

Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR)

Rachid Hmamouchi^{*1}, Mohammed Bouachrine² et Tahar Lakhli¹

¹Molecular Chemistry and Natural Substances Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco;

²Separation Process Laboratory, Faculty of Science, University Ibn Tofail, Kenitra, Morocco

³ESTM, University Moulay Ismail, Meknes, Morocco;

*Corresponding Author: E-mail: r.hmamouchi@gmail.com

Résumé : Différentes technique statistiques tel que la régression linéaire, non linéaire, ACP, PLS, et les réseaux de neurones artificiels (ANN) ont été utilisées pour mettre en place des modèles pour la prédiction des activités biologiques. Les descripteurs des modèles ont été sélectionnés dans un jeu étendu de plusieurs descripteurs (topologiques, géométriques et quantiques). La relation quantitative structure-activité/propriété (QSAR/QSPR) modélisation se rapporte à la construction de ces modèles prédictifs d'activités biologiques différentes, en fonction de l'information de structure moléculaire et d'une banque de composés. Cet avis vise à couvrir les concepts et techniques essentielles qui sont pertinents pour la réalisation d'études QSAR / QSPR grâce à l'utilisation d'exemples choisis dans nos travaux précédents.

Mots clés : Régression, Réseaux de Neurones Artificiels, Descripteurs, Activités Biologiques, Structure-Activité/Propriété, Prédiction.

Abstract: Different technical statistics such as linear regression, nonlinear PCA, PLS, and artificial neural networks (ANN) were used to develop models for predicting biological activities. The descriptors models were selected from an extensive set of several descriptors (topological, geometrical and quantum). Quantitative structure-activity / property (QSAR / QSPR) modeling refers to the construction of these predictive models of different biological activities, depending on the molecular structure information and a compound library. This review aims to cover the basic concepts and techniques that are relevant to the realization of QSAR / QSPR through the use of examples selected from our previous research studies.

Keywords: Regression, Artificial Neural Networks, Descriptors, Biological Activities, Structure-Activity / Property, Prediction.

1- Introduction

Depuis le milieu du 19^{ème} siècle, et, à partir de la fin de 20^{ème} siècle, une appréciation que l'activité biologique est liée à la structure chimique surtout aux propriétés physico- chimiques [1]. La Relations Quantitatives Structure-Activité (QSAR) se rapporte à la construction de modèles prédictifs d'activités biologiques, en fonction de l'information de structure moléculaire d'une série de composés. C'est un procédé qui relie de manière quantitative les propriétés moléculaires aussi bien électroniques que géométriques, appelées descripteurs, avec une activité biologique [2], suivant un modèle mathématique :

$$[\text{Activité biologique}] = f [\text{Descripteurs}]$$

Dans l'équation, l'activité biologique est normalement exprimée comme $\log [1/C]$, où C est généralement la concentration minimum requise pour causer une réponse biologique définie [3].

Le concept (QSAR) a généralement été utilisé pour la découverte de médicaments et développement, et a acquis une grande applicabilité pour corrélérer l'information moléculaire avec non seulement les activités biologiques, mais aussi avec d'autres propriétés physico-chimiques, qui a donc été appelées relations quantitatives structure-propriété (QSPR).

Les descripteurs moléculaires typiques qui sont utilisés, sont généralement classés en trois catégories ; physico-chimiques, topologiques ou électroniques qui peuvent être déterminé par une expérimentation soit empirique soit théorique via la chimie computationnelle. Ces descripteurs sont les caractéristiques de la structure bidimensionnelle (QSAR-2D) ou tridimensionnelle (QSAR-3D) de la molécule.

Plusieurs modèles QSAR / QSPR réussies ont été publiés au cours des années qui englobent un large durée de la diversité biologique et physico-chimique Kubinyi (2002) [4], Schultz Cronin et al. (2003) [1] (Cronin et Livingstone 2004,) [5]. QSAR / QSPR a un grand potentiel pour la modélisation et la conception de nouveaux composés avec des propriétés robustes en étant capable de prévoir les propriétés physico-chimiques d'un fonction de caractéristiques structurales.

Dans ce document nous avons décrit, les trois parties essentielles du QSAR qui sont l'activité biologique à modéliser, les descripteurs de propriétés structurales, et les techniques statistiques pour former une relation entre la structure chimique et activité biologique.

