Représentation des cycles d'une molécule sous forme d'hypergraphe

Benoit Gaüzère, Luc Brun, Didier Villemin

Représentation des cycles d'une molécule sous forme d'hypergraphe.

B. Gaüzère[1] L. Brun[1] D. Villemin[2]

[1] GREYC, UMR 6072 CNRS, Caen, France
[2] LCMT, UMR 6507 CNRS, Caen, France

6 Boulevard Maréchal Juin
14 050 CAEN CEDEX
{benoit.gauzere,luc.brun,didier.villemin}@ensicaen.fr

Résumé

La chémoinformatique utilise des méthodes issues de la théorie des graphes et de l'apprentissage automatique afin de classifier ou prédire des propriétés moléculaires. De ce point de vue, les noyaux sur graphes constituent une approche intéressante combinant les méthodes d'apprentissage et la représentation naturelle des molécules sous forme de graphes. Cependant, bien que les graphes moléculaires encodent l'ensemble de l'information structurelle des molécules, ils n'encodent pas explicitement l'information cyclique. Dans cet article, nous proposons de représenter une molécule par un hypergraphe encodant explicitement à la fois l'information cyclique et acyclique d'une molécule dans une même représentation. Nous proposons également une mesure de similarité sous forme de noyau afin d'utiliser cette représentation moléculaire dans des problèmes rencontrés en chémoinformatique.

Mots Clef

chémoinformatique, noyau sur graphe, cycles pertinents.

Abstract

Chemoinformatics aims to predict molecule's properties through informational methods. Some methods base their prediction model on the comparison of molecular graphs. Considering such a molecular representation, graph kernels provide a nice framework which allows to combine machine learning and graph theory. Despite the fact that molecular graphs encode most of the structural information of a molecule, it does not explicitly encode the cyclic information. In this paper, we propose a new molecular representation based on a hypergraph which explicitly encodes both cyclic and acyclic information into one molecular representation. We also propose a similarity measure defined as a kernel in order to resolve chemoinformatics prediction problems using our new molecular representation.

Keywords

graph kernel, chemoinformatics, relevant cycles.

1 Introduction

La chémoinformatique, et plus particulièrement les approches QSAR [10] (Quantitative Structure-Activity Relationship), consiste à prédire des propriétés moléculaires à partir de mesures de similarité entre molécules. Un grand nombre de méthodes, appelées empreintes digitales, représentent chaque molécule par un ensemble de descripteurs. La similarité entre deux molécules est alors déduite de la similarité de deux ensembles de descripteurs associés à chaque molécule. D'autres approches consistent à utiliser une représentation des molécules sous forme de graphe moléculaire $G = (V, E, \mu, \nu)$ où le graphe non-étiqueté (V, E) encode la structure de la molécule tandis que les fonctions d'étiquetage μ et ν associent respectivement chaque nœud et chaque arête à un élément chimique et un type de liaison atomique (simple, double, triple ou aromatique) connectant deux atomes.

En considérant cette représentation, la similarité entre deux molécules peut être déduite de la similarité de leurs graphes moléculaires. Les noyaux sur graphes correspondent à des mesures de similarité symétriques. En utilisant un noyau défini semi-positif, la valeur de $k(G, G')$, où G et G' correspondent à deux graphes, correspond à un produit scalaire entre deux projections $\psi(G)$ et $\psi(G')$ dans un espace de Hilbert. Les noyaux sur graphes permettent alors d'obtenir une connexion naturelle entre les méthodes d'apprentissage statistique et la reconnaissance de formes basée sur des graphes.

Une grande partie des noyaux sur graphes définis en chémoinformatique sont basés sur des sacs de motifs. Ces méthodes consistent à extraire un sac de motifs des graphes et à déduire la similarité entre les graphes de la similarité de leurs sacs de motifs. Un grand nombre de noyaux basés sur des sacs de motifs utilisent des motifs linéaires [8]. Bien que l'utilisation de motifs linéaires permet de limiter la complexité des algorithmes, elle ne permet de prendre en compte que partiellement la topologie du voisinage de chaque sommet. Afin d'encoder plus d'information structurelle, d'autres méthodes sont basées sur des structures d'arbres. Par exemple, le noyau de motifs d'arbres [9] est

basé sur une énumération implicite de motifs d'arbres, c.-à-d. des arbres où un même nœud peut apparaître plusieurs fois. Une autre approche, appelée noyau de treelets [5], consiste à calculer une énumération explicite d'un ensemble limité de sous arbres. Cette énumération explicite permet alors une pondération *a posteriori* de chaque motif énuméré [4]. D'autres noyaux sur graphes consistent à transformer un graphe moléculaire en un ensemble de groupes chimiquement pertinents [3] ou en un ensemble de cycles [7, 6] mais ces méthodes ne définissent pas de noyaux définis semi-positifs [14] ou n'encodent pas les relations d'adjacence entre les parties cycliques et acycliques d'une molécule.

