Graphes de stéréo sous-graphes pour prédire les propriétés moléculaires.

Pierre-Anthony Grenier[1] Luc Brun[1] Didier Villemin [2]

[1] GREYC UMR CNRS 6072
[2] LCMT UMR CNRS 6507

pierre-anthony.grenier@ensicaen.fr

Résumé

L'étude des relations quantitatives structure-activité (QSAR) ou structure-propriété (QSPR) visent à prédire des propriétés de molécules à l'aide de méthodes informatiques. 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. Dans notre article précédent nous proposions d'encoder la propriété de stéréoisomérie de chaque atome par un sous-graphe. Un noyau entre sacs de ces sous-graphes nous donnait alors une mesure de similarité prenant en compte la stéréoisomérie. Cependant, cette approche ne prend pas en compte les éventuels recouvrements entre ces sous-graphes. Nous proposons donc dans cet article, un nouveau noyau basé sur un codage explicite des relations de recouvrement entre sous-graphes.

Mots Clef

Noyau sur graphe, chémoinformatique, stéréoisomérie.

Abstract

Quantitative Structure Activity and Property Relationships (QSAR and QSPR), aim to predict properties of molecules thanks to computational techniques. In these fields, graphs provide a natural encoding of molecules. However some molecules may have a same graph but differ by the three dimensional orientation of their atoms in space. These molecules, called stereoisomers, may have different properties which cannot be correctly predicted using usual graph encodings. In a previous paper we proposed to encode the stereoisomerism property of each atom by a local subgraph. A kernel between bags of such subgraphs then provides a similarity measure incorporating stereoisomerism properties. However, such an approach does not take into account potential interactions between these subgrahs. We thus propose in this paper, a method to take these interactions into account hence providing a global point of view on molecules's stereoisomerism properties.

Keywords

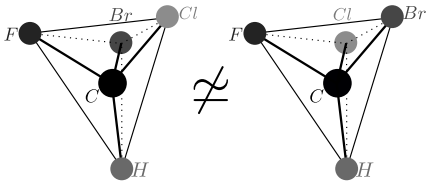
Graph kernel, Chemoinformatics, Stereoisomerism.

1 Introduction

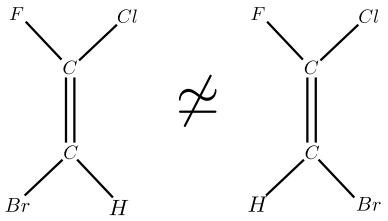
Les méthodes visant à prédire les propriétés de molécules sont basées sur le principe de similarité qui stipule que : « deux molécules similaires doivent avoir des propriétés similaires ». Une molécule est souvent représentée par son graphe moléculaire. Un graphe moléculaire est un graphe simple $G = (V, E, \mu, \nu)$, où chaque nœud $v \in V$ encode un atome, chaque arête $e \in E$ encode une liaison entre deux atomes, la fonction μ associe à chaque nœud un label identifiant la nature de l'atome (carbone, oxygène, ...) qu'il représente et la fonction ν associe à une arête le type de lien (simple, double, triple ou aromatique) de la liaison représentée.

Cependant, les graphes moléculaires ont une limite : ils ne codent 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. Nous pouvons imaginer par exemple, un atome de carbone avec quatre voisins possédant des labels différents, 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 1a). Un atome est appelé centre stéréogène si la permutation de deux de ses voisins crée un différent 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 voisinages, crée un stéréoisomère différent (Figure 1b). Parmi les molécules actuellement utilisées en chimie, 98% des centres stéréogène sont, soit des carbones avec quatre voisins, appelés carbones asymétriques (Figure 1a), soit des couples de deux carbones liés par une liaison double (Figure 1b) [7]. Nous limitons notre étude à ces deux cas.

Les noyaux sur graphes [8, 9, 10, 3], correspondent à une mesure de similarité entre graphes. Jusqu'à présent, seules



(a) Deux configurations spatiales différentes des voisins d'un carbone.



(b) Deux configurations spatiales différentes des voisins de deux carbones liés par une liaison double.