2- Bref historique de (QSAR)

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Elles ont été abondamment utilisées dans les industries pharmaceutiques, chimiques et cosmétiques, tout particulièrement pour la conception rationnelle de nouveaux principes actifs et de nouvelles entités chimiques.

Les premiers travaux utilisant la méthode (QSAR) ont commencé au début des années 60 avec Hansch [6] d'une part et de Free et Wilson [7] d'autre part, qui ont proposé un modèle mathématique pour corrélérer l'activité biologique et la structure chimique.

Les développements de cette étude (QSAR) sont très anciens, à partir de 1868, lorsqu'Alexander Crum-Brown et Thomas Fraser montrent l'existence d'une relation entre l'activité physiologiques et la structure chimique [8], suivis, avec Richet qui établit une relation entre la toxicité et les propriétés physicochimiques [9-10]. De manière indépendante, Meyer et Overton ont décrit une corrélation linéaire entre la lipophilie (coefficient de partage huile-eau) et les effets biologiques (narcotiques) [11-12].

Pendant les dernières décennies, ce domaine a largement été étudié et les données bibliographiques disponibles sur cette approche sont maintenant importantes [13].

3- Activité biologique

La construction d'un modèle QSAR est très dépendante des données expérimentales de référence. Le choix de la base de données est donc un point critique de son développement. Dans la plupart des cas, les données expérimentales sont issues de la littérature.

Habituellement, les données l'activité à modéliser peut être considéré comme étant un effet biologique (activité pharmacologique ou toxique etc.) ou d'une propriété physico-chimique (un coefficient de partage, point de fusion, etc.). Un effet biologique implique un certain type d'interaction entre un composé chimique et une cellule qui produit une réponse définie (exemple la perturbation de la fonction cellulaire, l'inhibition d'une enzyme, l'expression différentielle de gènes, etc.).

Dans cette partie nous allons citer quelques activités biologiques utilisés dans l'analyse QSAR et qui sont publiés dans certains journaux scientifiques.

3.1- Activité (pI_{50}) antifongique des N-phénylsuccinimides.

La plus grande variété des champignons parasites des animaux et de l'homme ainsi que l'abondance des structures chimiques potentiellement actives sur ces champignons, rendent difficile une appréciation précise des relations entre familles moléculaires de types d'activités biologiques. En effet, contrairement à certains grands domaines pharmacologiques pour lesquels, l'innovation se développe autour de l'archétype structural bien défini. La création de nouvelles molécules à activité antifongique (pI_{50}) fait le plus souvent appel à des structures chimiques très variées.

Depuis 1980, certains chercheurs ont développé de nouveaux N-phénylsuccinimides fongicides, qui contrôlent les champignons pathogènes tels que (*Botrytis cinerea*) [14].

Nous notons que pI_{50} est la réciproque du logarithme décimal de la concentration de croissance des *B. Cinerea* mesurée par la méthode de dilution.

3.2- Activité biologique (pI_{50}) du (2-méthyl-6-phényléthynylpyridine).

Des études antérieures de Fabrizio Micheli et al, en 2008, ont mis en place un certain nombre de dérivés de pyrazine 2-méthyl-6-phényléthynylpyridine (MPEP) qui sont utilisés dans différents domaines de la science et de la technologie et qui ont de nombreuses applications en biologie et médecine. La molécule de 2-méthyl-6-phényléthynylpyridine (MPEP) est un médicament expérimental qui a été l'un des premiers composés qui contient une activité biologique (pI_{50}) qui agit comme antagoniste sélectif du récepteur métabotrope du glutamate de sous-type mGluR5 [15].

3.3- Activité inhibitrice (DL_{50}) des composés coumarines.

Les coumarines sont une source naturelle d'antioxydants essentiels. Les molécules présentent une activité de radicaux libres dans les tissus humains par une variété de mécanismes principalement en raison de leurs flavonoïdes et des benzophénones d'équivalence structurelles.

Plusieurs études ont été effectuées montrant que les coumarines sont des molécules biologiquement actives, elles expriment diverses activités : ils peuvent prévenir la peroxydation des lipides membranaires et capturer des radicaux hydroxyles et superoxyde, peroxyde [16]. Ils ont démontré l'efficacité de blocage du cancer chimiquement induit par les rayonnements ultraviolets (activité anticancéreuse).

