Graph Kernels in Chemoinformatics

Luc Brun

Co authors: Benoit Gaüzère(MIVIA), Pierre Anthony Grenier(GREYC), Alessia Saggese(MIVIA), Didier Villemin (LCMT)

Bordeaux,
November, 20 2014

Plan

Plan

Motivation

Structural Pattern Recognition

- ✚ Rich description of objects
- ▬ Poor properties of graph's space does not allow to readily generalize/combine sets of graphs

Statistical Pattern Recognition

- ▬ Global description of objects
- ✚ Numerical spaces with many mathematical properties (metric, vector space, ...).

Motivation

Analyse large famillies of structural and numerical objects using a **unified** framework based on pairwise similarity.

Chemoinformatics

- Usages of computer science for chemistry
- **Prediction of molecular properties**
- Prediction of biological activities
  - Toxicity, cancerous ...
- Prediction of physico-chemical properties.
  - Boiling point, LogP, ability to capture CO_2...

Screening :

- Synthesis of molecule,
- Test and validation of the synthesized molecules

Virtual Screening :

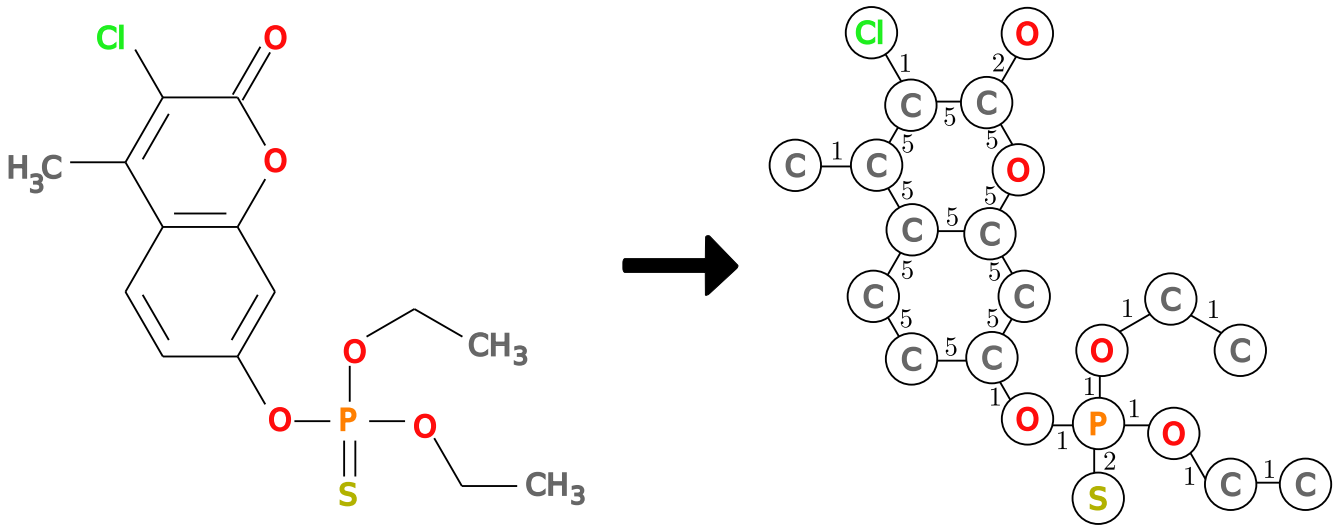
- Selection of a limited set of candidate molecules,
- Synthesis and test of the selected molecules,

→ Gain of Time and Money

Molecular representation

Molecular Graph:

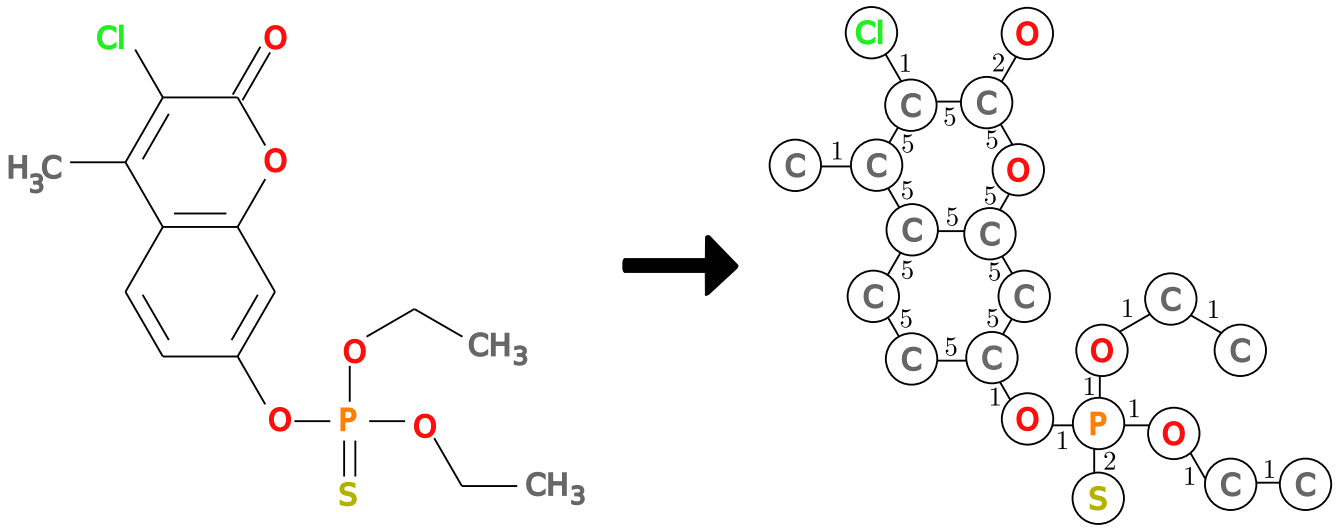
- Vertice ⟶ atoms.
- Edges ⟶ Atomic bonds.



Molecular representation

Molecular Graph:

- Vertice ⟶ atoms.
- Edges ⟶ Atomic bonds.



Similarity principle [Johnson and Maggiora, 1990]

- *"Similar molecules have similar properties."*

⟶ Design of similarity measures between molecular graphs.

Plan

Kernel Theory

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- Symmetric: $k(x_i, x_j) = k(x_j, x_i)$,
- For any dataset $D = \{x_1, \ldots, x_n\}$, Gram's matrix $K \in \mathbb{R}^{n \times n}$ is defined by $K_{i,j} = k(x_i, x_j)$.
- Semi-definite positive: $\forall \alpha \in \mathbb{R}^n, \alpha^T K \alpha \geq 0$

Kernel Theory

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- Symmetric: $k(x_i, x_j) = k(x_j, x_i)$,
- For any dataset $D = \{x_1, \ldots, x_n\}$, Gram's matrix $K \in \mathbb{R}^{n \times n}$ is defined by $K_{i,j} = k(x_i, x_j)$.
- Semi-definite positive: $\forall \alpha \in \mathbb{R}^n, \alpha^T K \alpha \geq 0$

Connection with Hilbert spaces, reproducing kernel Hilbert Spaces [Aronszajn, 1950]:

- Relationship scalar product $\Leftrightarrow$ kernel :

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$

Graph Kernels

Graph kernels in chemoinformatics

- Graph kernel $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$.
- Similarity measure between molecular graphs
- ✚ Natural encoding of a molecule
- ✚ Implicit embedding in $\mathcal{H}$,
- ✚ Machine learning methods (SVM, KRR, ...).

Graph Kernels

Graph kernels in chemoinformatics

- Graph kernel $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$.
- Similarity measure between molecular graphs
- ✚ Natural encoding of a molecule
- ✚ Implicit embedding in $\mathcal{H}$,
- ✚ Machine learning methods (SVM, KRR, ...).

