Comment modéliser les propriétés physico-chimiques de molécules à partir de leur structure

Aurélie Goulon, Abdelaziz Faraj, Marc Jacguin, Fabien Porcheron

COMMENT MODÉLISER LES PROPRIÉTÉS PHYSICO-CHIMIQUES DE MOLÉCULES À PARTIR DE LEUR STRUCTURE

Aurélie Goulon* – Abdelaziz Faraj* – Marc Jacquin** – Fabien Porcheron**
e-mails : prenom.nom@ifp.fr

Institut Français du Pétrole
* IFP-Rueil - F92852 Rueil-Malmaison Cedex
** IFP-Lyon Rond-point de l'échangeur de Solaize BP 3 - F69360 Solaize

Résumé : Dans le cas où les entrées d'un système se présentent sous la forme d'une structure, comme pour certaines applications dans le domaine de la chimie où les entrées sont des molécules, il est plus avantageux d'utiliser directement cette structure pour modéliser les réponses du système, qui peuvent être les propriétés physico-chimiques de ces molécules ou leur activité. Nous présentons, à cette fin, une méthode de modélisation par apprentissage statistique, que l'on désigne par *graph machines*, basée sur le codage des entrées en graphes acycliques orientés. Le modèle construit par la méthode des *graph machines* se présente sous forme d'une composition de fonctions paramétrées élémentaires (de type réseaux de neurones) qui partagent les mêmes paramètres. Les fonctions associées aux différentes observations sont différentes les unes des autres parce qu'elles reflètent la structure inhérente à chaque observation, i.e. à des observations distinctes sont associés des modèles distincts. Cette approche se positionne, de ce fait, en rupture avec les méthodes classiques de modélisation où les variables en entrée sont représentées par des vecteurs et le modèle construit est le même pour toutes les observations. Nous présentons une comparaison des résultats obtenus, pour la modélisation de propriétés physico-chimiques d'amines, d'une part, par une approche classique, appelée QSAR (Quantitative Structure-Activity Relationship), basée sur les descripteurs moléculaires, et, d'autre part, par les *graph machines*. Elle a été réalisée dans le cadre d'un projet sur la recherche de solvants chimiques pour la capture du CO_2 dans les fumées industrielles.

Mots-clefs : Apprentissage statistique – QSAR / QSPR – Propriétés physico-chimiques de molécules.

Summary: When the inputs of a system can be described as structured data – e.g. certain applications in the field of chemistry where inputs are molecules – it is more efficient to use directly this structure to model the output(s) of the system, which can be the physicochemical properties related to these molecules or their activity. We present, for this purpose, a statistical learning modelling method – called *graph machines* – where molecules, considered as structured data, are represented by graphs. For each individual of the data set, a mathematical function (*graph machine*) is built, whose structure reflects the structure of the molecule under consideration. It is the combination of identical parameterized functions (e.g. neural networks). The parameters of these "node functions", shared both within and across the graph machines, are adjusted during training with the "shared weights" technique. Model selection is then performed by cross-validation. A comparison of the results obtained by a classic approach, based on molecular descriptors, and *graph machines* for modelling physico-chemical properties of molecules is presented. This study was realized within the framework of a project on the research for chemical solvents for the capture of the CO2 in industrial smokes.

keywords: Statistical learning – QSAR / QSPR – physicochemical properties of molecules