Des études antérieures d'Andre Kimura Okamoto et al, ont mis en place un certain nombre de dérivés de coumarine comme activité inhibitrice (DL_{50}) à inhiber la mutagénicité quinolines dans *S. typhimurium*.

Nous notons que (DL_{50}) dose létale, est la quantité d'une matière, administrée en une fois, qui cause la mort de 50% (la moitié) d'un groupe d'animaux d'essai.

4- Descriptions structurels des molécules

Les descripteurs sont, en général, des représentations numériques des caractéristiques moléculaires spécifiques. Il existe un grand nombre de descripteurs moléculaire qui peuvent être utilisés dans une analyse QSAR pour prédire un événement biologique.

En général, les composés dans une analyse QSAR sont décrits soit par leurs propriétés physico-chimiques ou par une certaine forme de propriété structurelle. Ceux-ci peuvent se résumer en trois groupes généraux de descripteurs: Descripteurs 1D, Descripteurs 2D, et des Descripteurs 3D.

Descripteurs 1D

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

Descripteurs 2D

Les descripteurs 2D, sont des propriétés numériques qui peuvent être calculées à partir de la table de connectivité d'une molécule ou d'une représentation planaire (2D) de la structure. Ils sont basés sur les éléments présents, les charges partielles, la nature des liaisons, le nombre de cycles. La mesure la plus couramment utilisée est le coefficient de partage octanol-eau P , normalement utilisé dans sa forme logarithmique ($\log P$).

Descripteurs 3D

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

Les propriétés électroniques les plus pertinents calculés, qui sont utilisé dans la plus part de nos études sont, (E_{HOMO}) l'énergie de l'orbitale moléculaire la plus haute en énergie occupée par au moins un électron, (E_{HOMO}) L'énergie de l'orbitale la plus basse en énergie non occupée par un électron, (gap) déférence énergétique, (μ) le moment dipolaire, (E_T) l'énergie totale, (E_a) l'énergie d'activation, (λ_{max}) l'absorption maximale et le facteur d'oscillation $f_{(so)}$, sont calculé par la méthode DFT.

Pour le développement QSAR il n'y a pas de règles absolues et strictes, mais le but devrait toujours être d'utiliser le plus petit nombre de descripteurs possible.

Calcul DFT

La théorie de la fonctionnelle de densité (DFT) est une méthode qui devenue très populaire ces dernières années parce qu'ils peuvent atteindre une précision similaire à d'autres méthodes en moins de temps et moindre coût du point de vue informatique. L'énergie de l'état fondamental d'un système poly- électroniques peut être exprimée par la densité électronique totale, en fait, l'utilisation de la densité électronique à la place de la fonction d'onde pour calculer l'énergie, constitue la base fondamentale de DFT, en utilisant la fonctionnelle B3LYP et un ensemble de base 6-31G (d). Une autre version de la méthode DFT, utilise trois

paramètres fonctionnels de Becke (B3) et comprend un mélange d'HF avec DFT modalités d'échange associées au gradient de corrélation corrigée. La géométrie de toutes les espèces visées par l'enquête a été déterminée par l'optimisation de toutes les variables géométriques sans aucune contrainte de symétrie.

5- Méthodes statistique pour former la relation Structure -Activité

Un grand nombre de techniques statistiques sont disponibles pour réaliser une étude QSAR, allant de l'analyse de régression, telles que les moindres carrés partiels (PLS), analyse en composantes principales (ACP) et les Réseaux de Neurones Artificielle (RNA).

5.1- L'analyse de régression

L'analyse de régression est une technique statistique simple qui corrèle la variable dépendante (l'activité biologique) à une ou plusieurs variables indépendantes (physico-chimiques ou leurs propriétés structurales) [17]. Dans la modélisation QSAR, la régression prend généralement la forme suivante :

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Où β_0 est le modèle constant, $X_1 \dots X_k$ sont descripteurs moléculaires avec leur β_1 coefficients correspondants; \dots ; β_k (pour descripteurs moléculaires 1 à k).

En théorie, un certain nombre de propriétés peut être inclus, bien que pour des raisons pratiques de QSAR, la régression la plus valide seraient fondées sur trois ou moins de propriétés.

L'analyse de régression a un certain nombre d'avantages dans QSAR car il est simple à réaliser, et fournit des modèles clair, et transparente sans ambiguïté.