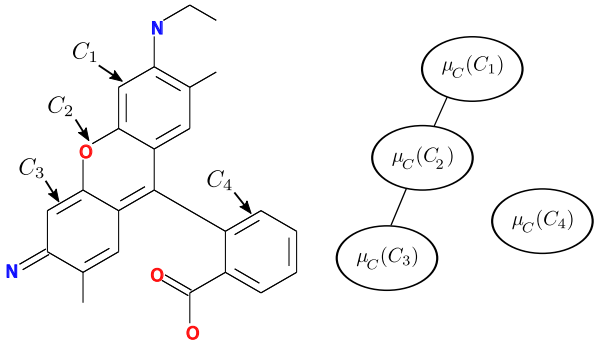
Dans cet article, nous proposons de définir une nouvelle représentation moléculaire sous forme d'hypergraphe afin d'encoder les relations d'adjacence entre les parties cycliques et acycliques d'une molécule. Après avoir présenté différents noyaux encodant l'information cyclique de manière explicite, nous définissons dans la section 3 notre nouvelle représentation moléculaire, appelée hypergraphe de cycles pertinents. Dans la section 4, nous présentons une méthode permettant d'appliquer le noyau de treelets sur cette nouvelle représentation. Cette approche permet alors d'utiliser notre hypergraphe de cycles pertinents pour prédire des propriétés moléculaires. La section 5 montre des résultats obtenus par notre contribution sur un problème de prédiction de toxicité de molécules.

2 Représentation de l'information cyclique

La plupart des noyaux sur graphes existants basés sur des sacs de motifs et appliqués en chémoinformatique sont basés sur les graphes moléculaires. Bien que cette représentation permette d'encoder une grande partie de l'information structurelle d'une molécule, elle ne permet pas de représenter explicitement des groupements particuliers d'atomes, tels que les cycles, qui peuvent avoir une influence importante sur certaines propriétés moléculaires.

Afin de prendre explicitement en compte l'information cyclique, une approche consiste à représenter chaque molécule par un ensemble de cycles la composant. Pour tout graphe $G = (V, E, \mu, \nu)$, un cycle simple est défini par un sous graphe $C = (V', E', \mu, \nu)$ de G où chaque nœud $v \in V'$ possède un degré égal à 2. Une première méthode, proposée par Horváth [7], consiste à déduire la similarité cyclique des ensembles de cycles simples extraits de chaque graphe moléculaire. La similarité entre deux molécules est alors définie par la somme de deux noyaux encodant respectivement les similarités cycliques et acycliques des deux molécules. La similarité cyclique est définie par le nombre de cycles simples communs et la similarité acyclique par le nombre de motifs d'arbres communs extraits de chaque partie acyclique des molécules.

Cependant, le nombre de cycles simples d'un graphe moléculaire peut croître exponentiellement avec le nombre de nœuds de ce dernier. Par conséquent, l'énumération de



(a) Graphe moléculaire. (b) Graphe de cycles pertinents.

FIGURE 1 – Un graphe moléculaire cyclique et le graphe de cycles pertinents associé.

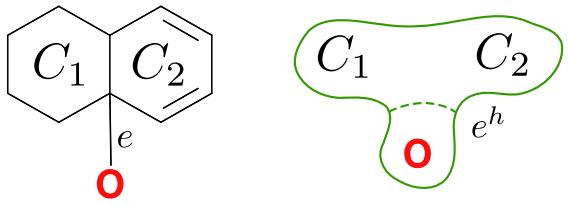
l'ensemble des cycles simples d'une molécule peut être extrêmement coûteux en temps de calcul. Afin de réduire le temps de calcul requis par l'énumération de l'information cyclique, Horvàth [6] a proposé d'utiliser l'ensemble des cycles pertinents tels que définis par [16]. Chaque cycle $C \subseteq G$ peut être représenté par un vecteur $\vec{C} \in \{0,1\}^{|E|}$ où $\vec{C}_i$ est égal à 1 si i est une arête de C et 0 sinon. En utilisant cette représentation vectorielle, l'ensemble des vecteurs encodant les cycles de G définit un espace vectoriel [15]. En considérant cet espace vectoriel, l'union de l'ensemble des bases de taille minimale définit l'ensemble des cycles pertinents, la longueur d'une base étant définie par la somme des tailles des cycles la composant. L'ensemble des cycles pertinents constitue alors un ensemble plus réduit que l'ensemble des cycles simples. Cependant, l'ensemble des cycles pertinents permet d'encoder l'ensemble des cycles constituant une molécule étant donné qu'il correspond à une union des bases de l'espace vectoriel associé aux cycles simples de la molécule.

