Relevant Cycle Hypergraph Representation for Molecules

Benoit Gaüzère†, Luc Brun†, and Didier Villemin‡

†GREYC UMR CNRS 6072, ‡LCMT UMR CNRS 6507,
Caen, France
{benoit.gauzere,didier.villemin,luc.brun}@ensicaen.fr,

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 techniques with graph theory. Despite the fact that molecular graph encodes all structural information of a molecule, it does not explicitly encode 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 called relevant cycle hypergraph. In addition, we propose a similarity measure in order to compare relevant cycle hypergraphs and use this molecular representation in a chemoinformatics prediction problem.

Keywords: Graph Kernel, Chemoinformatics, Relevant Cycles

1 Introduction

Chemoinformatics consists in predicting molecule's properties from their similarity. Most of existing methods, called fingerprint methods, encode molecules as collections of chemical descriptors and deduce similarity between molecules from the similarity of their collections of descriptors. Another approach consists in using the molecular graph $G = (V, E, \mu, \nu)$ representation associated to a molecule. Unlabeled graph (V, E) encodes molecular structural information while labelling function μ maps each vertex to an atom's label corresponding to its chemical element and labelling function ν characterizes each edge by the valency (single, double, triple or aromatic) of the corresponding atomic bond which connects two atoms. Hydrogen atoms are implicitly encoded into molecular graph representation using the valency of atoms.

Considering molecular graph representation, similarity between molecules can be deduced from the similarity of their molecular graphs. Graph kernels can be understood as symmetric graph similarity measures. Using a semi definite positive kernel, the value $k(G, G')$, where G and G' encode two graphs, corresponds to a scalar product between two vectors $\psi(G)$ and $\psi(G')$ in an Hilbert space. Graph kernels thus provide a natural connection between structural and statistical pattern recognition fields.

A large family of graph kernels defined in chemoinformatics is based on bag of patterns. These methods extract a bag of patterns from graphs and deduce similarity between graphs from similarity between their bags. Most of existing graph kernels based on bags of patterns are defined on linear patterns [8]. Such methods are generally limited by the lack of expressiveness of linear patterns to encode structural information of graphs. In order to encode more structural information, some methods are defined on non-linear patterns. For example, tree-pattern kernel [9] is based on an implicit enumeration of tree-patterns, i.e. trees where a vertex can appear more than once. Another approach, called treelet kernel [4], computes an explicit enumeration of a limited set of subtrees which allows to perform an a-posteriori feature weighting step [5]. Others graph kernels aim to transform a molecular graph into a set of chemical relevant groups [3] or a set of cycles [7, 6] but these methods do not define a valid kernel or do not allow to encode relationships between cyclic and acyclic parts of a molecule.

In this paper, we propose to define a new molecular representation encoded by an hypergraph which aims to encode adjacency relationships between cyclic and acyclic parts of a molecule. After a presentation of existing methods to encode molecular cyclic information in Section 2, we define in Section 3 our new molecular representation. In addition, we propose in Section 4 a method to apply treelet kernel on this new molecular representation. This method allows us to use our new molecular representation to predict molecule's properties. Section 5 shows results obtained by our contribution to a chemoinformatics problem.

2 Encoding Cyclic Information

Most of existing graph kernels based on bags of patterns applied to chemoinformatics are based on the molecular graph representation (Section 1). Whereas this representation allows to encode most of the structural information of a given molecule, it does not explicitly encode some special combinations of atoms, such as cycles, which may have a particular influence on molecule's properties. In order to highlight such particular groups of atoms, Frölich et al. [3] have proposed to encode a molecule by a set of predefined subgraphs composing the associated molecule. These predefined subgraphs correspond to chemical relevant groups of atoms and are generally defined by cycles or connected atom groups. Then, similarity between molecules is deduced by an optimal matching between two sets of relevant groups. Unfortunately, the kernel defined from this optimal assignment may lead to a non positive definite kernel [12], hence restricting the application field of this kernel.