5.2- L'analyse en composantes principales (ACP)

L'Analyse en Composantes Principales (ACP) (PCA, pour Principal Component Analysis), est une méthode d'analyse de données, qui cherche à synthétiser l'information contenue dans un tableau croisant des individus et des variables quantitatives [18]. C'est une technique qui est utilisée pour réduire la dimension de l'espace de représentation des données, il consiste à remplacer les variables initiales par de nouvelles variables, appelées composantes principales, deux à deux non corrélées, et telles que les projections des données sur ces composantes soient de variance maximale.

5.3- Régression des moindres carrées (PLS)

La régression des moindres carrées partielles (PLS) est la méthode d'estimation la plus communément utilisée en chimie, et à bien des égards la plus importante. Le modèle PLS (Partial Least Square) consiste à extraire des données originales des nouvelles variables non corrélées appelées variables latentes qui ne sont autres que des composantes principales des variables d'origine [19].

Ces méthodes se trouvent parfois limitées par la complexité des problèmes à traiter. Les chimistes se sont dirigés donc vers l'emploi de méthodes plus prometteuses qui surmontent ces problèmes. Parmi ces méthodes, les réseaux de neurones artificiels (RNA) qui présentent une option de choix privilégiée.

5.4- Réseaux de Neurones Artificielle

Les réseaux de neurones artificiels sont devenus en quelques années des outils précieux dans des domaines très divers de l'industrie et divers services, Ils sont particulièrement utilisés

pour résoudre des problèmes de classification, de prédiction, de reconnaissance des formes, de catégorisation, de mémoire associative et d'optimisation [20].

Réseau de neurones artificiels (ANN) est une technique de reconnaissance qui ressemble de près le fonctionnement interne du cerveau qui est essentiellement composé de neurones interconnectés. Ce phénomène est affecté par la conception architecturale de ANN où les unités neuronales sont interconnectés les uns aux autres.

Une architecture couramment utilisé comme représenté sur la figure 1 contient un réseau d'anticipation à trois couches qui est composé d'une couche d'entrée, une couche cachée, et une couche de sortie. La couche d'entrée passe essentiellement des informations des variables indépendantes dans le système ANN; par conséquent, le nombre d'unités de neurones présents dans la couche d'entrée est égal au nombre de variables indépendantes dans l'ensemble de données. Les connexions entre les neurones sont affectées des valeurs numériques connues comme poids.

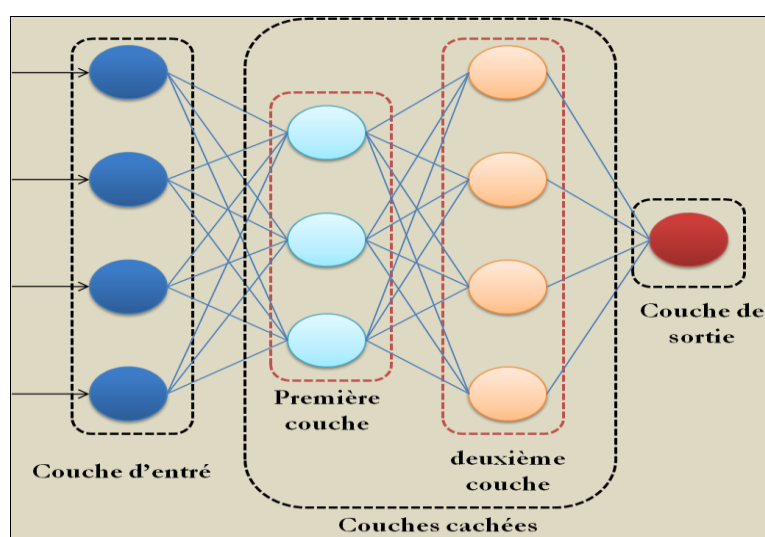


Figure 1 : Représentation d'un réseau de neurone

Les informations provenant de la couche d'entrée sont transmis à la couche cachée pour la configuration de traitement seront ensuite transmises de la couche cachée à la couche de sortie. Dans un algorithme de rétro-propagation, l'erreur calculée est dérivée de la différence entre la valeur prédite et la valeur réelle, et si elle est acceptable, le processus d'apprentissage ne cesse autrement signaux seront envoyés vers l'arrière de la couche cachée à un traitement ultérieur et réajustements de poids. Ceci est réalisé de manière itérative jusqu'à ce qu'une solution soit trouvée et l'apprentissage est terminé.

