A new hypergraph molecular representation

Benoit Gaüzère, Luc Brun, Didier Villemin

A new hypergraph molecular representation.

Benoît Gaüzère[†], Luc Brun[†], and Didier Villemin[‡]

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

Abstract. In this contribution, we define a new molecular representation together with a similarity measure which allows to encode adjacency relationships between cycles and their substituents.

Keywords: Chemoinformatics, Graph kernel, Hypergraph

1 Introduction

Within chemoinformatics research field, molecules are generally encoded by their molecular graphs. Such a molecular representation encodes molecules by a graph $G = (V, E, \mu, \nu)$ where the set of nodes V encodes the set of atoms and the set of edges E encodes the set of atomic bonds between atoms. The labeling function $\mu : V \to L_v$ associates to each node a label encoding the chemical element of the corresponding atom. The labeling function $\nu : E \to L_e$ associates to each edge the type of its corresponding atomic bond (single, double, triple or aromatic). This representation is widely used in chemoinformatics and particularly in QSAR/QSPR problems in conjunction with machine learning methods through graph kernels [6, 5, 2]. Graph kernels can be understood as graph similarity measures corresponding to scalar products between vectorial representations of graphs. This last point allows to use them in conjunction with machine learning methods such as SVM. Similarity measures between molecules encoded by graph kernels can be deduced from the similarity of bags of patterns extracted from molecular graphs. These patterns may be defined as linear patterns (trails, paths and random walks) or non linear patterns such as tree structures which allows to encode more structural information. However, these similarity measures do not take into account the cyclic similarity of molecular graphs.

In order to take into account molecular cycles, the optimal assignment kernel is based on a reduced representation obtained by collapsing some structural elements such as cycles into one single node [1]. However this kernel is not definite positive [8] which restricts its application within machine learning methods. Another approach aims to extract the set of simple cycles of molecular graphs and defines cyclic similarity from the number of common simple cycles. This approach, called cyclic pattern kernel [4], is combined with a tree pattern kernel in order to define a complete similarity measure between graphs. In order to reduce the complexity and to encode a more relevant set of cycles, this kernel has been improved by enumerating the set of relevant cycles [10] instead of the

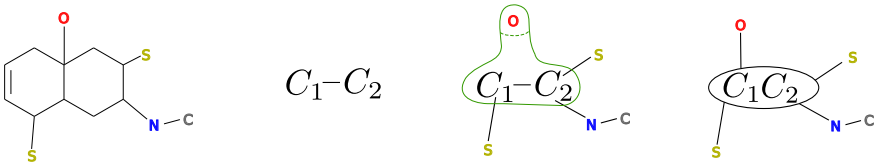
set of simple cycles. However, despite the fact that this kernel encodes cyclic similarity, it does not encode adjacency relationships between cycles. Therefore, cyclic information is only partially encoded by the cyclic pattern kernel.

In this contribution, we propose a new molecular representation which allows to encode adjacency relationships between molecular cycles. We also define a kernel based on this representation in order to resolve some QSAR/QSPR problems.

2 Relevant cycle hypergraph

We first encode adjacency relationships between cycles by the relevant cycle graph first introduced by Vismara and developed by [3] $G_{\mathcal{C}} = (\mathcal{C}_{\mathcal{R}}, E_{\mathcal{C}_{\mathcal{R}}}, \mu_{\mathcal{C}_{\mathcal{R}}}, \nu_{\mathcal{C}_{\mathcal{R}}})$ where each vertex $c \in \mathcal{C}_{\mathcal{R}}$ corresponds to a relevant cycle. An edge $e = (c_1, c_2) \in E_{\mathcal{C}_{\mathcal{R}}}$ iff cycles c_1 and c_2 share at least one vertex of the molecular graph. The labeling 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 shared by c and c'. This first step allows to encode cycles as single nodes and adjacency relationships between two relevant cycles. In order to define a complete molecular representation, we have to include acyclic parts to the relevant cycle graph.

In order to encode adjacency relationships between relevant cycles and acyclic parts, we propose to simply add acyclic parts to the relevant cycle graph by connecting an acyclic part to a cycle if it exists an edge connects the acyclic part and an atom of this cycle. However, a graph representation can not handle special cases where an acyclic part is connected to an atom included within two distinct cycles, such as atom O and cycles C_1 and C_2 in figure 1(a). On the other hand, hypergraphs allows to encode adjacency relationships between more than two nodes. Therefore, we propose to encode a molecule by the relevant cycle hypergraph $H_{RC}(G) = (V_{RC}, E_{RC}, \mu_{RC}, \nu_{RC})$ (figure 1(c)) in order to encode special cases as depicted in figure 1(a). The set of nodes V_{RC} is defined as the union of relevant cycles and atoms which are not included within any cycle. The set of hyperedges E_{RC} consists of a set of edges E^e_{RC} encoding adjacency relationships between relevant cycles, acyclic atoms or between one relevant cycle and one acyclic atom. Special cases involving more than two nodes are encoded by the set of hyperedges $e^h = (s_u, s_v) \in E^h_{RC} \subseteq E_{RC}$ where s_u, resp. s_v, encodes



(a) G. (b) $G_{\mathcal{C}}$. (c) $H_{RC}(G)$. (d) $G_{RCR}(G)$.