FIGURE 1 – Deux types de centres stéréogènes.

quelques méthodes ont essayé de construire des noyaux sur graphes prenant en compte la stéréoisomérie. Brown et al. [1] ont proposé d'inclure la stéréoisomérie dans une extension du noyau de motifs d'arbres [9]. Dans cette méthode, la similarité entre des molécules est déduite du nombre de motifs d'arbres commun à deux molécules.
Intuitivement, la stéréoisomérie est liée au fait que la permutation de deux voisins d'un centre stéréogène produit une configuration spatiale différente. Si ces deux voisins ont une même étiquette, l'influence de la permutation doit être cherchée au-delà du voisinage direct du centre stéréogène. En se basant sur cette constatation, nous avons proposé [5] de caractériser un centre stéréogène par un sous-graphe, appelé stéréo sous-graphe minimal. Ce sous-graphe est assez grand pour caractériser l'influence de chaque permutation des voisins du centre stéréogène, mais suffisamment petit pour garder une caractérisation locale. Nous avons ensuite proposé un noyaux basé sur ces sous-graphes.
Un inconvénient de cette approche est que chaque stéréo sous-graphe minimal est considéré indépendamment des autres. Ce qui n'est pas le cas de la méthode de [1], où un motif peut implicitement encoder un chemin qui relie plusieurs centres stéréogènes proches. Cependant, les motifs d'arbres ont une taille limitée par un paramètre. Ainsi le noyau de motif d'arbres peut ne pas capturer l'ensemble des informations définissant la stéréoisomérie.
Dans cet article, nous proposons une méthode basée sur [5], qui encode explicitement les recouvrements entre les stéréo sous-graphes minimaux. Dans la section 2 nous rappelons les deux points principaux de [5], la représentation des molécules par des graphes localement ordonnés et la construction des stéréo sous-graphes minimaux. Puis dans la section 3 nous présentons un nouveau modèle basé sur des graphes de recouvrements. Les résultats obtenus avec cette méthode sont présentés dans la section 4.

2 Graphes Localement Ordonnés et Stéréo sous-graphes minimaux

2.1 Graphes localement ordonnés appliqués aux molécules

La configuration spatiale des voisins de chaque atome peut être encodée par une séquence ordonnée de ses voisins [5]. Pour représenter cette informations, nous utilisons la notion de graphes localement ordonnés. Un graphe localement ordonné $G = (V, E, \mu, \nu, ord)$ est un graphe moléculaire $G_m = (V, E, \mu, \nu)$ avec une fonction $ord : V \to V^*$ qui associe à chaque sommet une liste ordonnée de ses voisins. Deux graphes localement ordonnés G et G' sont isomorphes ($G \underset{o}{\simeq} G'$) s'il existe un isomorphisme f entre leurs graphes moléculaires respectifs 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). Dans ce cas, f est appelé un isomorphisme localement ordonné entre G et G'.

Cependant, des graphes localement ordonnés différents peuvent représenter une même molécule. Nous devons donc définir une relation d'équivalence entre les graphes localement ordonnés, afin que deux graphes localement ordonnés soient équivalents s'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.

L'ensemble des fonctions de ré-ordonnancement, qui transforment un graphe localement ordonné en un graphe représentant la même configuration, est appelé famille valide de fonctions de ré-ordonnancement Σ [4]. Deux graphes localement ordonnés G et G' sont dit d'ordres équivalents selon Σ ($G \underset{\Sigma}{\simeq} G'$) s'il existe $\sigma \in \Sigma$ tel que $\sigma(G) \underset{o}{\simeq} G'$. Cette relation définit une relation d'équivalence [4] et deux stéréoisomères sont encodés par des graphes localement ordonnés non équivalents. On note $\text{IsomEqOrd}(G, G')$ l'ensemble des isomorphismes d'équivalence d'ordres entre G et G'.

Les carbones possédant quatre voisins et les carbones liés par une double liaison, ne sont pas forcément des centres stéréogènes. Dans ce cas, n'importe quelle permutation des séquences ordonnées associées aux carbones conduirait à un graphe localement ordonné d'ordres équivalents. On définit donc pour un graphe localement ordonné $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ et l'un de ses sommets $v \in V$ l'ensemble des isomorphismes $\mathcal{F}_G^v$ entre G et les graphes

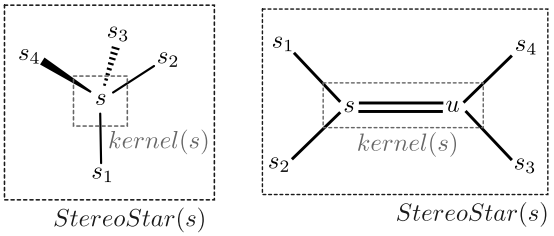


FIGURE 2 – Exemple d'ensembles de stéréo sommets.

obtenus aprés avoir permuté deux voisins de v :

$$\mathcal{F}_G^v = \bigcup_{\substack{(i,j)\in\{1,\dots,|N(v)|\}^2 \\ i\neq j}} \{f \mid f \in \text{IsomEqOrd}(G, \tau_{i,j}^v(G)) \text{ avec } f(v) = v\}$$

où $\tau_{i,j}^v$ est une fonction de réordonnancement égale à l'identité sur tous les sommets à l'exception de v pour laquelle elle permute les sommets d'indice i et j. Intuitivement, ces isomorphismes représentent une symétrie des voisins de v.
On définit alors un stéréo sommet comme un sommet pour lequel n'importe quelle permutation de ses voisins produit un graphe localement ordonné non équivalent :

Définition 1 (Stéréo sommet)**.** Soit $G = (V, E, \mu, \nu, ord)$ un graphe localement ordonné. Un sommet $v \in V$ est appelé stéréo sommet ssi $\mathcal{F}_G^v = \varnothing$.

