Étude lexicographique de sous-graphes pour l'élaboration de modèles structures à activité – cas de la chimie organique

Nicolas Bloyet[*,**,***], Pierre-François Marteau[*]
Emmanuel Frénod[**,***]

[*]IRISA, Université Bretagne Sud – 56000 VANNES - UMR 6074
{nicolas.bloyet, pierre-francois.marteau}@irisa.fr,
https://www.irisa.fr
[**]LMBA, Université Bretagne Sud – 56000 VANNES - UMR 6205
{nicolas.bloyet, emmanuel.frenod}@univ-ubs.fr
http://web.univ-ubs.fr/lmba/
[***]See-d SAS – Parc d'Innovation Bretagne Sud - 56000 VANNES
{nicolas.bloyet, emmanuel.frenod}@see-d.fr
https://www.see-d.fr

Résumé. Les modèles structure–activité (QSAR) cherchent à extraire de l'information utile dans des observations relatives à des structures, dans le but d'associer des éléments structurels à une activité d'ordre macroscopique. Un exemple typique est celui de la chimie organique, où certaines propriétés physiques et chimiques d'une molécule sont fonction de son agencement interne (conformation). On retrouve en particulier des sous-structures caractéristiques, nommées groupements fonctionnels ou fragments qui s'apparentent à des sous-graphes, ainsi que des structures de liaison. Nous proposons une analyse lexicographique de ces fragments et montrons que ceux-ci suivent approximativement des lois de puissance, proches des lois de Zipf observées dans le cadre des langues naturelles. En poursuivant cette analogie, nous développons la notion de "plongement" de fragment (fragment-embedding). Nous montrons l'intérêt de cette notion et en déduisons quelques perspectives.

1 Introduction

Les molécules chimiques s'apparentent bien plus à un réseau/graphe d'atomes qu'à une information de type vectorielle, contenant un nombre fini de descripteurs. Cette information structurelle renferme ainsi une grande partie de l'information caractéristique des molécules : elle définit la manière dont celles–ci vont se conformer dans l'espace, ce qui conditionnera considérablement leurs propriétés macroscopiques où leur réactivité potentielle en présence d'une autre espèce. Le traitement de cette information structurelle est spécifiquement dénommé *Quantitative Structure–Activity Relationship* (QSAR). On distingue deux cas d'usage des modèles QSAR : le premier est purement applicatif, il vise à estimer le comportement macroscopique d'une molécule vis-à-vis d'une propriété donnée. Le second cas d'usage est la restitution

d'information. Pouvoir interpréter directement un modèle QSAR permet d'identifier quelle serait la relation liant les éléments structurels aux propriétés macroscopiques de la molécule. Dans ce cas, il est nécessaire de s'appuyer sur des prédicteurs conservant un certain sens chimique.

2 Modèles QSAR *"Sac de Fragments"*

Parmi les nombreux types d'approche exploitant ces modèles (Wu et al., 2018), nous nous intéressons dans cet article aux approches orientées graphe. Les molécules sont assimilables à des graphes, dans lesquels les atomes tiennent le rôle de nœuds, tandis que les liaisons atomiques covalentes tiennent le rôle d'arêtes. Au sein même de cette famille d'approches, on distingue les méthodes de type fragmentation (Varnek et al., 2008), qui visent à séparer le graphe entier en plusieurs sous–graphes plus aisés à analyser individuellement. De part leur correspondance naturelle avec les groupements fonctionnels et structure bien connues, ces méthodes offrent une ré-interprétabilité accrue par rapport à des descripteurs plus artificiels.

2.1 Principe

On décrit ici un modèle de type *sac de fragments (Bag of Fragments)*, similaire à celui proposé dans (Baskin et Varnek, 2008). La méthodologie est la suivante : disposant d'une notation nous permettant d'assurer que deux graphes *isomorphes* [1] seront représentés par le même identifiant dans le contexte de l'étude, nous générons à partir de chaque nœud (atome) un sous–graphe centré sur lui–même, de longueur arbitraire. Ces sous–graphes pourront ainsi faire l'objet de *descripteurs topologiques*, en d'autres termes de nouvelles variables explicatives (prédicteurs), dont la valeur pour une molécule donnée correspondra au nombre d'occurrences de ce sous–graphe et de ses isomorphes dans cette molécule. On projette ainsi les molécules sur un espace de fragments (sous–graphes), d'où le terme de "sac" de fragments. Cette projection peut ensuite être exploitée par des méthodes d'analyse classiques (Varnek et al., 2008).