Introduction

La modélisation par apprentissage statistique consiste à construire, à partir d'un échantillon d'individus, des modèles mathématiques qui reproduisent le comportement d'un système, afin de pouvoir prédire – pour un ensemble plus grand d'individus – une ou plusieurs réponses du système à partir de ses variables d'entrée [3]. Dans de nombreux domaines, comme les sciences sociales, la chimie moléculaire ou le traitement de données textuelles, il arrive que les entrées du système se présentent sous forme de structures (réseaux sociaux, arrangements d'atomes, constructions grammaticales des phrases, ...). Il serait alors avantageux d'utiliser ces structures pour la modélisation des réponses étudiées. Ceci est souvent le cas dans le domaine de la chimie où, dans de nombreuses applications, les entités en entrée d'un procédé peuvent être des molécules dont on cherche à prédire les propriétés physico-chimiques (réponses du procédé), pour des réactions particulières, à l'aide de modèles construits à partir de données expérimentales. Il existe un certain nombre de méthodes, dans le domaine de la chimiométrie, qui s'appuient sur le principe que les propriétés physico-chimiques des molécules dépendent fortement de leur structure. Regroupées sous l'acronyme *QSAR* (pour Quantitative Structure-Activity Relationship), ce sont principalement des méthodes de régression linéaire ou non linéaires qui ont pour objectif de modéliser les propriétés (ou activités) physico-chimiques à partir de caractéristiques décrivant la structure des molécules [1,6,8]. Ces caractéristiques, appelées descripteurs moléculaires, sont générées par des techniques de modélisation moléculaire. On pourrait reprocher aux modèles obtenus par ces méthodes de ne pas être directement construits à partir de la structure des molécules, mais de s'appuyer sur des nouvelles variables, que sont les descripteurs moléculaires, qui sont en fait des représentations vectorielles de cette structure. Nous proposons, dans ce papier, une nouvelle méthode de modélisation basée sur un codage qui tient compte directement de la structure des molécules que nous désignons par *QSAR-GM* (GM pour *graph machines*). Dans ce codage, chaque molécule est représentée par un graphe acyclique orienté dont les noeuds sont associés aux atomes et les arêtes aux liaisons [4,5].

Pour distinguer QSAR-GM de la méthode QSAR classique, nous désignons cette dernière par *QSAR-DM* (DM pour descripteurs moléculaires). Nous comparons ensuite les résultats de ces deux méthodes en les appliquant successivement à un même jeu de données recueillies dans le cadre d'un projet de recherche de solvants chimiques pour la capture du CO_2 dans les fumées industrielles.

Méthodologie de la modélisation par les Graph Machines (QSAR-GM)

La prédiction de propriétés et d'activités pysico-chimiques de molécules présente un enjeu industriel important, car elle permet de réduire les délais et les coûts de développement. Deux disciplines de la chimiométrie se sont développées en réponse à ce besoin : la modélisation des relations structure-activité désignées par QSAR (pour Quantitative Structure-Activity Relationships), et la modélisation des relations structure-propriété désignées par QSPR (pour Quantitative Structure-Property Relationships). Elles consistent essentiellement en la recherche de similitudes entre molécules dans de grandes bases de données de molécules existantes dont les propriétés sont connues. La découverte de telles relations permettent de prédire les propriétés physiques et chimiques et l'activité biologique de composés, de développer de nouvelles théories ou d'expliquer les phénomènes observés. Elle permet également de guider les synthèses de nouvelles molécules, sans avoir à les réaliser, ou à analyser des familles entières de composés.

Nous proposons, de façon distincte mais complémentaire à l'approche QSAR-DM, une méthode que nous désignons par QSAR-GM qui permet de modéliser la propriété étudiée directement à partir de la structure des molécules codées par des graphes que nous appelons *graph machines* [5]. Les molécules sont représentées par des graphes acycliques qui tiennent compte des liaisons chimiques, de la nature des atomes ou encore de la stéréochimie du composé initial : à chaque atome non-

hydrogène est associé un nœud, et à chaque liaison entre deux atomes une arête entre les deux nœuds correspondants. Les nœuds peuvent de plus être caractérisés par des étiquettes, qui fournissent des informations sur la nature, le degré ou l'isomérie de l'atome en question. Il est également possible d'utiliser des descripteurs au sein même des *graph machines*, par l'intermédiaire des étiquettes. Enfin, le graphe est orienté, par le choix d'un noeud central. Un exemple de représentation de molécule par un graphe est donné à la Figure 1. Le noeud central est l'atome de carbone de degré 3.

