Graph kernel encoding substituents' relative positioning

Benoit Gaüzère, Luc Brun, Didier Villemin

Graph kernel encoding substituents' relative positioning

Benoît Gaüzère, Luc Brun
GREYC UMR CNRS 6072
Caen, France
{benoit.gauzere,luc.brun}@ensicaen.fr

Didier Villemin
LCMT UMR CNRS 6507,
Caen, France
didier.villemin@ensicaen.fr

***Abstract*—Chemoinformatics aims to predict molecular properties using informational methods. Computer science's research fields concerned by this domain are machine learning and graph theory. An interesting approach consists in using graph kernels which allow to combine graph theory and machine learning frameworks. Graph kernels allow to define a similarity measure between molecular graphs corresponding to a scalar product in some Hilbert space. Most of existing graph kernels proposed in chemoinformatics do not allow to explicitly encode cyclic information, hence limiting the efficiency of these approaches. In this paper, we propose to define a cyclic representation encoding the relative positioning of substituents around a cycle. We also propose a graph kernel taking into account this information. This contribution has been tested on three classification problems proposed in chemoinformatics.**

I. Introduction

Chemoinformatics aims to predict molecular properties from their structural similarities. Most of existing methods are based on fingerprints defined as collections of descriptors such as the boiling point, logP, molar refractivity, etc. An alternative strategy consists in computing a set of descriptors directly from the molecular graph G of a molecule. A molecular graph $G = (V, E, \mu, \nu)$ consists of an unlabeled graph (V, E) encoding the structure of a molecule and two labeling functions μ and ν mapping respectively each node to a label encoding atom's chemical element and each edge to a label identifying a type of atomic bond (single, double, triple or aromatic). Considering this molecular representation, the similarity between molecules can be directly deduced from the similarity of their molecular graphs instead of using a set of descriptors chosen a priori.

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. This latter property allows us to combine graph kernels with widely-used machine learning algorithms, such as SVM, thanks to the kernel trick which consists in replacing the scalar product $\langle \psi(G), \psi(G') \rangle$ by the value of $k(G, G')$ in these algorithms. Graph kernels provide thus a natural connection between molecular graphs and machine learning frameworks. A large family of kernels is based on bags of patterns. These methods consist in extracting a bag of patterns from a graph and deducing the similarity between two graphs from the similarity between their respective bags. The Hilbert space associated to the computed kernel is directly induced by the space associated to the set of extracted patterns. Therefore, the similarity measure encoded by a kernel mostly depends on the patterns extracted from molecular graphs. Most of existing methods are defined on linear patterns [1]. Such methods are generally associated to a low complexity but they are limited by the lack of expressivity of linear patterns. Nevertheless, in order to encode more structural information, some kernels are based on non-linear acyclic patterns [2], [3]. However, kernels defined directly on molecular graphs do not explicitly take into account the cyclic similarities of molecules. For example, treelet kernel [3] deduces the similarity of two molecules from the similarity of two bags of patterns corresponding to the set of all labeled subtrees composed of at most 6 nodes (figure 4). Obviously, considering a set of acyclic structures extracted from the molecular graph does not allow to encode the cyclic information. Nonetheless, cycles have an important impact on molecules' properties since they reduce atoms' degrees of freedom. Therefore, this information must be taken into account in order to define accurate similarity measures.

Considering this, Horváth proposed in [4] to combine a kernel encoding an acyclic similarity with an intersection kernel defined on the set of simple cycles of a graph. Despite the high complexity of the enumeration of all simple cycles of a graph, this method can be efficiently used when graphs have a low number of cycles. In order to tackle the complexity induced by the enumeration of all simple cycles, Horváth proposed in [5] to use a subset of simple cycles. This set is first initialized using the set of relevant cycles [6] of a graph. Then, additional simple cycles are iteratively enumerated by combining relevant cycles and newly discovered cycles. Horváth shown that a low number of iterations is sufficient to obtain results similar to the ones obtained using all simple cycles. In order to encode a finer cyclic information, Gaüzère et al. proposed in [7] to deduce the cyclic similarity of two molecules from the similarity of their relevant cycle graphs [8]. This approach allows to encode adjacency relationships between cycles in addition to isolated cycles as proposed by Horváth. However, this molecular representation does not include any acyclic information. Such information may be encoded by the relevant cycle hypergraph as defined in [9]. This molecular representation defines a global representation of a molecule while explicitly encoding cyclic information. Treelet kernel can then be adapted in order to define a similarity measure based on this molecular representation.

