

Méthodologie générale d'une étude RQSA/RQSP

Mounir Ghamali¹, Samir Chtita¹, Mohammed Bouachrine², Tahar Lakhlifi¹

¹CMSN, Université Moulay Ismail, Faculté des Sciences, Meknès, Maroc

²ESTM, Université Moulay Ismail, Meknès, Maroc

*Corresponding Author: E-mail: ghamalimounir86@gmail.com

Résumé: Le développement des modèles prédictifs des relations quantitatives structure-activité/propriété (RQSA/RQSP) joue un rôle important dans la conception des produits chimiques à usage fine, par exemple les produits pharmaceutiques. Compte tenu de large application de différents types de produits chimiques dans la vie humaine, la modélisation RQSA/RQSP est un outil utile pour la prédiction de l'activité biologique, propriété physicochimique et toxicologique des produits chimiques non testées. Les descripteurs moléculaires jouent un rôle critique dans le développement d'un modèle RQSA/RQSP car ils représentent quantitativement les informations chimiques codées. Ils aident non seulement dans la dérivation d'une corrélation mathématique entre la structure chimique et la réponse d'intérêt, mais ils permettent aussi l'exploration de l'aspect mécanistique impliqué dans un processus biochimique. L'analyse RQSA/RQSP est maintenant largement utilisée comme outil rationnel pour la découverte de médicaments et l'évaluation de risques environnementales.

Mots clés: RQSA, RQSP, Physicochimique, Descripteurs moléculaires.

1. Introduction

Le développement d'outils informatiques fiables couplée à la croissance de la puissance informatique a permis la mise en place des techniques de modélisation moléculaire, qui deviennent, actuellement des outils indispensables dans le domaine de la conception des médicaments.

La modélisation moléculaire a connu ces dernières décennies un essor très important dans de nombreux branches d'applications à savoir la structure électronique de l'atome, des molécules et des complexes organométalliques, l'évaluation de leurs propriétés spectroscopiques et magnétiques ou encore la structuration de molécules d'intérêts biologiques. Il s'agit de l'ensemble des techniques permettant d'étudier et de traiter les problèmes chimiques sur un ordinateur sans avoir besoin d'aller dans la salle de manipulation pour monter des expériences.

L'utilisation de méthodes alternatives à l'expérimentation, parmi lesquelles les relations quantitatives structure propriété/activité (RQSP/RQSA) sont devenues d'un grand intérêt et sont même recommandées dans les nouvelles réglementations [1,2].

L'élaboration des modèles mathématiques RQSP/RQSA reliant les propriétés physicochimiques et les activités biologiques à la structure moléculaire permet, d'une part, d'expliquer l'origine de ces activités/propriétés et, d'autre part, de les prédire pour des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

Les grandes phases de la mise en place d'un modèle RQSA/RQSP peuvent être décrites comme suit: Extraire les descripteurs à partir de la structure moléculaire, choisir les descripteurs adaptés à l'étude par rapport à l'activité/propriété analysée, et utiliser les descripteurs comme variables explicatives pour définir une relation qui les corrèle à l'activité/propriété en question. Chaque modèle doit être validé sur des jeux de données de test.

L'objectif de ce travail vise à présenter brièvement les différentes outils employées pour la mise en place des modèles RQSA/RQSP et leurs évaluations: bases de données expérimentales, descripteurs moléculaires, sélection des descripteurs pertinentes, méthodes

d'analyse de données, techniques de validation (interne et externe) et déterminer les domaines d'applicabilité. Les différentes méthodes présentées dans ce manuscrit sont celles employées au cours de notre étude [3-6].

2. Méthodologie générale d'une étude RQSP/RQSA

2.1 Base de données

Les études RQSA/RQSP étant avant tout des analyses statistiques, l'une des étapes absolument capitales est celle de la sélection des données initiales. En effet, un modèle RQSA/RQSP est très dépendant des données expérimentales de référence. Le choix de la base de données est décisif dans le développement de tel modèle. Dans la plupart des cas, les données expérimentales sont issues de la littérature. Pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible obtenues en suivant un protocole unique puisque les erreurs sur celles-ci se propageront dans le modèle final. Il y'a plusieurs éléments à vérifier dans les étapes de nettoyage d'une base de données. Il faut tout d'abord vérifier que les structures sont correctes d'un point de vue chimique (règle de valence, ...), des structures erronées entraînent la génération de mauvais descripteurs et donc de mauvais modèles.