Deux carbones liés par une liaison double peuvent être un centre stéréogènes et nous avons montrés dans [6] que si l'un des carbones est un stéréo sommet, alors l'autre carbone de la liaison double est aussi un stéréo sommet. Nous introduisons les notations suivantes (illustrées dans la Figure 2) :

Définition 2 (Ensemble de stéréo sommets liés)**.** Soit s un stéréo sommet. On définit son ensemble de stéréo sommets liés $kernel(s)$ par $\{s\}$ si s est un carbone avec quatre voisins et $\{s, n_=(s)\}$ si s est un carbone d'une liaison double, où $n_=(s)$ est l'autre carbone de la liaison double.

Définition 3 (Étoile de stéréo sommets)**.** Pour s un stéréo sommet on définit l'ensemble $StereoStar(s)$ par :

$$StereoStar(s) = kernel(s) \cup N(kernel(s))$$

2.2 Stéréo sous-graphes minimaux

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 grand pour caractériser le fait que s soit un stéréo sommet, mais suffisamment petit pour n'encoder que les informations pertinentes caractérisant la stéréo propriété de s. Un tel sous-graphe est appelé un stéréo sous-graphe minimal de s.
Nous présentons maintenant une définition constructive des stéréo sous-graphes minimaux. Soit s un stéréo sommet et H_s un sous-graphe de G contenant $kernel(s)$. On dit que la propriété de stéréoisomérie de s n'est pas capturée par H_s si (Definition 1) :

$$\mathcal{F}_{H_s}^s \neq \varnothing \quad (1)$$

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 (1) : $V(H_s^1) = StereoStar(s)$.
Si le sous-graphe induit actuel H_s^k ne caractérise pas le stéréo sommet s, nous savons par (1), qu'il existe des isomorphismes d'équivalence d'ordres $f \in \mathcal{F}_{H_s^k}^s$. On note $\mathcal{E}_f^k$ l'ensemble des sommets de H_s^k induisant l'isomorphisme f dans H_s^k :

$$\mathcal{E}_f^k = \{v \in V(H_s^k) \mid \exists p = (v_0, \dots, v_q) \in H_s^k \text{ avec } v_0 \in kernel(s) \text{ et } v_q = v \text{ s.t. } f(v_1) \neq v_1\} \quad (2)$$

Dans [6], nous montrons que pour n'importe quel isomorphisme f de $\mathcal{F}_{H_s^k}^s$, $\mathcal{E}_f^k$ n'est pas vide. Intuitivement, un sommet v appartient à $\mathcal{E}_f^k$ si ni son étiquette, ni son voisinage dans H_s^k ne permet de le différencier de $f(v)$. L'idée principale de notre algorithme est de vider les ensembles $\mathcal{E}_f^k$ en ajoutant à H_s^k les voisinages dans G de tous les sommets appartenant à un $\mathcal{E}_f^k$. L'ensemble des sommets du sous-graphe induit H_s^{k+1} est donc défini par :

$$V(H_s^{k+1}) = V(H_s^k) \cup \bigcup_{f\in\mathcal{F}_{H_s^k}^s} N(\mathcal{E}_f^k) \quad (3)$$

où $N(\mathcal{E}_f^k)$ désigne le voisinage de $\mathcal{E}_f^k$.
L'algorithme s'arrête quand $\mathcal{F}_s^k$ est vide. Nous avons prouvé dans [6] que le sous-graphe obtenu par cet algorithme capture la propriété de stéréoisomérie de s. La figure 3 illustre cet algorithme. Le calcul des stéréo sous-graphes minimaux demande le calcul d'isomorphismes entre graphes ce qui est un problème NP. Cependant, les stéréo sous-graphes minimaux correspondent à une caractérisation locale des sommets et ont par conséquent en pratique une petite taille [5].
Ainsi, pour chaque stéréo sommet, on peut construire un stéréo sous-graphe minimal qui le caractérise. On considère deux stéréo sommets comme similaires si leurs stéréo sous-graphes minimaux sont identiques. Pour le tester efficacement nous transformons un stéréo sous-graphe minimal S en un code c_S grâce à la méthode [12].