→ Natural connection between molecular graphs and machine learning methods

→ Definition of a similarity measure as a kernel

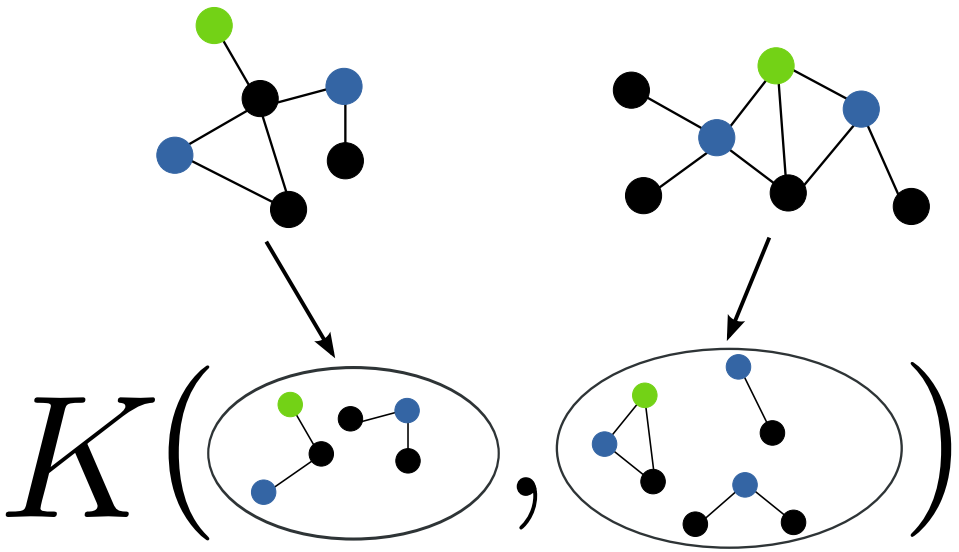
Kernels based on bags of patterns

Graph kernels based on bags of patterns

(1) Extraction of a set of patterns from graphs,

(2) Comparison between patterns,

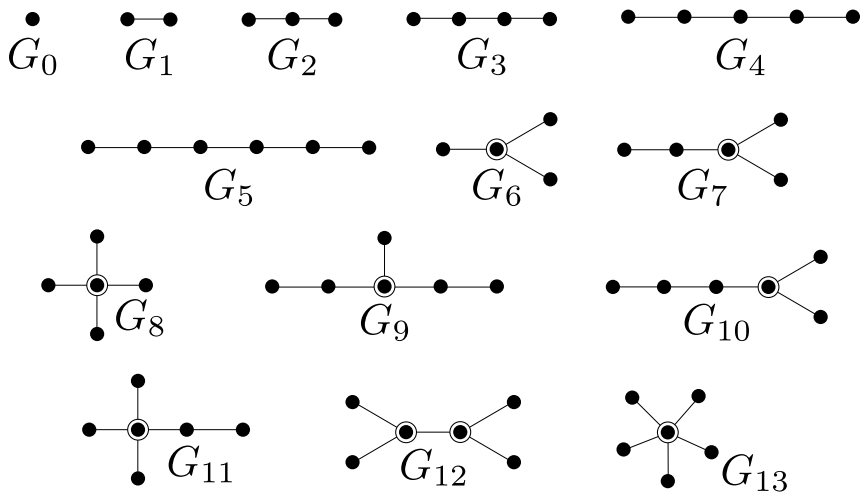
(3) Comparison between bags of patterns.



Plan

Treelets

Set of labeled subtrees with at most 6 nodes :

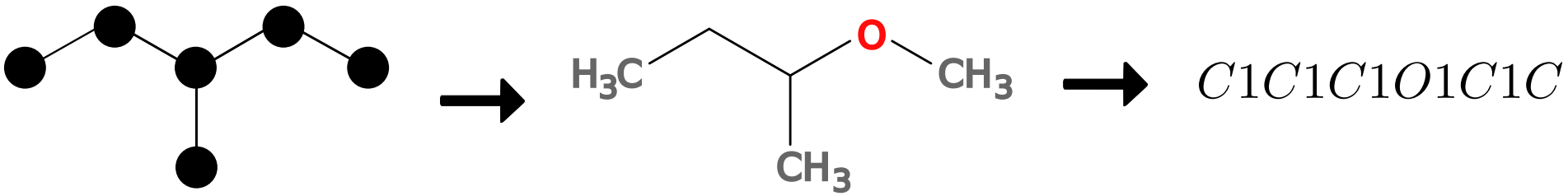


- Encode branching points
- Take labels into account.
- Pertinent from a chemical point of view
- limited set of substructures
- Linear complexity (in $\mathcal{O}(nd^5)$).

enumaration

Labelling

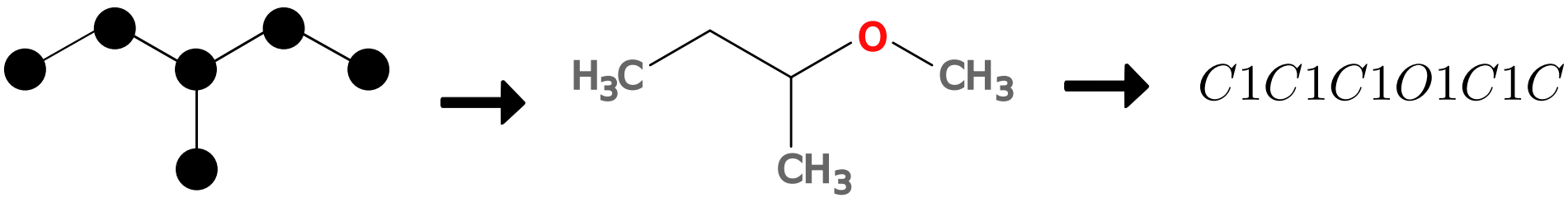
- Code encoding the labels,
- Defined on each of the 14 structure of treelet,
- ✚ Computation in constant time.



enumaration

Labelling

- Code encoding the labels,
- Defined on each of the 14 structure of treelet,
- ✚ Computation in constant time.



Canonical code :

- type of treelet,
- Labels.

} canonical code : G_9-$C1C1C1O1C1C$

$$t \simeq t' \Leftrightarrow \text{code}(t) = \text{code}(t')$$

Definition of the Treelet kernel

Counting treelet's occurences

- $f_t(G)$: Number of occurences of treelet t in G.

Definition of the Treelet kernel

Counting treelet's occurences

- $f_t(G)$: Number of occurences of treelet t in G.

Treelet kernel

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

- $\mathcal{T}(G)$: Set of treelets of G.
- $k(.,.)$: Kernel between real numbers.

Definition of the Treelet kernel

Counting treelet's occurences

- $f_t(G)$: Number of occurences of treelet t in G.

Treelet kernel

$$k_{\mathcal{T}}(G, G') = \sum_{t \in \mathcal{T}(G) \cap \mathcal{T}(G')} k_t(G, G')$$

- $\mathcal{T}(G)$: Set of treelets of G.
- $k_t(G, G') = k(f_t(G), f_t(G')$.
- $k_t(.,.)$: Similarity according to t.

Pattern selection

Prediction problems:

- Different properties ⟶ Different causes.

Pattern selection

Prediction problems:

- Different properties ⟶ Different causes.

Chemist approach:

- Pattern relevant for a precise property
- A priori selection by a chemical expert

Pattern selection

Prediction problems:

- Different properties ⟶ Different causes.

Chemist approach:

- Pattern relevant for a precise property
- A priori selection by a chemical expert

Machine learning approach:

- Parcimonious set of relevant treelets
- Selection *a posteriori* induced by data.

Multiple kernel learning

Combination of kernels [Rakotomamonjy et al., 2008] :

$$k(G, G') = \sum_{t \in \mathcal{T}} k_t(G, G')$$

Multiple kernel learning

Optimal Combination of kernels [Rakotomamonjy et al., 2008] :

$$k_W(G, G') = \sum_{t \in \mathcal{T}} w_t \, k_t(G, G')$$

- $\mathcal{T}$: Set of treelets extracted from the learning dataset.
- $w_t \in [0, 1]$: Measure the influence of t

✚ Weighing of each sub kernel

→ Selection of relevant treelets

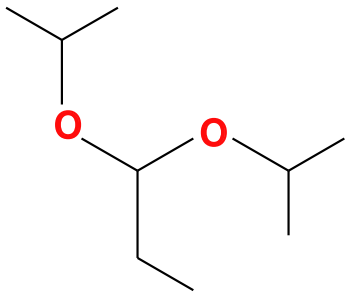
✚ Weighs computed for a given property.

Experiments

Prediction of the boiling point of acyclic molecules:

- 183 acyclic molecules :
  - Learning set: 80 %
  - Validation set: 10 %
  - Test set: 10 %

  } $10\times$
- Selection of optimal parameters using a grid search
- RMSE obtained on the 10 test sets.



Experiments

Prediction of the boiling point of acyclic molecules:

Méthode	**RMSE ($^\circ C$)**
Descriptor based approach [Cherqaoui et al., 1994]*	5,10
Empirical embedding [Riesen, 2009]	10,15
Gaussian kernel based on the graph edit distance [Neuhaus and Bunke, 2007]	10,27
Kernel on paths [Ralaivola et al., 2005]	12,24
Kernel on random walks [Kashima et al., 2003]	18,72
Kernel on tree patterns [Mahé and Vert, 2009]	11,02
Weisfeiler-Lehman Kernel[Shervaszide, 2012]	14,98
Treelet kernel	6,45
Treelet kernel with MKL	**4,22**

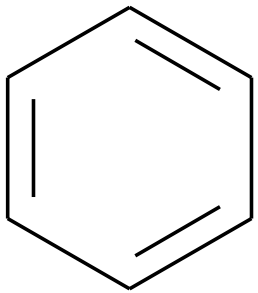
* leave-one-out

Plan

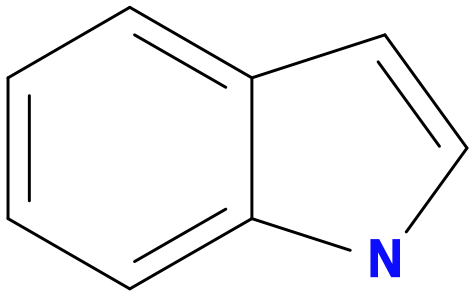
Cyclic molecular information

Cycles of a molecule

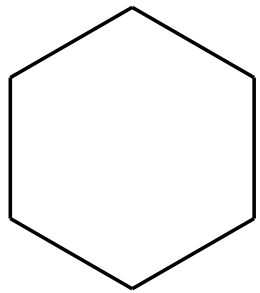
- Have a large influence on molecular properties.
- Identifies some molecular's families.



