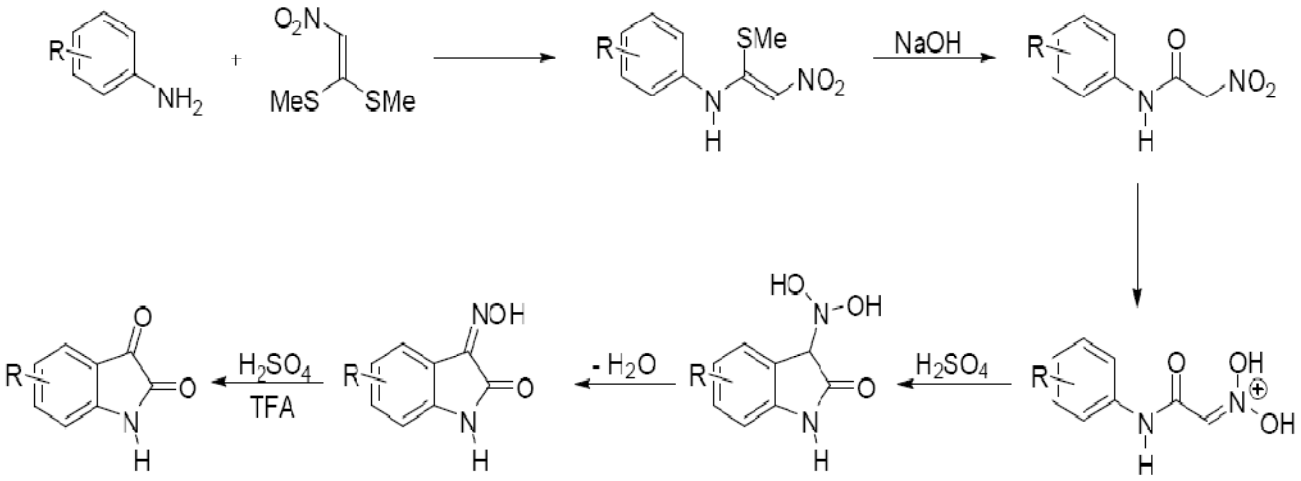


Schéma 2

V. Conclusion

Les modifications structurales de la structure de base de l'isatine, ont permis l'apparition de nouveau dérivés présentant un large spectre d'activité biologique. Ainsi, les variations structurales les plus importantes concernant les substituants en position (1), sur le carbone en position (3) et au substituant en position (5).

Les études antérieures ont montrées que la modification structurale sur les différentes positions de la molécule de base, permet d'améliorer son profil pharmacologique lui conférant des propriétés sédatives, anticonvulsivantes, anxiolytiques, antimicrobiennes, anti-VIH et anticancéreuses.

Références bibliographiques

[1] Weinberg RA 1994 Weinberg RA. Oncogenes and tumor suppressor genes. CA Cancer J Clin. 44,1994,160-70.

[2] Matirosyan AR et al.2004, Beer TM et Myrthue A 2004.

[3]Marie-Christine Meunier, Jean-Sébastien Delisle, Chantal Baron, Claude Perrault Immunothérapie anti-cancer sans dommages collatéraux 2007.

[4] Dr. Meunier ; 2002 : La place de la chirurgie dans le traitement du cancer. Cancérologie-Centre Eugène Marquis, CHU de Rennes, 2 rue Henri le Guilloux, 35033 Rennes Cedex.

[5]http://www.espacecancer.chuv.ch/ecc_home/ecc-maladie-traitement/ecc-traitement-chirurgie.html.

[6] V. Glover, J.M.Halket, P.J.Watkins, A. Clow, B.L.Godwin and M.J.Sandler Neurochem., 51, 1988, 656.

[7] J. Seidel and J.Wenzel, Pol. J. Pharmacol., 35, 1979, 407.

[8] I.M. Mc Intyre and T.R Norman, J. Neural Transm.79, 1990, 35.

[9] S.K. Bhattacharya, A. Chakraborti, Indian. J. Exp. Biol., 36, 1998, 118.

[10] F.D. Poop, H. Pajouhesh, J. Pharm. Sci., 72, 1983, 318.

[11] T.Kearney, P.A. Harris, A.Jackson, J.A. Joule, Synthesis, 1992, 769.

Chapitre II

Relation quantitative structure-activité (QSAR): Présentation des descripteurs et des méthodes statistiques d'analyses

I. Introduction

Une relation quantitative structure-activité **QSAR**, est le procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique **[1]**.

Ainsi par exemple l'activité biologique peut être exprimée de manière quantitative, comme pour la concentration de substance nécessaire pour obtenir une certaine réponse biologique. De plus, lorsque les propriétés ou structures physicochimiques sont exprimées par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure-activité, entre les deux. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de la réponse biologique pour des structures similaires.

La QSAR la plus commune est de la forme :

Activité = f (propriétés physico-chimiques et/ou structurales) **[2]**.

La relation mathématique entre un ou plusieurs paramètres physico-chimiques et l'activité biologique d'un composé peut être exprimée à travers l'équation de Hansch **[3]**.

$$\log\left(\frac{1}{C}\right) = a\,\pi + b\pi^2 + cE + dS + cte$$

Avec :

- C : représente la concentration.
- π : représente l'hydrophobicité.
- E : représente les propriétés électroniques.
- S : représente les effets stériques.
- a, b, c et d : représentes les constantes.

II. Descripteurs moléculaires

Les données utilisées pour former l'équation de QSAR sont représentées par un ensemble de matrice de chiffres dans lesquelles chaque ligne représente un composé et chaque colonne des propriétés physicochimiques (descripteurs).

Un descripteur moléculaire peut être considéré comme la conséquence d'un processus logique et mathématique, appliqué à l'information chimique codifiée à travers la représentation d'une molécule **[4]**. L'information codée d'un descripteur moléculaire dépend du type de représentation moléculaire employée et de l'algorithme défini pour son calcul. On distingue trois types de descripteurs :

- Des descripteurs moléculaires simples dérivés du nombre d'atome-type ou de fragments structuraux de la molécule ou descripteurs 1D.
- Des algorithmes appliqués à une représentation topologique (graphique moléculaire) et habituellement appelés topologique ou descripteurs (2D).
- Des descripteurs moléculaires dérivés d'une représentation géométrique, qui s'appellent descripteurs géométriques ou descripteurs (3D).

II.1. Descripteurs 1D

Les descripteurs 1D sont accessibles à partir de la formule brute de la molécule et ils dérivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est à dire les atomes qui la constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution **[5]**.

II.1.1. Poids Moléculaire

Désigne une quantité en gramme qui représente le poids d'une mole d'un composé.

II.2. Descripteurs 2D

Les analyses QSAR classiques (analyses de Hansch et Free Wilson) considèrent uniquement les structures en deux dimensions (2D).

Dans le (2D QSAR), il existe de nombreux descripteurs. Ceux qui sont le plus souvent utilisés sont des constantes (MR, Es,…). Un nombre important de valeur de constantes est collecté et un processus d'analyse statistique les exploitera pour trouver la relation entre les données biologiques et les descripteurs moléculaires **[6]**.

II.2.1. Réfraction molaire

La réfraction molaire est la déviation d'un rayonnement à la surface d'un corps, à l'interface entre deux milieux d'indices différents dans une direction fixée par la loi de Snell-Descartes définie comme suit :

$$MR = \frac{PM}{D} \times \frac{n^2 - 1}{n^2 + 2}$$

Avec :

- PM : représente le poids moléculaire.
- D : représente la densité.
- n : représente l'indice de réfraction **[7]**.

II.2.2. Paramètres stériques

Le calcul de la constante stérique $\mathbf{E_s}$ peut s'effectuer en utilisant la relation de Hansch et Charton **[8]**.

$$E_S = -1{,}839\, r_{vdw} + 3{,}484$$

II.3. Descripteurs 3D

Les relations structure activité quantitatives à trois dimensions (3D QSAR) sont des modèles qui établissent une relation entre une activité biologique et les paramètres structuraux (descripteurs moléculaires) calculés dans l'espace tridimensionnel pour un groupe de molécules.

