



Un noyau sur graphe prenant en compte la stéréoisométrie des molécules

Pierre-Anthony Grenier, Luc Brun, Didier Villemin

► To cite this version:

Pierre-Anthony Grenier, Luc Brun, Didier Villemin. Un noyau sur graphe prenant en compte la stéréoisométrie des molécules. Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014, Jun 2014, Rouen, France. <hal-00988762>

HAL Id: hal-00988762

<https://hal.science/hal-00988762v1>

Submitted on 9 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Un noyau sur graphe prenant en compte la stéréoisométrie des molécules

Pierre-Anthony Grenier¹

Luc Brun¹

Didier Villemin²

¹ GREYC UMR CNRS 6072, Caen, France

² LCMT UMR CNRS 6507, Caen, France

{pierre-anthony.grenier,luc.brun,didier.villemin}@ensicaen.fr

Résumé

L'étude des relations quantitatives structure-activité (QSAR) ou structure-propriété (QSPR) sont deux domaines de recherche actifs, où le but est la prédiction de propriétés de molécules. Dans ces domaines, les noyaux sur graphes permettent de combiner la représentation naturelle des molécules par des graphes avec des méthodes classiques d'apprentissage automatique tels que les machines à vecteurs de support. Malheureusement, le positionnement relatif des atomes dans l'espace peut être différent pour des molécules représentées par un même graphe, ces molécules peuvent donc avoir des propriétés différentes. Ces molécules sont appelées stéréoisomères. Les propriétés variant entre les stéréoisomères ne peuvent pas être prédites par les méthodes habituelles basées sur des graphes simples. Dans cet article, nous présentons une nouvelle représentation des molécules qui prend en compte la stéréoisométrie et nous proposons un noyau entre ces structures permettant de prédire des propriétés liées à la stéréoisométrie.

Mots Clef

Noyau sur graphe, chémoinformatique, stéréoisométrie.

Abstract

The prediction of molecule's properties through Quantitative Structure Activity (resp. Property) Relationships are two active research fields named QSAR and QSPR. Within these frameworks Graph kernels allow to combine a natural encoding of a molecule by a graph with classical statistical tools such as SVM or kernel ridge regression. Unfortunately some molecules encoded by a same graph and differing only by the three dimensional orientations of their atoms in space have different properties. Such molecules are called stereoisomers. These latter properties can not be predicted by usual graph methods which do not encode stereoisomerism. In this paper we propose a new graph encoding of molecules taking explicitly into account stereoisomerism and propose a new kernel between these structures in order to predict properties related to stereoisomerism.

Keywords

Graph kernel, Chemoinformatics, Stereoisomerism.

1 Introduction

Les méthodes visant à prédire les propriétés de molécules, sont basées sur le principe de similarité, qui stipule que : “deux molécules similaires doivent avoir des propriétés similaires”. La prédiction de propriétés de molécules implique donc de construire un modèle représentant celle-ci ainsi qu'une mesure de similarité entre ces modèles. Nous supposons implicitement que la similarité entre les modèles correspond à une similarité entre les molécules. Toutefois, différents modèles peuvent encoder différents niveaux d'information. Ainsi les mesures de similarité associées à différents modèles peuvent avoir différents degrés de pertinence.

Une molécule peut être représentée par sa formule brute (p. ex. CH_4). Cependant cette représentation ne prend pas en compte les liaisons entre les atomes. De ce fait, certaines molécules, appelées isomères de structures, peuvent être différentes, mais représentées par une même formule brute. Afin de remédier à cette limitation, une molécule est usuellement représentée par son graphe moléculaire. Un graphe moléculaire est un graphe simple $G = (V, E, \mu, \nu)$, où chaque nœud $v \in V$ encode un atome, où chaque arête $e \in E$ encode une liaison entre deux atomes, où la fonction μ associe à chaque nœud un label identifiant la nature de l'atome (carbone, oxygène, ...) qu'il représente et où la fonction ν associe à une arête le type de lien (simple, double, triple ou aromatique) de la liaison représentée. Les graphes moléculaires encodent les relations de voisinages entre les atomes. Ils permettent donc de différencier les isomères de structure.