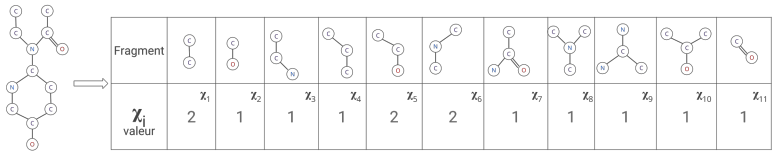


FIG. 1: Projection d'un graphe (molécule) sur un espace de fragments

Ces fragments sont utilisables en l'état pour la prédiction de certaines propriétés (densité, LogP [2] par exemple), mais se révèlent insuffisants pour d'autres propriétés telles que le point d'ébullition par exemple. On provoque en effet une perte d'information en ne tenant pas compte du contexte d'apparition de ces fragments.

1. identiques à un ensemble de permutations près de l'un à l'autre
2. rapport de solubilité eau/octanol, soit le caractère lipophile ou lipophobe de l'espèce chimique

Afin d'exploiter au mieux cette approche *Bag of Fragments*, et d'être le plus précis possible pour la suite de l'étude, il est nécessaire d'utiliser une notation qui permette de garantir au maximum que des fragments identiques, appartenant à la même classe d'isomorphisme, soient regroupés sous la même modalité de descripteur topologique. On utilise ici une méthode à base de projection vers un arbre canonique et de notation de Newick modifiée, pour exprimer les différentes modalités de fragments.

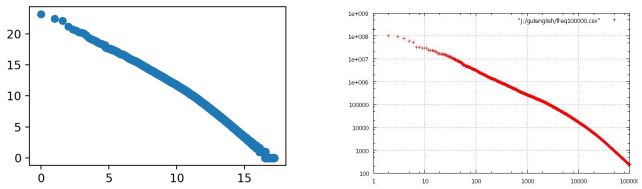
2.2 Vocabulaire généré

Selon les fragments que l'on considère (taille, prise en compte ou non du label, etc.), on génère un vocabulaire de fragments, dont la taille va être fonction du caractère précis ou au contraire approximatif de ceux-ci. En revanche, on constate un effet de la taille de ce vocabulaire sur la modélisation de QSAR, qui ne donne d'ailleurs pas nécessairement l'avantage aux fragments les plus précis. Dans (Ruggiu et al., 2010) notamment, on constate qu'en autorisant un niveau d'imprécision [3] sur les fragments considérés, ceux-ci peuvent s'avérer être plus discriminants sur certaines tâches de prédiction. Cela pourrait être révélateur d'une forme de *polysémie* à considérer, voire même d'une sémantique véhiculée par ces fragments.

3 Étude lexicographique des fragments moléculaires

L'observation précédente, qui révèle une apparente influence de polysémie, amène à formuler l'hypothèse d'une analogie avec ce que l'on observe au niveau des lexèmes en traitement automatique du langage naturel (TALN).

On propose pour étayer cette hypothèse l'étude suivante : considérant des fragments de taille t (un atome "feuille" est distant d'au plus t liaisons avec l'atome "racine"), et un corpus extrait de *PubChem* [4] recouvrant un espace chimique assez large (1 510 000 molécules), quels sont les écarts de fréquence entre les fragments les plus courants et les plus exotiques ? Le cas échéant, cette répartition se rapproche-t-elle d'une loi statistique connue ?