Fig. 1. Different encodings of a same molecule.

a set of subgraphs corresponding to the set of cycles including u, resp. v, or the atom itself if u, resp. v, is an acyclic part and $e = (u, v) \in E$. Labeling functions μ_{RC} and ν_{RC} correspond to labeling functions of nodes and edges either on molecular graph or relevant cycle graph.

The relevant cycle hypergraph encodes all atoms and edges included in a molecular graph since all cyclic and acyclic parts are encoded into our new molecular representation. However, similarity measures generally used in conjunction with machine learning methods are defined on graphs, not hypergraphs. In order to define a similarity measure between relevant cycle hypergraphs, we propose to adapt the treelet kernel [2] to the comparison of relevant cycle hypergraphs. Treelet kernel is a graph kernel based on a bags of patterns defined as all labeled sub trees having six nodes or less. In order to apply treelet kernel on hypergraph representation, we propose to define the bag of treelets $\mathcal{T}_{CH}$ as the union of two sets of treelets. The first subset $\mathcal{T}_1$ is composed of all sub trees having six nodes or less extracted from relevant the cycle hypergraph where hyperedges E^h_{RC} have been removed. This set of treelets encodes adjacency relationships between acyclic parts, cycles and between a cycle and an acyclic part. The second subset $\mathcal{T}_2$ is defined as the set of sub trees having six nodes or less extracted from a transformation G_{RCR} of the relevant cycle hypergraph defined by the contraction of sets $s_u \in E^h_{RC}$ into a single node (figure 1(d)). Since sets of nodes incident to any hyperedge have been contracted into one single node, hyperedges now correspond to edges and G_{RCR} corresponds to a graph. In order to avoid redundancy, $\mathcal{T}_2$ is restricted to the set of treelets containing at least one former hyperedge. Therefore, $\mathcal{T}_2$ encodes adjacency relationships corresponding to special cases where two or more relevant cycles are connected to an acyclic part. The set of treelets $\mathcal{T}_{CH} = \mathcal{T}_1 \cup \mathcal{T}_2$ is then defined as the bag of patterns used to compute treelet kernel. Therefore, this kernel allows to encode adjacency relationships between cycles and between cycles and their substituents.

3 Experiments and Conclusion

Table 1 shows the number of correctly classified molecules obtained by our contribution on the classification problem addressed by the PTC dataset [7]. These experiments shows the relevancy of encoding adjacency relationships between relevant cycles and their substituents. First, we can note that our new molecular representation together with a weighting step, which allows to only keep relevant sub trees [3], obtains the best results on two datasets over four. In addition, we can note that finer the cyclic information is encoded, better are the results (lines 2 to 4). Finally, best results are obtained by combining our relevant cycle hypergraph kernel, which encodes cyclic similarity, with a treelet kernel which only encodes acyclic similarity. The trade off between acyclic and cyclic contributions has to be tuned according to each chemoinformatics problem.

In conclusion, our contribution defines a new molecular representation, the relevant cycle hypergraph, which allows to encode adjacency relationships between cycles and their substituents. Thanks to the adaptation of a graph kernel

Table 1. Classification accuracy on PTC dataset.

Method	# correct predictions			
	MM	FM	MR	FR
(1) Treelet kernel (TK) [2]	208	205	209	212
(2) Cyclic pattern kernel [4]	209	207	202	228
(3) TK on relevant cycle graph (TC) [3]	211	210	203	232
(4) TK on relevant cycle hypergraph (TCH)	217	224	207	233
(5) TK with weighting step	217	224	223	250
(6) TC with weighting step	216	213	212	237
(7) TCH with weighting step	**225**	229	215	239
(8) TK + λTCH	**225**	**230**	**224**	**252**

to relevant cycle hypergraph comparisons, this molecular representation allows to obtain a better accuracy on QSAR/QSPR datasets. In order to encode finer cyclic information, future works will aim to encode the relative positioning of cycle substituents.

References

1. 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.
2. Benoit Gaüzère, Luc Brun, and Didier Villemin. Two New Graphs Kernels in Chemoinformatics. *Pattern Recognition Letters*, 33(15):2038–2047, 2012.
3. 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.
4. Tamás Horváth, Thomas Gärtner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, page 158. ACM Press, 2004.
5. Pierre Mahé and Jean-philippe Vert. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1)(September 2008):3–35, 2009.
6. Nino Shervaszide. *Scalable Graph Kernels*. PhD thesis, Universitt Tbingen, 2012.
7. 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.
8. Jean-Philippe Vert. The optimal assignment kernel is not positive definite. *CoRR*, abs/0801.4061, 2008.
9. Philippe Vismara. *Reconnaissance et représentation d'éléments structuraux pour la description d'objets complexes. Application l'élaboration de stratégies de synthèse en chimie organique*. PhD thesis, Universit Montpellier II, 1995.
10. Philippe Vismara. Union of all the minimum cycle bases of a graph. *The Electronic Journal of Combinatorics*, 4(1):73–87, 1997.