Afin de répondre à la problématique de cette recherche, telle que la prédiction des paramètres électroniques ((E_{HOMO}) , (E_{HOMO}) , (gap) , (μ) , (E_T) , (E_a) , (λ_{max}) , (f_{so})) d'un ensemble de structures chimiques de différentes familles qui contiennent une activité biologique recherchés, nous avons choisi l'architecture de perceptrons multicouches PMC [21] en utilisant l'algorithme de Levenberg-Marquardt avec un couple de fonction comme fonction de transfert.

6- Validation du modèle

Le traitement mathématique et statistique, permet la sélection des variables et la détermination d'une équation reliant de façon optimale les descripteurs structuraux et l'activité biologique.

Cette méthode utilise la probabilité critique (p-value), et les valeurs de test de **Student** (t) pour la sélection des meilleurs descripteurs pertinents. Les descripteurs sélectionnés sont obtenus dans un tableau, celui-ci regroupe les coefficients de régression, l'erreur standard, t -value et p -value.

➤ **Coefficient de corrélation, (R) :**

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

Où y_i et \hat{y}_i Étant les valeurs observées et calculées de la variable dépendante.

n : nombre de points expérimentaux considérés

Ces coefficients déterminent 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. Ces coefficients ne sont pas affectés par l'unité de mesure choisie et traduisent :

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.

➤ **Déviation standard, (S) :**

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

Où k est le nombre de restriction aux degrés de liberté

Elle mesure la variation de l'activité cible non expliquée par le modèle QSAR. En particulier, plus la déviation standard est petite et plus la corrélation est meilleure. Sa valeur est toujours fonction de l'unité de mesure de l'activité cible et tient également compte des erreurs expérimentales ce qui explique qu'une valeur trop petite n'ait aucune signification.

Les techniques de validation croisée ont été appliquées pour l'évaluation de la prédiction interne du modèle.

➤ **Test de *student***

Le test s'écrit :

$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

Si le coefficient de corrélation est différent de zéro on rejette l'hypothèse H_0 (l'hypothèse nulle) et on accepte H_1 donc le modèle est significatif.

Sous H_0 , la loi de Student à $(n-p-1)$ degrés de liberté t_{calc} s'écrit :
$$t_{calc} = \left[\frac{r}{\sqrt{\frac{1-r^2}{n-p-1}}} \right]$$

On rejette H_0 d'après (l'hypothèse nulle) lorsque : $t_{calc} > t(1-\alpha/2) ; (n-p-1)$

$t_{(1-\frac{\alpha}{2}), (n-p-1)}$ Est la valeur de la loi de student, à $(n-p-1)$ degré de liberté, à une Probabilité $(1 - \frac{\alpha}{2})$.

➤ **L'indice de Fisher F**

L'indice de Fisher F est également couramment employé afin de mesurer le niveau de signifiante statistique du modèle, c'est-à-dire la qualité du choix du jeu de paramètres.

$$F = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \hat{y}_i)^2} \frac{p - n - 1}{n}$$

En ce qui concerne la pertinence des descripteurs dans le modèle, elle est également évaluée par le t-test de Student. Il s'agit de tester l'hypothèse considérant le descripteur comme non significatif. Pour une régression multi-linéaire, cela revient à supposer le coefficient qui lui est associé comme nul.

$$|t_i| = \left| \frac{a_i}{s(a_i)} \right| > t_{1-\frac{\alpha}{2}}^{p-n-2}$$

Cette hypothèse est rejetée (avec un intervalle de confiance α) si le ratio t_i entre a_i et son erreur type $s(a_i)$ atteint la valeur du fractale d'ordre $(1 - \alpha/2)$ de la loi de Student à $(p-n-2)$ degrés de liberté.

➤ **Validation interne**

La validation croisée « cross-validation » est une des méthodes les plus employées pour déterminer la stabilité du modèle prédictif et de tester l'influence de chaque échantillon sur le modèle final. En fait, il y a au moins trois techniques de validation croisée : « test set validation » ou « hold out method », « k-fold cross-validation » et « leave-one-out cross-validation » (LOOCV).

Ce processus consiste à extraire un certain nombre n de molécules du jeu initial à k molécules et à construire un nouveau modèle avec les $n-k$ molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les n molécules retirées. Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement.