Dans ces méthodes, les propriétés des molécules sont décrites par différents champs. Par exemple :

- La surface moléculaire, accessible au solvant, de Connolly ou surface de contact.
- Le potentiel électrostatique (position des groupements chargés).
- La participation à des liaisons hydrogènes.
- Le potentiel de lipophile moléculaire.
- Les orbitales moléculaires.
- La forme de la molécule.

II.3.1. Surface de Van Der Waals

C'est la surface de l'enveloppe de Van der Waals.

❖ **Surface de Connolly**

C'est la surface de contact entre l'enveloppe van Der waals et une molécule sonde (en général l'eau).

II.3.2. Volume de Van Der Waals

C'est le volume occupé par l'enveloppe de Van Der Waals, ses valeurs numériques dépendent des rayons de Van der Waals atomiques et de la méthode de calcul.

II.3.3. Paramètres électroniques

Les paramètres électroniques permettant de caractériser les effets inductifs et résonnance des substituants sont principalement représentés par la **constante de Hammett**.

II.3.4. Constante de HAMMETT

L'effet électronique des substituants est dérivé de la comparaison de l'énergie libre de la dissociation des acides benzoïques substitués par rapport à la dissociation de l'acide benzoïque non substitué :

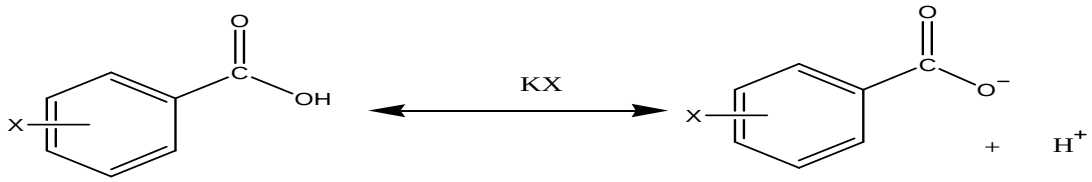


Schéma 3

$$\Delta G = \Delta H - T\Delta S = -RTln\ k$$

Donc, par définition l'effet de substituant est :

$$\Delta G_X - \Delta G_H = log\frac{K_X}{K_H} = \sigma_X$$

Or, $\sigma_H = 0$

L'effet électronique des substituants dépend de la position du substituant, il y a donc plusieurs constantes de HAMMETT pour le même substituant : σ_{para}, σ_{meta}, σ_{ortho}.

II.3.5. Paramètres de lipophile

La lipophile est mesurée par le **Log(P)** qui représente l'équilibre entre une phase polaire (aqueuse) et une phase non polaire **[9]**.

La lipophile est par conséquent une variable importante dans les équations de corrélations utilisées pour prédire l'activité biologique (QSAR).

II.3.5.1. Méthode de Hansch

On considère le partage d'un soluté entre l'eau et un solvant non miscible (hexane, éther, chloroforme…) d'après la relation suivante :

$$logP_{RX} = logP_{RH} + \Pi_X + \Pi_{CORR}$$

Avec :

- P_{RX} et P_{RH} : représentes les coefficients de partage des molécules RH et RX.
- Π_X : représente le paramètre de lipophile de substituant X.
- Π_{CORR} : représente le terme correctif tenant compte de l'effet par les ramifications, les doubles liaisons, les fermetures du cycle, les liaisons intramoléculaires.

II.3.5.2. Méthode de Rekker (méthode fragmentable)

Avec la méthode de Rekker la valeur de **logP** est calculée de la manière suivante :

$$\log P = \sum_{i=1}^{i=n} a_i f_i + f_{CORR}$$

Où a_i représente le nombre de fragments identiques dans la molécule, f_i représente la constante fragmentable hydrophobe, et f_{CORR} est un terme correctif qui décrit les caractéristiques structurales spécifiques (conjugaison Aryle-Aryle,…).

D'autres méthodes ont ainsi été introduites comme la méthode de Ghose et Viswanadhan.

II.3.5.3. Méthode de Ghose et Viswanadhan (méthode atomique)

$$\log P = \sum_{i=1}^{i=n} n_i \, a_i$$

Avec :

- n : représente le nombre d'atomes de type i.
- a_i : représente la constante atomique des atomes de type i.

III. Méthodes statistiques permettant de détecter une relation QSAR

Les principales méthodes utilisées dans le modèle **QSAR** sont :

1-La régression linéaire simple (**RL**).

2-La régression linéaire multiple (**RLM**).

3-Le réseau de neurones (**RN**).

4-La validation croisée (**VC**).

III.1. Régression linéaire simple (RL)

La régression linéaire (**RL**) est une méthode traditionnelle permettant de dériver un modèle QSAR en utilisant un seul paramètre. Elle corrèle ainsi l'activité cible à une variable explicative :

$$\hat{y} = a\,X + b$$

III.2. Régression linéaire multiple (RLM)

La régression linéaire multiple (**RLM**) est une généralisation, à *p* variables explicatives, de la régression linéaire simple.

Nous sommes toujours dans le cadre de la régression mathématique : étant donné un échantillon $(Y_i, X_{i1}, \ldots, X_{ip})$, i=1,…,n, nous cherchons à expliquer, avec le plus de précision possible, les valeurs prises par Y_i, dite variable endogène, à partir d'une série de variables explicatives $X_{i1}, \ldots, X_{ip}$. Le modèle théorique, formulé en termes de variables aléatoires, prend la forme :

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \ldots + a_p X_{ip} + \varepsilon_i,\ i=1,\ldots,n$$

Où $\boldsymbol{\varepsilon_i}$ est l'erreur du modèle qui exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des $X_{i1}, \ldots, X_{ip}$ (problème de spécifications, variables non prises en compte, etc.). a_0, a_1, …,a_p, sont les paramètres à estimer **[10]**.

Les objectifs de la régression multiple sont :

Identifier le modèle :

- Estimer les coefficients $\mathbf{a_j}$ avec j = 0,…p, à partir des observations disponibles.
- Estimer la précision de cette estimation.
- Décider de la nature de l'influence de chaque variable explicative $\mathbf{x_j}$ sur la variable à expliquer **y**.

Valider le modèle :

- Donner une mesure de la qualité globale de la régression.
- Détecter les points aberrants où hors épure.
- Prévoir pour une nouvelle observation **x** la valeur de **y**.

III.2.1. Méthode des moindres carrés

La méthode des moindres carrés ordinaire (**MCO**) est le nom technique de la régression mathématique en statistiques, et plus particulièrement de la régression linéaire. Il s'agit d'ajuster un nuage de points $\{Y_i, X_i\}_{i=1,\dots,n}$ selon une relation linéaire, prenant la forme de la relation matricielle $\boldsymbol{Y = X\beta + \varepsilon_i}$, où $\boldsymbol{\varepsilon_i}$ est un terme d'erreur **[11]**.

La régression linéaire multiple cherche à approximer une relation fonctionnelle trop complexe en général, par une fonction mathématique simple **Equation 1** :

$$Y_i = a_0 + a_1 x_{i,1} + \cdots a_p x_{i,p} + \varepsilon_i \qquad \textbf{(1)}$$

Avec :

- P : représente le nombre de variables explicatives.
- Yi : représente la variable réponse (à expliquer ou variable dépendante).
- Xp : représente la variable régression (explicative ou variable indépendante).
- i : représente l'indice de l'observation courante, i=1,….. , n.
- n : représente le nombre d'observation.

On fait d'abord une estimation des paramètres **Equation 2** :

$$\widehat{y_i} = \widehat{a_0} + \widehat{a_1} X_{i,1} + \cdots \widehat{a_p} X_{i,p} \qquad \textbf{(2)}$$

Afin de mesurer la qualité de l'ajustement, nous nous intéressons à mesurer l'erreur $\boldsymbol{\varepsilon_i}$:

$$\varepsilon_i = Y_i - \widehat{y_i}$$

Calculons la somme des carrés des erreurs individuelles afin de ne pas devoir nous occuper du signe des erreurs :

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

En divisant les deux côtés de l'égalité par **n**, on obtient la **variance de l'erreur** :

$$\sigma_e^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

Cette dernière expression représente la variance de l'erreur.