However, these approaches can not encode the relative positioning of the set of substituents around a given cycle. This information is particularly important in chemoinformatics since it characterizes the spatial configuration of a molecule which in turn strongly influences many biological properties. Therefore, we propose in this paper to define a graph kernel encoding this information. To this end, we first review in section II the different approaches to encode cyclic information

into graph kernels. Then, we propose in section III-A a representation, called augmented cycles, which allows to encode the relative positioning of substituents for each relevant cycle of a molecule. Then, we propose in section III-B a kernel based on the treelet kernel which includes the information encoded by augmented cycles. This approach relies on increasing the information carried by a treelet. Finally, section IV demonstrates the insight of our contribution.

II. Molecular cyclic information

A. Encoding cyclic information

Considering a molecular graph $G = (V, E, \mu, \nu)$, a simple cycle C is defined as a connected subgraph $C = (V', E', \mu, \nu)$ of G where each vertex $v \in V'$ has a degree equal to 2. In order to explicitly encode the cyclic information of a molecule, Horváth proposed to consider all simple cycles [4]. However, this first approach has an high complexity which restricts its use to molecules having a low number of cycles. A less complex approach consists in considering a subset of simple cycles such as the set of relevant cycles [6]. In order to define the notion of relevant cycle, let us first introduce the notion of cyclic vector space. 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. The set of vectors encoding the cycles of G defines a vector space where the addition of two cycles C and C' corresponds to a XOR bitwise [6]. The set of relevant cycles of a graph G, denoted $\mathcal{C}_{\mathcal{R}}(G)$, is defined by the union of all bases of the vector space of minimum length, the length of a base being defined as the sum of lengths of its cycles. Therefore, a relevant cycle can not be encoded by a combination of cycles having a lower size. Intuitively, the set of relevant cycles corresponds to a canonical set of elementary cycles which allows to retrieve all cycles included in a molecule. This set of cycles can be enumerated in polynomial time with the number of nodes of the graph. Therefore, enumerating the set of relevant cycles instead of the set of all simple cycles allows to obtain a reduced complexity.

The first approach which consists in encoding a molecule by a set of cycles allows to explicitly characterize an important part of the cyclic information. However, this approach is only based on isolated cycles which do not allow to identify the cyclic system of a molecule. This system is defined by the relevant cycle graph (figure 1(c)), initially proposed by [8] and modified by [7]. The relevant cycle graph defines a graph representation which encodes the set of relevant cycles and their adjacency relationships:

Definition 1 (Relevant cycle graph)**.** *Let us consider a graph $G = (V, E, \mu, \nu)$. The relevant cycle graph associated to G is defined by a graph $G_{\mathcal{C}}(G) = (V_{\mathcal{C}}, E_{\mathcal{C}}, \mu_{\mathcal{C}}, \nu_{\mathcal{C}})$ such that :*

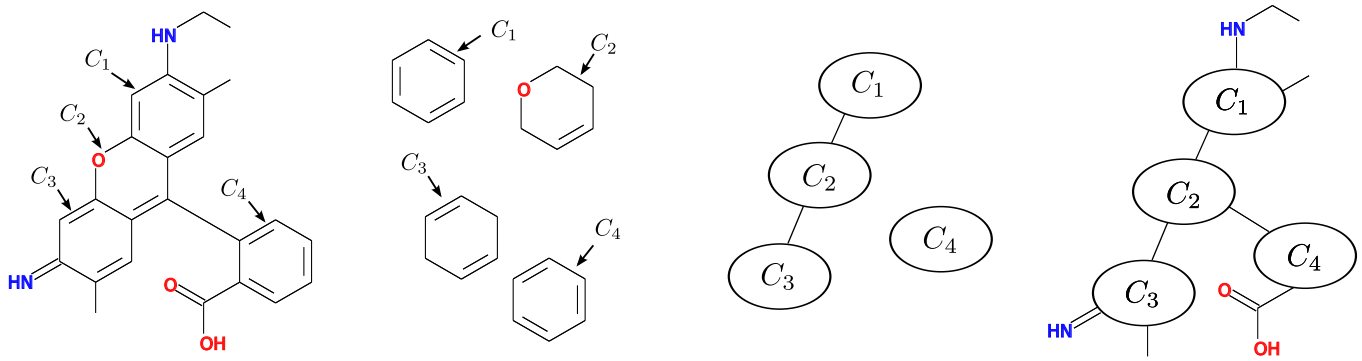
- *Each vertex $v \in V_{\mathcal{C}}$ corresponds to a relevant cycle $c_v \in \mathcal{C}_{\mathcal{R}}(G)$;*
- *two nodes $(u, v) \in V_{\mathcal{C}}^2$ are connected by an edge $(u, v) \in E_{\mathcal{C}}$ if the corresponding cycles c_u and c_v share at least one vertex in G.*

The relevant cycle graph allows to encode the cyclic system of a given molecule by mapping each relevant cycle to a node. Each node $v \in V_{\mathcal{C}}$ is then labeled by a label's sequence $\mu(v)$ identifying the set of atoms and atomic bonds included within the cycle c_v encoded by v. Similarly, each edge $e \in E_{\mathcal{C}}$ is labeled by a label's sequence identifying the set of nodes and edges common to the two cycles incident to e. Including cycles' adjacency relationships allows to encode a finer cyclic information than the one encoded by a set of isolated cycles. However, we can note that this representation encodes the cyclic system of a molecule by removing all acyclic parts. Therefore, an important part of the molecule, such as adjacency relationships between cycles and acyclic parts, is not encoded.