Some other approaches aim to encode a molecule by a subset of its cycles. A first approach, proposed by Horváth [7], consists in computing the set of simple cycles of a molecule. Then, similarity between two molecules is defined as a sum of two kernels encoding respectively the cyclic and acyclic similarities between both molecules. Similarity between cycles is defined by the number of common simple cycles and similarity between acyclic parts by a tree-pattern kernel [9]. An extension of this method only computes the set of relevant cycles [6], as defined

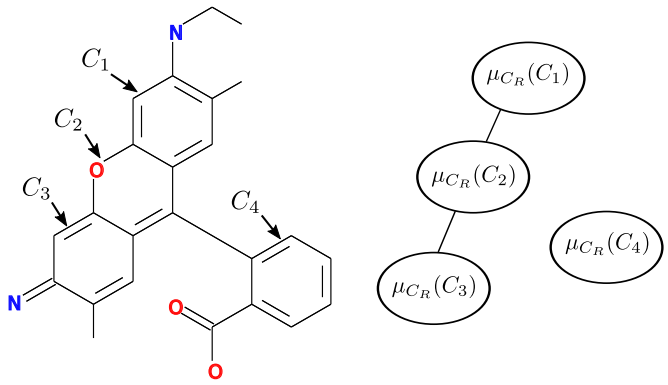
by Vismara [13], of the molecular graph hence providing a better computational efficiency. Whereas this approach provides an explicit encoding of cyclic information, the cyclic system is encoded by a set of cycles which does not encode relationships between cycles.

In order to encode additional information, Gaüzère et al. [5] have proposed to encode the set of relevant cycles and their adjacency relationships within the relevant cycle graph. The similarity between molecules can then be deduced by combining a kernel on relevant cycle graphs which encodes cyclic system similarity and a kernel on molecular graphs which encodes the similarity of molecules based on atom's relationships. Despite the fact that this approach leads to good results on experiments involving cyclic molecules, this representation, as the one of Horváth [6], separates cyclic and acyclic information by defining two different molecular representations. Then, global similarity between molecules is computed using two distinct similarity measures, each of them being applied on one representation. This separation induces a loss of adjacency relationships between cyclic and acyclic parts of molecules. In the following, we propose a new molecular representation which aims to merge cyclic and acyclic information into one molecular representation and hence encodes adjacency relationships between cyclic and acyclic parts.

3 Encoding Topological Relationships between Cyclic and Acyclic Parts

In order to encode adjacency relations between cyclic and acyclic parts of a molecule, we propose to define a molecular representation which aims to represent a set of atoms encoding a cycle as a single vertex. For any graph G, a simple cycle is defined as a subgraph $C = (V', E', \mu, \nu)$ of $G = (V, E, \mu, \nu)$ where each vertex $v \in V'$ has a degree equal to 2. Each cycle $C \subseteq G$ can be represented as a vector $\boldsymbol{C} \in \{0,1\}^{|E|}$ where $\boldsymbol{C}_i$ equals 1 if i is an edge of C and 0 otherwise. Using this vector representation, the set of vectors encoding cycles of G defines a vector space [13]. Given this vector space, the union of all bases of minimum length defines the set of relevant cycles, denoted $\mathcal{C}_\mathcal{R}$. The length of a base is defined as the sum of lengths of its cycles.

Adjacency relationships between relevant cycles can be encoded by the relevant cycle graph [5]. This graph is defined as $G_\mathcal{C} = (\mathcal{C}_\mathcal{R}, E_{\mathcal{C}_\mathcal{R}}, \mu_{\mathcal{C}_\mathcal{R}}, \nu_{\mathcal{C}_\mathcal{R}})$ where each vertex $c \in \mathcal{C}_\mathcal{R}$ corresponds to a relevant cycle. Each vertex c is associated to the set of vertices $V(c)$ corresponding to the set of atoms included within c and the set of edges $E(c)$ corresponding to the set of atomic bonds forming cycle c. By extension, $E(\mathcal{C}_\mathcal{R})$ denotes the set of atomic bonds belonging to a relevant cycle of $\mathcal{C}_\mathcal{R}$. An edge (c_1, c_2) is in $E_{\mathcal{C}_\mathcal{R}}$ if $V(c_1) \cap V(c_2) \neq \emptyset$, i.e. if c_1 and c_2 share at least one vertex of the molecular graph (Figure 1). The labelling function $\mu_{\mathcal{C}_\mathcal{R}}(c)$ is defined as a canonical code of the cyclic sequence of vertex and edge labels defining c. In the same way, the label function $\nu_{\mathcal{C}_\mathcal{R}}(e)$ of an edge $e = (c, c')$ is defined as a canonical code of the path common to c and c'. Despite the fact that this relevant cycle graph encodes adjacency information



(a) Molecular graph. (b) Relevant cycle graph.