Cependant, les graphes moléculaires ont aussi une limite : ils n'encodent pas la configuration spatiale des atomes. En effet certaines molécules, appelées stéréoisomères, sont représentées par un même graphe moléculaire, mais ont des positionnements relatifs de leurs atomes dans l'espace différents. Les propriétés variant entre les stéréoisomères ne peuvent donc pas être prédites seulement grâce au graphe moléculaire. La plupart des stéréoisomères sont caractérisés par l'orientation dans l'espace des voisins directs

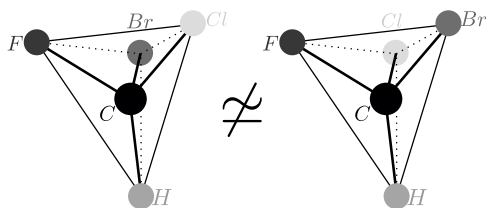


FIGURE 1 – Deux configurations spatiales différentes des voisins d'un carbone.

d'un atome ou d'un groupe de deux atomes connectés. Nous pouvons imaginer par exemple, un atome de carbone avec quatre voisins, chacun d'eux étant situé sur un des sommets d'un tétraèdre. Si l'on échange deux des voisins, on obtient alors une configuration spatiale différente (Figure 1). Un atome est appelé centre stéréogène si la permutation de deux de ses voisins crée un différent stéréoisomère. Il convient de souligner que, la stéréoisomérisation d'une molécule est indépendante (jusqu'à un certain point) de la position précise de chaque atome. En effet, dans la Figure 1, n'importe quelles coordonnées des atomes gardant le même positionnement relatif entre les atomes H, Cl, Br et F, correspond à un même stéréoisomère. De la même manière, deux atomes liés, forment un centre stéréogène si une permutation de la position de deux atomes appartenant à l'union de leur voisinage, crée un différent stéréoisomère (Figure 2). Parmi les molécules actuellement utilisées en chimie, 98% des centres stéréogènes sont, soit des carbones avec quatre voisins, appelés carbones asymétriques (Figure 1), soit des couples de deux carbones liés par une liaison double (Figure 2). Nous limitons notre étude à ces deux cas.

Les noyaux sur graphes [6, 3], correspondent à une mesure de similarité entre graphes. En supposant qu'un noyau k est symétrique et défini positif, la valeur $k(G, G')$, où G et G' sont deux graphes, correspond à un produit scalaire entre deux vecteurs $\Psi(G)$ et $\Psi(G')$ dans un espace de Hilbert. Cette propriété nous permet de combiner les noyaux sur graphes à des algorithmes classiques d'apprentissage automatique en utilisant l'astuce du noyau, qui consiste à remplacer les produits scalaires entre $\Psi(G)$ et $\Psi(G')$ par $k(G, G')$ dans ces algorithmes.

Jusqu'à présent, seules quelques méthodes ont essayé de construire des noyaux sur graphes prenant en compte

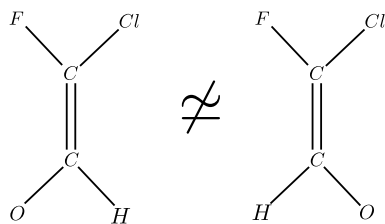


FIGURE 2 – Deux configurations spatiales différentes des voisins de deux carbones liés par une liaison double.

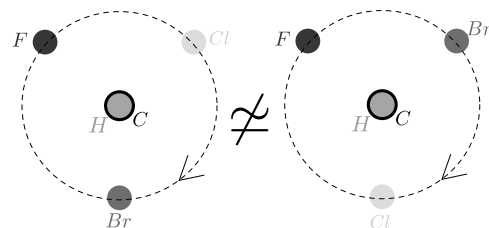


FIGURE 3 – Vues des molécules de la figure 1 depuis l'atome H.

la stéréoisomérisation. Brown et al. [2] ont proposé d'inclure la stéréoisomérisation dans une extension du noyau Tree-Pattern [6]. Un des inconvénients de cette méthode est que les motifs possédant une information sur la stéréoisomérisation et les motifs n'en possédant pas, sont combinés sans pondération dans la valeur finale du noyau. Donc pour une propriété uniquement liée à la stéréoisomérisation, les motifs ne possédant pas d'information sur la stéréoisomérisation peuvent être assimilés à du bruit et peuvent donc dégrader la prédiction. De manière intuitive, la stéréoisomérisation est liée au fait que, permuter deux voisins d'un atome donné produit une configuration spatiale différente. La stéréoisomérisation peut être facilement détectée si tous les voisins d'un centre stéréogène possèdent des labels différents (p. ex. Figure 1). Dans le cas contraire, l'influence d'une permutation entre deux voisins doit être cherchée au-delà du voisinage direct du centre stéréogène. En se basant sur cette constatation, Grenier et al. [5] ont introduit le sous-arbre minimal permettant de caractériser un centre stéréogène au sein d'une molécule acyclique. Ils ont aussi proposé un noyau basé sur ce sous-arbre, afin de prendre en compte la stéréoisomérisation. Ce noyau est cependant restreint aux molécules acycliques.