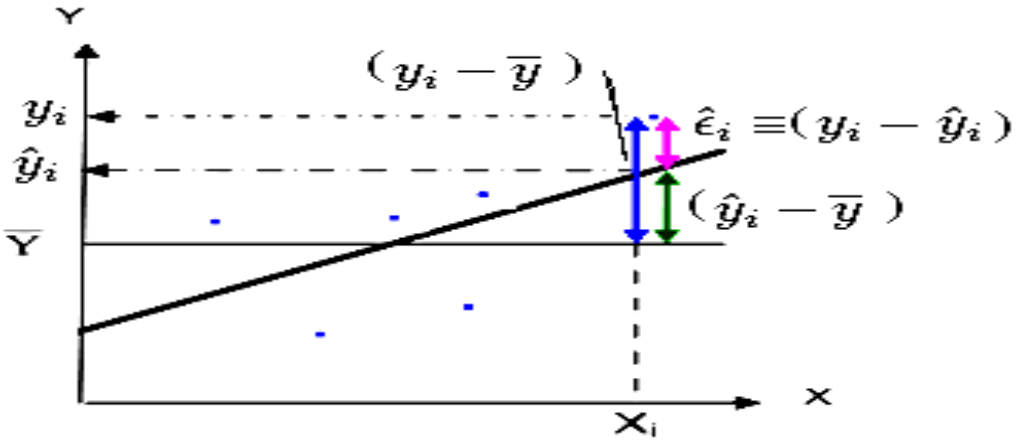


Figure 3 : Décomposition des différents écarts

- $y_i - \bar{y}$: représente l'écart total entre le point $\mathbf{P_i}$ et la moyenne $\overline{\boldsymbol{y}}$.
- $y_i - \widehat{y_i}$: représente l'écart entre le point $\mathbf{P_i}$ et l'estimation par la droite de régression, c'est $\boldsymbol{\varepsilon_i}$.
- $\widehat{y_i} - \bar{y}$: représente l'écart entre le point sur la droite de régression et la moyenne $\overline{\boldsymbol{y}}$.

La méthode des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des résidus.

III.2.2. Analyse de variance

L'analyse de la variance (ou test **ANOVA** de l'anglais ANalysis Of VAriance) est un test statistique qui permet de comparer globalement l'espérance mathématique de plusieurs

échantillons. Le nom de ce test s'explique par sa façon de procéder : on décompose la variance totale de l'échantillon en deux variances partielles, la variance interclasses ou **variation due au facteur** (**SCF**), et la variance résiduelle ou **variation due à l'erreur d'échantillonnage** (**SCE**), et on compare ces deux variances.

Ainsi : **SCT = SCF + SCE**

Ces variations se mesurent à l'aide de sommes de carrés (**SC**), **Tableau 1** et **Tableau 2**.

Tableau 1 : Sommes de carrés (**SC**)

Variation totale autour de la moyenne globale	$SCT = \sum_i (y_i - \bar{y})^2$
Variation due au facteur	$SCF = \sum_i (\hat{y}_i - \bar{y})^2$
Variation due à l'erreur	$SCE = \sum_i (y_i - \hat{y}_i)^2$

Tableau 2 : Analyse de variance

Source de variation	**Sommes des carrés**	**Degrés de liberté**	**Carrés moyens**
Facteur	SCF	p	CMF = SCF / p
Erreur	SCE	n- p- 1	CME = SCE / (n-p-1)
Totale	SCT	n – 1	----

III.2.3. Critères de validation du modèle RLM

La validation du modèle RLM est basée sur la vérification de deux hypothèses **H_0** et **H_1**. La formulation du test d'hypothèses qui permet d'évaluer globalement le modèle est la suivante :

- H_0 : $\mathbf{a_1}=\mathbf{a_2}=\ldots\ldots\ldots\ldots=\mathbf{a_p}=\mathbf{0}$, appelée hypothèse nulle.
- H_1 : un des coefficients au moins est non nul, appelée hypothèse non nulle.

III.2.3.1. Test de Student

La loi de Student à (**n-p-1**) degré de liberté ($\mathbf{t_{calc}}$) s'écrit :

$$\mathbf{t_{calc}}=\left(\frac{\mathbf{r}}{\sqrt{\frac{\mathbf{1-r^2}}{\mathbf{n-p-1}}}}\right)$$

On rejette **H_0** (l'hypothèse nulle) lorsque : $t_{calc} > t(1-\frac{\alpha}{2}), (n-p-1)$

D'où $t(1-\frac{\alpha}{2}), (n-p-1)$ est la valeur de la loi de Student à (n-p-1) degré de liberté, à une probabilité $(1-\frac{\alpha}{2})$.

III.2.3.2. Test de Fisher

La statistique « **F** » d'analyse de variance est utilisée pour le test sur l'égalité des moyennes, On l'appelle le **F** de Fisher.

III.2.3.2.1. Hypothèses du test

On pose l'hypothèse suivante, tel que :

- H_0 : $\mathbf{\mu_1} = \mathbf{\mu_2} = \ldots = \mathbf{\mu_K}$, appelée hypothèse nulle.
- H_1 : au moins deux moyennes sont différentes, appelée hypothèse non nulle.

Avec : $\mu_1, \mu_2, \ldots, \mu_K$ représentes les moyennes.

III.2.3.2.2. Conditions d'utilisation du test de Fisher.

Statistique du test : $$Fexp = \frac{CMF}{CME} = \frac{SCF/p}{SCE/n-p-1}$$

Si H_0 est vraie, on s'attend à ce que la variation due au facteur (c'est à dire entre les différents groupes) soit faible.

Par contre si les moyennes μ_1, μ_2,…, μ_K ne sont pas toutes égales, alors les différences entre ces moyennes font augmenter la valeur de SCF et donc celle de CMF. Une valeur très élevée de la statistique F suggère donc que les moyennes μ_1, μ_2,…, μ_K ne sont pas toutes égales.

Si H_0 est vraie, la statistique F obéit à une loi de **Fisher** à (p) et (n-p-1) degrés de liberté.

III.2.3.3. Coefficient de corrélation : r

Ce coefficient détermine la variance de l'activité cible qui est expliquée par le modèle de QSAR, c'est à dire par la régression de l'activité cible en fonction de l'activité initiale. Ce coefficient n'est pas affecté par l'unité de mesure choisie, d'après la relation suivante :

$$r = \sqrt{1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \sum_i \frac{\hat{y}_i}{n})^2}}$$

Il traduit :

-Une bonne corrélation entre l'activité cible et l'activité initiale si **r** est plus proche de 1.

-Une corrélation non linéaire entre l'activité cible et l'activité initiale si **r** est proche de 0.

III.2.3.4. Coefficient de détermination : r^2

Ce coefficient de détermination $\mathbf{r^2}$ donne le taux d'explication ou pourcentage de la variation de Y expliquée par la variation de X :

$$r^2 = \frac{varition\ expliquée}{variation\ totale} = \frac{SCF}{SCT}$$

Si r^2 = **0.80** ce chiffre signifie que 80% de la variable **Y** est attribuable à la variation de la variable **X**.

r^2 Proche de (**1**) n'est pas une qualité de l'ajustement.

III.3. Réseau de neurones (RN)

Un réseau neuronal (**RN**) s'inspire du fonctionnement des neurones biologiques et prend corps dans un ordinateur sous forme d'un algorithme. Le réseau neuronal peut se modifier lui-même en fonction des résultats de ses actions, ce qui permet l'apprentissage et la résolution de problèmes sans algorithme, donc sans programmation classique **[12]**.