➤ **Validation externe**

Afin de tester de manière fiable ce pouvoir prédictif, l'utilisation d'un jeu de validation externe, non employé pour le développement du modèle, est nécessaire. Pour peu que le jeu de données initial soit suffisamment large, ce dernier peut être aisément divisé en deux : un jeu d'entraînement sur lequel le modèle est développé et un jeu de validation utilisé pour caractériser son pouvoir prédictif.

7- Etudes de cas

Dans cette partie, nous présentons des exemples de nos précédentes travaux sur QSAR / QSPR qui sont publiés dans des journaux scientifique:

(1) "Combining DFT and QSAR result for predicting the biological activity of the phenylsuccinimide derivatives" (Hmamouchi et al., 2013) [22].

(2) “*QSAR Modeling of the toxicity of pI₅₀ Pyrazines Derived by Electronic Parameters Obtained by DFT*” (Hmamouchi et al., 2014) [23].

(3) “*Structure activity and prediction of biological activities of compound (2-methyl-6-phenylethynylpyridine) derivatives relationships rely on electronic and topological descriptors*” (Hmamouchi et al., 2014) [24].

(4) “*Predictive modelling of the LD₅₀ activities of coumarin derivatives using neural statistical approaches: Electronic descriptor-based DFT*” (Hmamouchi et al., 2015) [25].

Parmi ceux-ci, nous sélectionnons le dernier travail concerne modalisation d'activité inhibitrice des composées coumarine comme exemples représentatives des différentes techniques utiliser dans notre étude QSAR / QSPR.

7.1- Modélisation d'activité inhibitrice (DL₅₀) des composés coumarines.

Dans ce travail, nous avons appliqué une étude QSAR sur un ensemble de 30 molécules à base des coumarin, cette étude est réalisée en utilisant la méthode de régression linéaire multiple (MLR), Régressions multiples non linéaires (MNLr) et le réseau de neurones artificiels (ANN).

• Régression linéaire multiple (MLR)

De nombreuses tentatives ont été faites pour développer une relation avec la variable prédite DL₅₀, mais le meilleur rapport obtenu par ce procédé est que celui correspondant à la combinaison linéaire de plusieurs descripteurs tel que: l'énergie totale E_T , E_{HOMO} , E_{LUMO} , l'énergie d'activation E_a , le moment dipolaire μ , l'absorption maximale λ_{max} et le facteur d'oscillation $f_{(so)}$.

$$DI_{50} = -19,563 - 4,056 \times 10^{-4} \times E_T - 8,712 \times 10^{-3} \times E_{HOMO} + 0,507 \times 10^{-2} \times E_{LUMO} + 3,297 \times 10^{-2} \times \mu + 4141,438 \times E_a + 3,821 \times 10^{-2} \times \lambda_{max} + 1,531 \times f_{(so)}$$

Pour nos 30 composés, la corrélation entre la toxicité expérimentale et calculée basée sur ce modèle est assez importante comme il est indiqué par des valeurs statistiques:

$$N = 30 \quad R = 0,637 \quad R^2 = 0,406 \quad RMSE = 0,408$$

• Régressions multiples non linéaires (MNLr)

Nous avons également utilisé la technique du modèle de régression non linéaire afin d'améliorer la structure et la toxicité d'une manière quantitative, en tenant compte de plusieurs paramètres. Ceci est l'outil le plus commun pour l'étude des données multidimensionnelles. L'équation résultante est:

$$DI_{50} = -23933,109 - 2,812 \times 10^{-3} \times E_T + 0,187 \times E_{HOMO} - 9,337 \times E_{LUMO} - 0,507 \times \mu + 4141,438 \times E_a + 49,468 \times \lambda_{max} - 1,981 \times f_{(so)} - 6,201 \times 10^{-7} \times E_T^2 - 0,39 \times E_{HOMO}^2 - 1,179 \times E_{LUMO}^2 + 0,509 \times gap^2 + 3,624 \times 10^{-2} \times \mu^2 - 268,072 \times E_a^2 - 3,827 \times 10^{-2} \times \lambda_{max}^2 + 9,177 \times f_{so}^2$$

Les paramètres obtenus décrivant les aspects électroniques des molécules étudiées ont été les suivantes:

$$N = 30 \quad R = 0,755 \quad R^2 = 0,571 \quad RMSE = 0,451$$

La valeur de l'activité DI₅₀ prédite par ce modèle est assez semblable à celle observée qui montre une répartition très régulière des valeurs d'activité calculé en fonction de celle observées. Le coefficient obtenu de la corrélation de l'équation (2) est très intéressant

(0,571), qui a été développé on applique dans la prochaine partie les réseaux de neurones artificiels (ANN).