En se basant sur [5], nous présentons dans la section 2 un codage des molécules qui permet de distinguer les stéréoisomères. Dans la section 3 nous présentons la construction d'un sous-graphe qui caractérise localement un centre stéréogène. Puis dans la section 4, nous utilisons ce sous-graphe afin de construire un nouveau noyau sur graphes pouvant être utilisé aussi bien sur des molécules acycliques que sur des molécules cycliques, contrairement au noyau de [5]. Finalement, nous présentons dans la section 5 les résultats obtenus grâce à ce noyau et comparons ces résultats avec l'état de l'art.

2 Graphes Ordonnés

La configuration spatiale des voisins de chaque atome peut être encodée par une séquence ordonnée de ses voisins. Considérons, par exemple, la molécule à gauche dans la figure 1, depuis l'atome d'hydrogène (H). Les trois voisins restants, Cl, Br et F, peuvent être projetés dans un plan. En observant ce plan depuis H et en tournant dans le sens horaire, nous rencontrons les atomes dans l'ordre : Cl, Br puis F (figure 3). Ainsi, la séquence H, Cl, Br, F pourrait représenter cette configuration, et donc la séquence H, Br,

Cl, F représenterait la seconde configuration. Considérons à présent, la molécule à gauche dans la figure 2. Nous rencontrons F et Cl en tournant dans le sens horaire autour du carbone situé en haut de la représentation de la molécule, et H puis O pour le carbone situé en bas. Cette configuration peut donc être encodée par les deux séquences F, Cl et H, O, respectivement, pour le carbone situé en haut et celui situé en bas. Les séquences F, Cl et O, H permettent alors d’encoder la seconde configuration.

Pour ajouter cette information à un graphe, nous utilisons la notion de graphe ordonné. Un graphe ordonné $G = (V, E, \mu, \nu, ord)$ est un graphe moléculaire $G_m = (V, E, \mu, \nu)$ avec une fonction $ord : V \rightarrow V^*$ qui associe à chaque sommet une liste ordonnée de ses voisins. Deux graphes ordonnés G et G' sont isomorphes ($G \simeq_o G'$) si il existe un isomorphisme f entre leur graphe moléculaire respectif G_m et G'_m tel que $ord'(f(v)) = (f(v_1) \dots f(v_n))$ avec $ord(v) = (v_1 \dots v_n)$ (où $N(v) = \{v_1, \dots, v_n\}$ désigne le voisinage de v).

Cependant, des graphes ordonnés différents peuvent représenter une même molécule. Si l’on regarde le carbone central de la molécule située à gauche dans la figure 1 depuis un autre voisin, par exemple l’atome (F) nous pouvons obtenir une séquence différente (F, Br, Cl, H). De la même manière, si l’on considère la molécule située à gauche dans la figure 2, depuis l’autre côté du plan où elle est représentée, on obtient en tournant toujours dans le sens horaire les séquences Cl, F et O, H. Nous devons donc définir une relation d’équivalence entre les graphes ordonnés, afin que deux graphes ordonnés soient équivalents si ils représentent une même configuration. Pour cela, nous introduisons la notion de fonction de ré-ordonnancement σ qui associe à chaque sommet $v \in V$ une permutation $\sigma(v)$ sur $\{1, \dots, |N(v)|\}$, permettant de réordonner son voisinage. Le graphe avec un voisinage réordonné $\sigma(G)$ est obtenu depuis G en remplaçant pour chaque sommet v sa séquence ordonnée $ord(v) = v_1 \dots v_n$ par la séquence $v_{\sigma(v)(1)} \dots v_{\sigma(v)(n)}$, où $\sigma(v)$ désigne la permutation appliquée à v .

Afin de définir une permutation $\sigma(v)$ pour chaque sommet d’un graphe, nous commençons par introduire la notion de carbone potentiellement asymétrique, qui correspond à un carbone avec quatre voisins. Un tel sommet correspond à un centre stéréogène si une permutation de deux de ses voisins produit un stéréoisomère différent (Section 1). Les permutations associées à un carbone potentiellement asymétrique sont toutes les permutations paires de ses quatre voisins [7]. Nous pouvons aisément vérifier que les différents ordres obtenus par ces permutations encodent une même configuration, perçue depuis des points de vue différents. Par exemple les permutations paires (1, 4)(2, 3) et (2, 3)(3, 4) appliquées sur l’ordre $H.Cl.Br.F$ du carbone de la molécule située à gauche dans la figure 1, produisent respectivement les ordres $F.Br.Cl.H$ et $H.Br.F.Cl$ qui représentent la même configuration. Pour une double liaison entre deux carbones, les permutations associées à

