A new representation for molecule similarity

Stefi Nouleho,*,† Dominique Barth,*,† Franck Quessette,*,† Marc-Antoine
Weisser,*,‡ Dimitri Watel,*,¶,§ and Olivier David*,∥

†DAVID, Department of Computer Science, University of Versailles, France

‡LRI, CentraleSupelec, Paris-Saclay University

¶ENSIIE, 1 square de la Resistance, 91025, Evry, France

§SAMOVAR, Telecom SudParis, Evry, France

∥ILV, Department of Chemistry, University of Versailles, France

E-mail: stefi.nouleho-ilemo@uvsq.fr; dominique.barth@uvsq.fr; franck.quessette@uvsq.fr;
marc-antoine.weisser@centralesupelec.fr; dimitri.watel@ensiie.fr; olivier.david@uvsq.fr

Abstract

In order to define the process of restrosynthesis of a new organic molecule, it is
often necessary to be able to draw inspiration from that of a molecule similar to the
target one of which we know such a process. To compute such a similarity, an oftently
used approach is to solve a Maximum Common Edge Subgraph (MCES) problem[1] on
molecular graphs, but such an approach is limited by computation time and pertinence
of similarity measurement. In this paper, we define and analyse here a new graph
representation of molecules to algorithmically compare them. The purpose is to model
the structure of molecule by a graph smaller than the molecular graph and representing
the interconnexion of its elementary cycles. We provide an algorithm to efficiently
obtain such a graph of cycles from a molecular graph. Then by solving MCES problems
on those graphs, we evaluate the pertinence of using graphs of cycles for molecular
similarity on a select set of molecules.

Introduction

In organic chemistry, when a new molecule is designed, it is necessary to determine chemical reactions that can be used to synthesize this target molecule from available compounds. For this purpose, the goal we focus on here is to provide algorithms to help determine a sequence of reactions for the synthesis of new molecules, *i.e.* a tree of reactions whose root is the target molecule and leaves are the available possibly initial molecules (commercialised or easily synthesizable). In order to find such chemical reactions, chemists approach is to search in a reaction database (such as REAXYS[2] or CHEBI[3]) for a molecule that is structurally close to the target molecule. And then from a chemical reaction of this similar molecule they draw inspiration reaction tree to propose such a chemical reaction for the target molecule. To help such a processus, it is therefore a question of being able to algorithmically select molecules in a reaction database that are structurally similar to a target molecule.

Considering a modeling of molecules by graphs or hypergraphs, several definitions and similarity approaches between molecules have already been studied,[1] mainly due to the principle stating that structurally similar molecules are expected to display similar properties,[4,5] or to help virtual screening for drug design.[6] One of these approaches consists in measuring distances of weighted editions between two molecular graphs, an edition being an operation of adding or deleting a vertex or a link in such a graph, or the label change of a vertex. These approaches are notoriously used in the field of bioinformatics.[7,8] Another approach considers the kernel pattern of molecular graphs or hypergraphs, *ie* the presence or not of sub-graphs (also called "fingerprints"[9,10]) in a set of determined patterns, close to cycles or trees, related to the functional properties of molecules; this approach seems well suited to the classification of molecules according to the properties concerned,[11,12] but the choice of a significant set of substructures to compare molecules is ofently a difficult problem. Finally, a last approach considers the resolution of the problem of finding a Maximum Common Edge Subgraph[1] (MCES) between two graphs. This problem is NP-complete and is initially seen as a generalization of graph isomorphism, with different metrics evaluating the size of this

subgraph compared to those of the two graphs compared.[4,6,13,14] It is a variant of this last approach that we will considered in this article.

In the context of this paper, we focus on the similarity of the structural configurations of two molecules. Such a configuration is seen as the interconnexion of the cycles in the maximum 2-connected induced subgraph of the molecular graph. We assume that similar molecules may certainly have similar cyclic parts. A representation of the structure of a molecule based on the cycles it contains has already been proposed and used to classify and characterize sets of molecules.[15,16] Here, we propose a definition of a cycle graph of a molecule modeling not only a relevant subset of the molecule cycles but also their interconnection, whether they share vertices or not; such a representation can also be seen as the extension of a reduction of the Markush structure of a molecule into a ring/non-ring reduction scheme leading to express the core structure of a molecule,[17] for example to make classification.[18] Our objective is to confirm that this definition of cycle graph corresponds sufficiently to the intuitive approach followed by a chemist and that the comparison of the graphs of cycles, based on a specific MCES, corresponds well to the notion of similarity of molecules wished.