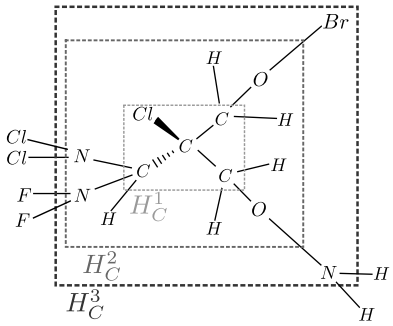


FIGURE 3 – Un carbone asymétrique et sa suite de sous-graphes induits $(H_C^k)_{k=1}^3$

3 Graphes de recouvrements

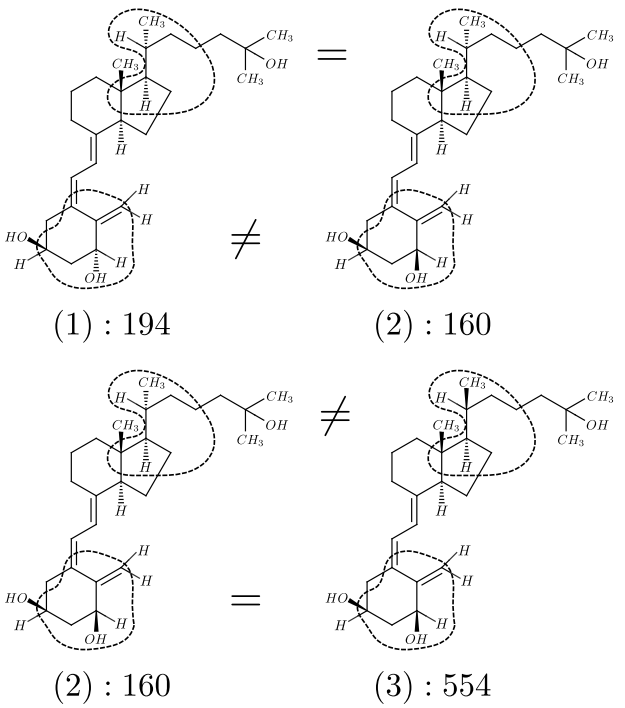


FIGURE 4 – Trois molécules avec leurs activités biologiques. Leurs stéréo sous-graphes minimaux différents sont entourés par des pointillés.

Dans la section précédente nous avons défini un moyen d'encoder les molécules et construit un sous-graphe permettant de caractériser un centre stéréogène. On associe alors à un graphe localement ordonné G son sac de stéréo sous-graphes minimaux $\mathcal{H}(G)$. Dans [5], nous avons proposé un noyau entre ces sacs qui donne une mesure de similarité entre graphes localement ordonnés :

$$k(G, G') = \sum_{H \in \mathcal{H}(G) \cap \mathcal{H}(G')} K(f_H(G), f_H(G')). \quad (4)$$

où $f_H(G)$ est le nombre d'occurences du stéréo sous-graphe minimal H dans G et K est un noyau entre valeur réelles (par exemples gaussien ou polynomial). Cependant, en utilisant ce noyau, chaque stéréo sous-graphe minimal est considéré indépendamment.

La figure 4 montre un exemple de trois molécules d'un jeu de donné utilisé dans la section 4. Dans cette figure, (2) a seulement un stéréo sous-graphe différent avec (1) et avec (3). Ainsi, si l'on considère la notion de distance associée au noyau de [5], (1) et (3) sont équidistants de (2). En effet, l'équation 4 nous donne $k(G_1, G_2) = k(G_2, G_3) = 1$ pour un noyau K polynomial (avec G_j le graphe associé à la molécule (j) de la figure 4). Cependant, l'activité biologique de (2) est plus proche de celle de (1) que de celle de (3). Une régression appliquée sur la molécule (2) en utilisant le noyau [5] fournirait donc a priori une moyenne non pondérée entre l'activité biologique de (1) et celle de (3). Le noyau [5] ne prend donc pas en compte l'ensemble des informations relatives à la stéréoisomérie. L'hypothèse formulé dans cet article et validée en Section 4, est qu'une partie de ces informations concerne les relations de recouvrement entre les stéréo sous-graphes.

Malheureusement, la définition de la quantité de recouvrement entre deux stéréo sous-graphes minimaux, pouvant influer sur une propriété, est un problème ouvert aussi bien en chimie qu'en informatique. Nous proposons donc de définir plusieurs fonctions de recouvrements.