In order to encode a finer cyclic information, Gaüzère et al. proposed in [9] a molecular representation, called relevant cycle hypergraph (figure 1(d)), which allows to add acyclic parts to the representation defined by the relevant cyclic graph. A relevant cycle hypergraph $H_{CH}(G) = (V_{CH}, E_{CH}, \mu_{CH}, \nu_{CH})$ can be defined from a molecular graph $G = (V, E, \mu, \nu)$ and its relevant cycle graph $G_{\mathcal{C}}(G) = (V_{\mathcal{C}}, E_{\mathcal{C}}, \mu_{\mathcal{C}}, \nu_{\mathcal{C}})$. The set of nodes $V_{\mathcal{C}}$ can be associated to the set $V_{C_{\mathcal{R}}} \subseteq V$ which corresponds to the set of nodes of G included within a cycle. Similarly, $E_{C_{\mathcal{R}}} \subseteq E$ encodes the set of edges included in at least one relevant cycle. The complement of $V_{C_{\mathcal{R}}}$ and $E_{C_{\mathcal{R}}}$ in V and E corresponds thus to the acyclic parts of the molecular graph G which are not encoded by its relevant cycle graph. In order to define a complete molecular representation including both cyclic and acyclic parts, the set of nodes V_{CH} of the relevant cycle hypergraph is defined by the union of two subsets. A first subset $V_{\mathcal{C}}$ corresponding to the set of relevant cycles and thus encoding an explicit cyclic information, and a second subset $V - V_{C_{\mathcal{R}}}$ corresponding to the set of atoms not included within a cycle. Considering the set of vertices V_{CH}, [9] defines a function $p : V \to \mathcal{P}(V_{CH})$ as $p(u) = \{u\}$ if $u \notin V_{C_{\mathcal{R}}}$ and $p(u) = \{c \in V_{\mathcal{C}} \mid u \in V(c)\}$ if not, where $V(c)$ corresponds to the set of nodes included within the relevant cycle encoded by c. In other words, if $u \in V_{C_{\mathcal{R}}}$, $p(u)$ is defined as all nodes $v \in V_{\mathcal{C}}$ whose associated relevant cycle includes u. This function p encodes thus the print of vertex $v \in V$ on V_{CH}. In the same way as for vertices, the set of hyperedges E_{CH} is composed of two subsets:

1) A set of edges E_{CH}^e composed of:
   - edges between relevant cycle vertices, corresponding to the set of edges $E_{\mathcal{C}}$ included within the relevant cycle graph,
   - 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 the molecular graph G connecting two acyclic atoms or two relevant cycles or a relevant cycle to an acyclic part of G,
2) and a set E_{CH}^h composed of oriented hyperedges $e = (p(u), p(v))$ 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. Such an edge connects two sets of vertices $s_1 = p(u)$ and $s_2 = p(v)$ and is thus encoded by an oriented hyperedge $e = (s_1, s_2) \in E_{CH}^h$.

Thanks to the use of hyperedges, the relevant cycle hypergraph allows to encode adjacency relationships between cycles and acyclic parts. These relationships may be useful to identify



(a) Molecular graph. (b) Set of relevant cycles. (c) Relevant cycle graph. (d) Relevant cycle hypergraph.

Fig. 1. Different molecular representations encoding explicitly the cyclic information.

some particular substructures which have an important influence on some molecular properties. Note that, by adding the acyclic parts to the cycles encoded by the relevant cycle graph, each node and each edge of the original molecular graph is encoded within the relevant cycle hypergraph, either by the node or the edge itself or by a node encoding a relevant cycle. Therefore, this representation corresponds to a complete molecular representation and acyclic, cyclic and acyclic/cyclic adjacency relationships are included into the relevant cycle hypergraph. In addition, since this representation is based on the relevant cycle graph, the cyclic information is explicitly encoded.