(a) Fragments de taille 2 (b) Mots de la langue anglaise

FIG. 2: Fréquence du terme (fragment/mot) selon le rang, en repère log-log

Cette étude, effectuée ici pour $t = 2$, montre qu'il existe un écart très important entre le fragment le plus courant et le fragment le plus rare [5]. Il est plus aisé de constater cette disparité

3. considérant les propriétés pharmacophoriques (plus générales) d'un atome au lieu de son symbole atomique (plus précis)
4. base de données publique de molécules `https://pubchem.ncbi.nlm.nih.gov`
5. 16'978'587 occurrences pour le fragment de rang 1, 1 seule occurrence pour le fragment de rang 153 004

dans un repère log-log, où l'on trace la fréquence d'un terme (fragment) en fonction de son rang (figure 2a). Apparaissent alors plusieurs lois de puissance, phénomène caractéristique d'une famille de distributions dites *Zipfienne* (Newman, 2005) , que l'on retrouve empiriquement sur les corpus textuels. On retrouve ainsi une distribution similaire sur le corpus *projet Gutemberg* [6] en langue anglaise, exception faite du nombre de termes (figure 2b).

On constate également que la taille du vocabulaire généré semble soumis à une explosion combinatoire quand t augmente [7], ce qui a pour conséquence de rendre très difficile une exploitation statistique de ces fragments pourtant plus précis sans disposer au choix d'un très grand nombre d'observations couvrant cet espace de fragments très vaste, ce qui est difficile à obtenir pour une propriété cible, ou d'une méthode permettant de rapprocher des fragments différents, mais porteurs d'une information similaire.

4 Vers une utilisation des fragments en contexte

Les observations empiriques précédentes indiquent que des fragments de graphes pris en tant que prédicteurs deviennent rapidement difficiles à analyser d'un point de vue statistique, au vu de l'explosion de dimensionalité que cela implique. Composer avec des observations caractérisées par des vecteurs creux à dimensionalité très élevée est une des problématiques avec lesquelles compose le traitement automatique des langages naturels (TALN), par troncature de vocabulaires (termes trop fréquents ou trop rares), ou bien grâce aux travaux plus récents dans le domaine des plongements *(embeddings)*, techniques de réduction de dimension basées sur une contextualisation des termes (Mikolov et al., 2013), (Le et Mikolov, 2014), (Pennington et al., 2014). Il est à noter que des apprentissages de représentation ont été récemment adaptés pour les graphes selon plusieurs variantes (Goyal et Ferrara, 2017) (Narayanan et al., 2017).

On propose ici de continuer à considérer des fragments de taille t comme des termes (lexèmes) qui composent un graphe, structure d'ordre supérieure. On embarque ainsi les informations portées par les arêtes dans cette structure de base que sont les fragments. On génère au préalable pour chaque nœud v de chaque graphe son fragment de taille t associé, noté φ_v. On considère ensuite une distance n de voisinage maximale, définissant I_v l'ensemble des voisins de ce nœud. Pour chaque $i \in I_v$, nous allons générer un ensemble de *n-grams* de fragments (figure 3).

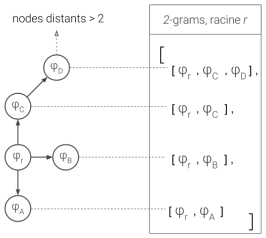


FIG. 3: Composition d'un *n-gram* comme énumération de fragments

6. source : `http://1.1o1.in/en/webtools/semantic-depth`
7. on passe ainsi de 153 004 à 1 386 918 fragments en passant t de 2 à 3