Bien que les deux approches présentées précédemment permettent de représenter explicitement l'information cyclique, cette dernière est simplement encodée par un ensemble de cycles disjoints et n'inclut pas leurs relations d'adjacences. Ces relations d'adjacence peuvent toutefois être encodées par le graphe des cycles pertinents [15, 5]. Ce graphe est défini par $G_{\mathcal{C}}(G) = (V_{\mathcal{C}}, E_{\mathcal{C}}, \mu_{\mathcal{C}}, \nu_{\mathcal{C}})$ où chaque nœud $c \in V_{\mathcal{C}}$ correspond à un cycle pertinent. L'ensemble des nœuds de G correspondant aux atomes composant un cycle $c \in V_{\mathcal{C}}$ est encodé par une fonction $\mathcal{V} : V_{\mathcal{C}} \rightarrow \mathcal{P}(V)$, où $\mathcal{P}(V)$ correspond à l'ensemble des parties de V. L'ensemble des nœuds inclus dans un cycle est alors noté $\mathcal{V}(V_{\mathcal{C}})$. De manière similaire pour les arêtes, la fonction $\mathcal{E} : V_{\mathcal{C}} \rightarrow \mathcal{P}(E)$ permet d'associer un cycle encodé par un nœud de $V_{\mathcal{C}}$ à l'ensemble des arêtes de G le composant. L'ensemble des arêtes contenues dans au moins un cycle pertinent est alors noté $\mathcal{E}(V_{\mathcal{C}})$. Une arête (c_1, c_2) appartient à $E_{\mathcal{C}}$ si $\mathcal{V}(c_1) \cap \mathcal{V}(c_2) \neq \emptyset$, c.-à-d. si les cycles encodés par c_1 et c_2 partagent au moins un nœud dans le graphe moléculaire (C_1 et C_2 dans les figures 1 et 3(a) par exemple). La fonction d'étiquetage $\mu_{\mathcal{C}}(c)$ est

définie de manière canonique par la séquence d'étiquettes de nœuds et d'arêtes composant le cycle encodé par c et ayant le plus faible ordre lexicographique. De la même manière, la fonction d'étiquetage $\nu_{\mathcal{C}}(e)$ d'une arête $e = (c, c')$ est définie par un code canonique identifiant les chemins communs aux deux cycles représentés par c et c'. Gaüzère et al. [5] ont proposé un noyau basé sur le graphe des cycles pertinents. La similarité entre les molécules est alors déduite de la combinaison d'un noyau sur le graphe des cycles pertinents, qui encode donc une similarité cyclique, et d'un noyau sur le graphe moléculaire encodant une similarité basée sur les relations d'adjacence entre atomes. Bien que cette approche permette d'obtenir de bons résultats sur des problèmes de prédiction incluant des molécules cycliques, cette représentation, de la même manière que celle proposée par Horváth [6], sépare les informations cyclique et acyclique en définissant deux représentations moléculaires distinctes. En effet, bien que le graphe de cycles pertinents encode les relations d'adjacence entre les cycles d'une molécule, il n'encode pas toutes les relations d'adjacence induites par les nœuds et les arêtes du graphe moléculaire. Par exemple, les parties acycliques adjacentes à C_1, C_2 et C_3 ainsi que la relation d'adjacence entre C_2 et C_4 dans la figure 1(a) ne sont pas encodées dans le graphe de cycle pertinents (figure 1(b)). La similarité globale des molécules est alors calculée en utilisant deux mesures de similarité distinctes, chacune d'elle étant appliquée sur une représentation moléculaire. Cette séparation implique donc la perte des relations d'adjacence entre les parties cycliques et acycliques des molécules. Par conséquent, nous proposons une nouvelle représentation moléculaire permettant d'encoder à la fois l'information cyclique de manière explicite et l'information acyclique dans une unique représentation et incluant donc les relations d'adjacence entre les parties cycliques et acycliques.

3 Représentation des relations d'adjacences entre cycles et substituants

Afin d'encoder les relations d'adjacence entre les parties cycliques et acycliques d'une molécule, nous proposons de définir une représentation moléculaire basée sur le graphe de cycles pertinents défini dans la section 2. Cette représentation permet d'encoder l'information cyclique d'une molécule en représentant un ensemble d'atomes composant un cycle pertinent par un nœud. Afin d'encoder l'information acyclique, une première approche consiste à ajouter les nœuds et arêtes manquants au graphe de cycles pertinents. Cependant, cette approche ne peut pas encoder les cas où un atome est connecté à deux cycles pertinents distincts. Comme montré dans la figure 2(a), l'atome étiqueté O est connecté par une unique arête à deux cycles distincts dans le graphe moléculaire. Cette relation d'adjacence ne peut pas être encodée par un graphe usuel où une arête définit une relation d'adjacence entre deux nœuds. Par conséquent, afin de pouvoir encoder ces relations d'adjacence, nous proposons de définir une représentation moléculaire sous forme d'hypergraphe.



(a) Atome acyclique connecté à deux cycles par une arête unique e.

(b) Hyperarête e^h représentant l'arête originale e. $e^h = (\{C_1, C_2\}, O)$.

FIGURE 2 – Cas spécial où un graphe ne peut pas encoder la relation d'adjacence entre une partie cyclique et acyclique.

