Chiral Kernel : Taking into account stereoisomerism

Pierre-Anthony Grenier†, Luc Brun†, and Didier Villemin‡

†GREYC UMR CNRS 6072, ‡LCMT UMR CNRS 6507,
Caen, France
{pierre-anthony.grenier,luc.brun,didier.villemin}@ensicaen.fr

Abstract. Graph kernels provides a framework combining machine learning and graph theory. However, kernels based upon the molecular graph, which can not distinguish stereoisomers, are unable to predict properties which differs among stereoisomers. This article presents a graph kernel which takes into account chirality, and is used (in combination with a classical graph kernel) to predict the optical rotation of molecules.

Keywords: Chemoinformatics, Graph kernel, Chirality

1 Introduction

Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR), aims to predict physical or biological properties of molecules. These fields of research are based upon the similarity principle which states that two structurally similar molecules should have similar properties. In order to represent a molecule, we can use its molecular graph $G = (V, E, \mu, \nu)$, where the unlabelled graph (V, E) encodes the structure of the molecule, each atoms being represented by a vertex and each chemical bond by an edge, μ associates to each vertex a label encoding the nature of the atom and ν associates to each edge a label encoding the type of the represented bond (single, double, triple or aromatic).

Two molecules are said to be stereoisomers if they have a same molecular formula, but a different relative positioning of their atoms. One drawback of the molecular graph is that it can not distinguish stereoisomers, therefore graph kernels based upon this representation [4, 3] can not predict properties varying between stereoisomers. The presence of an asymmetric carbon (a carbon connected to four different substituent) is one possible cause of stereoisomerism. Indeed any permutation of two neighbours of an asymmetric carbon leads to a different stereoisomer.

The goal of this paper is to present the association of a classical graph kernel [3], combined with a new graph kernel which takes into account chirality [1, 2], in order to predict the optical rotation of molecules. In Section 2, we present a graph kernel based upon an enumeration of subtrees. Then in Section 3, we introduce our graph kernel which distinguish stereoisomers. Finally Section 4 show results obtain by the combination of those kernels.

2 Treelet Kernel

The treelet kernel [3] is based upon the enumeration of substructures called treelets. Those treelets are defined as all subtrees of a molecular graph, having a number of nodes lower or equal to 6.

In order to define the treelet kernel, we must enumerate all treelets of a molecular graph G. The first step of this enumeration consists in finding all linear treelets. For each vertex of G, a depth first traversal is performed. This traversal is stopped when 5 edges are visited (to obtain all paths of size lower or equal to 6). The second step consists in an analysis of the neighbourhood of n-star, where n-star are vertices of degree n. This step allows to enumerate all non linear treelets of G.

In order to distinguish treelets with a same structure but different atoms and chemical bonds, a key encoding labels of vertices and edges is defined. This key is built by using Morgan numbering and the lexicographic order, so that two treelets with a same key are isomorphic.

Once treelets are enumerated, a vector $f(G)$ is associated to G, where each component $f_t(G)$ is equal to the number of occurrences of a treelet t in G. The treelet kernel between two graphs is then defined by:

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

where $\mathcal{T}(G)$ correspond to the set of treelets of G.

3 Chiral Kernel : Minimal chiral subtrees

In order to encode the relative positioning of atoms inside a molecule, we use the notion of ordered graphs. An ordered graph $G = (V, E, \mu, \nu, ord)$ is a molecular graph (V, E, μ, ν) with a function $ord : V \to V^*$ which maps each vertex to an ordered list of its neighbours. The configuration of an asymmetric carbon can be represented by an ordered list of its 4 neighbours. An asymmetric carbon can have two different configuration, but with 4 neighbours we can have $!4 = 24$ different ordered list. So two ordered list can encode a same configuration. We thus define a re-ordering function σ, which associates to each vertex of G a permutation changing an ordered list into an equivalent one encoding the same configuration. A valid family of re-ordering functions Σ is a set of re-ordering functions preserving the isomerism properties of a molecule [2], and two ordered graph G, G' are said to be equivalent according to Σ if:

$$G \underset{\Sigma}{\simeq} G' \Rightarrow \exists \sigma \in \Sigma,\ \sigma(G) \underset{o}{\simeq} G'$$

where $\underset{o}{\simeq}$ is the isomorphism between ordered graphs, thus respecting the function ord.

Two stereoisomers have two non equivalent ordered graphs, thus this representation allows to distinguish them. From a local point of view, this representation is used to define chiral vertices :

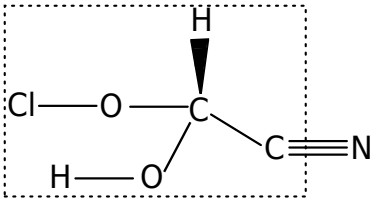

Fig. 1. Molecule with one asymmetric carbon. Its minimal chiral subtree is surrounded by dotted line.