Les fonctions de recouvrements sont définies à l'aide d'un ensemble de conditions $(c_1, \ldots, c_n)$. Ces conditions sont de plus en plus contraignantes :

$$\forall i \in \{1, \ldots, n-1\}\, c_{i+1} \Rightarrow c_i$$

Soit S_1 et S_2 deux stéréo sous-graphes minimaux ayant respectivement pour stéréo sommets s_1 et s_2. On considère l'ensemble de conditions :

$$\begin{array}{l} c_1(S_1, S_2) : S_1 \bigcap S_2 \neq \emptyset \\ c_2(S_1, S_2) : kernel(s_1) \subset S_2 \\ c_3(S_1, S_2) : StereoStar(s_1) \subset S_2 \\ c_4(S_1, S_2) : S_1 \subset S_2 \end{array} \quad (5)$$

On considère dans cet article trois fonctions de recouvrement F_i. Chaque fonction F_i est définie de manière à être plus restrictive que F_j (avec $j < i$). Pour cela F_i est définie en utilisant uniquement les conditions c_j avec j appartenant à $\{i, \ldots, 4\} \cup \{0\}$, c_0 est définie par $\neg c_i$. Cette condition décrit l'absence de recouvrement. La valeur $F_i(H_1, H_2)$ est obtenue en prenant l'indice maximal j tel que la condition $c_j(S_1, S_2)$ soit vraie :

$$F_i(S_1, S_2) = \max\{j \in \{i, \ldots, n\} \cup \{0\} \mid c_j(S_1, S_2)\}$$

On peut remarquer que les fonctions de recouvrements ne sont pas symétriques. Ainsi, pour caractériser les recouvrements entre deux stéréo sous-graphes minimaux S_1 et S_2, il faudra considérer $F_i(S_1, S_2)$ et $F_i(S_2, S_1)$.

On utilise les fonctions de recouvrements pour construire trois graphes de recouvrements G_i. Dans ces graphes, chaque sommet $v \in V_i$ représente un stéréo sous-graphe minimal, et les arcs représentent les recouvrements entre les stéréo sous-graphes.

Définition 4 (Graphe de recouvrements)**.** Soit $G = (\hat{G} = (V, E, \mu, \nu), ord)$ un graphe moléculaire localement ordonné. Un graphe de recouvrements non orienté $G_i = (V_i, E_i, \mu_i, \nu_i)$ de G est un graphe non orienté tel que :

— Il existe une bijection S qui associe à chaque sommet u de V_i un unique stéréo sous-graphe minimal $S(u) \in \mathcal{H}(G)$.
— $\forall u \in V_i, \mu_i(u) = c_{S(u)}$.
— $E_i = \{\{u_1, u_2\} \in \mathcal{P}_2(V_i) | F_i(S(u_1), S(u_2)) \neq 0 \vee F_i(S(u_2), S(u_1)) \neq 0\}$.
— $\forall e = (u_1, u_2) \in E_i$, $\nu_i(e) = \min(F_i(S_1, S_2), F_i(S_2, S_1)) \odot \max(F_i(S_1, S_2), F_i(S_2, S_1))$. Avec $S_1 = S(u_1)$ et $S_2 = S(u_2)$.

où $c_{S(u)}$ est le code obtenu par l'algorithme de dénomination [12] et utilisé afin d'identifier le stéréo sous-graphe minimal $S(u)$. L'opérateur $\odot$ désigne la concaténation.

Vérifier si un sommet appartient à un stéréo sous-graphe minimal est fait en temps constant. Ainsi, les complexités des temps de calcul des quatre conditions (5) de $F_1(S_1, S_2)$ sont respectivement $\mathcal{O}(max(|S_1|, |S_2|))$, $\mathcal{O}(|kernel(s_1)|)$, $\mathcal{O}(|StereoStar(s_1)|)$ et $\mathcal{O}(|S_1|)$. La complexité en temps de calcul de la construction des graphes de recouvrement est donc égale à :

$$\mathcal{O}(|\mathcal{H}(G)|^2 \max_{H \in \mathcal{H}(G)} |H|)$$

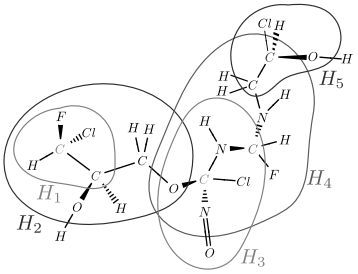
En pratique, cette valeur est petite (pour le jeu de données des dérivés synthétiques de la vitamine D présenté dans la section 4, nous avons au plus $|\mathcal{H}(G)| = 9$ et $\max_{H \in \mathcal{H}(G)} |H| = 24$).