The rest of the paper is organised as follow. In the next section, we give some preliminar definitions about graph theory and molecular graphs. Then in Section 3, we define the graph of cycle of a molecule and we propose an algorithm to efficiently obtain it for any given molecule. Finally, in Section 4, we evaluate the perfomances of using such graph of cycles (in terms of time computation and pertinence) to measure the similarity of pairs of molecules.

Molecular graph

In this paper, we consider definitions and notations on graph theory form Berge.[19] An usual representation of a molecule is a molecular graph.[20] A molecular graph is an undirected labeled graph $G = (V, E)$ encoding the structural and fonctionnal information of the molecule.

The set of vertices V of G encodes atoms and the set of edges E encodes the adjacency relationship between atoms in the molecule. Each vertex is labeled by the corresponding chemical element (for example C = Carbon, H = Hydrogen) and each edge is labeled by its type of covalent bond (single $-$, double $=$, triple, aromatic).

Since hydrogen atoms can be connected at least to one atom, they can be omitted in the representation of a molecule (see Figure 1). A molecular graph does not encode neither the relative spatial arrangement of atoms nor the distance between atoms.
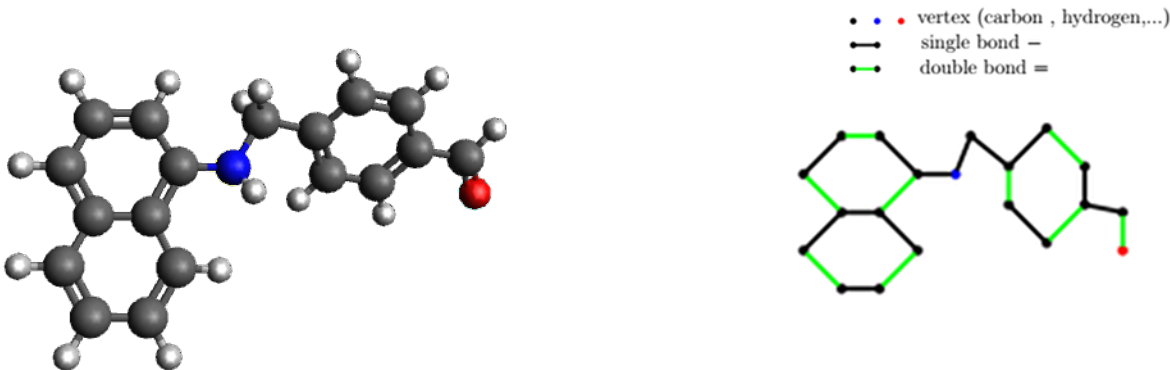

Figure 1: Example of a molecule and its corresponding molecular graph.

Preliminary

We consider a simple and undirected labeled graph $G = (V, E)$ with $n = |V|$ the number of vertices and $m = |E|$ for the number of edges in $E = \{e_1, e_2, ..., e_m\}$.

An elementary cycle c can be represented by a vector $v_c = (e_1^c, e_2^c, ..., e_m^c)$ where $e_i^c = 1$ iff the edge e_i^c belongs to c otherwise $e_i^c = 0$. The **length of a cycle** c is the number of edges that belongs to the cycle $|c| = \sum e_i^c$.

Definition 1. *Let us consider two cycles c_1 and c_2 of vectors $v_{c_1} = (e_1^{c_1}, e_2^{c_1}, ..., e_m^{c_1})$ and $v_{c_2} = (e_1^{c_2}, e_2^{c_2}, ..., e_m^{c_2})$. The **union of cycles** c_1 and c_2 with the boolean operator XOR (symbol $\oplus$) is $c_{12} = c_1 \oplus c_2$ such that the vector $v_{c_{12}} = (e_1^{c_1} \oplus e_1^{c_2}, e_2^{c_1} \oplus e_2^{c_2}, ..., e_m^{c_1} \oplus e_m^{c_2})$.*

Since c_1 and c_2 are elementary cycles, then the union of c_1 and c_2 is an union of edge-disjoint cycles by definition of $\oplus$.

A $2-$connected component is a maximal (in terms of inclusion) $k-$connected induced subgraph with $k \geq 2$.