(a) Benzène



(b) Indole

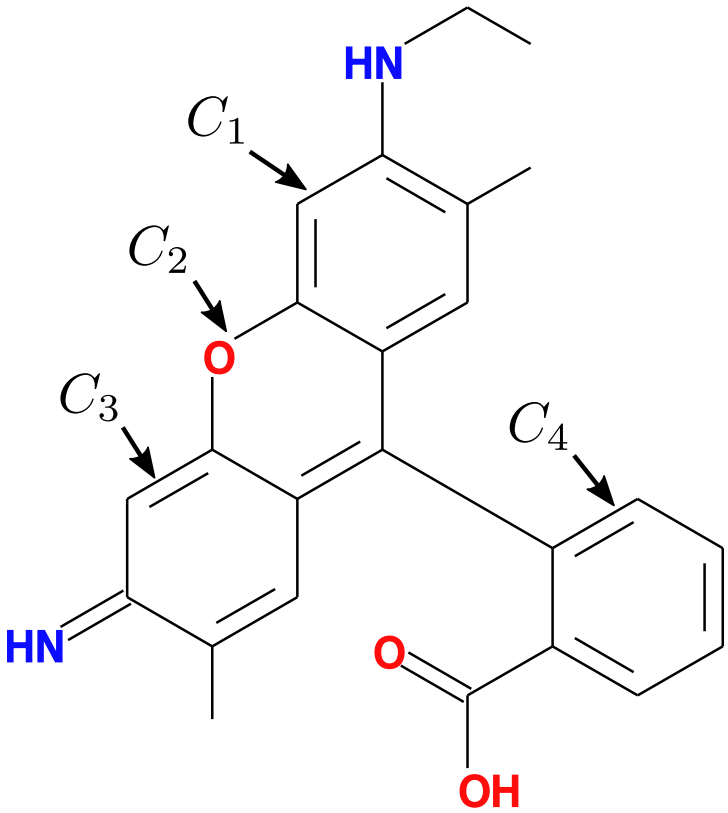


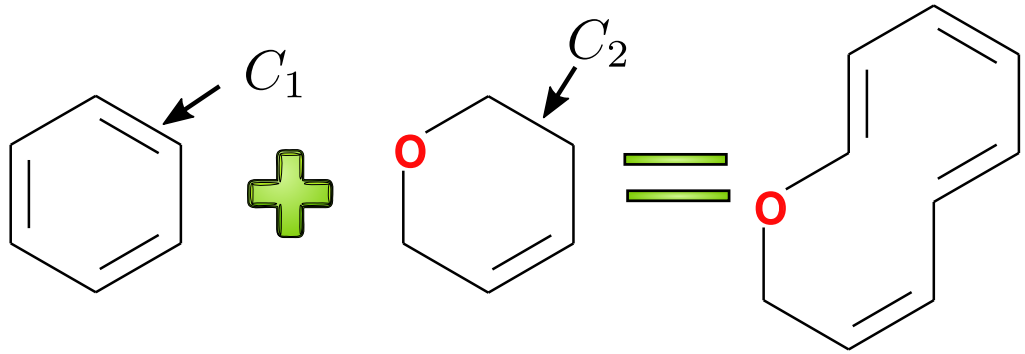
(c) Cyclohexane

→ **Cyclic information must be taken into account.**

Relevant Cycles

- Enumeration of all cycles is NP complete
- Cycles of a graph form a vector space on $\frac{\mathbb{Z}}{2\mathbb{Z}}$.

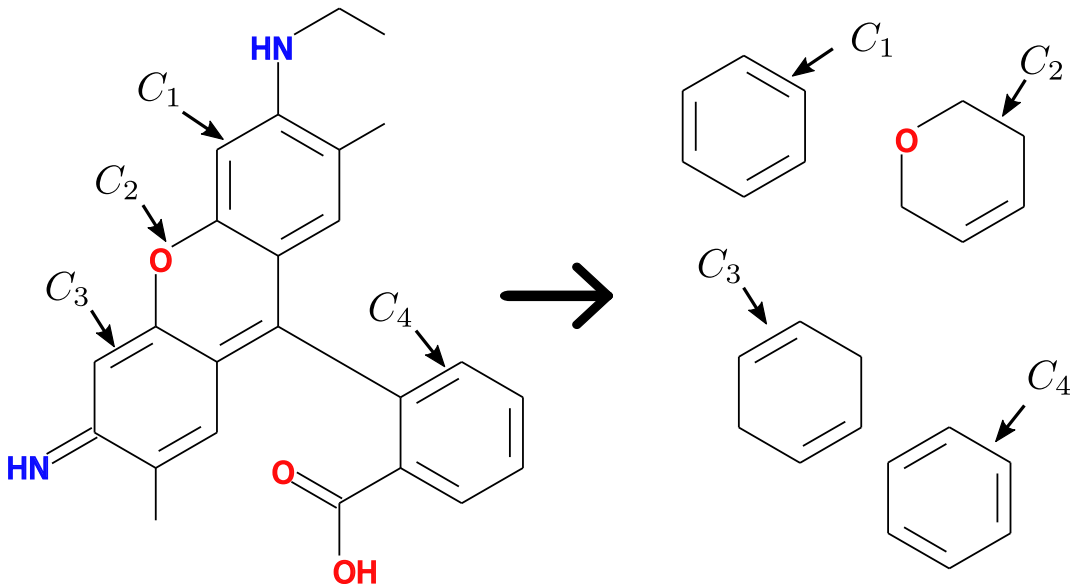




Relevant Cycles

Enumeration of relevant cycles [Vismara, 1995]:

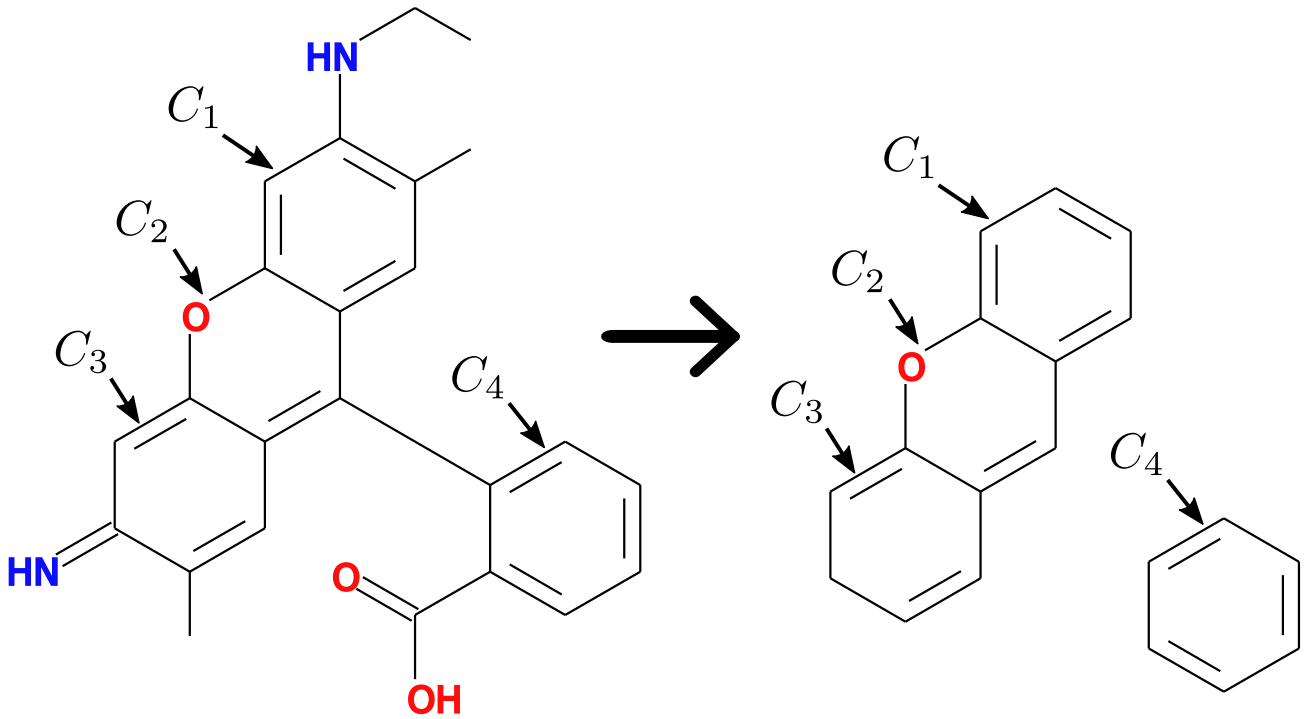
- Union of all cyclic bases with a minimal size,
- ✚ Unique for each graph,
- ✚ Enumeration in polynomial time,
- Kernel on relevant cycles [Horváth, 2005].