Un hypergraphe orienté [1, 2] $H = (V, E)$ peut être défini par un ensemble de nœuds V et un ensemble $E = E^e \cup E^h$ encodant l'union d'un ensemble d'arêtes $E^e \subset V \times V$ et un ensemble d'hyperarêtes orientées $E^h \subset \mathcal{P}(V) \times \mathcal{P}(V)$. Une hyperarête orientée $e = (s_u, s_v)$ avec $s_u = \{u_1, \ldots, u_i\}$ et $s_v = \{v_1, \ldots, v_j\}$ définit une relation d'adjacence entre les ensembles $\{u_1, \ldots, u_i\}$ et $\{v_1, \ldots, v_j\}$, comme illustré dans la figure 2(b). Dans la suite de cet article, nous considérerons que si $\exists e = (s_1, s_2) \in E$ alors $\exists e' = (s_2, s_1) \in E$ et e et e' sont considérés comme une seule et même hyperarête. Cette définition permet alors de représenter les relations d'adjacence entre un atome et un ensemble de cycles, chaque cycle étant encodé par un nœud. Un graphe moléculaire $G = (V, E, \mu, \nu)$ et son graphe de cycles pertinents associé $G_{\mathcal{C}}(G) = (V_{\mathcal{C}}, E_{\mathcal{C}}, \mu_{\mathcal{C}}, \nu_{\mathcal{C}})$, peuvent alors être représenté par un hypergraphe de cycles pertinents $H_{CH}(G) = (V_{CH}, E_{CH}, \mu_{CH}, \nu_{CH})$. Le graphe de cycles pertinents $G_{\mathcal{C}}(G)$, représentant le système cyclique d'une molécule, permet d'encoder l'ensemble des nœuds correspondant à l'ensemble des atomes $\mathcal{V}(V_{\mathcal{C}})$ et l'ensemble des liaisons atomiques $\mathcal{E}(V_{\mathcal{C}})$ inclues dans un cycle. En considérant cette représentation, l'information du graphe moléculaire manquante correspond aux atomes et liaisons atomiques n'appartenant pas à un cycle. Ces ensembles sont respectivement définis par le complément de $\mathcal{V}(V_{\mathcal{C}})$ et $\mathcal{E}(V_{\mathcal{C}})$ dans V et E. Par conséquent, afin de représenter l'ensemble des atomes du graphe moléculaire dans l'hypergraphe de cycles pertinents, V_{CH} est défini par l'union de deux sous ensembles. Un premier sous ensemble $V_{\mathcal{C}}$ correspondant à l'ensemble des cycles pertinents (C_1, C_2, C_3 et C_4 dans la figure 3(c)) et un second sous ensemble $V - \mathcal{V}(V_{\mathcal{C}})$ correspondant à l'ensemble des atomes non inclus dans un cycle (atomes d'oxygène et atomes de carbone acycliques dans la figure 3(c)). En considérant l'ensemble des nœuds V_{CH}, nous définissons une fonction $p : V \rightarrow \mathcal{P}(V_{CH})$ définie par $p(u) = \{u\}$ si $u \notin \mathcal{V}(V_{\mathcal{C}})$ et $\{c \in V_{\mathcal{C}} \mid u \in \mathcal{V}(c)\}$ sinon. Pour chaque nœud $v \in \mathcal{V}(V_{\mathcal{C}})$, la valeur de $p(v)$

correspond donc à l'ensemble des nœuds de l'hypergraphe encodant un cycle incluant l'atome v. De la même manière que pour les nœuds, l'ensemble des hyperarêtes E_{CH} est composé de deux sous ensembles :

1. Un ensemble d'arêtes E^e_{CH} composé des :
   — arêtes entre des nœuds représentant des cycles pertinents. Cet ensemble correspond à l'ensemble des arêtes $E_{\mathcal{C}}$ (par exemple les arêtes entre C_1 et C_2, C_2 et C_3 ainsi que C_3 et C_4 dans la figure 3(c)),
   — arêtes $e = (p(u), p(v))$ telles que $(u, v) \in E - \mathcal{E}(V_{\mathcal{C}})$, $|p(u)| = |p(v)| = 1$. Cet ensemble d'arêtes correspond aux arêtes du graphe moléculaire qui ne sont pas inclues dans un cycle (C_1 et O dans la figure 3(c)),
2. et un ensemble d'hyperarêtes $e = (p(u), p(v)) \in E^h_{CH}$ défini tel que $(u, v) \in E - \mathcal{E}(V_{\mathcal{C}})$, $|p(u)| > 1$ ou $|p(v)| > 1$. Cet ensemble d'hyperarêtes correspond aux cas où une arête connecte au minimum deux cycles pertinents distincts à une autre partie de la molécule. Par exemple les relations d'adjacence entre les deux atomes de carbone connectés à $C_1 - C_2$ et $C_3 - C_4$ dans la figure 3(c) seront respectivement représentées par deux hyperarêtes $e_1 = (\{C_1, C_2\}, C)$ et $e_2 = (\{C_3, C_4\}, C)$.