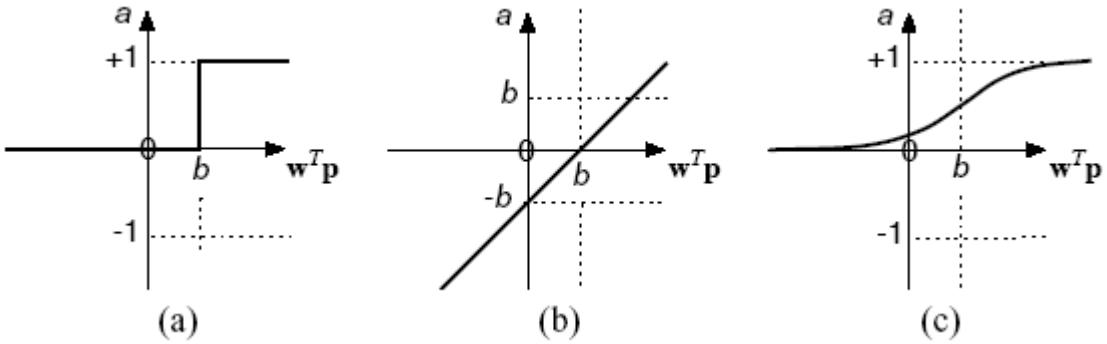
III.3.1. Fonctions de transfert

Jusqu'à présent, nous n'avons pas spécifié la nature de la fonction d'activation **a=f(n)** de notre modèle. Il se trouve que plusieurs possibilités existent et celles-ci sont quasiment empiriques et à adapter en fonction des situations. Les plus courantes et les plus citées dans la littérature **[13]** sont énumérées dans le **Tableau 3** ci-dessous.

Tableau 3 : Fonctions de transfert

Nom de la fonction	Relation d'entrée/sortie	Icône
seuil	$a = 0 \quad \text{si } n < 0$ $a = 1 \quad \text{si } n \geq 0$	
seuil symétrique	$a = -1 \quad \text{si } n < 0$ $a = 1 \quad \text{si } n \geq 0$	
linéaire	$a = n$	
linéaire saturée	$a = 0 \quad \text{si } n < 0$ $a = n \quad \text{si } 0 \leq n \leq 1$ $a = 1 \quad \text{si } n > 1$	
linéaire saturée symétrique	$a = -1 \quad \text{si } n < -1$ $a = n \quad \text{si } -1 \leq n \leq 1$ $a = 1 \quad \text{si } n > 1$	
linéaire positive	$a = 0 \quad \text{si } n < 0$ $a = n \quad \text{si } n \geq 0$	
sigmoïde	$a = \frac{1}{1+\exp^{-n}}$	
tangente hyperbolique	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$	
compétitive	$a = 1 \quad \text{si } n \text{ maximum}$ $a = 0 \quad \text{autrement}$	C

Les trois les plus utilisées dans le domaine de l'ingénierie sont les fonctions "**seuil**" (**a**), "**linéaire**" (**b**) et "**sigmoïde**" (**c**) comme représentées ci-dessous :



Comme son nom l'indique, la fonction seuil applique un seuil sur son entrée.

Plus précisément, une entrée négative ne passe pas le seuil, la fonction retourne la valeur 0 (faux), alors qu'une entrée positive ou nulle dépasse le seuil, et la fonction retourne 1 (vrai). Il

est évident que ce genre de fonction permet de prendre des décisions binaires.

La fonction linéaire est quant à elle très simple, elle affecte directement son entrée à sa sortie selon la relation ***a=f(n)=n***. Il est évident que la sortie du neurone correspond alors à son niveau d'activation dont le passage à zéro (l'ordonnée à l'origine) se produit lorsque $\boldsymbol{w}.\overrightarrow{\boldsymbol{p}} = \boldsymbol{b}$.

La fonction de transfert sigmoïde est définie par la relation mathématique :

$$a = \frac{1}{1 + e^{-n}}$$

Elle ressemble soit à la fonction seuil, soit à la fonction linéaire, selon que nous somme loin ou près de **b** respectivement. La fonction seuil est très non linéaire car il y a une discontinuité lorsque $\mathbf{w}.\overrightarrow{\mathbf{p}} = \mathbf{b}$. De son côté, la fonction linéaire est tout à fait linéaire. Elle ne comporte aucun changement de pente. La sigmoïde est un compromis intéressant entre les deux précédentes. Notons finalement que la fonction tangente hyperbolique est une version symétrique de la sigmoïde **[14]**.

III.3.2. Architecture d'un RN

Un réseau de neurones (**RN**) est composé d'unités de calculs disposées en couches et reliées entre elles pour échanger de l'information**.** Trois types de couches constituant le RN : la couche d'entrée, la couche cachée et la couche de sortie. L'information circule des neurones d'entrée vers les neurones de sortie sans retour en arrière possible via des fonctions de transfert **[15]**.

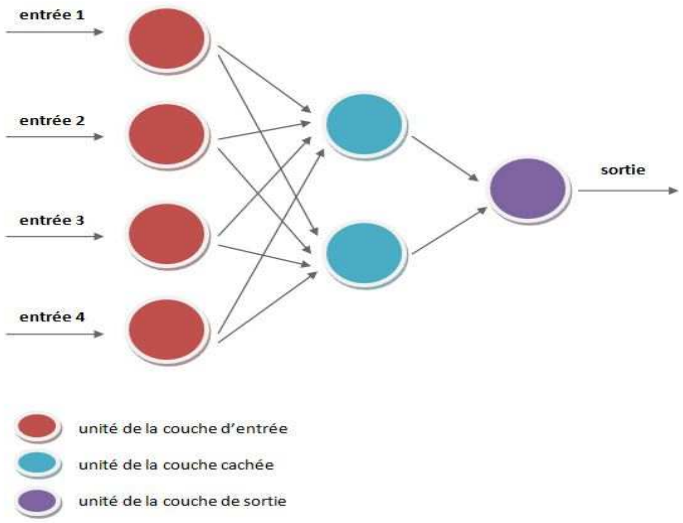


Figure 4 : Architecture générale d'un réseau de neurones à trois couches

- **Les nœuds d'entrée ou couche d'entrée**

La première couche est appelée couche d'entrée. Elle recevra les données source que l'on veut utiliser pour l'analyse. Dans le cas de l'aide au diagnostic médical. Cette couche recevra les symptômes. Sa taille est donc directement déterminée par le nombre de variables d'entrées.

- **Les nœuds cachés ou couche cachée**

La seconde couche est une couche cachée, en ce sens qu'elle n'a qu'une utilité intrinsèque pour le réseau de neurones et n'a pas de contact direct avec l'extérieur. Les fonctions d'activation sont en général non linéaires sur cette couche mais il n'y a pas de règle à respecter. Le choix de sa taille n'est pas implicite et doit être ajusté. En général, on peut commencer par une taille moyenne des couches d'entrée et de sortie mais ce n'est pas toujours le meilleur choix. Il sera souvent préférable pour obtenir de bon résultats, d'essayer le plus de tailles possible.

Le nombre de nœuds dans la couche cachée est un facteur important déterminant la performance du réseau. Il a été trouvé que beaucoup de nœuds causent une mémorisation des données de la série par le réseau (overfitting). Cependant, les réseaux avec peu de nœuds pourraient être insuffisants pour utiliser toutes les informations des données de la série (underfitting) et la généralisation est pauvre. Des études antérieures ont conduit à la

détermination du nombre approprié des unités cachées suggérant que ρ, le rapport du nombre des individus au nombre des poids dans le réseau de neurones, devrait avoir une valeur entre **1,8 et 2,3 [16]**.

❖ **Les nœuds de sortie ou couche de sortie**

La troisième couche est appelée couche de sortie. Elle donne le résultat obtenu Après compilation par le réseau des données entrées dans la première couche. Dans le cas de l'aide au diagnostic médical, cette couche donne le diagnostic. Sa taille est directement déterminée par le nombre de variables qu'on veut en sortie.

III.4. Méthodes de validation

III.4.1. Validation croisée (VC)

La validation croisée (**VC**). **[17]** (« cross-validation ») est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. En fait, il y a au moins trois techniques de validation croisée : « testset validation » ou « holdout method », « k-fold cross-validation » et « leave-one-out cross-validation » (**LOOCV**).

- La première méthode est très simple, il suffit de diviser l'échantillon de taille **n** en échantillon d'apprentissage (>60% de l'échantillon) et échantillon de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'erreur est estimée en calculant l'erreur quadratique moyenne.
- Dans la seconde, on divise **k** fois l'échantillon en échantillon d'apprentissage/échantillon de test, et on applique la première méthode aux k couples d'échantillons. La moyenne des erreurs est calculée.
- Dans la troisième, $\mathbf{k = n^2}$.