chaque carbone doivent avoir la même parité. Nous pouvons vérifier que ces permutations correspondent à différentes représentations d’une même configuration. Finalement, pour n’importe quel sommet restant, nous ne cherchons pas à caractériser la configuration spatiale de ses voisins et, associons donc à ces sommets, toutes les permutations possibles de leurs voisins. L’ensemble des fonctions de ré-ordonnancement, qui transforment un graphe ordonné en un graphe représentant la même configuration, est appelé famille valide de fonctions de ré-ordonnancement Σ [4]. Deux graphes ordonnés G et G' sont dit d’ordres équivalents selon Σ ($G \simeq_\Sigma G'$) si il existe $\sigma \in \Sigma$ tel que $\sigma(G) \simeq_o G'$. Cette relation définit une relation d’équivalence [4] et deux stéréoisomères sont encodés par des graphes ordonnés non équivalents. On note $\text{IsomEqOrd}(G, G')$ l’ensemble des isomorphisme d’équivalence d’ordres entre G et G' .

Les carbones potentiellement asymétriques et les carbones liés par une double liaison, ne sont pas forcément des centres stéréogènes. Par exemple, si les labels des sommets Br des deux molécules de la figure 1 étaient remplacés par des Cl, ces molécules seraient identiques. Dans ce cas, n’importe quelle permutation des séquences ordonnées associées aux carbones conduirait à un graphe ordonné d’ordres équivalents. On définit donc un stéréo sommet, comme un sommet pour lequel n’importe quelle permutation de deux de ses voisins produit un graphe ordonné d’ordres non équivalents :

Définition 1 (Stéréo sommet). Soit $G = (V, E, \mu, \nu, ord)$ un graphe ordonné. Un sommet $v \in V$ est appelé stéréo sommet ssi :

$$\forall (i, j) \in \{1, \dots, |N(v)|\}^2, i \neq j, \\ \nexists f \in \text{IsomEqOrd}(G, \tau_{i,j}^v(G)) \text{ avec } f(v) = v. \quad (1)$$

où $\tau_{i,j}^v(G)$ correspond à un graphe ordonné obtenu à partir de G , en permutant les sommets d’indice i et j dans $ord(v)$.

3 Stéréo Sous-Graphe Minimal

La définition 1 se base sur tout le graphe G afin de tester si v est un stéréo sommet. Cependant, si l’on considère un stéréo sommet s , on peut observer que, dans certains cas, la suppression de sommets éloignés de s ne change pas le fait que s soit un stéréo sommet. Dans le but d’obtenir une caractérisation plus locale d’un stéréo sommet, nous devons déterminer un sous-graphe induit par sommet H de G , contenant s , assez gros pour caractériser le fait que s soit un stéréo sommet, mais suffisamment petit pour n’encoder que les informations pertinentes caractérisant le stéréo sommet s . Un tel sous-graphe est appelé un stéréo sous-graphe minimal de s .

Nous présentons maintenant une heuristique, utilisée afin de calculer un stéréo sous-graphe minimal d’un stéréo sommet. Nous nous concentrons d’abord sur le cas des carbones asymétriques. Soit H un sous graphe de G contenant un stéréo sommet s , correspondant à un carbone

asymétrique. On dit que la propriété de stéréoisomérisation de s n'est pas capturée par H si (Définition 1) :

$$\begin{aligned} \exists (i, j) \in \{1, \dots, |N(s)|\}^2, i \neq j, \\ \exists f \in \text{IsomEqOrd}(H, \tau_{i,j}^s(H)) \text{ avec } f(s) = s. \end{aligned} \quad (2)$$

Afin de définir un stéréo sous-graphe minimal de s , on considère une suite finie $(H_s^k)_{k=1}^n$ de sous-graphes induits par sommet de G . Le premier élément de cette suite H_s^1 est le plus petit sous-graphe induit de G pour lequel on peut tester (2) :

$$V(H_s^1) = \{s\} \cup N(s)$$

où $V(H_s^1)$ et $N(s)$ désignent respectivement l'ensemble des sommets de H_s^1 et l'ensemble des voisins de s .

Si le sous-graphe induit actuel H_s^k ne caractérise pas le stéréo sommet s , nous savons par (2), qu'il existe des isomorphismes d'équivalence d'ordres entre H_s^k et $\tau_{i,j}^s(H_s^k)$, avec $i \neq j$. Soit f un de ces isomorphismes. Par définition des isomorphismes d'équivalence d'ordres, il existe $\sigma \in \Sigma$ tel que f soit un isomorphisme entre les graphes ordonnés H_s^k et $\sigma(\tau_{i,j}^s(H_s^k))$. Par définition des isomorphismes entre graphes ordonnés, et comme $f(s) = s$, nous avons :

$$\forall l \in \{1, \dots, |N(s)|\}, f(v_l) = v_{\sigma(s) \circ \tau_{i,j}^s(l)}.$$

avec $\text{ord}(s) = v_1, \dots, v_n$.