La Figure 5 montre les différents graphes de recouvrements obtenus en prenant les différentes fonctions de recouvrements.

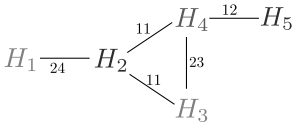
Pour le premier graphe de recouvrements G_1 on considère que quatre niveaux de recouvrements sont possibles. Pour ce graphe, la condition minimale pour que deux stéréo sous-graphes minimaux interagissent est que leur intersection ne soit pas vide. Le recouvrement d'un stéréo sous-graphe minimal S_1 par un autre stéréo sous-graphe minimal S_2 est considéré comme plus fort si le stéréo sommet de S_2 est inclus dans S_1. Le troisième niveau de recouvrement a lieu si le stéréo sommet de S_2 et son voisinage sont inclus dans S_1. Finalement, le niveau de recouvrement est considéré comme maximal lorsque S_2 est inclus dans S_1.

Cependant, certains de ces niveaux de recouvrements peuvent être considérés comme peu pertinents. On peut supposer que le fait d'avoir une intersection non vide n'est pas suffisant pour dire que deux stéréo sous-graphes minimaux interagissent. Le graphe G_2 est donc construit en considérant que la condition minimale pour que deux stéréo sous-graphes minimaux interagissent est que le stéréo sommet de l'un soit inclus dans le stéréo sous-graphe minimal de l'autre.

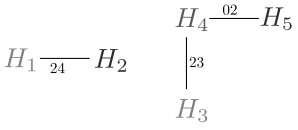
Un atome est un centre stéréogène à cause du positionnement relatif de ses voisins. Si un stéréo sommet est inclus



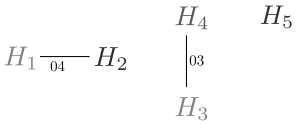
(a) Un graphe moléculaire localement ordonné et ses stéréo sous-graphes minimaux



(b) G_1



(c) G_2



(d) G_3

FIGURE 5 – Exemple de graphe moléculaire localement ordonné et des trois graphes de recouvrements obtenus en utilisant les trois fonctions de recouvrements.

dans un stéréo sous-graphe minimal ($kernel(s_2) \in S_1$), mais pas son voisinage ($StereoStar(s_2) \notin S_1$), on peut alors supposer que ce stéréo sommet a la même influence qu'un sommet qui n'est pas stéréo. Ainsi le graphe de recouvrements G_3 est construit sans considérer que la condition « le stéréo sommet de S_2 est inclus dans S_1 »est différente de la condition « l'intersection de S_1 et S_2 est non vide ». Ainsi pour ce graphe, la condition minimale de recouvrement entre deux stéréo sommets est que le premier stéréo sommet et son voisinage soient inclus dans le stéréo sous-graphe minimal du second.

Les graphes de recouvrements ne sont pas des graphes localement ordonnés, on peut donc utiliser n'importe quel noyau usuel sur graphes (par exemple [9, 10, 3]) afin de mesurer leurs similarités. En considérant le noyau de treelets [3], les treelets de taille 1 correspondent aux sommets des graphes de recouvrements. Ils encodent donc exactement la même information que le noyau stéréo présenté dans [5].

Notons que les graphes de recouvrements ne prennent en compte que le la stéréoisomérie. Les atomes qui ne sont présent dans aucun stéréo sous graphe minimal ne sont pas pris en compte par cette représentation.

TABLE 1 – Moyennes du nombre de sommets ($\overline{|V|}$), nombre d'arêtes ($\overline{|E|}$), nombre d'étiquettes différentes ($\overline{|\mathcal{L}_V|}$,$\overline{|\mathcal{L}_E|}$), des degrés moyens ($\overline{d}$) des graphes de recouvrements.

(a) Jeu de données des stéréoisomères de la périndoprilate.

	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	$\overline{\lvert\mathcal{L}_V\rvert}$	$\overline{\lvert\mathcal{L}_E\rvert}$	$\overline{d}$
Graphe 1	5	7	4.5	3	2.8
Graphe 2	5	2	4.5	2	0.8
Graphe 3	5	1	4.5	1	0.4

(b) Jeu de données des dérivés synthétiques de la vitamine D.