Definition 2. *An **isthmus** is an edge of G whose deletion increases its number of connected components. An edge is an isthmus if it is not contained in any cycle of G.*

An *isthmus-free graph* is a graph that does not have any isthmus. If a graph G has p isthmus then its number of $2-$connected components K is such that $K - 1 \leq p$; each connected component of a bridgeless graph is $2-$edge-connected. The $2-$connected components in a graph are connected in G by isthmus-chains (a chain which all edges are isthmus).

Cycle structure of a molecular graph

In this section, the goal is to define a molecular representation which encodes the interconnection between the cyclic parts of the molecule. We assume that the cyclic part ($k-$connected component with $k \geq 2$) describes the structure and the acyclic part describes chemical properties of the molecule. So, similar molecules may certainly have similar cyclic parts.

This cyclic structure of molecular graph is based on the interconnection of its induced cycles. However, we do not compute and represent in the graph of cycles all the elementary cycles as there can be an exponential number of such cycles. In order to get a compact representation of the molecule cycles, we can use minimum cycle bases[21] of the graphs.

Definition 3. *A generator ζ is a set of cycles such that for each cycle c of G there is a set of cycles $c_1, c_2, ..., c_k$ in ζ such that $c = c_1 \oplus c_2 \oplus ... \oplus c_k$.*

The weight of a generator is the sum of the lengths of its cycles. We denote ζ^i the generator of cycles with weight equal to or lower than i.

Definition 4. *A cycle basis of a graph is a minimal generator in terms of inclusion.*

A minimum cycle basis is a cycle basis with a minimum weight. Note that, for a graph we can have more than one minimum cycle basis. It is difficult to choose a canonical cycle basis to represent the interconnection of cycles because of the non uniqueness of cycle basis in a graph (see Figure 2). This means that depending on the choosen algorithm to compute a minimum cycle basis and the vertex labelling, two isomorphic graphs may have different cycle basis as results. Thus, we cannot only refer to one cycle basis to decide on the similarity between molecules.
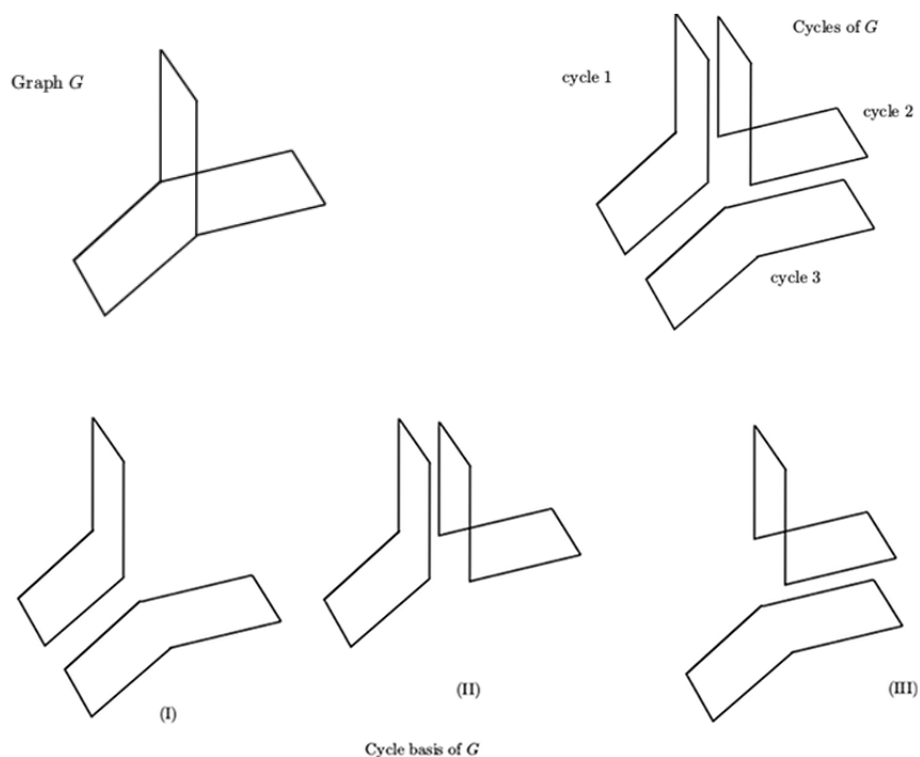

Figure 2: Differents cycle basis

Vismara[22] reported that the union of minimum cycle basis of the graph is a canonical generator. In the same paper, it is said that the union of minimum cycle basis is the smallest canonical set of cycles which computes the cyclic structure of a graph and the number of cycles of the union of cycle basis can be exponential. Referring to the definition of a cycle

basis, the union of minimum cycle basis in a graph is a generator. Although the algorithm proposed by Vismara computes a compact representation of the potentially exponential-sized set, there is no algorithm to list all the cycles of the union of minimum cycle basis. In the following section we will introduce an algorithm to compute a canonical generator of a molecular graph.