B. Kernels computing a cyclic similarity

The different representations presented in the previous section allow to encode different levels of cyclic information. These different representations can thus be used to define different similarity measures. First graph kernels introduced in chemoinformatics consists in extracting a bag of patterns composed of acyclic sub-structures directly from the original molecular graph [10], [11], [3]. The similarity of the two molecules being compared is then deduced from the similarity of the two extracted bags of patterns. For example, the treelet kernel [3] is based on a bag of patterns $\mathcal{T}(G)$, called treelets, defined as all labeled subtrees composed of at most 6 nodes (figure 4). Given two bags of treelets $\mathcal{T}(G)$ and $\mathcal{T}(G')$ extracted from two molecular graphs G and G', the similarity between G and G' is encoded by a kernel $k(G, G')$ defined as:

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

where each term of the sum corresponds to a subkernel k' encoding a similarity measure between the number of occurrences of treelet t in G (encoded by $f_t(G)$) and G' (encoded by $f_t(G')$). This kernel allows to take into account most of the information encoded within a molecular graph by considering labeled and nonlinear patterns. However, since the similarity is computed from acyclic structures extracted from the molecular graph, such kernels can only partially and implicitly encode the cyclic information included in molecular graphs.

In order to compute a graph kernel taking into account the cyclic similarity of molecules, Horváth proposed in [5] to define a kernel on a set of cycles extracted from both graphs to be compared. This kernel is defined as an intersection kernel on the set of relevant cycles, hence computing the number of common relevant cycles of two molecular graphs. Acyclic similarity is then included by a second kernel computing the number of common subtrees extracted from the bridges of the two molecules to be compared. While providing a first kernel including cyclic similarity, this approach is only based on a set of isolated cycles and can not take into account adjacency relationships between cycles.

An alternative graph kernel, defined in [7], consists in applying the treelet kernel on relevant cycle graphs instead of original molecular graphs in order to compute a cyclic system similarity measure. This kernel consists thus in extracting the set of treelets from the relevant cycle graphs associated to the molecules. This set of patterns allows us to encode most of the adjacency relationships encoded within relevant cycle graphs and thus the cyclic system of molecules. Considering this bag of patterns, similarity between relevant cycle graphs is defined analogously as the case of the original treelet kernel:

$$k_{\mathcal{C}}(G, G') = \sum_{\substack{t \in \mathcal{T}(G_{\mathcal{C}}(G)) \cap \\ \mathcal{T}(G_{\mathcal{C}}(G'))}} k'(f_t(G_{\mathcal{C}}(G)), (f_t(G_{\mathcal{C}}(G'))) \qquad (2)$$

The main difference with the kernel defined in equation 1 is that treelets are extracted from the relevant cycle graph, hence encoding an explicit cyclic information and cycle's adjacency relationships.

Similarly to the method proposed by Horváth, the cyclic system similarity measure can be combined with an acyclic similarity measure, such as the treelet kernel applied on the molecular graph, in order to define a global similarity measure between molecules including both cyclic and acyclic similarities. Despite the fact that this approach leads to good results on experiments involving cyclic molecules, this representation, as the one proposed by Horváth [5], separates cyclic and acyclic information by using two different molecular representations. Then, global similarity between molecules is computed by the combination of two distinct similarity measures, each of them being applied on one representation. This separation induces thus a loss of adjacency relationships between cyclic and acyclic parts in the similarity measure.

Adjacency relationships between a cycle and its substituents can be encoded within the relevant cycle hypergraph (section II-A). In the same way as the adaptation of the treelet

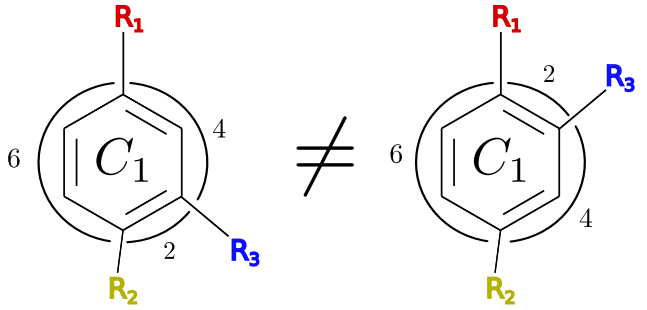


Fig. 2. Two cycles and their respective sets of substituents having a different relative positioning.