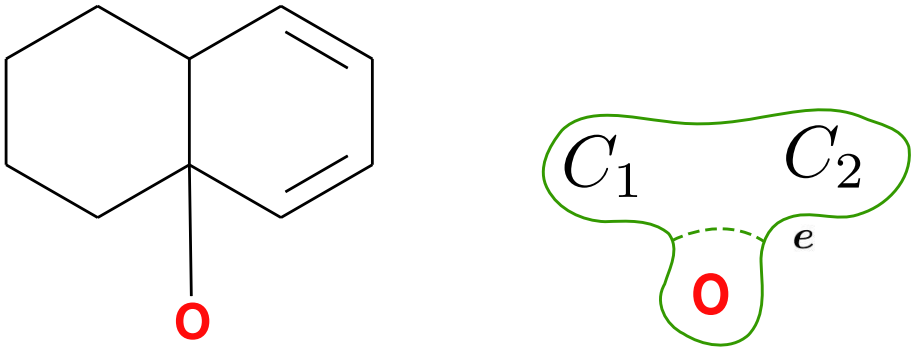
Fig. 1. A cyclic molecular graph and its relevant cycle graph representation.

of the molecular cyclic system, all adjacency information involving vertices and edges of a molecular graph which are not included within a cycle is missing. For example, acyclic parts connected to C_1, C_3 and C_4 and connection between C_2 and C_4 in Figure 1(a) are not encoded within the associated relevant cycle graph representation (Figure 1(b)).

Therefore, in order to add this information into our molecular representation, a first approach consists in adding missing vertices and edges to our relevant cycle graph. Unfortunately, such an approach can not handle the case where an atom is connected to two distinct relevant cycles. As shown in Figure 2(a), the atom labeled O is connected by an unique edge to two distinct cycles in the molecular graph representation. This adjacency relationship can not be encoded by a simple graph where an edge connects only two vertices. Therefore, in order to handle such relationships, we propose to define a new hypergraph representation of the molecular graph.

A directed hypergraph [1, 2] $H = (V, E)$ is defined as a set of vertices V and a set $E = E^e \cup E^h$ encoding the union of a set of edges $E^e \subset V \times V$ and a set of hyperedges $E^h \subset \mathcal{P}(V) \times \mathcal{P}(V)$ where $\mathcal{P}(V)$ denotes the set of all subsets of V. An ordered hyperedge $e = (s_u, s_v)$ with $s_u = \{u_1, \ldots, u_i\}$ and $s_v = \{v_1, \ldots, v_j\}$ defines an adjacency relation between sets $\{u_1, \ldots, u_i\}$ and $\{v_1, \ldots, v_j\}$, as illustrated in Figure 2(b). In the following, we assume that if $\exists e = (s_1, s_2) \in E$ then $\exists e' = (s_2, s_1) \in E$ and e and e' are considered as a same unique hyperedge. Such a definition allows us to represent relationships between an acyclic atom and a set of cycles, each cycle being encoded as a vertex.

A molecular graph $G = (V, E, \mu, \nu)$ can now be encoded as a relevant cycle hypergraph $H_{RC}(G) = (V_{RC}, E_{RC})$. Within relevant cycle graph representation, the set of vertices $C_{\mathcal{R}}$ encodes the set of atoms $V(C_{\mathcal{R}})$ and the set of atomic bonds $E(C_{\mathcal{R}})$ which belong to a cycle. Considering such a representation, missing molecular graph information corresponds to atoms and atomic bonds not included within a cycle. These sets are respectively defined by the complement



(a) Acyclic atom connected to two cycles by an unique edge

(b) Hyperedge e representing original edge. $e = (\{C_1, C_2\}, O)$.