  On peut utiliser la technique de la validation croisée qui consiste en un processus contrôlé par le nombre de composés dans la table **QSAR**.

Principe de VC

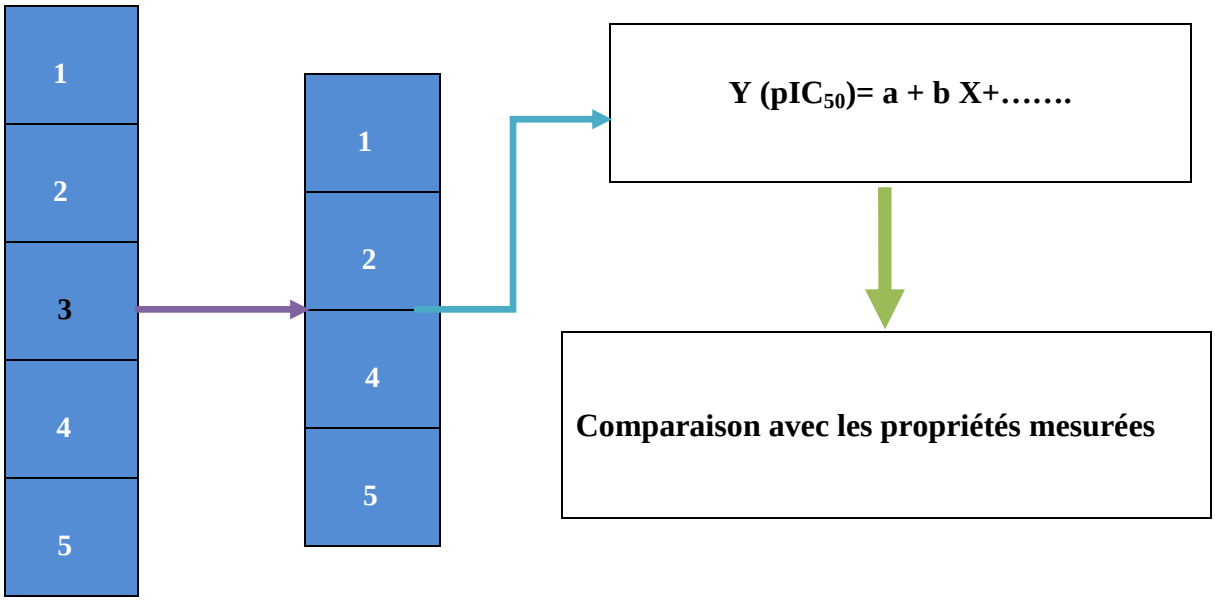


Figure 5 : Principe et procédure de la validation croisée

Références bibliographiques

[1] G. A. Patani, E. J. LaVoie, Bioisosterism: A Rational Approach in Drug Design. Chem. Rev., 1996, 96, 3147-3176.

[2] Danail Bonchev, D.H. Rouvray: Chemical Graph Theory: Introduction and Fundamentals. Gordon and Breach Science Publishers, 1990.

[3] G. Thomas, "Fundamentals of Medicinal Chemistry". "The SAR and QSAR approach to drug design". 2003: John Wiley & Sons, Ltd.

[4] H. Gonzalez-Diaz, E. Olazabal, L. Santana, E. Uriarte, Y. Gonzalez-Diaz, N. Castanedo, "QSAR study of anticoccidial activity for diverse chemical compounds: Prediction and experimental assay of trans-2-(2-nitrovinyl)furan." Bioorg. Med. Chem., 2007. 15: 962-968.

[5] R. B. Silverman, "The Organic Chemistry of Drug Design and Drug Action." 2nd éd. 2004, USA: Elsevier

[6] www.viadeo.com/fr/profile/francois.legoff.Francois le GOFF-expert ingénieur. Finland

[7] www.viadeo.com/fr/profile/francois.legoff.Francois le GOFF-expert ingénieur. Finland

[8] Pro-chemist .online .fr/cours/lipoO.htm

[9] Et-serveur.univ-lyon1.fr/immédiato/…./ Statistiques _index.htm Henri IMMEDIATO. Cours Statistique, licence scientifique (2008), p 3.

[10] Régis Bourbonnais, *Économétrie*, Dunod, 1998.

[11] H. Waterbeemd, S. Rose, "Quantitative approaches to structure- activity relationships", in Book "Quantitative approaches to structure-activity relationships". 2003, Elsevier. 351-367.

[12] Artificial neural networks – a review of applications in the atmospheric sciences – MW GARDNER ET SR DORLING

[13] Modelling hourly diffuse solar-radiation in the city of Sao Paulo using a neural technique - Jacyra Soares, Amauri P. Oliveira, Marija Zlata Bo_znar, Primo Mlakar, Joao F. Escobedo, Antonio J. Machado

[14] http://www.sciences.ch/htmlfr/infotheorique/infomethnum02.php

[15] G. Hinton Apprentissage et réseaux de neurons .pour la science, 1992, 181, 124,132;

[16] Payam Refaeilzadeh, Lei Tang, Huan Liu,Cross-Validation vol 2009, p. 532-538

[17] So S.-S. ; W. G.Richards, j. Med.,vol 1995, 35, p 3201.

Chapitre III

Etude QSAR d'une série de dérivés de l'isatine par les méthodes statistiques

I. Introduction

L'isatine est un composé endogène identifié chez l'homme comme dérivé métabolique de l'adrénaline. Elle a été trouvée dans le tissu de certains mammifères, comme modulateur de processus biochimique, et dans les plantes de genre Isatis **[1-3]**. L'isatine et ses dérivés ont montré des propriétés anxiolytiques, sédatives et anticonvulsivantes. Ils se sont avérés de bons antagonistes agissant comme anticancéreuses **[4-5]**.

Dans cette partie de ce travail nous allons établir une relation linéaire structure-activité quantitative entre l'activité anticancéreuse d'une série de molécules bioactives dérivées de l'isatine et leurs descripteurs structuraux, par conséquent nous proposons un modèle quantitatif, et nous essayons d'interpréter l'activité de ces molécules en se basant sur les différentes méthodes statistiques multi-variées suivantes :

- Régression linéaire multiple (**RLM**) : qui permet à la sélection des descripteurs utilisés comme paramètres de la couche d'entrée dans le réseau de neurones.
- Réseau de neurones (**RN**) : qui est une méthode non linéaire et qui permet la prédiction des activités.
- Validation croisée (**VC**) : utilisée pour tester la performance du modèle.

II. Présentation de la série de molécule et des descripteurs

II.1. Série de molécule

Nous avons étudié et analysé la série de la molécule d'isatine constituée de (**47**) dérivés **Figure 6**. Ceci dans le but de trouver une relation quantitative entre l'activité anticancéreuse et la structure de ces molécules qui sont décrites par leurs substituants R_1, R_2, R_3, R_4, R_5, R_6 **[6]**.

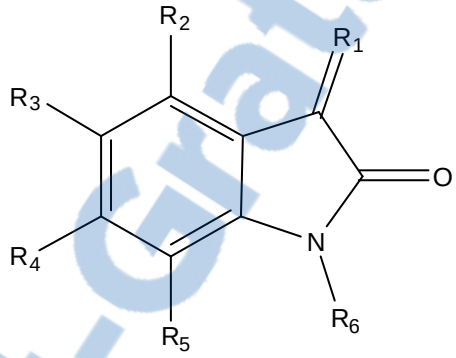


Figure 6 : Structure géométrique de l'isatine

II.2. Descripteurs

Dans ce travail, (**7**) descripteurs ont été choisis pour décrire la structure des molécules constituant la série à étudier**:**

1. Le poids moléculaire (**PM**)
2. L'électronégativité (**X**)
3. La réfraction molaire (**RM**)
4. Les accepteurs des liaisons hydrogènes (**ALH**)
5. Les donneurs des liaisons hydrogènes (**DLH**)
6. La lipophile (**LogP**)

7. L'énergie de répulsion (**NRE**)

III. Méthodes et matériels

Les descripteurs sont calculés pour les molécules entières grâce au logiciel «ChemBioDraw 13.0». Ce logiciel comporte deux fenêtres :

- **ChemBioDraw Ultra 13.0** : permet de dessiner les molécules.
- **ChemBio3D Ultra 13.0** : permet de minimiser l'énergie de la molécule, de sélectionner les descripteurs et ensuite de les calculer.