- **Réseau de neurones artificiels de type PMC.**

L'apprentissage des PMC se fait dans la classification ou la prédiction de l'activité antioxydant des coumarines, de façon supervisée, la variable de classement ou la variable à prédire doivent être connus. Dans le cas de l'estimation de l'activité antioxydant, les collections à observer sont celles pour lesquelles on possède cette information.

La détermination de l'architecture du réseau de neurones non récurrents de type PMC pose une question au niveau de choix de nombre de couche cachée, de nombre de neurone caché, de nombre d'itération et des fonctions de transfert. Pour cela nous avons divisé aléatoirement notre base de données en trois parties : 70% pour l'apprentissage, 15% pour le test et 15% pour la validation.

Choix de nombre de couches cachées

La réalisation de plusieurs essais de calcul, on a constaté que l'augmentation du nombre des couches cachées augmente la charge des calculs sans aucun rapport de performance.

Nous pouvons donc, assurer que l'utilisation d'une seule couche cachée est préférable pour le modèle de type PMC.

Choix des fonctions de transferts et le nombre d'itération

Notre étude utilise Levenberg-Marquardt (LM) comme algorithmes d'apprentissage, qualifiés de haute performance.

Dans ce cas, nous avons essayé de changer le nombre de neurones dans la couche cachée ainsi que les couples de fonctions de transfert. Les performances ont été évaluées grâce à l'erreur quadratique moyenne (MSE) et le coefficient de corrélation (R).

Les résultats obtenus montrent que le couple de fonctions de transfert efficace est celui de **Tansig -Purlin** qui a donné un coefficient de corrélation $R = 0,908$ et d'une erreur quadratique moyenne de $MSE = 2,93 \times 10^{-2}$, avec un réseau d'architecture [8-4-1].

Avec cette configuration nous arrivons à une meilleure performance pour l'algorithme d'apprentissage LM. Cette performance a été rencontrée au bout de 6 itérations.

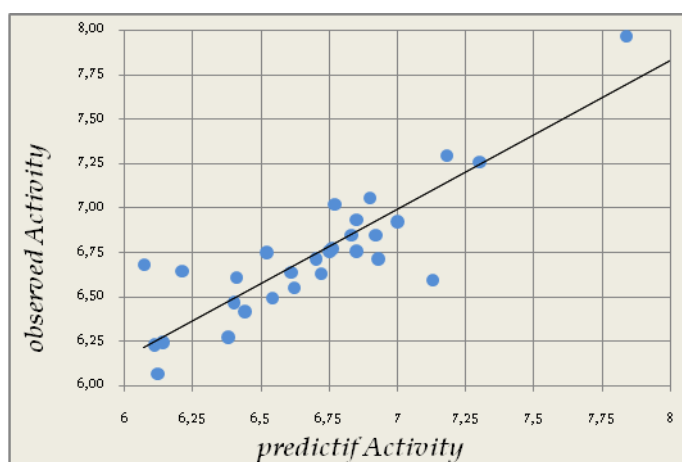


Figure 2: représente la relation entre les valeurs calculées et observées de toxicité DI_{50} , établies par (ANN)

Nous pouvons affirmer d'après ces résultats que le modèle le plus performant de la prédiction de l'activité antioxydante des coumarines, est celui qui utilise comme fonctions de transfert, la fonction **Tansig** dans la couche cachée et la fonction **Purelin** dans la couche de sortie, tout en utilisant un algorithme d'apprentissage LM, de type PMC de configuration [8-4-1] et renfermant trois couches :

- 8 neurones dans la couche d'entrée, représentant les variables électroniques indépendantes ;
- 4 neurones dans la couche cachée ;
- Un seul neurone de la couche de sortie, qui représente l'activité antioxydante des coumarines.

8- Conclusion

Dernièrement on a observé un développement parfait dans le calcul des modèles pour la prédiction des activités biologiques et chimiques ayant des propriétés robustes.