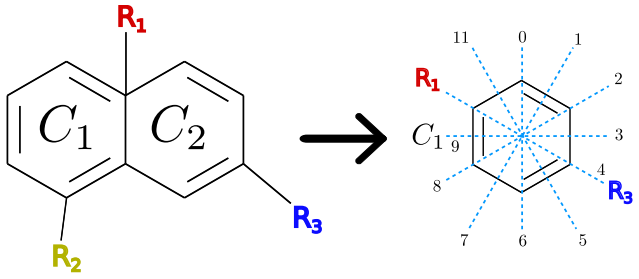


Fig. 3. Encoding of positions around a cycle.

kernel to relevant cycle graph comparisons, [9] proposed to adapt the treelet kernel to the comparison of relevant cycle hypergraphs. The set of treelets extracted from the relevant cycle hypergraph then encodes adjacency relationships between a cycle and its substituents in addition to the information encoded by the relevant cycle graph. This additional information allows to define a finer similarity measure which leads to an increasing accuracy on several molecular property prediction problems [9]. This last contribution defines thus a global similarity measure based on an unique molecular representation encoding the whole molecule and an explicit cyclic information.

III. Kernel based on augmented cycles

A. Definition of augmented cycles

Previous approaches allow to encode an important part of the cyclic information by defining different molecular representations encoding different levels of cyclic information. While these representations encode most of the cyclic information, they do not allow to distinguish the relative positioning of substituents around a given cycle. Indeed, the two patterns displayed in figure 2 are composed of a same cycle, denoted C_1, surrounded by a same set of substituents R_1, R_2 and R_3. However, we can note that these substituents are not positioned in the same way around C_1. Using a molecular representation encoding cycles adjacency relationships such as the relevant cycle hypergraph, these two patterns can not be distinguished since their representations are isomorphic. However, a molecular cycle is a planar structure and the relative positioning of its substituents characterizes the spatial configuration of the whole pattern. Since different spatial configurations may lead to different molecular properties, the relative positioning of the substituents should be taken into account in order to define an accurate similarity measure.

Let us consider a molecular graph G and its associated relevant cycle hypergraph $H_{CH}(G)$. Given a relevant cycle C of G encoded by c in $H_{CH}(G)$, the set of substituents of C is defined as the neighbourhood $\Gamma(c)$ of c. For example, N and C_2 are two substituents of C_3 in figure 1(a). Since a cycle is a planar structure, we can define a coherent orientation on its boundary and hence a cyclic order on the neighbourhood $\Gamma(c)$ of each vertex $c \in V_{CH}$ encoding a cycle.

Our basic idea to encode the relative positioning of the substituents of a cycle C consists in associating each substituent to a position on the boundary of C. Given the node $c \in V_{CH}$ associated to C, edges encoding the adjacency between c and nodes belonging to $\Gamma(c)$ may either correspond to an edge of the original molecular graph G or to an edge of the relevant cycle graph encoding a cycle's adjacency relationship. The position of substituents incident to edges of the original molecular graph may be easily defined since these edges are incident to a single atom of C (e.g. (C_3, N) in figure 1(a)). However, edges of the relevant cycle graph correspond to a common path between two cycles (e.g. (C_3, C_2) in figure 1(a)). In such cases we fix the position of the substituents incident to such edges to the middle of the shared path. Since this last position may not correspond to a cycle's atom but to the middle of an edge, the number of possible positions is equal to twice the number of edges of the cycle (figure 3). Using an arbitrary origin and orientation on C, these positions may be efficiently encoded by a function $\text{pos}_C : \Gamma(c) \to \mathbb{N}$ counting the number of half-edges between our origin and any substituent.

The relative positioning of two substituents according to C may be encoded by their angle. The angle between two substituents of C is defined as the number of half-edges separating them using the orientation fixed on C. This number of half-edges is directly retrieved from the substituents' positions encoded by the function pos_C:

Definition 2 (Angle between substituents)**.** *Using the above notations and hypothesis, given two nodes s_1 and s_2 of $\Gamma(c)$, the angle between s_1 and s_2 according to c is defined by:*

$$\text{ang}_c(s_1, s_2) = (\text{pos}_C(s_2) - \text{pos}_C(s_1)) \bmod (2|C|) \quad (3)$$

where $|C|$ denotes the number of edges of the cycle C of G.

We can note that equation 3 does not depend on the origin but only on the orientation defined on C. The combination of a cycle c and a function $\text{ang}_c : \Gamma(c)^2 \to \mathbb{N}$ defines the notion of augmented cycle.