Pour réaliser cette étude nous disposons des logiciels suivants :

- **CHEMBIODRAW 13.0** : logiciel permettant de calculer les descripteurs.
- **SYSTAT 13** : permettant d'effectuer la régression linéaire multiple (**RLM**).
- **MATLAB** : version (**7**) nous a permis de réaliser le réseau de neurones (**RN**).
- La validation croisée (**VC**) a été effectuée en utilisant le réseau de neurones disponible dans MATLAB.

Le **Tableau 4** regroupe les formules chimiques et les valeurs de l'activité anticancéreuse observée (**$pIC_{50\ Obs}$**), et calculée par la **RLM** (**$pIC_{50\ RLM}$**) et par le **RN** (**$pIC_{50\ RN}$**) ainsi que celle prédite par la **VC** (**$pIC_{50\ VC}$**). On note que l'activité a été utilisée sous forme logarithmiques (**pIC_{50}**).

On note sur le **Tableau 4** par **N** le nombre de molécules et par R_1, R_2, R_3, R_4, R_5, R_6, les substituants.

Tableau 4 : Formules chimiques des (**47**) molécules et activités observées et calculées

N	R_1	R_2	R_3	R_4	R_5	R_6	$pIC_{50\ Obs}$	$pIC_{50\ RLM}$	$pIC_{50\ RN}$	$pIC_{50\ VC}$
1	O	H	Br	H	Br	$H_2CCH=CH_2$	5,18	5,27	5,20	5,26
2	O	H	Br	H	Br	$H_2CCH_2OCH_3$	5,46	5,07	5,42	5,38
3	O	H	Br	H	Br	$H_2CCH_2CH(CH_3)_2$	5,62	5,52	5,62	5,83
4	O	H	Br	H	Br	$H_2CC_6H_5$	5,94	5,63	5,76	5,87
5	O	H	Br	H	Br	$H_2CC_6H_4CH_3$	6,31	5,84	5,93	6,07
6	O	H	Br	H	Br	$H_2CC_6H_4OCH_3$	5,74	5,61	5,72	5,68
7	O	H	Br	H	Br	$H_2CC_6H_4OCH_3$	5,75	5,50	5,61	5,66
8	O	H	Br	H	Br	$H_2CC_6H_4NO_2$	6,05	5,85	5,98	6,03
9	O	H	Br	H	Br	$H_2CC_6H_4NO_2$	5,64	5,63	5,55	5,86
10	O	H	Br	H	Br	$H_2CC_6H_4Cl$	6,01	5,86	5,90	6,04
11	O	H	Br	H	Br	$H_2CC_6H_4Br$	6,20	6,07	5,94	5,88
12	O	H	Br	H	Br	$H_2CC_6H_4I$	5,64	6,33	6,01	5,29
13	O	H	Br	H	Br	$H_2CC_6H_4CF_3$	6,10	5,94	5,94	6,06
14	O	H	H	Br	H	$H_2CC_6H_4CF_3$	5,28	5,56	5,79	5,09
15	O	H	Br	H	Br	$H_2CC_6H_4COOCH_3$	5,92	5,45	5,52	5,93
16	O	H	Br	H	Br	$H_2CC_6H_4C(CH_3)_3$	5,95	6,16	6,09	5,98
17	O	H	Br	H	Br	$H_2CCH=CHC_6H_5$	5,63	5,63	5,79	5,28
18	O	H	Br	H	Br	$H_2CC_6H_4C_6H_5$	6,12	6,44	6,14	6,08
19	O	H	H	H	H	H	3,25	3,45	3,07	3,85
20	O	Br	H	H	H	H	3,67	4,00	3,89	3,93
21	O	H	Br	H	H	H	4,19	4,09	4,25	3,92
22	O	H	H	Br	H	H	4,13	4,00	3,93	4,15
23	O	H	H	H	Br	H	4,08	4,04	4,11	3,94
24	O	H	F	H	H	H	4,01	3,72	3,91	4,07
25	O	H	I	H	H	H	4,27	4,33	4,45	3,83
26	O	H	NO_2	H	H	H	3,88	3,67	3,73	3,69
27	O	H	OCH_3	H	H	H	3,38	3,59	3,50	3,57
28	O	H	Br	H	Br	H	4,98	4,54	4,72	4,34
29	O	H	Br	Br	H	H	4,94	4,53	4,71	4,29
30	O	H	I	H	I	H	5,11	5,06	5,30	5,05
31	O	H	Br	H	NO_2	H	3,59	4,20	3,90	3,08
32	O	H	Br	Br	Br	H	5,17	5,00	5,28	5,15
33	N-C_6H_5	H	H	H	H	H	4,12	4,13	3,87	4,08
34	N-C_6H_5	H	Br	H	Br	H	4,86	5,08	5,08	4,35
35	O	H	H	H	H	CH_3	3,62	4,29	3,88	3,97
36	O	H	Br	H	Br	$H_2CCH_2C_6H_5$	6,11	5,75	5,86	6,04
37	O	H	Br	H	Br	$H_2CCH_2C_6H_4Br$	6,11	6,13	5,98	6,03
38	O	H	Br	H	Br	$H_2CCH_2C_6H_4Br$	6,06	6,18	6,00	5,88

39	O	H	Br	H	Br	$H_2CCH_2C_6H_4OCH_3$	5,97	5,65	5,77	6,07
40	O	H	Br	H	Br	$H_2CCH_2C_6H_4OCH_3$	5,63	5,71	5,82	5,83
41	O	H	Br	H	Br	$CH_2C_{10}H_7$	6,72	6,41	6,54	5,96
42	O	H	Br	H	Br	$CH_2C_{10}H_7$	6,13	6,39	6,36	5,95
43	O	H	Br	H	Br	$CH_2COC_6H_5$	5,00	5,28	5,23	5,06
44	O	H	Br	H	H	$CH_2COC_6H_4Br$	5,20	5,71	5,58	5,79
45	O	H	Br	H	Br	$CH_2COC_6H_4Br$	5,04	5,69	5,57	5,28
46	O	H	Br	H	Br	$CH_2COC_6H_4OCH_3$	5,33	5,19	5,05	5,15
47	O	H	Br	H	Br	$CH_2COC_6H_4OCH_3$	5,27	5,19	5,06	5,17

IV. Résultats et Discussion

L'exploitation des données expérimentales observées par l'utilisation des outils mathématiques et statistiques est une méthode efficace pour trouver de nouveaux composés chimiques ayant une activité anticancéreuse élevée.

IV.1. Régression linéaire multiple (RLM)

L'objectif général de la régression linéaire est de proposer un modèle mathématique exprimant l'activité en fonction des descripteurs. Cette méthode se base sur la valeur critique (**p-value**), et les valeurs de test de student (**t**) pour sélectionner les meilleurs descripteurs, (**5**) descripteurs sont alors sélectionnés par la RLM pour un seuil de signification (p-value) inférieur a (**0,05**).

Le **Tableau 5** rassemble les descripteurs sélectionnés, leurs coefficients, leurs erreurs standard, t-value et p-value.

Tableau 5 : Descripteurs sélectionnés par la RLM

	Coefficient	Erreur Standard	t-Value	p-Value
CONSTANT	10,035	2,526	3,972	0,000
MP	0,003	0,001	2,434	0,020
X	-1,535	0,650	-2,360	0,023
MR	-0,013	0,004	-3,188	0,003
DLH	-0,497	0,185	-2,690	0,010
LogP	0,370	0,111	3,322	0,002

Donc l'équation de la régression linéaire obtenue est la suivante :

$$\mathbf{pIC_{50\ RLM}} = 10{,}035 + 0{,}003\ \mathbf{MP} - 1{,}535\ \mathbf{X} - 0{,}013\ \mathbf{MR} - 0{,}497\ \mathbf{DLH} + 0{,}370\ \mathbf{LogP}$$

Avec : **N = 47** **r = 0.94** et **r^2 = 0.88**

On constate d'après l'équation de régression que l'effet de l'électronégativité (**X**) et l'effet donneur des liaisons hydrogènes (**DLH**) et la lipophile (**LogP**) ont une grande influence sur la concentration d'inhibition et une contribution très importante par rapport aux autres descripteurs qui ont des représentations très faible à l'explication du modèle.