Dans ce modeste travail, nous avons fourni une brève introduction aux concepts de QSAR avec des exemples de nos précédentes recherches sur la diversité biologique et chimique des systèmes. Il est également intéressant de noter qu'il y a beaucoup de chemins pour les chercheurs dans le domaine de QSAR / QSPR dans leur recherche pour établir certaines relations entre la structure et activités / propriétés. La nature intellectuelle dispose la beauté du domaine car les possibilités infinies pour atteindre les mêmes objectifs de concevoir de nouvelles molécules avec des propriétés souhaitables.

Acknowledgements

We are grateful to the "Association Marocaine des Chimistes Théoriciens" (AMCT) for its pertinent help concerning the programs.

References

- [1]: T.W, Cronin, M.T.D., Walker, J.D., and Aptula, A.O. (2003). Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective, *Journal of Molecular Structure (Theochem)*, 622, 1–22.
- [2]: V. N. Viswanadhan, M. R. Reddy, R.J. Bacquet et M.D. Erion, *Journal of computational chemistry*, 1993, 14, 10-19.
- [3]: G.Thomas, "Fundamentals of Medicinal Chemistry". "The SAR and QSAR approaches to drug design". John Wiley & Sons, Ltd, (2003).
- [4]: H Kubinyi, (2002). From narcosis to hyperspace: the history of QSAR, *Quantitative Structure–Activity Relationships*, 21, 348–356.
- [5]: M.T.D Cronin, (2005). Toxicological information for use in predictive modelling: quality, sources, and databases, in C. Helma (ed.), *Predictive Toxicology*, New York: Dekker, pp. 93–133.
- [6]: C. Hansch, T. Fujita, *J. Am. Chem. Soc.*, **1964**, 86, 1616-1626.

- [7]: S.M. Free, J.W. Wilson, *J. Med. Chem.*, **1964**, 7, 395-399.
- [8]: A. Crum Brown, T.R. Fraser, *Trans. Roy.Soc. Edinburgh*, **1868**, 25, 151-203.
- [9]: A. Crum-Brown, T.R. Fraser, *Trans. R. Soc. Edinburgh*, **25**, 151 (1868).
- [10]: C. Richet, C. R. Séances Soc. Biol. Ses. Fil., **9**, 775 (1893).
- [11]: H. Meyer, Zur Theorie der Alkohalnarkose, *Arch. Exp. Pathol. Pharmacol.*, **42**, 109 (1899).
- [12]: E. Overton, Studien über die Narkose. Zugleich ein Beitrag zur allgemeinen Pharmakologie. Jena: Gustav Fischer, Germany, 1901.
- [13]: A. K. Debnath, "Quantitative Structure-Activity Relationship (QSAR) Paradigm Hansch Era to New Millenium". Mini Reviews in Medicinal Chemistry, 2001. I: 187-195.
- [14]: C.Takayma, M.Youshita, Y.Miyashita, H.Imajo, S.Sasaki, *Anal.Sci.*, 1988,4, p 21.
- [15]: F Micheli et al., *Bioorganic and Medicinal Chemistry Letters*, 2008, 18, 1804-9 ;PTH Epstein, *Agric. Food Chem.*, 1973, 27, 714-716.
- [16]: C.M. Anderson, A Hallberg, T Hogberg, 1996. Advances in the developpement of pharmaceutical antioxidant drug. *Food Chem*, 28: 65-180.
- [17]: H. Van de Waterbeemd, (ed.). (1995). *Chemometric Methods in Molecular Design*, Weinheim, Germany: VCH.
- [18]: C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis*, Chapman and Hall, 1980.
- [19]: H. Wold, *Multivariate analysis*, Academic Press, 1966.
- [20]: P.J. DREW, et J.R.T. MONSON, (2000). Artificial neural networks, *Surgery* 127: 3-11.
- [21]: S Haykin, (1994) *Neural Networks. A Comprehensive Foundation*. Macmillan, New York, NY.
- [22]: R Hmamouchi; A Taghki. I; M Larif; A Adad; A Abdellaoui; M Bouachrine and T Lakhlifi, *Journal of Chemical and Pharmaceutical Research*, 5(9): 198-202, 2013.
- [23]: R Hmamouchi; M Larif; A Adad; M Bouachrine; T Lakhlifi, *Int. J. Adv. Res. Comp. Sci. Soft. Eng.*, 2014, 4 (2), 241-251.
- [24]: R Hmamouchi; M Larif; A Adad; M Bouachrine; T Lakhlifi, *Journal of Computational Methods in Molecular Design*, 2014, 4 (3):61-71.