2.2 Descripteurs moléculaires

Avant toute modélisation, il est nécessaire de calculer ou de mesurer un grand nombre de descripteurs différents, car les mécanismes qui déterminent l'activité d'une molécule ou une de ses propriétés sont fréquemment mal connus. Il faut ensuite sélectionner parmi ces descripteurs ceux qui sont les plus pertinentes pour la modélisation. En fait, chaque modélisation repose sur le nombre de descripteurs pertinents k utilisés par ce dernier. Une règle empirique apparaît dans la littérature, selon laquelle le nombre maximal de descripteurs utilisés devrait idéalement être de l'ordre du cinquième du nombre de composés dans le jeu d'apprentissage [7].

Les descripteurs moléculaires sont généralement classés en trois catégories; les descripteurs physicochimiques, topologiques et électroniques. Ces descripteurs sont caractéristiques de la structure bidimensionnelle ou tridimensionnelle de la molécule.

Parmi les descripteurs les plus utilisées dans notre étude RQSA/RQSP [3-6], on trouve l'énergie totale (E_T), l'énergie de l'orbitale la plus haute occupée (E_{HOMO}), l'énergie de l'orbitale la plus basse vacante (E_{LUMO}) et le moment dipolaire (DM),... comme descripteurs électroniques qui sont extraites du programme Gaussian 03 [8]. Il ya aussi les descripteurs physicochimiques et topologiques comme la réfractivité molaire (MR), le volume molaire (MV), le poids moléculaire (MW), la densité (D) en utilisant le logiciel ACD/ChemSketch [9].

2.3 Sélection des descripteurs

Lorsqu'une grande quantité de descripteurs est introduite, certains d'entre eux peuvent contenir des informations redondantes, entraînant un problème de colinéarité. De plus, les descripteurs calculés n'ont pas nécessairement une influence sur l'activité à modéliser.

Les descripteurs employés doivent être, autant que possible, porteurs de sens et facilement interprétables d'un point de vue chimique, aussi les modèles RQSA/RQSP devraient être simples, transparents et compréhensibles d'un point de vue phénoménologique. Il est nécessaire donc d'éliminer les descripteurs dont l'influence est inférieure à celle de l'erreur, et de sélectionner uniquement les plus pertinents d'entre eux. De manière générale, pour qu'un descripteur soit retenu, il faut que son retrait entraîne une décroissance significative de performance du modèle. Il faut donc être attentif à ne pas perdre de l'information essentielle.

Finally, the chemical sense of the descriptors used must be taken into consideration since, the more descriptors are chemically related to the phenomenon, the more the probability of random descriptors is reduced.

The selection and reduction procedure of descriptors can be performed in two steps:

- *Selection objective*
- *Selection subjective*

2.3.1 Sélection objective

It consists in the selection of variables by reducing the number of descriptors without making the dependent variable (biological activity) participate. The first step of this procedure consists in excluding all descriptors having a high percentage of identical values for the set of compounds (variance non significant). This allows to ensure that such descriptors are not included by chance in the final model. Moreover, when two descriptors are strongly correlated and their combination has a determination coefficient superior to the required threshold ($R^2 > 0,95$), only the one presenting the highest variance is retained. Not only these procedures avoid the introduction, in the model, of inappropriate descriptors but they also make the continuation of the analysis less costly in terms of calculation time, since they reduce the number of descriptors remaining to be treated.

2.3.2 Sélection subjective

2.3.2.1 Introduction progressive

This method consists in incorporating, one by one, the variables into the model by selecting at each step the variable whose partial correlation with the modeled property is the highest.

2.3.2.2 Elimination progressive

This method consists in establishing the model with the set of descriptors and then keeping only those that allow the obtaining of a model with a good correlation.

2.3.2.3 Sélection pas à pas

It is the combination of the two methods cited previously. The variables are incorporated one by one into the model by progressive selection. However, at each step, one verifies that the partial correlations of the variables previously introduced are still significant.

2.4 Division de la base des données

There are several possible approaches for the selection of the training series (on which a model RQSA/RQSP is established) and the test series (to test the predictive power of this model) of a family of compounds [10]. The approach most used by practitioners of RQSA and the most used in our study is the random division which consists in dividing the data base by simple random selection.

2.5 Méthodes d'analyse des données

To elaborate a model RQSA/RQSP we need a data analysis method, this method allows to quantify the relation that exists between the Property/Activity and the Structure (descriptors).