Fig. 2. Special case where a graph can not encode the representation based on relevant cycle graph.

of $V(C_{\mathcal{R}})$ and $E(C_{\mathcal{R}})$ in V and E. Therefore, in order to include all atom information into our relevant cycle hypergraph, V_{RC} is defined by the union of two subsets:

1. A first subset $C_{\mathcal{R}}$ corresponding to the set of relevant cycles,
2. and a second subset $V - V(C_{\mathcal{R}})$ corresponding to the set of atoms not included within a cycle.

Considering set of vertices V_{RC}, we define a function $p : V \to \mathcal{P}(V_{RC})$ defined as $p(u) = \{u\}$ if $u \notin V(C_{\mathcal{R}})$ and $\{c \in C_{\mathcal{R}} \mid u \in V(c)\}$ if not. This function p encodes the print of vertex $v \in V$ on V_{RC}. In the same way as for vertices, the set of hyperedges E_{RC} is composed of two subsets:

1. A set of edges E^e_{RC} composed of:
   - edges between relevant cycle vertices, corresponding to the set of edges $E_{\mathcal{C}_{\mathcal{R}}}$,
   - edges $e = (p(u), p(v))$ such that $(u, v) \in E - E(\mathcal{C}_{\mathcal{R}})$, $|p(u)| = 1$ and $|p(v)| = 1$. This set of edges corresponds to edges of molecular graph G connecting two acyclic atoms or connecting a single relevant cycle to another single relevant cycle (C_2 and C_4 in Figure 1) or an acyclic part of G (C_3 and N in Figure 1),
2. and a set of hyperedges $e = (p(u), p(v)) \in E^h_{RC}$ such that $(u, v) \in E - E(\mathcal{C}_{\mathcal{R}})$, $|p(u)| > 1$ or $|p(v)| > 1$. This set of hyperedges corresponds to special cases where an edge connects at least two distinct relevant cycles to another part of the molecule (Figure 2). This edge is thus encoded by an hyperedge which connects the two sets of vertices $p(u)$ and $p(v)$.

This molecular hypergraph representation (Figure 3(c)) encodes all atoms $v \in V$ either by a vertex encoding a cycle or by v itself if $v \notin V(C_{\mathcal{R}})$. In the same way,

each atomic bond $e \in E$ is encoded within our molecular hypergraph representation. In addition, we note that set of vertices incident to an hyperedge defines a clique:

Theorem 1. *Let be a graph $G = (V, E)$ and its associated relevant cycle hypergraph $H_{RC}(G) = (V_{RC}, E_{RC})$. If $\exists e = (s_1, s_2) \in E^h_{RC}$ and c_1, $c_2 \in V_{RC}$ such that $\{c_1, c_2\} \subseteq s_1$ or $\{c_1, c_2\} \subseteq s_2$, then $(c_1, c_2) \in E^e_{RC}$, i.e. c_1 is adjacent to c_2.*

Proof. If $c_1 \in s_1$ and $c_2 \in s_1$, then by construction of E^h_{RC}, $\exists e = (u, v) \in E$ such that $\{c_1, c_2\} \subseteq p(u) = s_1$. By definition of function p and since $c_1, c_2 \in C_{\mathcal{R}}$, it holds that $u \in V(c_1) \cap V(c_2)$. By definition of relevant cycle graph, $(c_1, c_2) \in E_{\mathcal{C}_{\mathcal{R}}} \subset E^e_{RC}$. The proof for $c_1 \in s_2$ and $c_2 \in s_2$ is similar.

Algorithm 1 describes the different steps required to transform molecular graph G into its associated relevant cycle hypergraph H_{RC}. The first step consists in computing the relevant cycle graph of G, as described in [5], and initializing our hypergraph by this graph (Algo. 1, Lines 3 and 4). Then, the set of acyclic parts is included to the current graph representation (Algo. 1, Lines 6 and 7). Finally, hyperedges are included into our relevant cycle hypergraph (Algo. 1, Line 9).