	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	$\overline{\lvert\mathcal{L}_V\rvert}$	$\overline{\lvert\mathcal{L}_E\rvert}$	$\overline{d}$
Graphe 1	8.55	17.4	8.38	5.71	4.07
Graphe 2	8.55	11.3	8.38	4.71	2.62
Graphe 3	8.55	6.14	8.38	2.71	1.43

4 Expérimentations

Afin de prouver l'intérêt des graphes de recouvrements, nous les avons utilisés pour deux problèmes, un de classification et un de régression. Pour les deux jeux de données nous utilisons le même protocole : deux validations croisées imbriquées afin de sélectionner les paramètres et d'estimer les performances de chaque noyau. La validation croisée externe est une procédure de « leave-one-out », servant à calculer une erreur pour chaque molécule du jeu de données. Nous utilisons une autre procédure de « leave-one-out »sur les molécules restantes, afin de calculer une erreur de validation. Les paramètres obtenant la plus petite erreur quadratique sur l'ensemble de validation sont sélectionnés. Nous utilisons un SVM pour chacun des problèmes.

Le tableau 1 contient plusieurs informations concernant les graphes de recouvrements obtenus avec ces jeux de données. On peut notamment remarquer que le degré moyen ($\overline{d}$) des sommets varie beaucoup selon les fonctions de recouvrements utilisées pour construire le graphe de recouvrements.

Le premier problème est fondé sur un jeu de données contenant tous les stéréoisomères de la périndoprilate et issu de [2]. Comme cette molécule contient 5 centres stéréogènes, le jeu de données est composé de $2^5 = 32$ molécules. Dans ce jeu de données nous tentons de prédire si une molécules inhibe l'enzyme de conversion de l'angiotensine.

Pour ces molécules, deux stéréo sommets ont des stéréo sous-graphes minimaux identiques ou opposés (c'est-à-dire qui n'ont comme différence que l'ordre des voisins de leurs stéréo sommets). Ces stéréo sommets, lorsqu'ils sont identiques sont donc confondus par le noyau [5]. Toutefois, les stéréo sous-graphes minimaux associés à ces deux stéréo sommets ont des voisinages différents. Le noyau de motifs d'arbre adapté à la stéréoisomérie [1] est capable de distinguer ces deux stéréo sommets par leur voisinage,

TABLE 2 – Classification des stéréoisomères de la périndoprilate en fonction de leur action sur l'enzyme de conversion de l'angiotensine.

	Méthodes	Précision
1 -	Noyau de motifs d'arbres stéréo [1]	*96.875*
2 -	Noyau stéréo [5]	87.5
Graphes de recouvrements avec [3]		
3 -	Graphes de recouvrements 1	93.75
4 -	Graphes de recouvrements 2	93.75
5 -	Graphes de recouvrements 3	84.375
Graphes de recouvrements avec [9]		
6 -	Graphes de recouvrements 1	93.75
7 -	Graphes de recouvrements 2	62.5
8 -	Graphes de recouvrements 3	62.5
Graphes de recouvrements avec [10]		
9 -	Graphes de recouvrements 1	93.75
10 -	Graphes de recouvrements 2	62.5
11 -	Graphes de recouvrements 3	62.5
Graphes de recouvrements avec [3] et [11]		
12 -	Graphes de recouvrements 1	**100**
13 -	Graphes de recouvrements 2	87.5
14 -	Graphes de recouvrements 3	90.625

et obtient donc une meilleure précision que le noyau stéréo de [5]. Comme le noyau stéréo, le graphe de recouvrements G_3 n'est pas capable de discerner les deux stéréo sous-graphes minimaux et obtient donc une mauvaise précision. En appliquant le noyau de treelets aux deux autres graphes de recouvrements G_1 et G_2 nous obtenons de bons résultats, mais cependant moins bons que les résultats obtenus par [1]. Cela peut être expliqué par le fait que les treelets de taille 1 donnent la même information que le noyau stéréo. En utilisant un algorithme d'apprentissage à noyaux multiples [11], on peut calculer un poids optimal pour chaque treelet lors de l'apprentissage ce qui nous permet d'obtenir une précision parfaite avec le graphe de recouvrements G_1. On peut voir dans le Tableau 1a que les seconds graphes de recouvrements ne possèdent que deux arêtes pour cinq sommets et donc un petit degré. Ce point explique pourquoi on obtient de mauvais résultats avec le second graphe de recouvrements et les noyaux [3, 9, 10].

Le second jeu de données utilisé pour valider l'approche utilisant les graphes de recouvrements est un jeu de données des dérivés synthétiques de la vitamine D issu de [1]. Ce jeu de données est composé de 69 molécules avec en moyenne 8.55 centre stéréogène par molécule. Ce jeu de données est associé à un problème de régression où l'on doit prédire l'activité biologique des molécules.