Definition 1. Chiral vertices

Let $G = (V, E, \mu, \nu, ord)$ be an ordered graph. A vertex $v \in V$ of degree n is said to be chiral iff:

$$\forall (i,j) \in \{1,\dots,n\}^2 \text{ with } i \neq j, \nexists f \in \text{IsomEqOrd}(G, \tau_{i,j}(G)) \text{ with } f(v) = v.$$

where $\tau_{i,j}$ is a re-ordering function equals to the identity on all vertices except v for which it permutes the vertices of index i and j in $ord(v)$.

Thus a vertex is chiral if a permutation of its neighbours leads to a non-equivalent ordered graph.

Definition 1 is global since we use the whole ordered graph to characterize the chirality of a vertex. In order to characterize locally the chirality of a vertex, we search for the minimal subgraph of G for which Definition 1 still holds.

For acyclic graphs, this subgraph is a subtree, rooted in v, called the minimal chiral subtree of v. To compute it, we first check for the subtree of G rooted on v and of height 1. If Definition 1 is valid, this tree is the minimal chiral subtree, else we increment the height of the tree, until Definition 1 becomes true.

Figure 1 represent the minimal chiral subtree of an asymmetric carbon.

Therefore, we can associate to each chiral vertex in an acyclic ordered graph, a minimal chiral subtree describing it. This subtree is then encoded by a string, obtained by a depth-first traversal [2]. This string allows us to distinguish different chiral vertices. As for the treelet kernel, we associate to G a vector $f^c(G)$, where each component $f^c_a(G)$ is equal to the number of occurrence of the minimal chiral subtree a in G. The chiral kernel between ordered graphs is thus defined by:

$$k_c(G, G') = \sum_{a \in \mathcal{T}_a(G) \cap \mathcal{T}_a(G')} K(f^c_a(G), f^c_a(G')).$$

where $\mathcal{T}_a(G)$ is the set of minimal chiral subtree of G.

4 Experiments

In order to test our kernel, we use it to predict the optical rotation of molecules. Our dataset is composed of 35 acyclic and chiral molecules, each of them having one or two asymmetric carbon. The standard deviation of the optical rotation

Table 1. Prediction of the optical rotation of acyclic chiral molecules.

	Average Error	RMSE
Treelet Kernel [3]	26.0	33.9
Chiral Kernel [1]	11.6	16.3
Combining of the two kernels	11.5	**15.1**

is 38.25 with values ranging from -89 to 78. We present in Table 1, the results obtained by the treelet kernel alone (Section 2), the chiral kernel alone (Section 3), and a combination of both kernels. To combine these two kernels, we first use the chiral kernel in a classification problem to predict the sign of the optical rotation, then we use the treelet kernel in a regression problem to predict the absolute value of the optical rotation. SVM is used both for classification and regression.

For the combination of kernels, the classification step predicts a right sign for all molecules. We can see that the treelet kernel alone obtain poor results (a rooted mean squared error close to the standard deviation of the dataset). This clearly shows that without distinguishing stereoisomers, we can not predict the optical rotation. However, the combination of both kernel obtains the best results hence showing that the chirality impact essentially the sign of optical rotation while its absolute value could be predicted thanks to the treelet kernel.

5 Conclusion

Ordered graph representation allows to distinguish stereoisomers. From this representation, we have constructed a graph kernel, which can be used to predict properties related to the chirality of molecules. We have shown that this chiral kernel can predict the sign of optical rotation of molecules, and that the absolute value of optical rotation can be predicted without taking into account the chirality. Our future work will consists of extend the chiral kernel to molecules including cycles.

References

1. Grenier, P. A., Brun, L., & Villemin, D. (2013). *Treelet Kernel Incorporating Chiral Information.* In Graph-Based Representations in Pattern Recognition (pp. 132-141). Springer Berlin Heidelberg.
2. Grenier, P. A., Brun, L., & Villemin, D. (2013). *Incorporating chirality within the graph kernel framework* Technical Report. (http://hal.archives-ouvertes.fr/hal-00809066/)
3. Gaüzère, B., Brun, L., & Villemin, D. (2012). *Two new graphs kernels in chemoinformatics.* In Pattern Recognition Letters.
4. Mahé, P., & Vert, J. P. (2009). *Graph kernels based on tree patterns for molecules.* In Machine learning, 75(1), 3-35.