Relevant cycles Graph

Graph of relevant cycles $G_{\mathcal{C}}(G)$ [Vismara, 1995] :

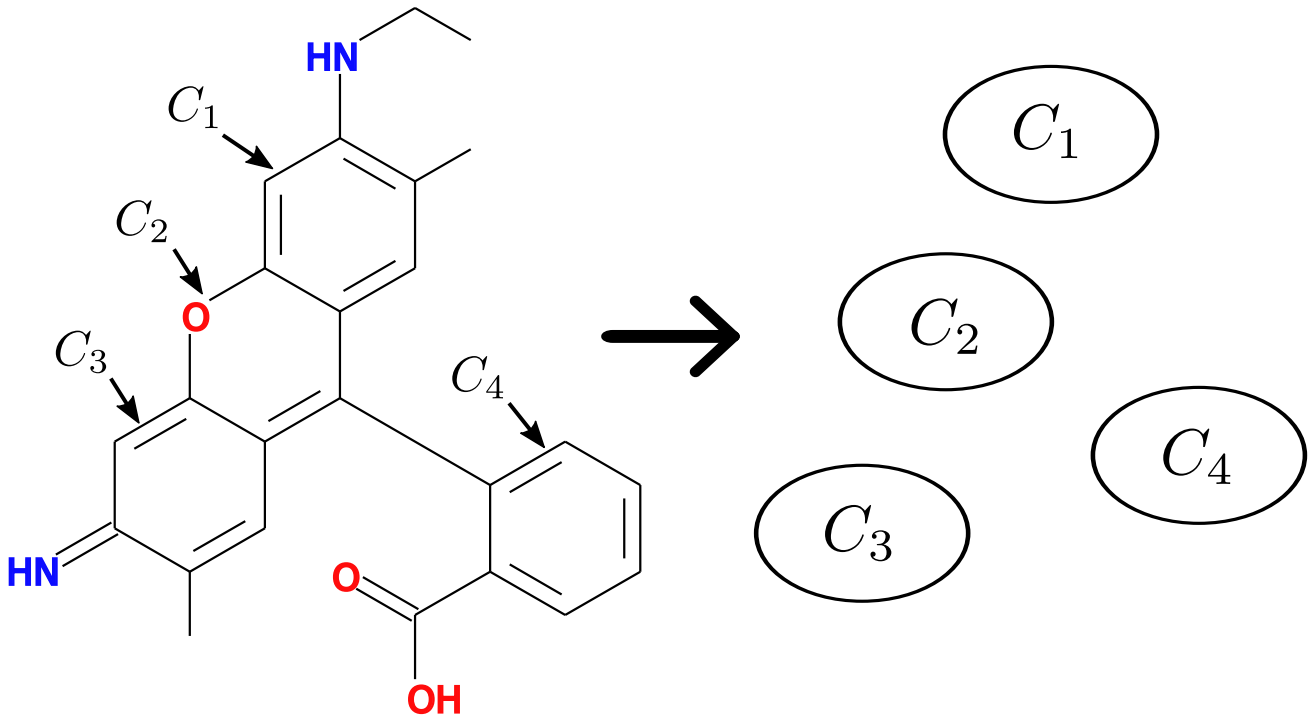
- Representation of the cyclic system,
- Vertices: Relevant cycles,
- Edges: intersection between relevant cycles.



Relevant cycles Graph

Graph of relevant cycles $G_{\mathcal{C}}(G)$ [Vismara, 1995] :

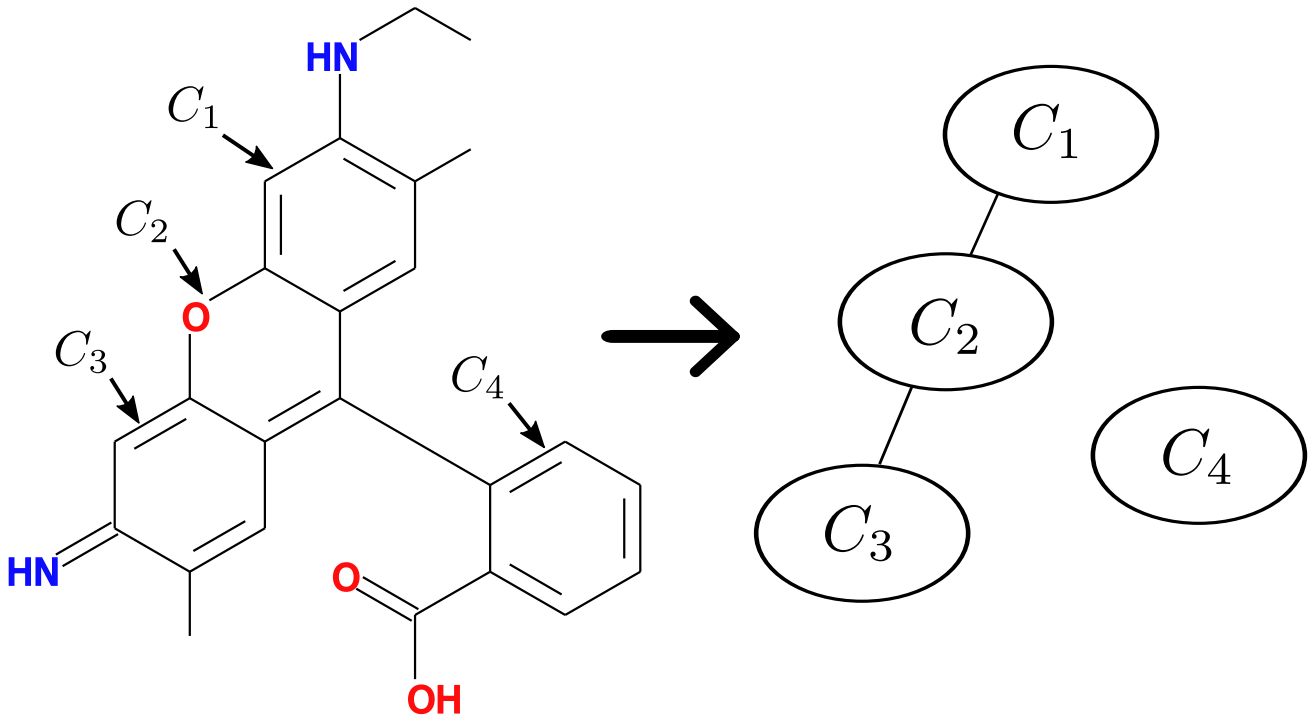
- Representation of the cyclic system,
- Vertices: Relevant cycles,
- Edges: intersection between relevant cycles.



Relevant cycles Graph

Graph of relevant cycles $G_{\mathcal{C}}(G)$ [Vismara, 1995] :

- Representation of the cyclic system,
- Vertices: Relevant cycles,
- Edges: intersection between relevant cycles.

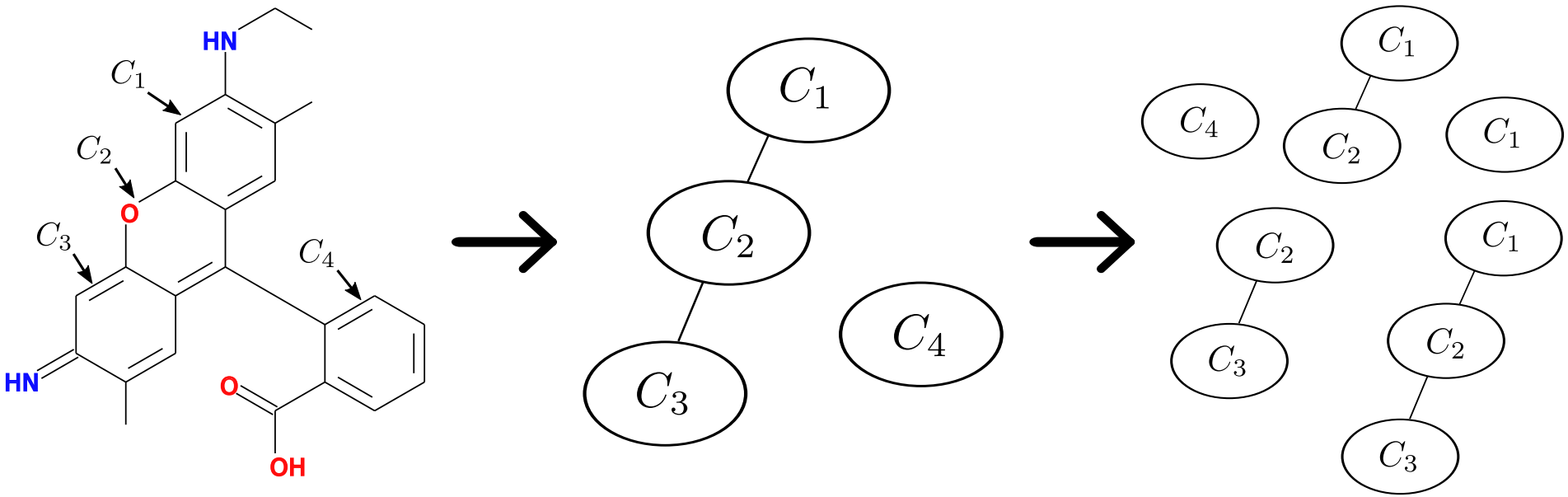


Relevant cycles Graph

Similarity between relevant cycle graphs

- Enumeration of relevant cycle graph's treelets.
- ✚ Adjacency relationship between relevant cycles.