La fonction d'étiquetage μ_{CH} correspond à la fonction d'étiquetage de nœuds $\mu_{\mathcal{C}}$ du graphe de cycles pertinents pour chaque nœud $v \in V_{\mathcal{C}}$ correspondant à un cycle pertinent, et à la fonction d'étiquetage de nœud μ du graphe moléculaire original dans le cas contraire. De manière similaire, la fonction d'étiquetage d'arête ν_{CH} est définie par la fonction d'étiquetage $\nu_{\mathcal{C}}$ du graphe de cycles pertinents pour chaque arête encodant une relation d'adjacence entre deux cycles et à la fonction d'étiquetage d'arêtes ν du graphe moléculaire original sinon. Les hyperarêtes de l'hypergraphe de cycles pertinents correspondent aux arêtes connectant au moins deux cycles pertinents à une autre partie du graphe moléculaire. Par conséquent, l'étiquette associée à une hyperarête $e \in E^h_{CH}$ correspond à l'étiquette de l'arête correspondante dans le graphe moléculaire original. Cette représentation moléculaire sous forme d'hypergraphe (figure 3(c)) encode l'ensemble des atomes $v \in V$ soit par un nœud encodant un cycle ou par v lui-même si $v \notin \mathcal{V}(V_{\mathcal{C}})$. De la même manière, chaque liaison atomique $e \in E$ est encodée dans notre hypergraphe de cycles pertinents. De plus, chacun des deux ensembles de nœuds incidents à une hyperarête est une clique du graphe de cycles pertinents :

Proposition 1 *Soit un graphe $G = (V, E)$ et l'hypergraphe de cycles pertinents associé $H_{CH}(G) = (V_{CH}, E_{CH})$. Si $\exists e = (s_1, s_2) \in E^h_{CH}$ et $c_1, c_2 \in V_{CH}$ tel que $\{c_1, c_2\} \subseteq s_1$ ou $\{c_1, c_2\} \subseteq s_2$, alors $(c_1, c_2) \in E^e_{CH}$, c.-à-d. c_1 est adjacent à c_2.*

Preuve 1 *Si $c_1 \in s_1$ et $c_2 \in s_1$, alors par construction de E^h_{CH}, $\exists e = (u, v) \in E$ tel que $\{c_1, c_2\} \subseteq p(u) = s_1$. Par définition de la fonction p et étant donné que $c_1, c_2 \in V_{\mathcal{C}}$, nous avons $u \in \mathcal{V}(c_1) \cap \mathcal{V}(c_2)$. Par définition du graphe de cycles pertinents, $(c_1, c_2) \in E_{\mathcal{C}} \subset E^e_{CH}$. La preuve pour $c_1 \in s_2$ et $c_2 \in s_2$ est similaire.*

L'algorithme 1 décrit les différentes étapes nécessaires pour transformer un graphe moléculaire G en son hypergraphe de cycles pertinents H_{CH}. La première étape consiste à calculer le graphe des cycles pertinents de G, tel que décrit dans [5], et initialiser H_{CH} par ce graphe (algo. 1, lignes 3 et 4). Ensuite, l'ensemble des parties acycliques est ajoutée (algo. 1, lignes 6 et 7). Enfin, l'ensemble des hyperarêtes est inclu dans notre hypergraphe de cycles pertinents (algo. 1, ligne 9).

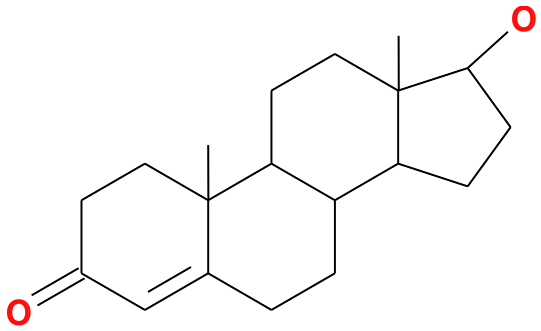
Algorithme 1 Calcul de l'hypergraphe de cycles pertinents à partir d'un graphe moléculaire.

Require: $G = (V, E)$
Ensure: $H_{CH} = (V_{CH}, E_{CH}), E_{CH} = E^e_{CH} \cup E^h_{CH}$
1: $G_{\mathcal{C}}(V_{\mathcal{C}}, E_{\mathcal{C}}) = G_{\mathcal{C}}(G)$ {Graphe de cycles pertinents}
2: {Ajout de l'information inclue dans les cycles}
3: $V_{CH} = V_{\mathcal{C}}$
4: $E^e_{CH} = E_{\mathcal{C}}$
5: {Ajout de l'information non inclue dans un cycle}
6: $V_{CH} = V_{CH} \cup \{v \notin \mathcal{V}(V_{\mathcal{C}})\}$
7: $E^e_{CH} = E^e_{CH} \cup \{(p(u), p(v)) \mid (u, v) \in E, |p(u)| = 1 \wedge |p(v)| = 1\}$
8: {Cas spéciaux (figure 2).}
9: $E^h_{CH} = \{(p(u), p(v)) \mid (u, v) \in E, |p(u)| > 1 \vee |p(v)| > 1\}$
10: **return** H_{CH}