Comme $\sigma(s)$ est une permutation paire, $\sigma(s) \circ \tau_{i,j}^s$ est impaire. Il existe donc l dans $\{1, \dots, |N(s)|\}$ tel que $l \neq \sigma(s) \circ \tau_{i,j}^s(l)$.

Autrement dit, n'importe quel isomorphisme d'équivalence d'ordre de l'équation (2) associe au moins un des voisins de s dans H_s^k à un autre de ses voisins. Notons \mathcal{E}_f^k l'ensemble des sommets de H_s^k connecté à s par un chemin dont tous les sommets ne sont pas associés à eux mêmes par f :

$$\begin{aligned} \mathcal{E}_f^k &= \{v \in V(H_s^k) \mid \exists c = (v_0, \dots, v_q) \in H_s^k \\ &\text{avec } v_0 = s \text{ et } v_q = v \text{ t.q.} \\ &\forall r \in \{1, \dots, q\}, f(v_r) \neq v_r\} \end{aligned} \quad (3)$$

Pour tout isomorphisme d'équivalence d'ordres f satisfaisant (2), l'ensemble \mathcal{E}_f^k n'est pas vide car il contient au moins deux sommets. Un sommet v appartient à \mathcal{E}_f^k si, ni son label, ni son voisinage dans H_s^k , ne permet de le différencier de $f(v)$. Le principe de base de notre algorithme consiste à vider les ensembles \mathcal{E}_f^k en ajoutant à H_s^k les voisinages dans G de tous les sommets appartenant à un \mathcal{E}_f^k , avec f satisfaisant (2). L'ensemble des sommets du sous-graphe induit H_s^{k+1} est donc défini par :

$$V(H_s^{k+1}) = V(H_s^k) \cup \bigcup_{f \in \mathcal{F}_s^k} N(\mathcal{E}_f^k) \quad (4)$$

où \mathcal{F}_s^k désigne l'ensemble des isomorphismes d'équivalence d'ordres satisfaisant (2).

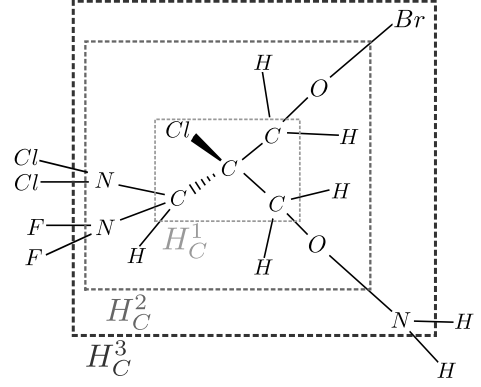


FIGURE 4 – Un carbone asymétrique et sa suite de sous-graphes induits $(H_C^k)_{k=1}^3$

Comme $f \in \mathcal{F}_s^k$ implique que \mathcal{E}_f^k n'est pas vide, ajouter itérativement des contraintes sur l'existence des sommets de \mathcal{E}_f^k permet de retirer f de \mathcal{F}_s^k . L'algorithme s'arrête quand \mathcal{F}_s^k est vide. Le sommet s étant un stéréo sommet, G ne satisfait pas (2), et donc l'algorithme s'arrêtera.

Algorithm 1 Construction d'un stéréo sous-graphe minimal

Input : un stéréo sommet s et un graphe ordonné G

Output : un stéréo sous-graphe minimal

```

 $H_s^1 \leftarrow \{s\} \cup N(s)$ 
 $(\mathcal{F}_s^1, \mathcal{E}_f^1) \leftarrow \text{getIsomorphism}(H_s^1)$ 
 $k \leftarrow 1$ 
while  $\mathcal{F}_s^k \neq \emptyset$  do
   $k \leftarrow k + 1$ 
   $V(H_s^k) \leftarrow V(H_s^{k-1}) \cup N(\mathcal{E}_f^{k-1})$ 
   $(\mathcal{F}_s^k, \mathcal{E}_f^k) \leftarrow \text{getIsomorphism}(H_s^k, \mathcal{F}_s^{k-1})$ 
end while

```

Les principales étapes de notre méthode sont résumées dans l'algorithme 1. La fonction `getIsomorphism` utilise un algorithme d'isomorphisme rapide [1] afin de calculer les isomorphismes f entre H_s^k et $\tau_{i,j}^s(H_s^k)$ ainsi que les ensembles \mathcal{E}_f^k , pour chaque $(i, j) \in \{1, \dots, |N(s)|\}^2$. De plus, dans le but de réduire les temps de calculs, les isomorphismes de \mathcal{F}_s^{k-1} trouvés à l'itération $k-1$ entre H_s^{k-1} et $\tau_{i,j}^s(H_s^{k-1})$, sont utilisés pour initialiser l'algorithme d'isomorphisme à l'itération k .