Extracted Patterns :



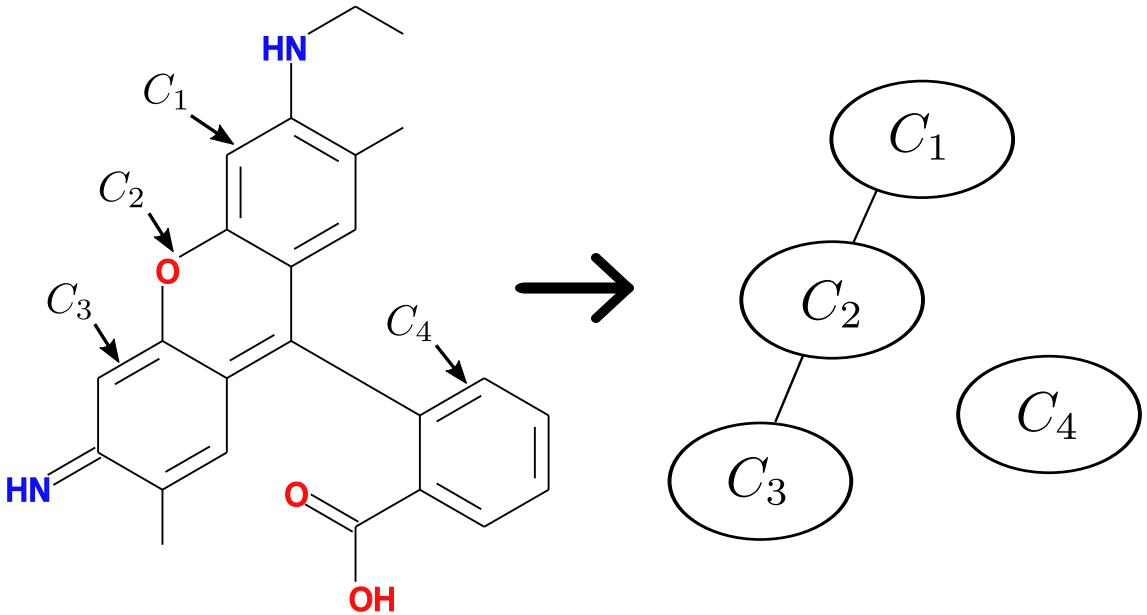
- Solely cyclic information.

Relevant cycle Hypergraph

Connexion between cycles and acyclic parts

- **Idea** : Molecular Representation:
  - Global,
  - Explicit cyclic Information

→ Addition of acyclic parts to the relevant cycle graph.

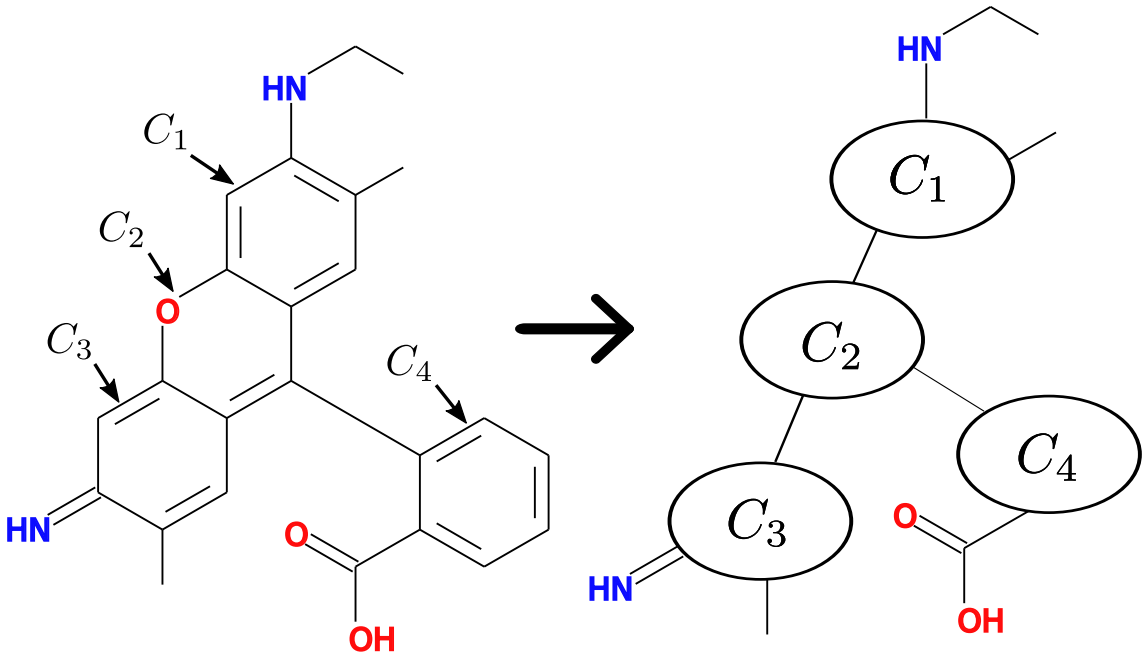


Relevant cycle Hypergraph

Connexion between cycles and acyclic parts

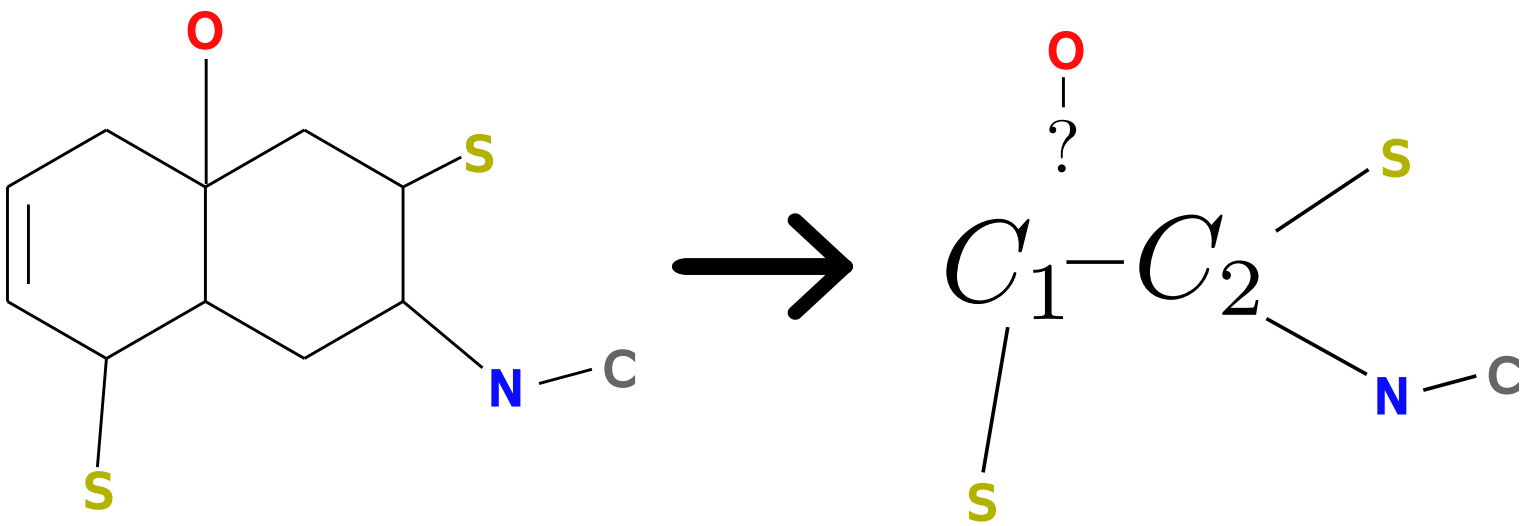
- **Idea** : Molecular Representation:
  - Global,
  - Explicit cyclic Information

→ Addition of acyclic parts to the relevant cycle graph.