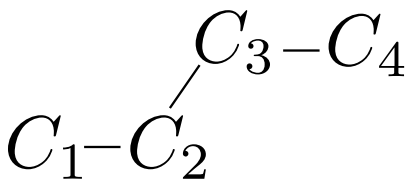
4 Similarité d'hypergraphes de cycles pertinents

La précédente section définit une représentation moléculaire qui permet d'encoder les relations d'adjacence entre les parties cycliques et acycliques d'une molécule. Afin d'appliquer les méthodes QSAR utilisant cette représentation, nous devons définir une mesure de similarité entre hypergraphes de cycles pertinents. Les noyaux sur graphes, tel que le noyau de treelets [5], sont seulement définis sur les graphes et ne peuvent pas être appliqués directement sur les hypergraphes de cycles pertinents. Dans cette section, nous proposons d'adapter le noyau de treelets à la comparaison d'hypergraphes de cycles pertinents.

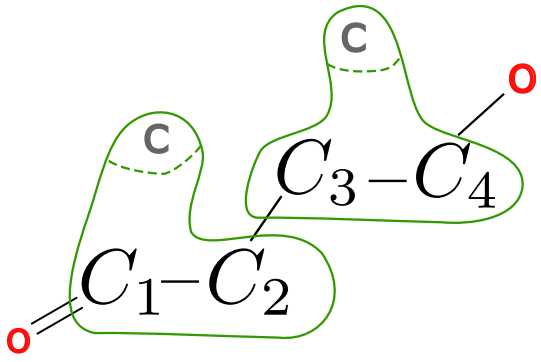
Le noyau de treelet est basé sur un sac de motifs. L'ensemble des motifs considérés, noté $\mathcal{T}$ et appelé treelets, est composé de l'ensemble des arbres étiquetés ayant au plus 6 nœuds. Le nombre d'occurrences de chaque treelet $t \in \mathcal{T}(G)$ extrait d'un graphe G est encodé par une fonction $f_t(G)$ égale à $|(t \trianglelefteq G)|$, où $\trianglelefteq$ encode l'isomorphisme de sous graphe. La similarité de deux graphes G et G' est alors définie par une somme de sous noyaux comparant les



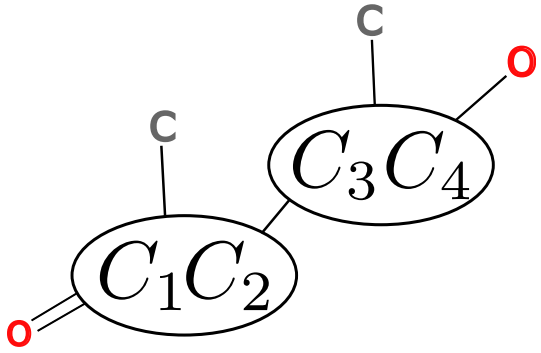
(a) Graphe moléculaire G représentant la molécule de testostérone incluant des cycles.



(b) Graphe de cycles pertinents $G_{\mathcal{C}}(G)$.



(c) Hypergraphe de cycles pertinents $H_{CH}(G)$.



(d) Graphe de cycles pertinents réduit $G_{RC}(G)$.

FIGURE 3 – Différentes représentations d'une même molécule.

nombres d'occurrences de chaque treelet :

$$K_{\mathcal{T}}(G,G') = \sum_{t\in\mathcal{T}(G)\cap\mathcal{T}(G')} k(f_t(G), f_t(G')) \quad (1)$$

où $k(.,.)$ correspond à un noyau entre réels. Bien que ce noyau soit applicable sur différents types de graphes, il ne peut pas être utilisé directement pour comparer des hypergraphes. Un hypergraphe permet d'encoder des relations d'adjacences globales entre ensembles de nœuds. Inversement, le noyau de treelets est défini sur des graphes usuels où les relations d'adjacence sont définies localement entre deux nœuds élémentaires.

Une hyperarête $e = (s_1, s_2) \in E^h_{CH}$ encode une relation d'adjacence globale entre les deux ensembles de nœuds s_1 et s_2. Cette relation globale peut être représentée par une relation locale entre deux nœuds n_1 et n_2 encodant respectivement les deux ensembles de nœuds s_1 et s_2. Ces nœuds peuvent être naturellement obtenus par la fusion de l'ensemble s_1 d'une part et la fusion de l'ensemble s_2 d'autre part. Intuitivement, deux cycles partageant une même hyperarête seront fusionnés en un seul et même nœud. Cette opération de fusion permet alors de transformer une hyperarête entre deux ensembles de nœuds en une arête usuelle entre les deux nœuds obtenus par fusion. Notez que dans le cas où un nœud appartient à l'intersection de plusieurs ensembles s^i_1 avec $(s^i_1, s^i_2) \in E^h_{CH}$, la relation d'adjacence globale définie précédemment nous conduit à créer un nœud représentant l'union des s^i_1. Nous définissons le graphe de cycles pertinents réduit (figure 3(d)) $G_{RC}(G) = (V_{RC}, E_{RC}, \mu_{RC}, \nu_{RC})$ associé à un graphe G par :

- un ensemble de nœuds V_{RC} obtenu par la fusion des nœuds de l'hypergraphe adjacents à une même hyperarête ;
- un ensemble d'arêtes E_{RC} correspondant intuitivement à l'union des arêtes usuelles $E^e_{CH} \in H_{CH}$ et la transformation des hyperarêtes $E^h_{CH} \in H_{CH}$ en arêtes usuelles.