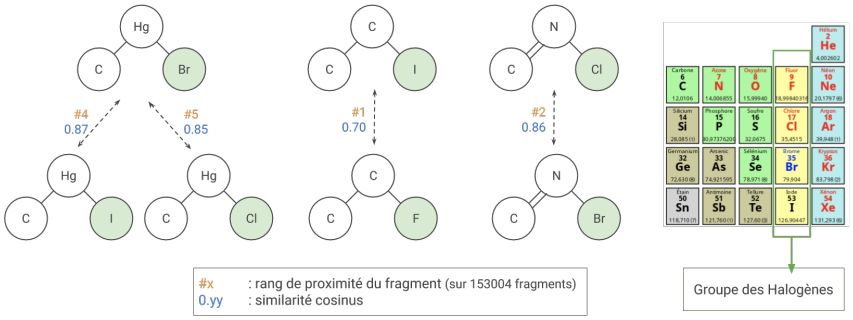


FIG. 4: Observation de substitutions dans des fragments similaires

Étude de similarités

Suivant le protocole précédent, les *embeddings* des 153 004 fragments de taille $t = 2$ sont calculés sur un corpus de 1 510 000 molécules avec un modèle Skip-Gram de paramètre $n = 4$. Les fragments étant projetés sur un nombre de dimensions arbitraire $m = 100$, on peut définir une similarité entre deux fragments. La figure 4 illustre une tendance qui semble se retrouver en opérant une recherche de fragments les plus proches d'un fragment de référence au sein du corpus étudié, au vu de la similarité cosinus. Dans le cas où l'on soumet un fragment de référence comportant des atomes appartenant au groupe des halogènes [8], on constate bien souvent la présence parmi les fragments les plus proches, de fragments identiques à la référence, à une substitution d'halogène près. Ce type de substitution concerne donc deux atomes aux propriétés similaires, ce qui du point de vue chimique fait sens. Les embeddings de fragments en contexte, contrairement aux fragments bruts, semblent donc capable, sur cet exercice, d'inférer des proximités sémantiques intéressantes, tâche nécessitant pourtant une certaine connaissance de ce domaine. Des études complémentaires sont requises pour confirmer l'intérêt de tels embeddings sur des tâches de régression et de classification de molécules.

5 Conclusion

Dans cet article nous avons présenté un état de l'art sur l'utilisation des modèles *Bag of Fragments* dans l'élaboration de QSAR. Des études lexicographiques menées sur ces fragments sur un corpus conséquent exposent empiriquement des similarités étonnantes vis-à-vis des lois distributionnelles typiques des langages naturels, et nous encourage à formuler une analogie entre ces domaines, notamment pour résoudre des problématiques bien adressées en TALN. En ce sens, nos premières expérimentations (application requête-réponse par exemple) portant sur des *plongements* de fragments *(embeddings)*, pris dans leur contexte d'apparition, paraissent très encourageantes, et nous amènent à approfondir cette analogie en élaborant de nouvelles méthodes de caractérisation de graphe, dans l'espoir d'améliorer notamment la prédiction de propriétés physico-chimiques des molécules.

8. Fluor `F`, Chlore `Cl`, Brome `Br`, Iode `I`, tous situés dans la 17-ième colonne du tableau périodique, et présentant des propriétés chimiques très homogènes

Références

Baskin, I. et A. Varnek (2008). Building a chemical space based on fragment descriptors. *Combinatorial chemistry & high throughput screening 11*(8), 661–668.

Goyal, P. et E. Ferrara (2017). Graph embedding techniques, applications, and performance : A survey. *arXiv preprint arXiv :1705.02801*.

Le, Q. et T. Mikolov (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pp. 1188–1196.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Narayanan, A., M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, et S. Jaiswal (2017). graph2vec : Learning distributed representations of graphs. *arXiv preprint arXiv :1707.05005*.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics 46*(5), 323–351.

Pennington, J., R. Socher, et C. Manning (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Ruggiu, F., G. Marcou, A. Varnek, et D. Horvath (2010). Isida property-labelled fragment descriptors. *Molecular informatics 29*(12), 855–868.

Varnek, A., D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko, et G. Marcou (2008). Isida-platform for virtual screening based on fragment and pharmacophoric descriptors. *Current Computer-Aided Drug Design 4*(3), 191.

Wu, Z., B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, et V. Pande (2018). Moleculenet : a benchmark for molecular machine learning. *Chemical science 9*(2), 513–530.

Summary

The development of structure-activity models (QSAR) consists in being able to extract useful information in observations relating to molecular structures, in order to associate structural elements with an macroscopic activity. A typical example is that of organic chemistry, where certain physico-chemical properties of a molecule are a function of its internal arrangement (conformation). In particular, we find characteristic substructures, called functional groups or fragments that are similar to subgraphs, as well as structural links. We describe in this paper a distributional analysis of these fragments and show that they follow approximately power laws, close to the Zipf laws well known for natural languages. Pursuing this analogy, we develop the concept of "fragment-embedding" that we evaluate on classification/regression tasks by comparing our results to traditional "bag-of-fragments" approaches. We show the interest of this concept and deduce some perspectives.