Relevant cycle Hypergraph

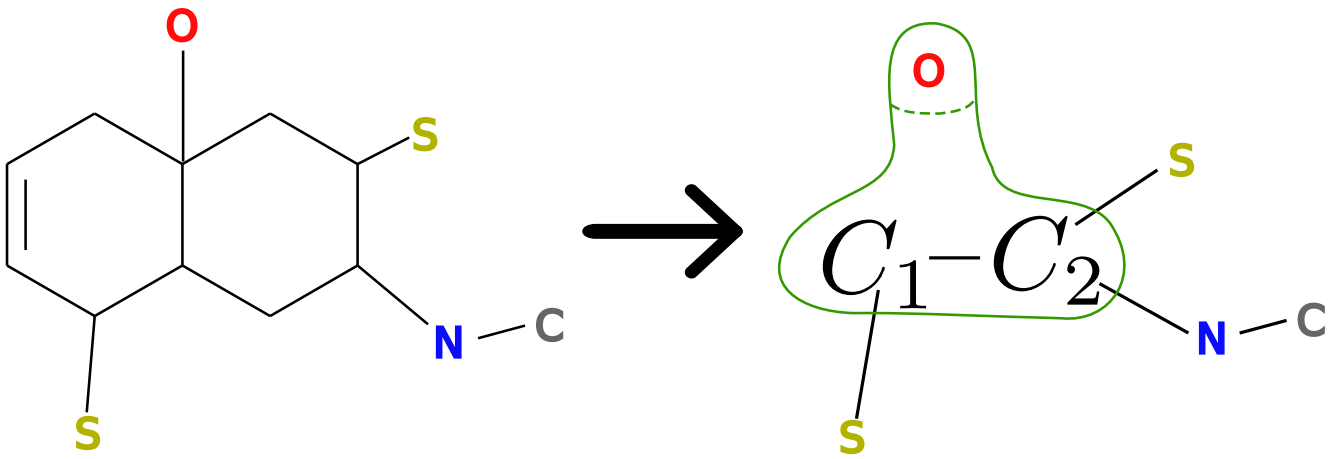
Problem : An atomic bound can connect an atom to two cycles.



Relevant cycle Hypergraph

Connexion between cycles and acyclic parts

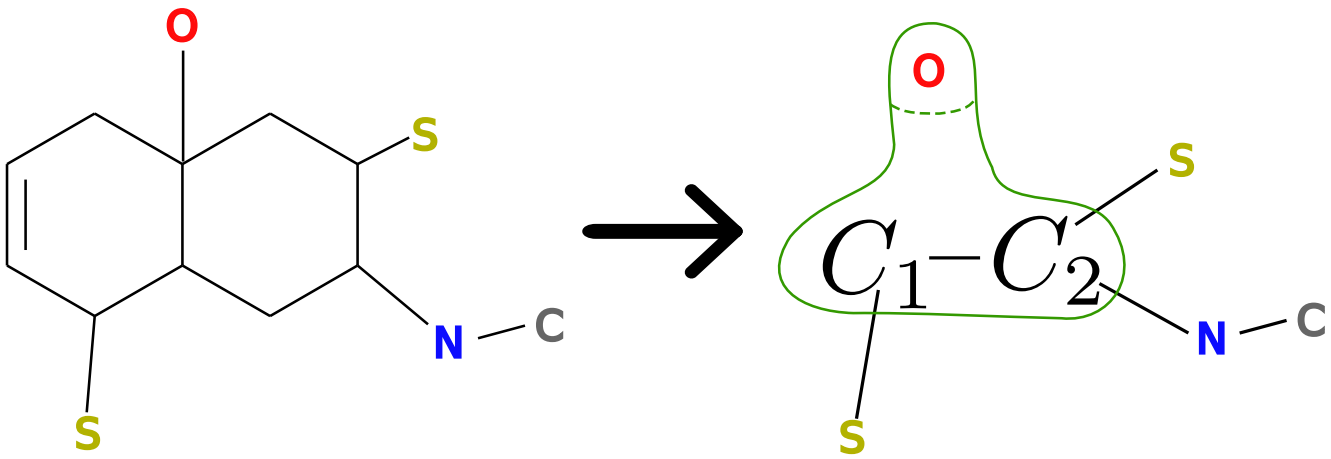
- hypergraph representation of the molecule.
- Adjacency relationship: oriented hyper-edges.



Relevant cycle Hypergraph

Connexion between cycles and acyclic parts

- hypergraph representation of the molecule.
- Adjacency relationship: oriented hyper-edges.

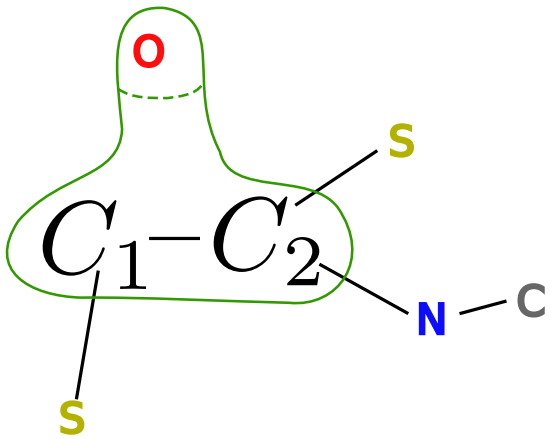


→ Hypergraph comparison.

Treelets on relevant cycle hypergraphs

Adaptation of the treelet kernel:

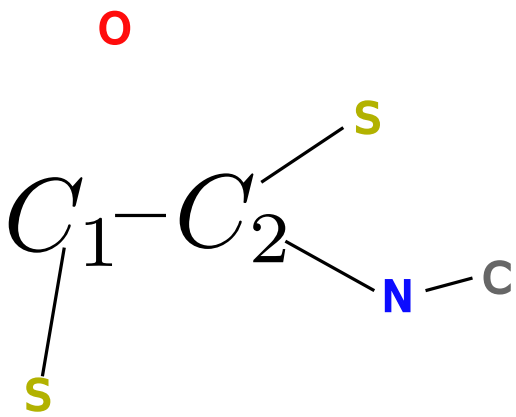
(1) enumeration of the relevant cycle hypergraph's treelets without hyper-edges.



Treelets on relevant cycle hypergraphs

Adaptation of the treelet kernel:

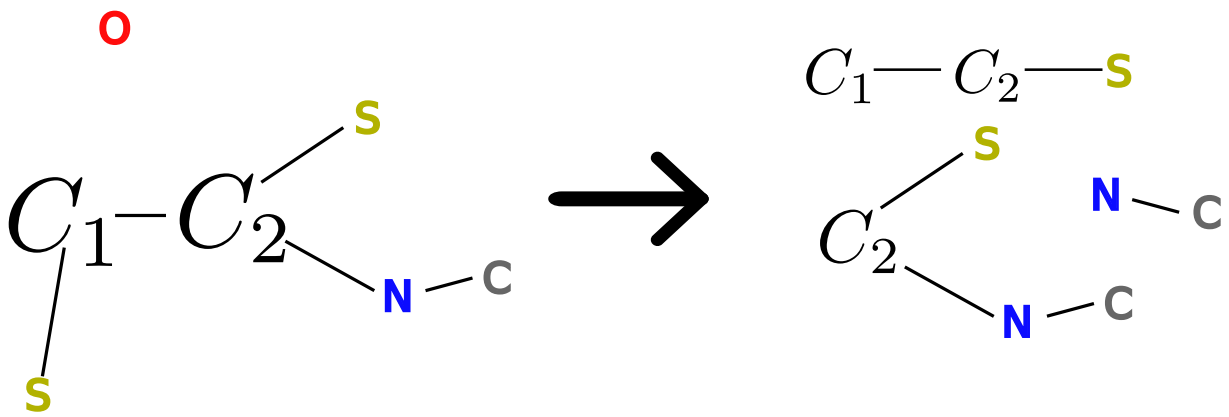
(1) enumeration of the relevant cycle hypergraph's treelets without hyper-edges.



Treelets on relevant cycle hypergraphs

Adaptation of the treelet kernel:

(1) enumeration of the relevant cycle hypergraph's treelets without hyper-edges.