Let us denote by $\pi(\Gamma(c))$ the set of possible orderings of $\Gamma(c)$. Using equation 3, any ordering $\sigma \in \pi(\Gamma(c))$ may be associated to a sequence of angles, each angle being defined between two successive elements of σ. Such a sequence is encoded by the function $\text{angles}(\sigma)$:

Definition 3 (Angle's sequence)**.** *Given a cycle encoded by $c \in V_{CH}$ and an ordering $\sigma = (s_1, \ldots, s_n)$ on $\Gamma(c)$, the angle's sequence $\text{angles}(\sigma) = \alpha_c^1 \ldots \alpha_c^{n-1}$ is defined by*

$$\alpha_c^i = \text{ang}_c(s_i, s_{i+1}), \forall i \in \{1, \ldots, n-1\}. \quad (4)$$

For example in figure 3, the ordering $\sigma = (R_1, R_3, C_1)$ of $\Gamma(C_2)$ corresponds to the sequence of angle's $(6, 5)$. Choosing

the alternative orientation on C_2 or a different origin leads to another ordering associated to a different angle's sequence. Let us consider a sequence of substituents $\sigma = (s_1, \ldots, s_n)$ and its associated sequence $\text{angles}(\sigma)$. We can define a coherent function pos'_c from both sequences by setting $\text{pos}'_c(s_1) = 0$ and $\text{pos}'_c(s_i) = \sum_{j=1}^{i-1} \alpha_c^j \mod 2|C|$ for each $i \in \{2, \ldots, n\}$. Such a function corresponds thus simply to a shift of the original function pos_c where $\text{pos}'_c(s_1)$ is set to the origin.

B. Augmented treelets

A treelet is defined by a subtree $T = (V_T, E_T, \mu_T, \nu_T)$ extracted from a graph $G = (V, E, \mu, \nu)$. Each treelet extracted from a graph is canonically identified by a function $\text{key}(T)$ defined as the combination of an index encoding its structure (V_T, E_T) (figure 4) and a code encoding the labels associated to each node and edge of T given a particular treelet's structure. This function is defined such as two isomorphic treelets T and T' are associated to the same key, i.e. $T \simeq T' \Leftrightarrow \text{key}(T) = \text{key}(T')$ [3]. The code identifying the labeling information of a treelet is defined as a node and edges' label sequence. When applied to relevant cycle hypergraph, each node of an extracted treelet may encode either an atom or a cycle and this node is labeled by the labeling function μ_{CH} defined on the relevant cycle hypergraph (section II-A). Two treelets encoding the two configurations shown in figure 2 are associated to a same structure and the same set of node labels. Therefore, these two treelets are isomorphic since the label of a node encoding a relevant cycle does not include the relative positioning of its adjacent nodes.

In order to encode this information, we propose to introduce the notion of augmented treelets by encoding the relative positioning of substituents into the notion of treelet. Given a treelet $T = (V_T, E_T, \mu_T, \nu_T)$ let us consider an order $\{v_1, \ldots, v_n\}$ on V_T induced by our key construction scheme [3]. This order is unique up to the automorphisms of T and induces an order on the neighborhood $\Gamma(c)$ of each vertex $c \in V_T$ encoding a cycle. Let us denote by $\Sigma_c(T)$ the sequence of subtrees incident to c, the order being deduced from the one defined on the neighborhood $\Gamma(c)$. Each automorphism $\tau_i \in Aut(T)$ of T induces a new numbering of the vertices of T and hence a new order on the neighborhood of vertices c corresponding to a relevant cycle. Let us denote by $\Sigma_c(\tau_i(T))$ the new ordering of $\Sigma_c(T)$ induced by τ_i. By definition of automorphism if $\Sigma_c(T) = t_1 \ldots . t_p$ and $\Sigma_c(\tau_i(T)) = t'_1 \ldots . t'_p$ we have $t_i \simeq t'_i$ for all $i \in \{1, \ldots, p\}$. In other words, if a tree automorphism maps two branches of a tree one onto the other, these branches should be isomorphic. However, these trees may be attached differently on the cycle encoded by c and thus correspond to different sequences of angles. We thus define the canonical sequence of angles $\text{angles}^\star(c)$ for each node $c \in V_T$ encoding a relevant cycle as:

$$\text{angles}^\star(c) = \min_{\tau_i \in \text{Aut}(T)} \text{angles}(\sigma_c(\tau_i(T))) \quad (5)$$

where $\sigma_c(\tau_i(T))$ corresponds to the order on $\Gamma(c)$ induced by $\Sigma_c(\tau_i(T))$. Let us consider a canonical sequence $\text{angles}^\star(c) = \alpha_c^1 \ldots \alpha_c^p$ and the initial order of subtrees $\Sigma_c(T) = t_1 \ldots . t_p$ defined on $\Gamma(c)$ in T. Each angle α_c^i encodes an angle between a subtree isomorphic to t_i and a subtree isomorphic to t_{i+1}. Hence, the sequence $\text{angles}^\star(c)$ encodes the angles between the different subtrees of T incident to c. Using Section II-A, the position of each subtree on the cycle associated to c may be retrieved up to a shift.