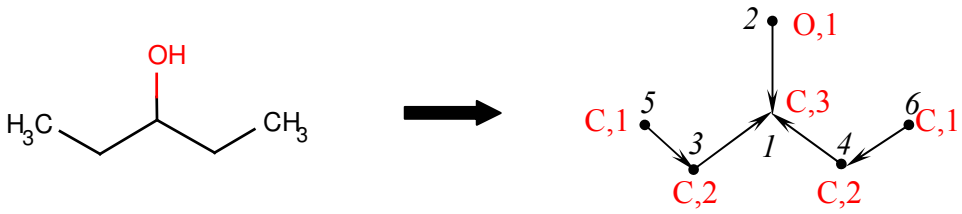

Figure 1 : Représentation d'une molécule par un graphe étiqueté. Les étiquettes du graphe (police rouge), indiquent la nature de l'atome (C ou O) ainsi que son degré (1, 2 ou 3) i.e. le nombre de liaisons avec les atomes voisins. Les numéros en gras italique sont les indices de chacun des noeuds.

La méthode QSAR-GM consiste alors à faire correspondre à chaque graphe de la base de données une fonction de même structure mathématique que le graphe associé, de la façon suivante :

- A chaque noeud du graphe est associée une fonction paramétrée $f_{\boldsymbol{\theta}}$, appelée pour cette raison *fonction de noeud*, où $\boldsymbol{\theta}$ est le vecteur des paramètres, identique pour tous les noeuds. Les fonctions paramétrées sont, par exemple, des réseaux de neurones.
- Pour chaque graphe G_i, on construit une fonction $g_{\boldsymbol{\theta}}^{i}$ par composition des fonctions $f_{\boldsymbol{\theta}}$, de façon à refléter la structure du graphe : si s_a et s_b sont deux sommets du graphe, tels que a est parent de b (i.e. un arc part de s_a et arrive en s_b), alors le résultat de la fonction associée au nœud s_b est argument de celle associée au nœud s_a.

La fonction de nœud paramétrée $f_{\boldsymbol{\theta}}$ associée au noeud z est donc de la forme :

$$f_{\boldsymbol{\theta}}(z) = f(\boldsymbol{u},\boldsymbol{v})$$

où :

- $\boldsymbol{u}$ est un vecteur dont les composantes sont égales aux arguments de sorties des fonctions associées aux nœuds parents du nœud z en question
- $\boldsymbol{v}$ est un vecteur optionnel dont les composantes fournissent l'information localisée au nœud : ce sont les étiquettes du nœud pouvant être une valeur qualitative (comme la nature du noeud, exemple le type d'atome associé au noeud, codée en disjonctif complet) ou quantitative (comme le nombre total d'arêtes qui sont reliées au noeud).

Ainsi, la fonction *graph machine* associée à la molécule représentée sur la Figure 1 est :

$$g_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(f(z_2), f(z_3), f(z_4), (0,1,0,0), 3)$$

Sorties des fonctions des nœuds 2, 3 et 4 parents du noeud 1 — Atome : C — Degré : 3

À N molécules correspondent ainsi N fonctions composées, appelées *graph machines*, partageant le même jeu de paramètres. La modélisation d'une propriété consiste ensuite à estimer ces paramètres par apprentissage statistique [3,7]. Cet apprentissage diffère de l'apprentissage traditionnel, pour lequel le modèle est unique, et la base d'apprentissage constituée de N couples entrées/sorties. Lors

de l'apprentissage des *graph machines*, la base d'apprentissage est constituée de N couples structures/sorties, et le modèle n'est plus unique. Cependant, puisque ces modèles partagent le même jeu de paramètres, il est possible d'utiliser les techniques traditionnelles d'apprentissage pour estimer ces paramètres.

La modélisation par apprentissage statistique consiste à estimer les paramètres qui conduisent à la meilleure approximation de la fonction de régression, à partir des couples entrées/sorties constituant l'*ensemble d'apprentissage*. Dans le cadre des méthodes classiques d'apprentissage, les paramètres d'un modèle $g_{\boldsymbol{\theta}}$ sont estimés à l'aide d'un ensemble de N couples $\left\{\left(\boldsymbol{x}^i, y^i\right), i = 1, ..., N\right\}$, où les vecteurs $\boldsymbol{x}^i$ sont les entrées du modèle, et y^i les valeurs mesurées de la réponse à modéliser. Le modèle est le même pour toutes les observations, et la fonction de coût minimisée peut se mettre sous la forme :

$$\mathrm{J}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(y^i - g\left(\boldsymbol{x}^i, \boldsymbol{\theta}\right)\right)^2$$