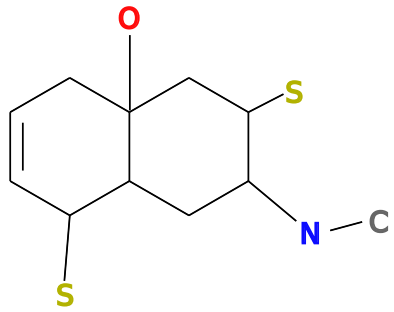
4 Similarity between Relevant Cycle Hypergraphs

The previous section defines a molecular representation which provides a new way to encode adjacency relations between cyclic and acyclic parts of a molecule. In order to apply QSAR methods on this molecular representation, we have to define a similarity measure between relevant cycle hypergraphs. Graph kernels, such as treelet kernel [4], are only defined on molecular graphs and can not be applied directly on an hypergraph representation of a molecule. In this section, we propose to adapt treelet kernel to the comparison of relevant cycle hypergraphs.

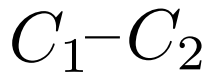
Treelet kernel is a graph kernel defined as a kernel between two sets of patterns extracted from both graphs to be compared. The set of extracted patterns,

Algorithm 1 Computing relevant cycle hypergraph from molecular graph.

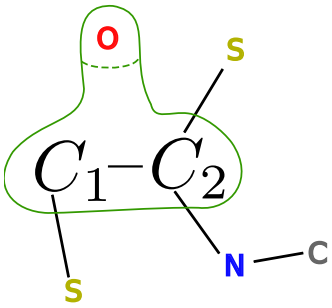
Require: $G = (V, E)$
Ensure: $H_{RC} = (V_{RC}, E_{RC}), E_{RC} = E^e_{RC} \cup E^h_{RC}$
1: $G_{\mathcal{C}}(\mathcal{C}_{\mathcal{R}}, E_{\mathcal{C}_{\mathcal{R}}}) = G_{\mathcal{C}}(G)$ {Relevant cycle graph}
2: {Adding all information included within cycles}
3: $V_{RC} = \mathcal{C}_{\mathcal{R}}$
4: $E^e_{RC} = E_{\mathcal{C}_{\mathcal{R}}}$
5: {Adding information not included within a cycle}
6: $V_{RC} = V_{RC} \cup \{v \notin V(\mathcal{C}_{\mathcal{R}})\}$
7: $E^e_{RC} = E^e_{RC} \cup \{(p(u), p(v)) \mid (u, v) \in E, |p(u)| = 1 \text{ AND } |p(v)| = 1\}$
8: {Special case (Figure 2).}
9: $E^h_{RC} = \{(p(u), p(v)) \mid (u, v) \in E, |p(u)| > 1 \text{ OR } |p(v)| > 1\}$
10: **return** H_{RC}



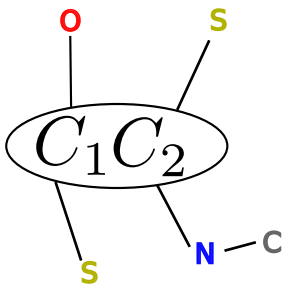
(a) Molecular graph G including cycles.



(b) Relevant cycle graph $G_{\mathcal{C}}$.



(c) Relevant cycle hypergraph $H_{RC}(G)$.



(d) Reduced relevant cycle graph $G_{RCR}(G)$.

Fig. 3. Different encodings of a same molecule.

denoted $\mathcal{T}$ and called treelets, is composed of all labeled trees with a number of vertices less than or equal to 6. Based on the explicit enumeration of this set of substructures, each graph G is associated to a vector $f(G)$ encoding the number of occurrences of each treelet t in G:

$$f(G) = (f_t(G))_{t\in\mathcal{T}(G)} \text{ with } f_t(G) = |(t \trianglelefteq G)| \tag{1}$$

where $\mathcal{T}(G)$ denotes the set of treelets extracted from G and $\trianglelefteq$ the subgraph isomorphism relationship. Using this vector representation, similarity between treelet distributions is computed using a sum of subkernels between treelet's number of occurrences:

$$K_{\mathcal{T}}(G, G') = \sum_{t\in\mathcal{T}(G)\cap\mathcal{T}(G')} k(f_t(G), f_t(G')) \tag{2}$$

where $k(.,.)$ defines any positive definite kernel between real numbers such as linear kernel, Gaussian kernel or intersection kernel. Despite the fact that this method may be applied on many kinds of graphs, it can not be directly applied to hypergraphs.