La fonction d'étiquetage $\mu_{RC}(v_{RC})$, pour tout $v_{RC} \in V_{RC}$, est définie par une séquence d'étiquettes de nœuds et d'arêtes rencontrées durant un parcours en profondeur d'un arbre couvrant du sous graphe de $H_{CH}(G)$ correspondant à l'ensemble des nœuds fusionnés pour obtenir v_{RC}. Étant donné que les nœuds c et c' partageant une hyperarête sont adjacents (proposition 1), un tel arbre couvrant existe obligatoirement. Afin de définir une fonction d'étiquetage canonique, la séquence d'étiquettes est définie comme la séquence ayant le plus petit ordre lexicographique parmi toutes les séquences possibles.

En considérant le graphe de cycles pertinents réduit, notre mesure de similarité basée sur le noyau de treelets est définie en deux temps : Une première étape consiste à extraire l'ensemble des treelets $\mathcal{T}_1 = \mathcal{T}(V_{CH}, E^e_{CH})$. Le graphe (V_{CH}, E^e_{CH}) correspond à un sous hypergraphe de H_{CH} qui inclut l'ensemble des arêtes $e \in E^e_{CH}$. Par conséquent, l'ensemble des treelets $\mathcal{T}_1$ encode l'information ne correspondant pas aux cas spéciaux décrits dans la figure 2. Cette

TABLE 1 – Précision de la classification sur le jeu de données PTC.

	Méthode	MM	FM	MR	FR
1	Noyau de treelets (TK)	61.9%	58.7%	**60.8%**	60.4%
2	Noyau de treelets sur graphe de cycles pertinents (TC)	62.8%	60.2%	59%	66.1%
3	Noyau de treelets sur hypergraphe de cycles pertinents (TCH)	64.6%	**64.2%**	60.2%	66.4%
4	Noyau de motifs cycliques [6]	62.2%	59.3%	58.7%	65%
5	Noyau sur la distance d'édition [11]	**66.4 %**	60.7%	56.4%	**66.7%**
6	TK + MKL	64.9%	64.2%	**65.1%**	**71.2%**
7	TC + MKL	64.3%	61%	61.6%	67.5%
8	TCH + MKL	**67%**	**65.6%**	62.5%	68.1%

information, encodée par les hyperarêtes $e \in E^h_{CH}$, est incluse dans notre mesure de similarité par l'ensemble des treelets $\mathcal{T}_2$ extraits du graphe de cycles pertinents réduit G_{CH} construit à partir de la transformation des hyperarêtes en arêtes usuelles. Afin d'éviter une redondance des motifs extraits, nous réduisons l'ensemble des treelets $\mathcal{T}_2$ aux treelets contenant au moins une arête encodant une hyperarête $e_h \in E^h_{CH}$. Enfin, nous définissons l'ensemble des treelets $\mathcal{T}_{CR}(G)$ associé à un graphe moléculaire G par $\mathcal{T}_1 \cup \mathcal{T}_2$. La similarité des molécules est alors définie par une somme de sous noyaux comparant les nombres d'occurrences de chaque treelet $t \in \mathcal{T}_{CR}(G)$ (équation 1). Cette approche nous permet d'utiliser un ensemble de motifs encodant les relations d'adjacence entre parties acycliques, entre cycles mais aussi entre cycles et parties acycliques.