Lors de l'apprentissage des *graph machines*, l'ensemble d'apprentissage est constitué de N couples structures/sorties $\left(\left\{\mathrm{G}_i, y^i\right\},\ i = 1, ..., N\right)$, où G_i est la fonction mathématique paramétrée associée au graphe i, et y^i la valeur de la réponse modélisée pour ce même graphe. Il n'y a plus un modèle unique pour toutes les observations : à chaque exemple i correspond une fonction particulière $g^i_{\boldsymbol{\theta}}$, composée de la fonction paramétrée $f_{\boldsymbol{\theta}}$, associée la structure de l'individu i. Une fonction de coût similaire à la fonction de coût des moindres carrés traditionnelle peut être définie. Cette fonction mesure les écarts entre les observations et les valeurs prédites par le modèle :

$$\mathrm{J}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(y^i - g^i_{\theta}\right)^2$$

La minimisation de cette fonction de coût s'effectue de la même manière que lors d'un apprentissage classique, en modifiant les paramètres de façon itérative en fonction de son gradient. Lorsque la fonctions $f_{\boldsymbol{\theta}}$ est un réseau de neurones, ce gradient peut être calculé par rétropropagation, de la manière usuelle.

Les techniques habituelles de sélection de modèle, par validation croisée par exemple, peuvent également être appliquées aux *graph machines* [7]. En effet, la modélisation vise à fournir un modèle qui soit non seulement ajusté aux données d'apprentissage, mais aussi capable de prédire la valeur de la sortie sur des molécules n'appartenant pas à l'ensemble d'apprentissage, c'est-à-dire de généraliser.

Application : captage de CO_2 par les amines

L'application que nous présentons s'inscrit dans un projet visant à capturer le CO_2 dans les fumées industrielles en post-combustion. Un procédé a été mis en place via un lavage des fumées par une solution appelée solvant constituée en général par un mélange d'eau et d'amines. Dans ce procédé, les fumées sont mises en contact avec le solvant qui va absorber le CO_2 sélectivement par rapport aux autres composés des fumées.

L'approche proposée est de développer une méthodologie permettant de modéliser les propriétés physico-chimiques de la molécule d'amine et sa réactivité vis-à-vis du CO_2 à partir de sa structure géométrique. Cette méthodologie devrait permettre de mettre en évidence des molécules adaptées à la capture du CO_2 dans les fumées, i.e. présentant des propriétés particulières (faible enthalpie de réaction avec le CO_2 , forte capacité d'absorption du CO_2, faible tension de vapeur, faible taux de dégradation).

Un travail est alors nécessaire pour identifier ces molécules parmi celles disponibles dans une liste commerciale particulièrement élevée (plus de 500 candidats). Pour éviter une expérimentation longue et coûteuse, un échantillon de molécules représentatives a été judicieusement sélectionné. Nous appliquons comparativement les 2 méthodes QSAR-DM et QSAR-GM pour modéliser les propriétés à partir des données expérimentales relatives à ces molécules.

Bibliographie

[1] Bevan D.R. QSAR and Drug Design, http://www.netsci.org/Science/Compchem/feature12.html

[2] De Bruin T., Faraj A. (2007). Descripteurs moléculaires pour caractériser des molécules de solvants aminés, NT F110/EV-LC/07-034, IFP Rueil-Malmaison

[3] Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M.B., Badran, F., et Thiria, S. (2008) Apprentissage statistique, Eyrolles, .

[4] Goulon, A., Duprat, A., et Dreyfus, G. (2005). From Hopfield nets to recursive networks to graph machines: numerical machine learning for structured data. Theoretical Computer Science, 2005, **344** (2-3), p. 298-334.

[5] Goulon A.(2008). Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments, thèse, Université Pierre et Marie Curie (Paris 6)

[6] Hansch, C., Leo, A., Hoekman, D. (1995) Exploring QSAR - Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, D.C.

[7] Monari, G. (1999). Sélection de modèles non linéaires par leave-one-out : étude théorique et application des réseaux de neurones au procédé de soudage par points thèse [en ligne], Université Pierre et Marie Curie (Paris 6), 1999.

[8] Wold S. (1991). Validation of QSARs, QSAR, **10**, 191-193