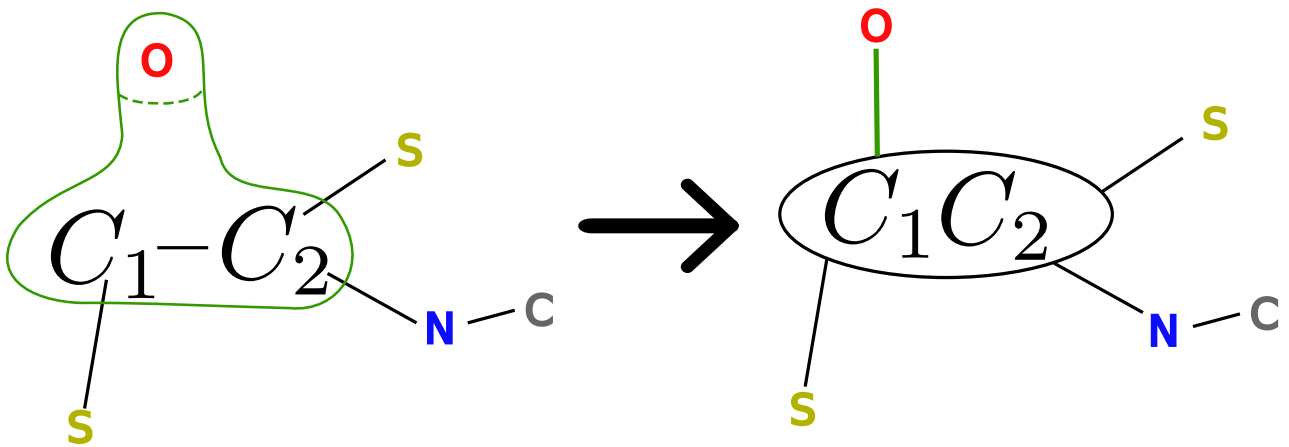
Treelets on relevant cycle hypergraphs

Adaptation of the treelet kernel:

(1) enumeration of the relevant cycle hypergraph's treelets without hyper-edges.

(2) Contraction of cycles incident to each hyper-edge.

→ Construction of the reduced relevant cycle hypergraph $G_{RC}(G)$.



Treelets on relevant cycle hypergraphs

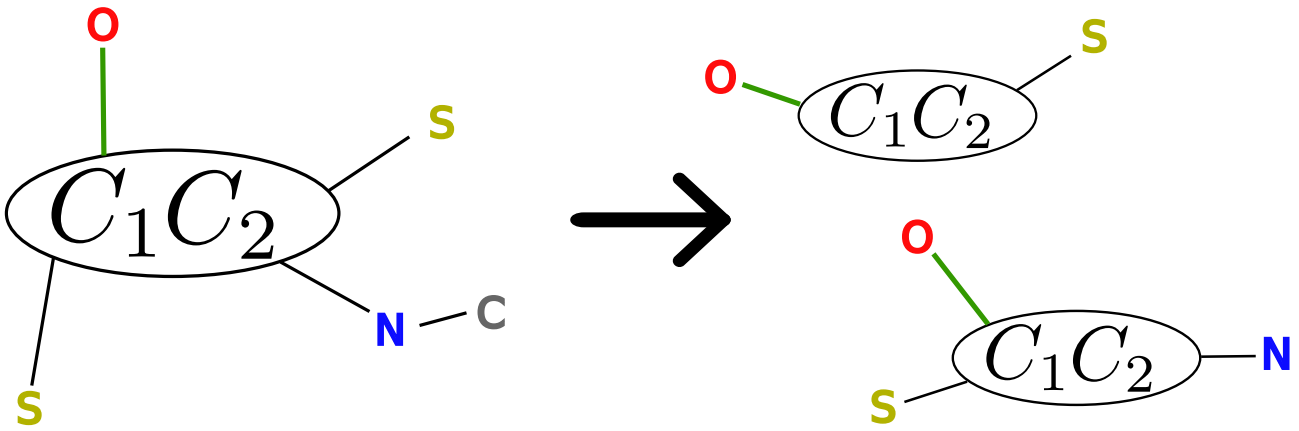
Adaptation of the treelet kernel:

(1) enumeration of the relevant cycle hypergraph's treelets without hyper-edges.

(2) Contraction of cycles incident to each hyper-edge.

→ Construction of the reduced relevant cycle hypergraph $G_{RC}(G)$.

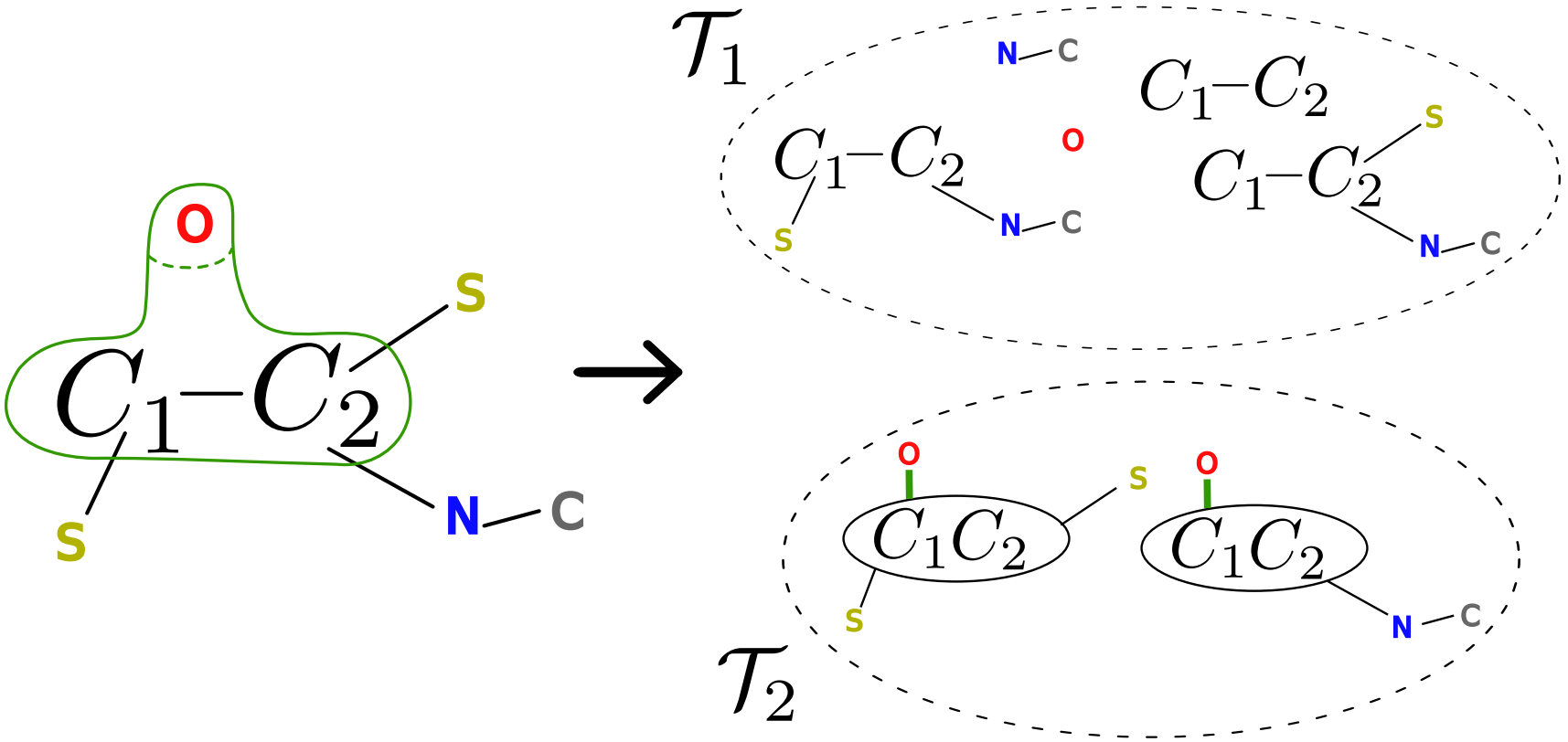
(3) Extraction of treelets containing an initial hyper-edge



Kernel on relevant cycle hypergraphs

Bag of patterns: $\mathcal{T}_{CR}(G) = \mathcal{T}_1 \cup \mathcal{T}_2$

- $\mathcal{T}_1$: Set of treelets obtained from the hypergraph without hyper-edges.
- $\mathcal{T}_2$: Set of treelets extracted from $G_{RC}(G)$.



Kernel on relevant cycle hypergraphs

Bag of patterns: $\mathcal{T}_{CR}(G) = \mathcal{T}_1 \cup \mathcal{T}_2$

- $\mathcal{T}_1$: Set of treelets obtained from the hypergraph without hyper-edges.
- $\mathcal{T}_2$: Set of treelets extracted from $G_{RC}(G)$.

Kernel design:

$$k_{\mathcal{T}}(G, G') = \sum_{t \in \mathcal{T}_{CR}(G) \cap \mathcal{T}_{CR}(G')} w_t \; k_t(f_t(G), f_t(G'))$$

Experiments

Predictive Toxicity Challenge

- Molecular graph labelled and composed of cycles.
- $\simeq$ 340 molecules for each dataset
- 4 classes of animals (MM, FM, MR, FR) :
  - Learning set $\simeq$ 310 molécules,
  - Test set $\simeq$ 34 molecules,

  $\Big\}\ 10\times$
  - Number of molecules correctly classified on the 10 test sets.