La figure 4 montre un exemple de stéréo sous-graphe minimal, ainsi que les sous-graphes intermédiaires, trouvés par notre algorithme. On peut remarquer, qu'à l'itération 2, il existe un isomorphisme d'équivalence d'ordres $f \in \mathcal{F}_C^2$ associant le chemin CCO (en bas à droite de la figure) au même chemin situé en haut à droite de la figure. Dans ce cas \mathcal{E}_f^2 contient les carbones de ces chemins ainsi que les deux oxygènes. Les oxygènes appartiennent à \mathcal{E}_f^2 car leur voisinage dans H_C^2 ne permet pas de les différencier. À l'itération suivante, les voisinages dans G de ces atomes sont ajoutés à H_C^3 , ajoutant ainsi un N et un Br , qui per-

mettent de différencier les deux chemins et donc de retirer f de \mathcal{F}_C^3 . On pourra remarquer que les voisinages des carbones sont aussi ajoutés, mais que cela n’a aucune influence car ces voisinages appartenait déjà à H_C^2 .

Pour deux carbones v et w , liés par une double liaison, nous devons calculer un seul stéréo sous-graphe minimal. En effet l’un des carbones ne peut être un stéréo sommet que si le second en est aussi un [4]. Un stéréo sous-graphe minimal de v et w est défini de la même manière que pour un carbone asymétrique, l’unique différence venant de l’initialisation de la suite $(H_{v,w}^k)_{k=1}^n$. En effet le plus petit sous-graphe induit pour lequel on peut tester (2) est défini par l’ensemble de sommets :

$$V(H_{v,w}^1) = \{v, w\} \cup N(v) \cup N(w)$$

4 Noyau Stéréo

Soit G un graphe ordonné, nous pouvons associer à chacun de ses centres stéréogènes un stéréo sous-graphe minimal. Un même stéréo sous-graphe minimal peut être présent plusieurs fois dans une molécule, nous devons donc associer à chaque sous-graphe un code permettant de tester rapidement si deux sous-graphes sont identiques (au sens de l’isomorphisme d’équivalence d’ordres). Pour cela, nous utilisons [8], ce qui nous permet de calculer l’ensemble des stéréo sous-graphes minimaux $\mathcal{H}(G)$ ainsi que le spectre $spec(G) = (f_H(G))_{H \in \mathcal{H}(G)}$ qui contient la fréquence $f_H(G)$ de chaque stéréo sous-graphe minimal H dans G . Contrairement à [1] qui permet de trouver efficacement tous les isomorphismes entre deux graphes, [8] associe à chaque molécule un code unique permettant de tester efficacement si il existe un isomorphisme d’équivalence d’ordres entre deux sous-graphes. L’ensemble $\mathcal{H}(G)$ et le spectre $spec$ fournissent une caractérisation de l’ensemble des centres stéréogènes de G et permettent donc de décrire la stéréoisométrie de G . Afin d’obtenir un noyau entre deux molécules qui prend en compte la stéréoisométrie, nous comparons les spectres des graphes ordonnées associés à ces molécules :

$$k(G, G') = \sum_{H \in \mathcal{H}(G) \cap \mathcal{H}(G')} K(f_H(G), f_H(G')). \quad (5)$$

où K est un noyau entre valeur réelles (p. ex. gaussien ou polynomial). Le choix de ce noyau ainsi que ses paramètres est effectué par une validation croisée.

5 Expérimentations

Nous avons évalué notre noyau sur deux problèmes de régression, où les propriétés à prédire sont liées à la stéréoisométrie.

Le premier jeu de données est composé de molécules acycliques [9]. Il contient 90 molécules ainsi que leur pouvoir rotatoire. Le pouvoir rotatoire d’une molécule est une propriété physique qui mesure l’angle de déviation d’une lumière polarisé passant à travers une solution de

TABLE 1 – Résultats des problèmes de régression.

Méthodes	RMSE	
Noyau Tree pattern [6]	34.1	
Noyau de Treelets [3]	26.2	
Stéréo Tree pattern [2]	24.2	
Noyau Stéréo	15.7	
Méthodes	Moyennes des RMSE	Écarts types des RMSE
Noyau Tree pattern [6]	0.274	0.007
Noyau de Treelets [3]	0.278	0.013
Stéréo Tree pattern [2]	0.193	0.003
Noyau Stéréo	0.210	0.006