Une bonne corrélation à été montrée entre l'activité observé et celle obtenue par la **RLM** tels que ***r*** = **0.94** et ***r^2*** = **0.88**.

La **Figure 7** ci-dessous montre la corrélation entre l'activité observée et l'activité obtenue par la **RLM.**

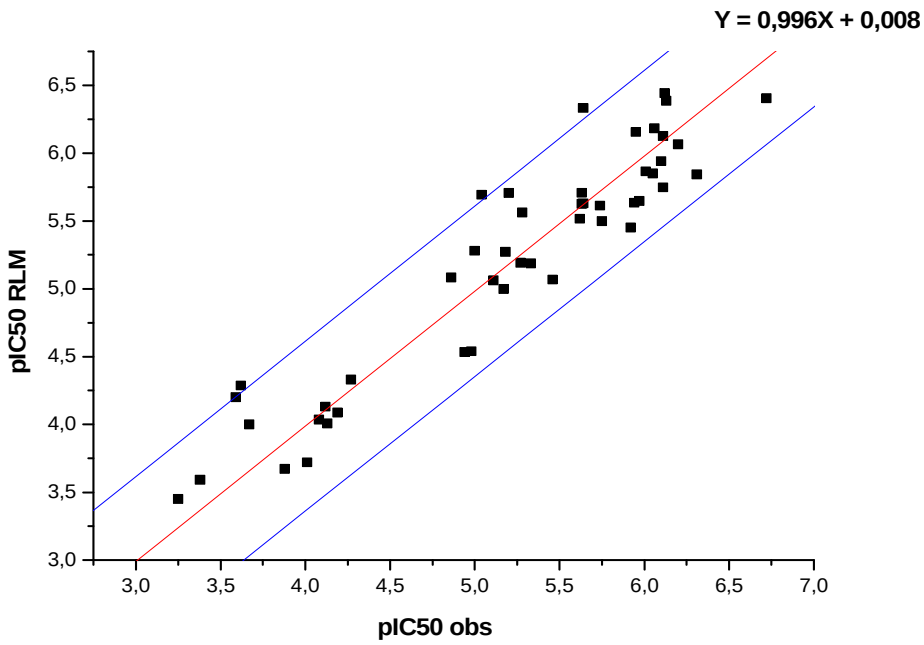


Figure 7 : Corrélation entre les **$pIC_{50\ Obs}$** et les **$pIC_{50\ RLM}$**

Pour avoir une idée sur la corrélation entre l'activité observée et celle obtenue à partir de la **RLM**, on se base sur les valeurs du coefficient directeur de la droite **a=0.996** et son ordonnée à l'origine **b=0.008**. Plus la valeur de **a** est proche de (**1**) et **b** est proche de (**0**), plus la corrélation entre l'activité observée et celle obtenue par la RLM est meilleure.

IV.1.1. Evaluation globale de la régression

Le **Tableau 6** regroupe les variances, les degrés de liberté, les sommes des carrés, la valeur de **F** de Fisher et la valeur de p-value globale du modèle.

Tableau 6 : Analyse de variance

Source	SC	Ddl	Variance	F-exp	p-Value
Régression	**33.730**	**7**	**4.819**	**42.588**	**0.000**
Résiduel	**4.413**	**39**	**0.113**	**-**	**-**
Totale	**38.143**	**46**	**4.932**	**-**	**-**

- La variabilité non expliquée par le modèle est la somme des carrés résiduels du modèle **SCE= 4.413** avec un dégré de liberté égal à **39** (**47-7-1**).
- La variabilité expliquée par le modèle est la somme des carrés de la régression **SCF= 33.730** avec un degré de liberté égal à **7** (**47-39-1**).
- Les résultats paraissent excellents et le modèle est significatif parce que nous avons obtenu des bons résultats pour $\mathbf{F_{exp}}$ de Fisher (**42.588**) et p-value globale inferieure au seuil **α = 0.05.**

IV.1.2. Test de significativité

Le premier test qui vient à l'esprit est la significativité de la corrélation c.à.d. le coefficient de corrélation **r** est-il significativement différent de (**0**) ?

Le test s'écrit : $\begin{cases} \mathbf{H_0} : r = 0 \\ \mathbf{H_1} : r \neq 0 \end{cases}$

Si le coefficient de corrélation est différent de zéro, on rejette l'hypothèse $\mathbf{H_0}$ (l'hypothèse nulle) et on accepte $\mathbf{H_1}$. Donc le modèle est significatif.

IV.1.2.1. L'intervalle de confiance (IC)

L'intervalle de confiance (**IC**) à **1-α** est un intervalle de valeurs qui a **1-α** de chance de contenir la vraie valeur du paramètre estimé.

Si la valeur de p-value dépasse (**0.05**), on rejette H_1 et on accepte H_0. Donc le modèle n'est pas significatif.

1. Si α > p-value, rejet de H_0.
2. Si α < p-value, acceptation de H_0.

IV.1.2.2. Test de student

La loi de Student à (**n-p-1**) degré de liberté ($\mathbf{t_{calc}}$) s'écrit :

$$t_{calc} = \left(\frac{r}{\sqrt{\frac{1-r^2}{n-p-1}}} \right)$$

On rejette H_0 (l'hypothèse nulle) lorsque : $\mathbf{t_{calc}} > \mathbf{t}(1-\frac{\alpha}{2}),(\mathbf{n\text{-}p\text{-}1})$

D'où $t(1-\frac{\alpha}{2})$,(n-p-1) est la valeur de la loi de Student à (n-p-1) degré de liberté, à une probabilité $(1-\frac{\alpha}{2})$.

Dans notre cas on a **n = 47** et **r = 0.94**. Ceci correspond a $\mathbf{t_{calc}}$ **= 18.8**, on rejette H_0 (l'hypothèse nulle) lorsque : $t_{calc} > t(1-\frac{\alpha}{2})$,(n-p-1)

D'après la table de Student à $\left(1-\frac{\alpha}{2}\right) = 0.975$ et n = 47 on obtient $\mathbf{t_{(0.975,\ 39)}}$ **= 2.023**.

$\mathbf{t_{calc}} > \mathbf{t_{(0.975,\ 39)}}$ alors on rejette l'hypothèse nulle H_0.

IV.1.2.3. Test de Fisher

L'analyse de variance est utilisée pour tester l'égalité des moyennes, On l'appelle le **F** de **Fisher**.

-Hypothèse **H_0** : SCF = SCE

-Contre hypothèse **H_1** : SCF > SCE

Le F de Fisher est calculé selon l'équation suivante :

$$F_{exp} = \frac{CMF}{CME} = \frac{SCF/p}{SCE/n-p-1}$$

Pour un seuil de (**0.05**) on compare $\mathbf{F_{exp}}$ obtenue par le calcul théorique et celle obtenue à partir de la table de Fisher pour un degré de liberté (**p, n-p-1**) avec **p=7** et **n=47** tel que (**n-p-1**) = **39**.

-On accepte H_1 si $\mathbf{F_{exp} > F_{(7, 39)}}$.

-On trouve alors $\mathbf{F_{(7, 39)} = 2.255}$ et $\mathbf{F_{exp} = 42.588}$, ainsi on accepte H_1 et on rejette H_0.

IV.1.3. Coefficient de détermination : r^2

Le coefficient de détermination $\mathbf{r^2}$, donne le taux d'explication ou pourcentage de la variation de **Y** (variables endogènes) expliqué par la variation de **X** (variable exogènes), est donné par r^2 :

$$r^2 = \frac{varition\ expliquée}{variation\ totale} = \frac{SCF}{SCT}$$

On a $\mathbf{r^2=0.88}$, ce chiffre signifie que **88%** de la variable Y est attribuable à la variation de la variable X.