Méthode	**MM**	**FM**	**MR**	**FR**
Treelet kernel (TK)	208	205	**209**	212
Relevant cycle kernel	209	207	202	228
TK on relevant cycle graph (TC)	211	210	203	232
TK on relevant cycle hypergraph (TCH)	**217**	**224**	207	**233**

Experiments

Predictive Toxicity Challenge

Méthode	MM	FM	MR	FR
Treelet kernel (TK)	208	205	209	212
Relevant cycle kernel	209	207	202	228
TK on relevant cycle graph (TC)	211	210	203	232
TK on relevant cycle hypergraph (TCH)	217	224	207	233
TK + MKL	218	224	**224**	**250**
TC + MKL	216	213	212	237
TCH + MKL	**225**	**229**	215	239

Experiments

Méthode	**MM**	**FM**	**MR**	**FR**
Treelet kernel (TK)	208	205	209	212
Relevant cycle kernel	209	207	202	228
TK on relevant cycle graph (TC)	211	210	203	232
TK on relevant cycle hypergraph (TCH)	217	224	207	233
TK + MKL	218	224	224	250
TCH + MKL	**225**	229	215	239
Combo TK - TCH	**225**	**230**	224	**252**
Gaussian kernel on graph edit distance [Neuhaus and Bunke, 2007]	223	212	194	234
Laplacien kernel	207	186	209	203
Kernel on paths [Ralaivola et al., 2005]*	223	225	**226**	234
Weisfeiler-Lehman Kernel [Shervaszide, 2012]	138	154	221	242

* leave-one-out

Plan

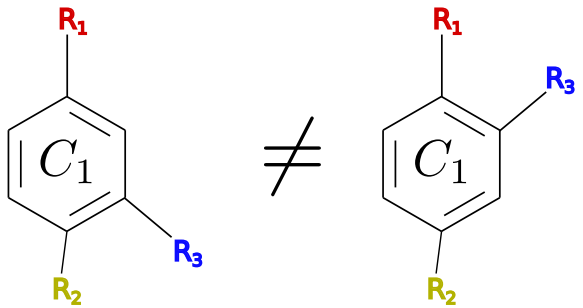
Conclusion

General problem:

Kernel between molecular graphs including cyclic and acyclic patterns in order to predict molecular properties.

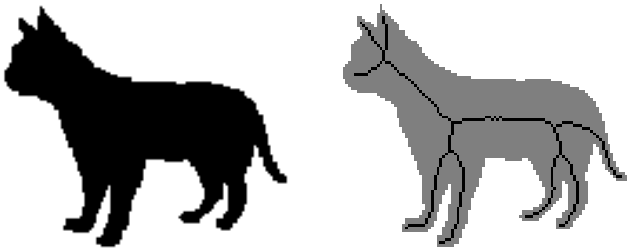
Extensions :

- Cyclic Information:
  - Encoding the relative positioning of atoms
- Taking into account stereoisometry.

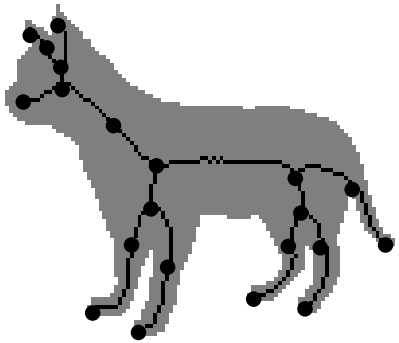


Plan

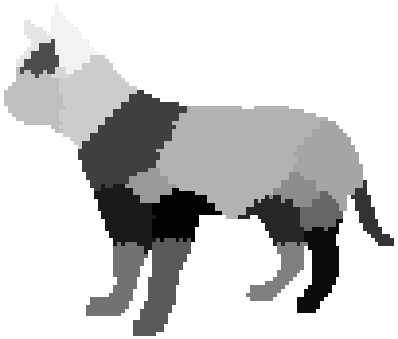
Kernel between shapes



- Compute skeletons
- Encode skeletons by graphs,
- Kernel between bags of treelets.

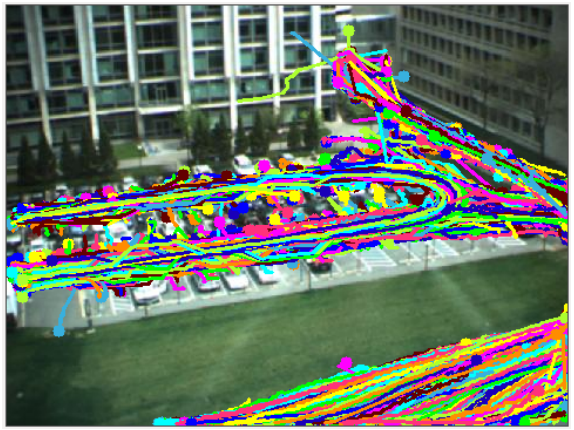


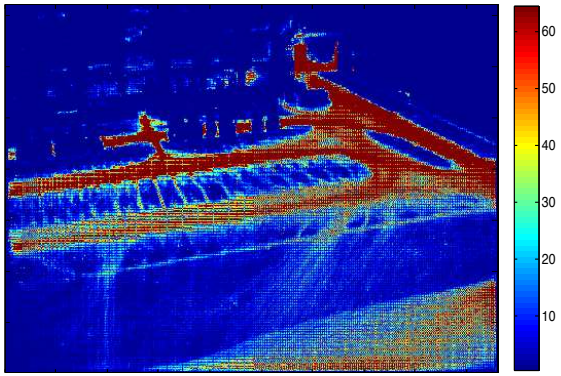
a) le graphe

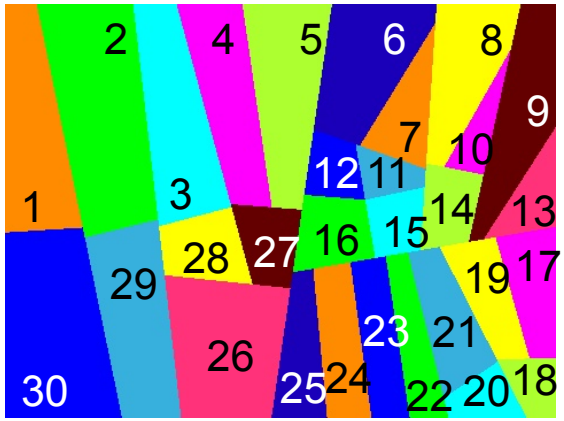


b) la segmentation induite

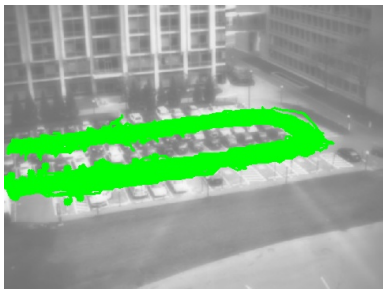
Kernel on strings

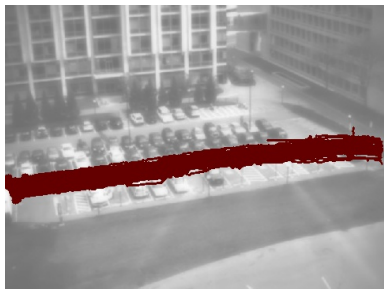






- Decompose the scene into zones,
- Compute a string encoding the sequence of traversed zones,
- Define a kernel between strings,
- Cluster strings.





Bibliographie I


Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68(3):337–404.


Cherqaoui, D., Villemin, D., Mesbah, A., Cense, J. M., and Kvasnicka, V. (1994).
Use of a Neural Network to Determine the Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals and their Sulfur Analogues.
J. Chem. Soc. Faraday Trans., 90:2015–2019.


Horváth, T. (2005).
Cyclic pattern kernels revisited.
In Springer-Verlag, editor, *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, volume 3518, pages 791 – 801.

Bibliographie II


Johnson, M. A. and Maggiora, G. M., editors (1990).
Concepts and Applications of Molecular Similarity.
Wiley.


Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized Kernels Between Labeled Graphs.
Machine Learning.


Mahé, P. and Vert, J.-p. (2009).
Graph kernels based on tree patterns for molecules.
Machine Learning, 75(1)(September 2008):3–35.


Neuhaus, M. and Bunke, H. (2007).
Bridging the gap between graph edit distance and kernel machines, volume 68 of *Series in Machine Perception and Artificial Intelligence*.
World Scientific Publishing.

Bibliographie III


Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2008).
SimpleMKL.
Journal of Machine Learning Research, 9:2491–2521.


Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005).
Graph kernels for chemical informatics.
Neural networks : the official journal of the International Neural Network Society, 18(8):1093–110.


Riesen, K. (2009).
Classification and Clustering of Vector Space Embedded Graphs.
PhD thesis, Institut für Informatik und angewandte Mathematik Universität Bern.


Shervaszide, N. (2012).
Scalable Graph Kernels.
PhD thesis, Universität Tübingen.

Bibliographie IV

Vismara, P. (1995).
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.