There are several methods to build a model and analyze statistical data of the last one, some are linear such as multiple linear regression (MLR), the regression on the partial least squares (PLS), others are non linear such as the multiple non-linear regression (MNL), the artificial neural networks (RNA)... these methods are available in software such as, Excel, Systat, Exstat, Minitab, Statistica, SPSS, R,...

Parmi Les méthodes utilisées dans notre étude sont la Régression Linéaire Multiple (MLR) et la Régression Non-Linéaire Multiple (MNL) implémentées sur Exstat et le réseau de neurone artificiel (RNA) implémenté sur Matlab.

2.6 Interprétation et validation d'un modèle RQSA/RQSP

Une fois développé, le modèle doit être interprété en analysant tous les paramètres statistiques de ce modèle, sa qualité doit être aussi étudiée, cette qualité est vérifiée par ce que l'on appelle validation. Sa robustesse, c'est-à-dire l'influence des composés de la série d'apprentissage sur le modèle, est estimée par des méthodes de validation interne. Afin d'estimer son pouvoir prédictif, des données expérimentales supplémentaires sont nécessaires afin de déterminer la capacité du modèle à prédire ces valeurs c'est ce que l'on appelle validation externe. Enfin, il est important de savoir quel type de molécules utilisées avec quel modèle. On parle alors de domaine d'applicabilité.

2.6.1 Validation interne

La validation interne d'un modèle RQSA/RQSP a été réalisée en utilisant la validation croisée LOO (Leave-One-Out) ou LMO (Leave-Many-Out) qui est quantifiée par le coefficient R^2_{cv} . Ce processus consiste à extraire un certain nombre k de molécules du jeu initial à N 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 processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules de la série d'apprentissage. En fonction du nombre de molécules retirées à chaque itération, on parlera de Leave-One-Out (LOO) ou de Leave-Many-Out (LMO) selon qu'une ou plusieurs molécules est (sont) retirée(s) [11].

Cependant, la validation interne est insuffisante pour étudier le pouvoir prédictif d'un modèle. Pour cette raison la validation externe du modèle est devenue une norme et une partie obligatoire dans la modélisation RQSA/RQSP [12,13].

2.6.2 Validation externe

Cette méthode consiste à prédire la propriété/activité d'une série de molécules appelée généralement série de test qui ne sont pas dans la série de développement du modèle, cette validation est caractérisée par le paramètre R^2_{test} . Récemment plusieurs études [14,15] ont montré l'insuffisance des paramètres R^2 , R^2_{cv} pour vérifier le pouvoir prédictif des modèles RQSA/RQSP. Par conséquent, d'autres paramètres doivent être vérifiés pour cet objectif. Ces paramètres sont connus sous le nom "critères de validation externe" ou souvent appelés "critères de Trophsa" [14].

2.7 Domaine d'applicabilité (DA)

Un modèle RQSA/RQSP ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique. Par conséquent l'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable [16,17]. Le domaine d'applicabilité (DA) permet de définir la zone dans laquelle un composé pourra être prédit avec confiance. Le DA correspond donc à la région de l'espace chimique incluant les composés de la série d'apprentissage et les composés similaires, proches dans ce même espace [18]. Ils existent plusieurs méthodes pour la détermination de domaine d'applicabilité d'un modèle RQSA/RQSP parmi ces méthodes on trouve la méthode de "leverage". Cette méthode est basée sur la variation des résidus de prédiction standardisés en fonction des valeurs des leviers h_i pour chacun des composés, pour lesquels le modèle RQSA/RQSP est utilisé pour prédire l'activité:

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

Où x_i est le vecteur ligne des descripteurs du composé i et X est la matrice du modèle déduite des valeurs des descripteurs de la série d'apprentissage. L'indice T se réfère à la matrice/vecteur transposé. La valeur critique du levier h^* est, en général, fixée à $3(k+1)/N$, où N est le nombre de composés de la série d'apprentissage, et k est le nombre de descripteurs du modèle. Si un composé a un résiduel et un leverage qui dépasse la valeur critique h^* , ce composé est considéré en dehors du domaine d'applicabilité du modèle élaboré.

3. Conclusion

La modélisation RQSA/RQSP de l'activité biologique ou des propriétés physicochimiques de molécules constitue ces dernières années, un champ de recherche très important et une phase innovante qui peut guider la synthèse de médicaments.