5 Expériences

Nous avons testé notre contribution sur un jeu de données extrait du Predictive Toxicity Challenge [13] qui consiste à prédire la carcinogènécité de composés chimiques appliqués sur des femelles (F) et mâles (M) rats (R) et souris (M). Pour chaque classe d'animal, l'ensemble des données est découpée en 10 jeux de données. Chacun des 10 jeux de données est alors prédit en utilisant les 9 restants comme ensemble d'apprentissage. L'ensemble des molécules constituant l'ensemble des données est égal à 336 pour les souris mâles, 349 pour les souris femelles, 344 pour les rats mâles et 351 pour les rats femelles. Le tableau 1 montre le pourcentage de composés chimiques correctement classifiés sur l'ensemble des 10 jeux de données pour chaque noyau et chaque classe d'animal. Les trois premières lignes du tableau 1 montrent les résultats obtenus par un noyau de treelets appliqué sur différentes représentations moléculaires. La ligne 1 correspond au noyau de treelets appliqué sur le graphe moléculaire, la ligne 2 sur le graphe de cycles pertinents et la ligne 3 correspond au noyau défini dans la section 3. Premièrement, nous pouvons noter que l'information cyclique ajoutée par notre proposition permet d'améliorer la précision de la classification obtenue par le noyau sur graphe de cycles pertinents (lignes 2 et 3). De plus, l'utilisation des hypergraphes de cycles pertinents permet d'améliorer les résultats obtenus en utilisant le graphe moléculaire sur 3 jeux de données parmi 4 (lignes 1 et 3). Ces observations permettent de valider notre hypothèse sur l'importance des relations d'adjacence entre les parties cycliques et acycliques. La ligne 4 montre les résultats obtenus par le noyau défini par Horváth et basé sur l'ensemble des cycles pertinents communs à deux molécules. Comme nous pouvons le voir, la précision de classification diminue lorsque nous ne prenons pas en compte les relations d'adjacence de cycles. La ligne 5 correspond à un noyau sur graphe basé sur la distance d'édition [11] entre graphes moléculaires. Ce noyau obtient de meilleurs résultats que le noyau de treelet appliqué sur l'hypergraphe de cycles pertinents pour deux classes d'animaux parmi les 4. La deuxième partie du tableau 1 montre les résultats obtenus par les noyaux de treelets après une étape de pondération des noyaux associés à chaque treelet [5, 12]. Le noyau de treelets pondéré appliqué sur notre nouvelle représentation moléculaire (tableau 1, ligne 8) obtient alors les meilleurs résultats sur les deux jeux de données de souris et obtient les seconds meilleurs résultats sur les deux jeux de données correspondant aux rats. On peut noter que les meilleurs résultats sur les deux autres classes d'animaux sont obtenus par notre noyau de treelets appliqué sur le graphe moléculaire. Cette observation nous conduit à émettre l'hypothèse que la toxicité des molécules sur les rats est plus liée à des sous structures acycliques des molécules qu'à des sous structures cycliques, et inversement chez les souris.

6 Conclusion

Dans cet article, nous avons défini une nouvelle représentation moléculaire basée sur un hypergraphe, ce qui permet d'encoder les relations d'adjacence entre les parties acycliques et cycliques d'une molécule. De plus, nous avons proposé une méthode afin d'appliquer le noyau de treelets sur l'hypergraphe de cycles pertinents. Les expériences effectuées montrent que l'information encodée par l'hypergraphe de cycles pertinents permet d'obtenir une meilleure précision de classification que les méthodes directement basées sur le graphe moléculaire. Une perspective intéressante de ces travaux est d'encoder une information plus fine en encodant les positions relatives des liaisons atomiques connectant les parties acycliques à un cycle.

Références

[1] Claude Berge. *Graphs and hypergraphs*, volume 6. Elsevier, 1976.

[2] Aurélien Ducournau. *Hypergraphes : clustering, réduction et marches aléatoires orientées pour la segmentation d'images et de vidéo.* PhD thesis, École Nationale d'Ingénieurs de Saint-Étienne., 2012.

[3] Holger Fröhlich, Jörg K. Wegner, Florian Sieker, and Andreas Zell. Optimal assignment kernels for attributed molecular graphs. In *Proceedings of ICML '05*, pages 225–232. ACM Press, 2005.

[4] Benoit Gaüzère, Luc Brun, and Didier Villemin. Two New Graphs Kernels in Chemoinformatics. *Pattern Recognition Letters*, 33(15) :2038–2047, 2012.

[5] Benoit Gaüzère, Luc Brun, and Didier Villemin. Noyau de treelets appliqué aux graphes étiquetés et aux graphes de cycles. *Revue d'Intelligence Artificielle*, 27(1) :121–144, 2013.

[6] Tamás Horváth. Cyclic pattern kernels revisited. *Proceedings of AKDD 2005*, pages 791–801, 2005.

[7] Tamás Horváth, Thomas Gartner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. *Proceedings of AKDD 2004*, pages 158–167, 2004.

[8] Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. *Kernels for graphs*, chapter 7, pages 155–170. MIT Press, 2004.

[9] Pierre Mahé and Jean-Philippe Vert. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1) :3–35, 2009.

[10] Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, Thanakorn Naenna, and Virapong Prachayasittikul. A practical overview of quantitative structure-activity relationship. *EXCLI J*, 8 :74–88, 2009.

[11] Michel Neuhaus and Horst Bunke. *Bridging the gap between graph edit distance and kernel machines*. World Scientific Pub Co Inc, 2007.

[12] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9 :2491–2521, 2008.

[13] Hannu Toivonen, Ashwin Srinivasan, Ross King, Stefan Kramer, and Christoph Helma. Statistical evaluation of the predictive toxicology challenge 2000-2001. *Bioinformatics*, 19(10) :1183–1193, 2003.

[14] Jean-Philippe Vert. The optimal assignment kernel is not positive definite. http ://hal.archives-ouvertes.fr/hal-00218278.

[15] Philippe Vismara. *Reconnaissance et représentation d'éléments structuraux pour la description d'objets complexes. Application à l'élaboration de stratégies de synthèse en chimie organique*. PhD thesis, Université Montpellier II, 1995.

[16] Philippe Vismara. Union of all the minimum cycle bases of a graph. *The Electronic Journal of Combinatorics*, 4(1) :73–87, 1997.