IV.2. Réseau de neurones (RN)

La forme générale du réseau de neurones est constituée de trois couches reliées entre elles par des neurones.

La couche d'entrée de notre réseau de neurones est constituée de (**5**) descripteurs proposés par la RLM, la couche cachée contient (**3**) neurones, et la couche de sortie est un neurone linéaire d'où la configuration du réseau est (**5-3-1**).

On note que **ρ**, le rapport du nombre des individus (**47**) au nombre de poids (**22**) dans le réseau de neurones, devrait avoir une valeur entre **1,8 et 2,3**, avec une architecture (**5-3-1**), **ρ=47/22=2,13**.

Les valeurs de l'activité calculées par la configuration du réseau de neurones (**5-3-1**) sont affichées dans le **Tableau 4** et la corrélation entre l'activité observée et celle obtenue à partir du réseau de neurones est illustrée dans la **Figure 8**.

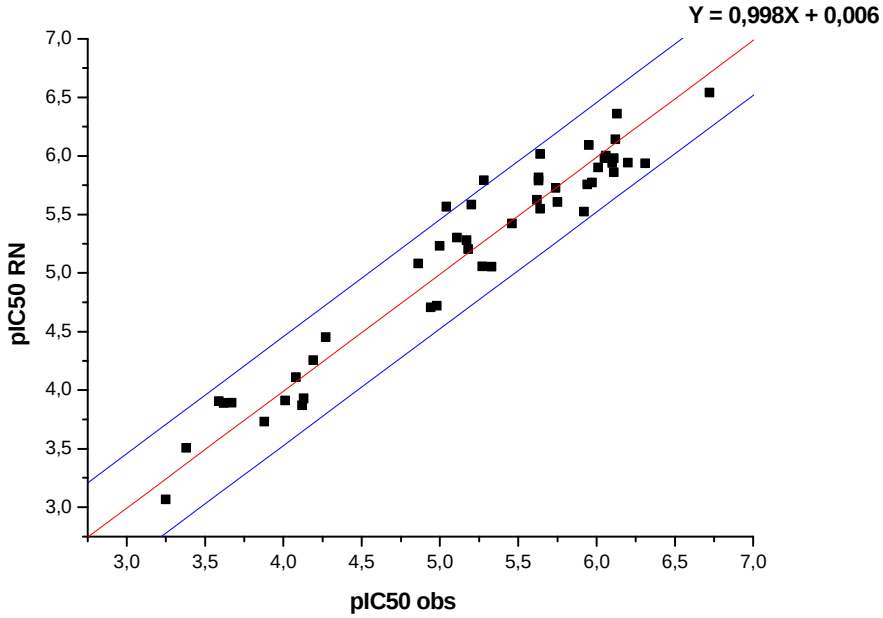


Figure 8 : Corrélation entre les **$pIC_{50\ Obs}$** et les **$pIC_{50\ RN}$**

Avec : **N = 47** ***r* = 0.97** **r^2 = 0.94**

Une bonne corrélation à été observée entre les activités observées et celles calculées à partir du réseau de neurones avec un coefficient de corrélation ***r* = 0.97.**

IV.3. Validation croisée (VC)

Nous avons utilisé la méthode de validation croisée avec la procédure (leave one out) dans le but de tester la performance du réseau de neurones et la validité du choix de nos descripteurs sélectionnés par la **RLM** et entrainés par le **RN**.

Les valeurs de l'activité obtenues par le calcul de validation croisée sont données dans le **Tableau 4**, et leur corrélation avec l'activité observée est illustrée dans la **Figure 9**.

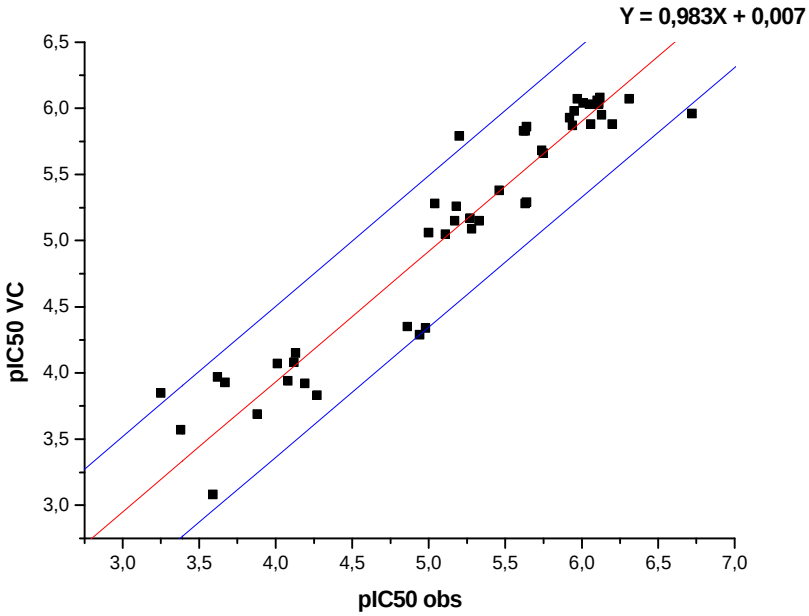


Figure 9 : Corrélation entre les $\mathbf{pIC_{50\ Obs}}$ et les $\mathbf{pIC_{50\ VC}}$

Avec : $r = 0.95$ $r^2 = 0.90$

Une bonne corrélation a été obtenue avec la validation croisée, donc l'activité calculée par ce modèle est très significative.

V. Conclusion

Les trois méthodes d'analyses **RLM**, **RN** et **VC** utilisées dans cette partie pour étudier la série de dérivés de l'isatine convergent pour conclure que les descripteurs : l'effet de l'électronégativité (**X**), l'effet donneur des liaisons hydrogènes (**DLH**) et la lipophile (**LogP**) ont une grande influence sur l'activité. Le coefficient de corrélation obtenue à partir de la **RLM** (r=0.94) est important et ce coefficient s'améliore par l'utilisation du **RN** (r=0.97), donc le modèle proposé est très significatif et sa performance est testée par la méthode de validation croisée **VC** (**r=0.95**).

Références bibliographiques

[1] W.C. Sumpter, Chem. Rev, 34, 1954, 407.

[2] F.D.Popp, Adv. Heterocycl. Chem, 18, 1975, 1.

[3] Y. Guo, F. Chen, Zhongcaoyao, 17, 1986, 8.

[4] Manjari, S. Pandey, A. Chakrabarti, L.K. Pandey, S.K. Bhattacharya, Stress and Health, 18, 2002, 133.

[5] T.J. Singh, P.K. Gujral, IND. J. Pharmac., 3 (4), 1971,187.

[6] R.Sabet, M. Mohammadpoura, A. Sadeghi, Fassihi'' QSAR study of isatine analogues as in vitro anti-cancer agents'' .2009.

Conclusion générale

Dans ce travail nous avons effectué une étude **QSAR** sur une série de molécules de l'isatine par l'utilisation des méthodes statistiques d'analyses. Le but de cette étude est d'établir une relation linéaire quantitative structure-activité. Cette étude à montré que l'activité de l'isatine est liée à l'effet de l'électronégativité (**X**), l'effet donneur des liaisons hydrogènes (**DLH**) et la lipophile (**LogP**) qui ont une grande influence sur l'activité anticancéreuse.

Ces résultats sont confirmés par les trois méthodes d'analyses statistiques : la régression linéaire multiple (**RLM**), le réseau de neurones (**RN**) et la validation croisée (**VC**).

Des études antérieures de QSAR, déjà effectuées sur la même série de l'isatine, en utilisant la **régression linéaire**, ont obtenu un coefficient de corrélation égal à (**0.92**). Dans cette étude le coefficient de corrélation obtenu avec la **régression linéaire** est égal à (**0.94**) par l'utilisation d'une variété des descripteurs, ce résultat a pu être amélioré avec le **réseau de neurones** (**r=0.97**).

Ainsi, grâce aux études QSAR, surtout avec le réseau de neurones qui nous a permis d'améliorer la corrélation entre l'activité biologique observée et celle prédite, nous pouvons profiter de la performance du pouvoir prédictif de ce modèle pour découvrir et proposer de nouvelles molécules qui pourraient être actives.