La mise en place de modèles RQSA/RQSP n'est pas une chose aisée. Un des problèmes importants réside également dans le traitement de données en grande quantité. Un grand nombre de descripteurs et de molécules peuvent être à analyser, mais aucune règle stricte n'existe quant au choix des paramètres structuraux les plus importants parmi le jeu complet de ceux disponibles.

Une bonne pratique de la modélisation RQSA/RQSP, si l'utilisation des lignes directrices recommandées de l'organisation de coopération et de développement économiques (OCDE) peut développer des bons modèles prédictifs avec des applications pratiques démontrées dans divers domaines biologiques et chimiques qui peuvent encore renforcer son acceptabilité par la communauté scientifique.

Références

- [1] Règlement (CE) n° 1907/2006 du Parlement Européen et du Conseil du 18 décembre 2006 concernant l'enregistrement, l'évaluation et l'autorisation des substances chimiques, ainsi que les restrictions applicables à ces substances (REACH), instituant une agence européenne des produits chimiques, modifiant la directive 1999/45/CE et abrogeant le règlement (CEE) n° 793/93 du Conseil et le règlement (CE) n° 1488/94 de la Commission ainsi que la directive 76/769/CEE du Conseil et les directives 91/155/CEE, 93/67/CEE, 93/105/CE et 2000/21/CE de la Commission.
- [2] N. Margossian, Le règlement REACH - La réglementation européenne sur les produits chimiques, Dunod / L'Usine Nouvelle, Paris, 2008.
- [3] M. Ghamali, S. Chtita, R. Hmamouchi, A. Adad, M. Bouachrine, T. Lakhli, The inhibitory activity of aldose reductase of flavonoids compounds. Combining DFT and QSAR calculations, J. of Taibah Univ. for Sci. (2016), In Press, <http://dx.doi.org/doi:10.1016/j.jtusci.2015.09.006>.
- [4] S. Chtita, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhli, Quantitative structure-activity relationship studies of di-benzo[a,d]cycloalkenimine derivatives for non-competitive antagonists of N-methyl-D-aspartate based on density functional theory with electronic and topological descriptors, J. of Taibah Univ. for Sci., 9 (2015) 143-154. <http://dx.doi.org/10.1016/j.jtusci.2014.10.006>.
- [5] S. Chtita, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhli, QSAR Studies of Toxicity Towards Monocytes with (1,3-benzothiazol-2-yl) amino-9-(10H)-acridinone Derivatives Using Electronic Descriptors, Orbital: Electron. J. Chem. 7 (2) (2015) 176-184. <http://dx.doi.org/10.17807/orbital.v7i2.677>.
- [6] S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhli, QSPR studies of 9-anilinoacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors, J. of Taibah Univ. for Sci. (2016), In Press, <http://dx.doi.org/10.1016/j.jtusci.2015.04.007>.
- [7] J.G. Topliss and R.P. Edwards, Chance factors in studies of quantitative structure-activity relationships, Journal of Medicinal Chemistry, (1979) 22 (10):1238-1244.
- [8] M.J. Frisch, Gaussian 03 Revision B.01, Gaussian, Inc., Pitts-burgh, PA, 2003.
- [9] Advanced Chemistry Development, Inc., Toronto, Canada, 2009.

(www.acdlabs.com/resources/freeware/chemsketch/).

- [10] K. Roy, S. Kar, R.N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, 2015.
- [11] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds, *Pharm. Res.* 25 (2008) 1902–1914.
- [12] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, *J. Mol. Graph. Model.* 23 (2005) 503-523.
- [13] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* 22 (2003) 69-77.
- [14] P.P. Roy, S. Paul, I. Mitra, K. Roy, On Two Novel Parameters for Validation of Predictive QSAR models, *Molecules* 14 (2009) 1660-1701.
- [15] A. Golbraikh, A. Tropsha, Beware of q^2 !, *J. Mol. Graph. Model.* 20 (2002) 269–276.
- [16] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection, *J. Chem. Inf. Model.* 48 (2008) 1733.
- [17] J. Jaworska, N.N. Jeliaskova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* 33 (2005) 445-459.
- [18] T. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. Cronin, P. Gramatica, J. Jaworska, S. G. Kahn, C. Klopman, G. Marchant, N.N. Myatt, G. Jeliaskova, R. Patlewicz, D. Roberts, T. Schultz, D. Stanton, J. Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (Quantitative) Structure–Activity Relationships, *Altern. Lab. Anim.* 33 (2005) 1-19.