Les méthodes ne prenant pas en compte la stéréoisomérie [9, 3] obtiennent de mauvais résultats comme on peut le voir dans la Table 3 (lignes 1-2). L'adaptation du noyau de motifs d'arbres à la stéréoisomérie [1] et notre méthode précédente [5] (lignes 3-4) améliorent ces résultats, montrant l'intérêt de prendre en compte la stéréoisomérie.

TABLE 3 – Prédiction de l'activité biologique des dérivés synthétiques de la vitamine D.

	Méthodes	RMSE
1 -	Noyau de motifs d'arbres [9]	0.251
2 -	Noyau de treelets [3]	0.271
3 -	Noyau de motifs d'arbres stéréo [1]	0.184
4 -	Noyau stéréo [5]	0.194
Graphes de recouvrements avec [3]		
5 -	Graphes de recouvrements 1	0.177
6 -	Graphes de recouvrements 2	0.169
7 -	Graphes de recouvrements 3	0.171
Graphes de recouvrements avec [9]		
8 -	Graphes de recouvrements 1	0.185
9 -	Graphes de recouvrements 2	*0.162*
10 -	Graphes de recouvrements 3	**0.161**
Graphes de recouvrements avec [10]		
11 -	Graphes de recouvrements 1	0.201
12 -	Graphes de recouvrements 2	0.166
13 -	Graphes de recouvrements 3	*0.162*

La prise en compte des recouvrements entre stéréo sous-graphes minimaux (lignes 5-13) nous permet d'obtenir des résultats meilleurs que ceux obtenus avec notre méthode précédente [5]. Pour ce jeu de données, l'utilisation des graphes de recouvrements G_2 et G_3 permet d'obtenir les meilleurs résultats. Contrairement au premier jeu de données, on peut voir que les résultats obtenus en utilisant le graphe de recouvrements G_1 (lignes 5, 8 et 11), sont moins bons que ceux obtenus avec les deux autres graphes de recouvrements. Dans le Tableau 1b, on peut voir que les graphes de recouvrements G_1 ont un degré moyen de 4 pour 8.55 sommets en moyenne. Ainsi les sommets de ces graphes de recouvrements sont liés en moyenne à la moitié du graphe. De plus il y a en moyenne 8.38 labels différents pour les 8.55 sommets. Ainsi quasiment tous les sommets ont des labels différents. Ceci crée de nombreux motifs uniques pour chaque molécule, ce qui explique pourquoi les résultats sont moins bons en utilisant ce graphe.
En conclusion, l'utilisation des graphes de recouvrements permet d'obtenir de meilleurs résultats que notre précédente méthode. De plus, nos expériences montrent que le choix du graphe de recouvrement à utiliser dépend du degré moyen de ses sommets. Un degré moyen de 2 est un bon compromis entre un degré moyen trop petit, correspondant à un ensemble de sommets sans structure de graphe, et un degré moyen trop élevé correspondant à un graphe complet également sans structure.

5 Conclusion

Dans cet article, nous avons proposé une extension de notre précédente méthode [5] qui permet de prendre en compte les recouvrements entre les stéréo sous-graphes minimaux. Au lieu de comparer des ensembles de stéréo sous-graphes minimaux, on compare des graphes de stéréo sous-graphes minimaux. Ainsi, un stéréo sous-graphe minimal n'est plus considéré indépendamment de sa position par rapport aux autres stéréo sous-graphes minimaux. La pertinence de cette approche est démontrée sur deux jeu de données.

Remerciements

Ce travail a été effectué en utilisant les ressources informatiques partiellement financé par le CPER Normandie.

Références

[1] 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.

[2] J. A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, and R. Rotondo. Atom-based stochastic and non-stochastic 3d-chiral bilinear indices and their applications to central chirality codification. *Journal of Molecular Graphics and Modelling*, 26(1) :32–47, 2007.

[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. A graph kernel incorporating molecule's stereisomerism information. *Proceedings of ICPR 2014*, 2014.

[6] P.-A. Grenier, L. Brun, and D. Villemin. Taking into account interaction between stereocenters in a graph kernel framework. Technical report, CNRS UMR 6072 GREYC, 2014. https ://hal.archives-ouvertes.fr/hal-01103318.

[7] J. Jacques, A. Collet, and S. Wilen. *Enantiomers, racemates, and resolutions*. Krieger Pub. Co., 1991.

[8] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *ICML*, volume 3, pages 321–328, 2003.

[9] P. Mahé and J.-P. Vert. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1) :3–35, Oct. 2008.

[10] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-lehman graph kernels. *The Journal of Machine Learning Research*, 12 :2539–2561, 2011.

[11] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.

[12] W. T. Wipke and T. M. Dyott. Stereochemically unique naming algorithm. *Journal of the American Chemical Society*, 96(15) :4834–4842, 1974.