An hypergraph encodes global relationships defined between sets of vertices. At the opposite, treelet kernel is defined on graphs where relationships are defined locally between elementary vertices. Therefore, in order to apply treelet kernel to our hypergraph representation, we have to transform global relationships defined within our hypergraph representation to local relationships between elementary vertices. This transformation is performed by merging all sets of vertices incident to an hyperedge. This merge operation relies to transform hyperedges to edges.

An equivalence relation $\sim$ between vertices $c \in V_{RC}$ is defined such that $c_1 \sim c_2$ if and only if $\exists e = (s_1, s_2) \in E_{RC}$ such that $\{c_1, c_2\} \subseteq s_1$ or $\{c_1, c_2\} \subseteq s_2$. Using equivalence relation $\sim$ previously defined, we can now define the equivalence class $\bar{c} = \{c'; c \sim c'\}$ of a vertex c. Intuitively, two cycles sharing a common hyperedge belong to the same equivalence class. Then, by applying a contraction kernel on each class $\bar{c}$, we define a reduced relevant cycle graph $G_{RCR} = (V_{RCR}, E_{RCR})$ with:

- $V_{RCR} = \{\bar{c}, c \in V_{RC}\}$,
- $E_{RCR} = \{e = (\bar{c_1}, \bar{c_2}), (c_1, c_2) \in E_{RC}, c_1 \nsim c_2\}$. Intuitively, the set of edges E_{RCR} corresponds to the union of the usual edges E^e_{RC} of H_{RC} and the transformation of hyperedges E^h_{RC} into usual edges.

Labelling function $\mu_{RCR}(\bar{c}), c \in V_{RC}$, is defined in a canonical way by the sequence of atom and edge labels encountered during a depth first traversal of the spanning tree covering $\bar{c}$ and having the lowest lexicographic order. Such a spanning tree exists since any pair of vertices $\{c, c'\}$ sharing a same hyperedge is connected (Theorem 1).

Given this second representation of a molecule defined by the reduced relevant cycle graph, our new similarity measure based on treelet kernel is defined in two parts. A first step aims to extract the set of treelets $\mathcal{T}_1 = \mathcal{T}(V_{RC}, E^e_{RC})$. (V_{RC}, E^e_{RC}) corresponds to a sub hypergraph of H_{RC} which does not include any hyperedge $e \in E^h_{RC}$. Therefore, the set of treelets $\mathcal{T}_1$ encodes information which does not include special cases depicted in Figure 2. Information corresponding to these special cases, encoded by hyperedges $e \in E^h_{RC}$, is included into our similarity measure by the set of treelets $\mathcal{T}_2$ extracted from the reduced relevant cycle Graph G_{RCR} built from the transformation of hyperedges into edges. In order to avoid redundancy, we reduce the set of treelets $\mathcal{T}_2$ to treelets containing at least one edge corresponding to an hyperedge $e_h \in E^h_{RC}$. Finally, we define the set of treelets $\mathcal{T}_{CR}(G)$ associated to a molecular graph G by $\mathcal{T}_1 \cup \mathcal{T}_2$. Similarity between molecules is then defined as a sum of subkernels comparing number of occurrences of each treelet $t \in \mathcal{T}_{CR}(G)$ (Equation 2). This approach allows us to use a set of patterns which encodes most of the adjacency relations between cyclic and acyclic parts.

5 Experiments

We have tested our new molecule representation on an experiment defined as a classification problem. This dataset is taken from the Predictive Toxicity Chal-

Table 1. Classification accuracy on PTC.