cette molécule. En pratique nous ne sélectionnons que 35 molécules, car presque toutes les molécules ne possèdent qu’un seul centre stéréogène, et pour 55 d’entre elles ce centre stéréogène est unique dans le jeu de données, le pouvoir rotatoire de ces molécules ne peut donc pas être prédit correctement à partir du jeu de données. L’écart type du pouvoir rotatoire est égal à 38,25 pour les 35 molécules sélectionnées. À cause du nombre limité de molécules, certaines molécules ont une similarité non nulle avec un nombre restreint de molécules. Ainsi une division en ensembles de test, validation et apprentissage n’est pas possible, car des molécules pourraient avoir une similarité nulle avec toutes les molécules de l’ensemble d’apprentissage. Chaque molécule est donc prédite en utilisant le reste du jeu de données comme ensemble d’apprentissage. Les différents paramètres, tel que le C du SVM ou le type de sous noyau utilisé dans (5), sont choisis grâce à une grille de recherche. Comme il n’y a pas d’ensemble de validation, les paramètres sélectionnés pour chaque méthode sont ceux obtenant la plus petite erreur quadratique. Le haut tableau 1 montre les racines carrées des erreurs quadratiques (RMSE) obtenues par notre noyau, ainsi que celles obtenues par trois autres noyaux [6, 3, 2].

Le second jeu de données contient des dérivés synthétiques de la vitamine D et a été utilisé par [2]. Ce jeu de données est composé de 69 molécules contenant des cycles, avec une moyenne de 9 centres stéréogènes par molécules. La valeur à prédire est leur activité biologique. Ce jeu de données a été choisi car, pour de nombreuses propriétés biologiques, la stéréoisométrie est une caractéristique importante des molécules. Cependant, d’autres propriétés moléculaires, non liées à la stéréoisométrie, peuvent également déterminer partiellement les propriétés biologiques de ce jeu de données.

Après normalisation des valeurs du jeu de données, l’écart type des activités biologiques est de 0.258. Les résultats du bas du tableau 1 correspondent aux moyennes et écarts types des racines carrées des erreurs quadratiques obtenues par chaque méthode pour 3 expérimentations. Pour chacune, nous avons divisé aléatoirement le jeu de données en dix ensembles contenant un même nombre de molécules.

TABLE 2 – Taille des graphes ($|G|$) et des stéréo sous-graphes minimaux ($|H_s|$) pour les deux jeux de données.

	Jeu de données 1		Jeu de données 2	
	$ G $	$ H_s $	$ G $	$ H_s $
Taille minimale	14	6	68	10
Taille maximale	32	13	88	24
Taille moyenne	21.3	10.4	76.9	13.7

Chaque ensemble est ensuite utilisé comme ensemble de test, un autre étant utilisé comme ensemble de validation et finalement les 8 derniers ensembles sont utilisés pour l'apprentissage. Les paramètres des méthodes, tel que le choix d'un sous noyau pour notre méthode, ou la taille des arbres pour le noyau Tree-Pattern sont choisis en utilisant uniquement les ensembles d'apprentissage et de validation et en utilisant une grille de recherche. Cette opération est répétée dix fois pour chaque expérimentations, afin de prédire la valeur de chaque molécule.

Pour les deux jeux de données, les sous noyaux testés pour notre méthode sont les noyaux linéaire, binaire, intersection et gaussien (avec un sigma allant de 0.5 à 4 avec un pas de 0.5). Les différentes valeurs testées pour le paramètre epsilon du SVM sont proportionnelles à l'écart type σ des valeurs à prédire de chaque jeu de données ($\sigma/5$, $\sigma/10$ et $\sigma/20$). Enfin le paramètre C du SVM est testé avec les valeurs 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, . . . , 1000.

On peut voir dans les tableaux 1 que les erreurs les plus élevées sont obtenues par les méthodes n'incluant aucune information concernant la stéréoisométrie [6, 3]. L'adaptation du noyau Tree-Pattern à la stéréoisométrie [2], ainsi que notre méthode, obtiennent de meilleurs résultats que les deux méthodes précédentes. Pour le premier jeu de données, où la propriété à prédire est uniquement liée à la stéréoisométrie, le fait de mélanger sans pondération des informations liées et non liées à la stéréoisométrie pénalise la méthode [2]. Cependant, pour le second jeu de données, où la propriété n'est pas uniquement liée à la stéréoisométrie, l'ajout d'informations non liées à la stéréoisométrie permet à la méthode [2] d'obtenir de meilleurs résultats que ceux obtenus par notre méthode.

On peut voir dans le tableau 2 que sur les deux jeux de données, les stéréo sous-graphes minimaux calculés par notre méthode, sont petits, mais en général plus grand que le voisinage direct d'un centre stéréogène, que se soit un carbone asymétrique (de taille 5) ou une double liaison entre carbones (de taille 6). Les molécules du second jeu de données sont environ trois fois plus grandes que celles du premier (tableau 2). Cependant, la taille moyenne des stéréo sous-graphes minimaux n'augmente que légèrement, passant de 10.4 à 13.7.