In order to include the relative positioning of substituents into our augmented treelets, we define a new labeling function $\mu'_T(v)$ as the concatenation of $\mu_T(v)$ and $\text{angles}^\star(v)$. The key associated to an augmented treelet is then computed in the same way as in [3] according to the augmented treelet $T_a = (V_T, E_T, \mu'_T, \nu_T)$. Two augmented treelets with a same key are thus isomorphic. Hence two vertices c and c' corresponding to relevant cycles and mapped one onto the other have the same label and hence the same set of connected subtrees and the same sequence of angles. According to our previous considerations, the positions of each subtree may be retrieved without ambiguities from the sequence of angles in both graphs. Both vertices encode thus cycles with the same connected subtrees separated by the same angles. Considering such a labeling function, the two configurations depicted in figure 2 may thus be distinguished. The set of augmented treelets allows us to extract a set of patterns encoding cyclic, acyclic and cyclic/acyclic adjacency relationships. In addition, our bag of treelets allows to distinguish configurations having a different subtituents' relative positioning, hence encoding a finer cyclic information than the set of treelets extracted from the relevant cycle hypergraph.

IV. Experiments

Our extension has been tested on three classification experiments involving molecular graphs composed of cycles. Results obtained by different kernels using SVM are displayed in table I. The first experiment, denoted PAH, consists in 94 molecules only composed of carbons. This experiment aims to predict if a molecule is carcinogenic or not. Results displayed in table I correspond to the best results obtained by using a 10-fold cross classification over a grid of parameters including both machine learning and kernel parameters. The second experiment, denoted Mutag, is taken from [12] and is composed of 4 337 molecules identified as mutagen or nonmutagen. This set of molecules is divided into 3 distinct subsets: 1 500 molecules are used to train our model, 500 as a validation set used to tune the parameters and the accuracy is evaluated on the remaining 2 337 molecules, hence composing the testset. Finally, the last experiment corresponds to the PTC [13] dataset which consists in predicting the toxicity of different molecules on four classes of animals: male mice (MM), female mice (FM), male rats (MR) and female rats (FR). Each class of animal is composed of ten trainsets and ten testsets and, similarly to the first experiment, parameters are tuned over the testset for each trainset for all methods. Table I displays the mean accuracy over the ten testsets of each class of animals.

In order to evaluate our contribution, we compare it with different kernels. Treelet kernel (Line 1) is a kernel based on a bag of patterns (section II-B) which do not explicitly encode cyclic information. The second kernel (Line 2) is computed by applying a gaussian kernel on an approximate edit distance [14] which thus do not explicitly encode cyclic information. Note that this kernel may need some regularization step in order to define a positive definite kernel. Next kernels explicitly include different levels of cyclic information (sections II-A and II-B). Line 3 corresponds to the kernel

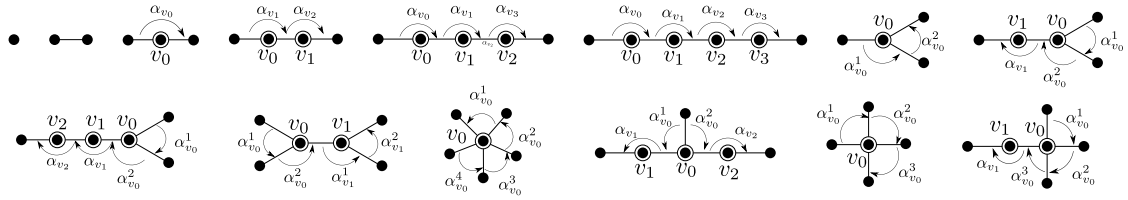


Fig. 4. Augmented treelets and the set of angles $(\alpha_1, \ldots, \alpha_n)$ associated to each potential relevant cycle having at least one pair of substituents.

TABLE I. ACCURACY ON MUTAG, PAH AND PTC DATASETS.

Method	Mutag	PAH	PTC			
			MM	FM	MR	FR
Treelet Kernel [3]	77.1	71.3	61.9	58.7	60.8	60.4
Graph edit distance [14]	71.5	72	66.4	60.7	56.4	66.7
Cycles [5]	64.5	63.0	62.2	59.3	58.7	65.0
Relevant cycle graph [7]	67.8	77.7	62.8	60.2	59.0	66.0
Relevant cycle hypergraph [9]	78.3	76.3	64.6	64.2	60.2	66.4
Augmented cycle	**80.2**	**80.7**	**67.9**	**64.8**	**61.3**	**68.7**

defined by Horváth in [5] which only compares the set of relevant cycles extracted from molecules to be compared. The kernel based on the relevant cycle graph (Line 4) allows to encode cycle adjacency relationships and the one based on the relevant cycle hypergraph (Line 5) encodes in addition the acyclic parts of molecules. Finally, line 6 corresponds to our contribution which encodes the position of each substituent along the boundary of each relevant cycle.