	Method	MM	FM	MR	FR
1	Treelet Kernel (TK)	208	205	209	212
2	TK on cycles (TC)	211	210	203	232
3	Treelet on relevant cycle hypergraph (TCH)	217	224	207	233
4	Cyclic Pattern Kernel [6]	209	207	202	228
5	Gaussian Edit Distance Kernel [10]	223	212	194	234
6	TK + MKL	218	224	224	250
7	TC + MKL	216	213	212	237
8	TCH + MKL	225	229	215	239
9	Combo TK - TC	219	226	**226**	251
10	Combo TK - TCH	**225**	**230**	224	**252**

lenge [11] which aims to predict carcinogenicity of chemical compounds applied to female (F) and male (M) rats (R) and mice (M). This experiment is based on ten different datasets for each class of animal, each of them being composed of one train set and one test set. The amount of predicted molecules is equals to 336 for male mice, 349 for female mice, 344 for male rats and 351 for female rats. Table 1 shows the amount of correctly classified molecules over the ten test sets for each method and for each class of animal. The first three lines of Table 1 shows results obtained by a treelet kernel applied on differents molecular representations. Line 1 corresponds to treelet kernel applied on molecular graph, Line 2 to relevant cycle graph and Line 3 corresponds to kernel defined in Section 3. First, we can note that our new molecular representation obtains the best results among the three tested representations. This observation validates our hypothesis on the importance of relationships between cyclic and acyclic parts. This results can be compared with two other graph kernels. Line 4 shows results obtained by the kernel defined by Horváth based on the set of relevant cycles common to two molecules. As we can see, omitting relevant cycles relationships and adjacency relationships between cyclic and acyclic parts decreases the accuracy of this kernel. Line 5 corresponds to a graph kernel based on the notion of edit distance [10] between molecular graphs. This kernel obtains better results than treelet kernel applied on relevant cycle hypergraph for two classes over four. The second part of Table 1 shows results obtained by treelet kernels after a feature weighting step as defined in [5]. After this weighting step, treelet kernel applied on our new representation (Table 1, Line 8) obtains best results on two classes of animals and obtains second best results on the two other classes when only considering Lines 6 to 8. Note that our sparse feature weighting step reduces the number of treelets extracted from relevant cycle hypergraphs from 5700 to 25 relevant treelets. In comparison, treelet weighting step applied on molecular graph reduces the set of treelets from 3500 to 150 treelets. Note that this optimal weighting step selects both non linear treelets and treelets having 6 nodes which validates the relevance of using such substructures. Finally, treelet kernel has been combined with relevant cycle graph (Table 1, Line 9) and our new representation (Table 1, Line 10). This combination of two molecular representations

obtains the best results on three classes over four of animals when compared to combination of relevant cycle graph and molecular graph representations, hence showing the relevance of our molecular representation.

6 Conclusion

In this article, we have defined a new molecular representation based on hypergraphs which is able to encode adjacency relationships between cyclic and acyclic parts of a molecule. In addition, we have proposed a method to apply treelet kernel on our hypergraph representation. Our experiments show that the adjacency information encoded by this molecular representation can lead to better results than methods applied on classic molecular graphs. One outlook of this work consists in including the relative positioning of bonds connecting acyclic parts of a molecule on a same cycle.

References

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 the 22nd international conference on Machine learning - 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 Gaüzére, Luc Brun, Didier Villemin, and Myriam Brun. Graph kernels based on relevant patterns and cycle information for chemoinformatics. In *Proceedings of ICPR 2012*, pages 1775–1778. IAPR, IEEE, November 2012.
6. Tamás Horváth. Cyclic pattern kernels revisited. *PAKDD 2005*, pages 791–801, 2005.
7. Tamás Horváth, Thomas Gartner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 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)(September 2008):3–35, 2009.
10. Michel Neuhaus and Horst Bunke. *Bridging the gap between graph edit distance and kernel machines.* World Scientific Pub Co Inc, 2007.
11. 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.
12. Jean-Philippe Vert. The optimal assignment kernel is not positive definite. http://hal.archives-ouvertes.fr/hal-00218278.
13. Philippe Vismara. Union of all the minimum cycle bases of a graph. *The Electronic Journal of Combinatorics*, 4(1):73–87, 1997.