Les temps d'exécutions nécessaire pour calculer les matrices de gram de chaque noyau, pour le second jeu de données, sont affichés dans le tableau 3. En effet, comme le premier jeu de données est composés de peu de molécules, ne contenant qu'un centre stéréogène, le temps de calcul

TABLE 3 – Temps d'exécution pour le calcul des matrices de gram de taille 69×69 pour le second jeu de données.

Méthodes	Temps de calcul
	des matrices de gram (s)
Noyau Tree-Pattern [6]	230
Noyau de Treelets [3]	7
Noyau Stéréo	86

de notre méthode est petit, mais peu significatif. La première ligne du tableau 3 montre que pour le noyau Tree-Pattern [6] et son adaptation à la stéréoisométrie [2], il faut 4 minutes afin de calculer les matrices de gram. Ce temps d'exécution élevé peut être expliqué par la complexité polynomiale du calcul du noyau et par le fait que ces noyaux sont basés sur une énumération implicite des sacs de motifs. Ces sacs doivent être calculés à chaque évaluation des noyaux, ils sont donc calculés 69 fois par molécules. À l'inverse, le noyau de treelets [3] est basé sur un algorithme linéaire d'extraction de sacs de motifs, et le sac de treelets associé à chaque molécule est stocké explicitement. Il n'est donc calculé qu'une seule fois par molécules. Cette méthode obtient donc le temps de calcul le plus bas. Finalement, notre noyau stéréo a un temps de calcul intermédiaire de 83 secondes. Nous utilisons pour le calcul du noyau un algorithme d'isomorphisme, cependant nous pouvons voir dans le tableau 2, que les graphes, sur lesquels cet algorithme est utilisé, sont petits. Ceci permet d'obtenir un temps de calcul raisonnable. De plus, comme pour le noyau de treelets, notre noyau est basé sur une énumération explicite de motifs. L'ensemble des stéréo sous-graphes minimaux caractérisant une molécule n'est donc calculé qu'une seule fois par molécule. Ce temps de calcul est en moyenne de 0.65 secondes par molécule pour le second jeu de données.

6 Conclusion

L'étude et la définition de nouveaux stéréoisomères constitue un sous-champ important de la chimie et donc un défi majeur pour la chimoinformatique. Nous avons proposé dans cet article, un noyau sur graphe basé sur une énumération explicite des stéréo sous-graphes d'une molécule. Chaque stéréo sous-graphe est associé à un centre stéréogène et contient la partie du graphe qui confère à un sommet sa propriété de stéréoisométrie. La similarité de deux molécules est alors définie par la similarité de leur sacs de stéréo sous-graphes. Les résultats obtenus par notre méthode, sur deux jeux de données liés à la stéréoisométrie, valident la pertinence de notre approche. La suite de notre travail consistera à étudier ce que peut apporter l'ajout de sous-graphes plus grand que les stéréo sous-graphes afin de représenter les relations entre les stéréo sous-graphes minimaux et le reste de la molécule.

Références

- [1] V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro. A subgraph isomorphism algorithm and its application to biochemical data. *BMC Bioinformatics*, 14(Suppl 7) :S13, 2013.
- [2] J. Brown, T. Urata, T. Tamura, M. A. Arai, T. Kawabata, and T. Akutsu. Compound analysis via graph kernels incorporating chirality. *Journal of Bioinformatics and Computational Biology*, 8(1) :63–81, 2010.
- [3] B. Gaüzère, L. Brun, and D. Villemin. Two New Graphs Kernels in Chemoinformatics. *Pattern Recognition Letters*, 33(15) :2038–2047, 2012.
- [4] P.-A. Grenier, L. Brun, and D. Villemin. Incorporating stereo information within the graph kernel framework. Technical report, CNRS UMR 6072 GREYC, 2013. <http://hal.archives-ouvertes.fr/hal-00809066/>.
- [5] P.-A. Grenier, L. Brun, and D. Villemin. Treelet kernel incorporating chiral information. In *Graph-Based Representations in Pattern Recognition*, pages 132–141. Springer, 2013.
- [6] P. Mahé and J.-P. Vert. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1) :3–35, Oct. 2008.
- [7] M. Petitjean. Chirality in metric spaces. *Symmetry, Culture and Science*, 21 :27–36, 2010.
- [8] W. T. Wipke and T. M. Dyott. Stereochemically unique naming algorithm. *Journal of the American Chemical Society*, 96(15) :4834–4842, 1974.
- [9] H.-J. Zhu, J. Ren, and C. U. Pittman Jr. Matrix model to predict specific optical rotations of acyclic chiral molecules. *Tetrahedron*, 63(10) :2292–2314, 2007.