First, we can observe that the similarity measures which do not include explicitly the cyclic information are generally outperformed by the treelet kernel on relevant cycle hypergraph (line 5) which explicitly includes the cyclic information. This observation allows to highlight the relevancy of explicitly encoding cyclic information. However, we can note that considering only the cyclic system (Horváth and relevant cycle graph kernels) do not allow to obtain a good accuracy. Second, we can note that finer the cyclic information is encoded, better are the results. This observation is systematically observed on PTC and Mutag dataset whereas we observe a lower accuracy for the relevant cycle hypergraph on PAH dataset. This can be explained by the fact that the molecules included within PAH dataset are mainly cyclic. In this case, the addition of the acyclic information without including the relative positioning of substituents induces some irrelevant features. Conversely, adding acyclic parts together with the relative positioning of substituents lead to the best results on this dataset. In addition, we can note that our contribution allows to systematically outperform the results obtained by other kernels, hence showing the usefulness of our contribution. From a computational point of view, computing a Gram matrix on 150 molecules of kernel including the relative positioning of substituents requires about 1 s whereas it requires about 0.5 s for the kernel on relevant cycle hypergraph. These experiments have been performed thanks to the resources of computer center CRIHAN.

V. CONCLUSION

In this paper, we proposed a new cycle representation, called augmented cycles, which allows to encode cyclic information of molecules in a finer way. Our contribution consists in including the relative positioning of substituents around a cycle. This new representation can be used in molecular representations where the cyclic information is explicitly encoded, such as the relevant cycle hypergraph. In order to use this additional information, we also proposed to adapt the treelet kernel to encode the relative positioning of substituents by defining the augmented treelets. The different experiments show the relevancy of this additional information to solve classification problems in chemoinformatics. Our contribution defines only locally this orientation for a given cycle. Further works will aim to study the insight of using a coherent orientation, when it exists, between adjacent cycles. To that end, we will have to determine for each cycle its maximal planar connected component.

REFERENCES

[1] H. Kashima, K. Tsuda, and A. Inokuchi, *Kernels for graphs*. MIT Press, 2004, ch. 7, pp. 155–170.

[2] P. Mahé and J.-p. Vert, "Graph kernels based on tree patterns for molecules," *Machine Learning*, vol. 75(1), no. September 2008, pp. 3–35, 2009.

[3] B. Gaüzère, L. Brun, and D. Villemin, "Two new graphs kernels in chemoinformatics," *Pattern Recognition Letters*, vol. 33, no. 15, pp. 2038–2047, 2012.

[4] T. Horváth, T. Gärtner, and S. Wrobel, "Cyclic pattern kernels for predictive graph mining," in *Proceedings of the 2004 ACM SIGKDD*. ACM Press, 2004, p. 158.

[5] T. Horváth, "Cyclic pattern kernels revisited," in *Proceedings of KDD'04*, Springer-Verlag, Ed., vol. 3518, 2005, pp. 791 – 801.

[6] P. Vismara, "Union of all the minimum cycle bases of a graph," *The Electronic Journal of Combinatorics*, vol. 4, no. 1, pp. 73–87, 1997.

[7] B. Gaüzére, L. Brun, D. Villemin, and M. Brun, "Graph kernels based on relevant patterns and cycle information for chemoinformatics," in *Proceedings of ICPR 2012*, vol. 7626, November 2012, pp. 1775–1778.

[8] P. Vismara, "Reconnaissance et reprsentation d'lments structuraux pour la description d'objets complexes. application l'laboration de stratgies de synthse en chimie organique," Ph.D. dissertation, Universit Montpellier II, 1995.

[9] B. Gaüzère, L. Brun, and D. Villemin, "Relevant cycle hypergraph representation for molecules," in *Graph-Based Representations in Pattern Recognition*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 7877, pp. 111–120.

[10] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized Kernels Between Labeled Graphs," *Machine Learning*, 2003.

[11] P. Mahé and J.-P. Vert, "Graph kernels based on tree patterns for molecules," *Machine Learning*, vol. 75, no. 1, pp. 3–35, Oct. 2008.

[12] K. Riesen and H. Bunke, "Iam graph database repository for graph based pattern recognition and machine learning," in *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, ser. SSPR & SPR '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 287–297.

[13] H. Toivonen, A. Srinivasan, R. D. King, S. Kramer, and C. Helma, "Statistical evaluation of the predictive toxicology challenge 20002001," *Bioinformatics*, vol. 19, no. 10, pp. 1183–1193, 2003.

[14] M. Neuhaus and H. Bunke, *Bridging the gap between graph edit distance and kernel machines*, ser. Series in Machine Perception and Artificial Intelligence. World Scientific Publishing, 2007, vol. 68.