Given a molecular graph G and a canonical generator ζ of G, our goal is to compute a graph G^ζ representing the cyclic part of G and describing the interconnection between cycles of the molecular graph G.

Graph of cycles for molecular graph

Before defining formally the graph of cycles, we illustrate and explain it on one example.

Example 1. Let us consider the molecular graph of quinine, with $\{c_1, c_2, c_3, c_4, c_5\}$ a canonical generator containing 5 cycles (see Figure 3). These cycles are the vertices of the corresponding graph of cycle. In terms of similarity between molecules, when considering interaction between cycles in a molecular graph, its is important to distinguish cycles sharing some vertices (like cycles c_1 and c_2) and cycles linked by a path (like c_2 and c_3). It is why we consider two types of edges in the graph of cycle of a molecule. Firstly , the type 1 is used for closed cycles *i.e.* for cycles sharing at least one vertex in the molecular graph. Edges of type 1 have as label value the number of shared bonds. For instance, the plain blue edges on Figure 3 are of type 1. The edge between c_1 and c_2 is equals to 1 because they have one bond in common. Secondly, the type 2 is for cycles with a relationship than can be easily broken (more often the cycles are not closed in the molecular graph). Edges of type 2 have as label value the length of a shortest path between the corresponding cycles in the molecular graph. For example, the dashed green edges on Figure 3 are of type 2.
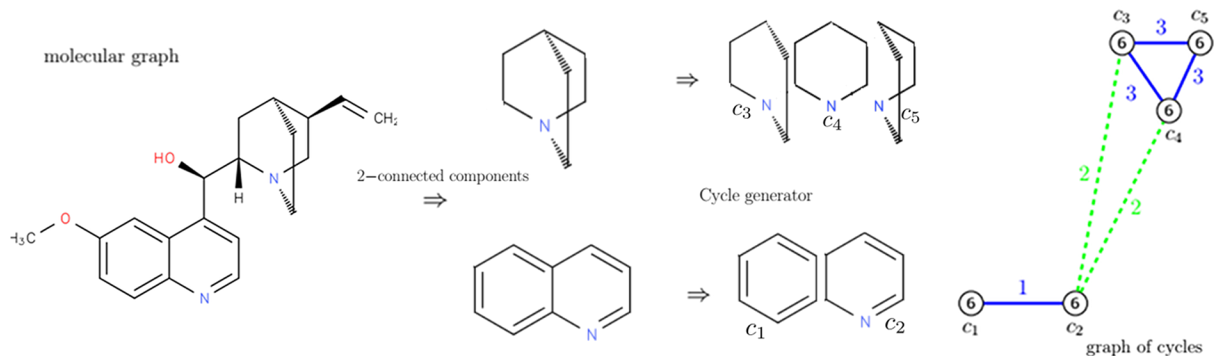
Figure 3: Molecular graph, 2−connected components and cycles of the generator.

In terms of similarity measurement, we will also have to upper bound the size of the considered cycles to be considered in the target molecular graph. Let us focus on two molecules considered as structurally similar : strychnine and vomicine. Indeed, as it is illustrated in Figure 4, if we consider all the sizes of cycles in the vomicine, the two molecular graphs appear to be not similar. But, if we do not consider the cycles of size 9 in the vomicine molecular graph, then the two obtained graphs of cycles are similar. In fact these cycles of size 9 aren't cycles involved in the structure of the molecule but rather a connection between the structural part of the molecule and an azote atom. When in this case, reducing the graphs of cycles to cycles with size lower than or equal to 7 is relevant, and it will be the case in most cases. It is why we introduce parameter j in the next definition in order to allow or remove cycles for similarity.

8

Figure 4: Similar molecules : Strychnine and Vomicine with their graph of cycles

Definition 5. *Let us consider a generator ζ and an integer j. The generator ζ is $j-$hierarchical if the subset of cycles of ζ with length equal or lower than j can generate all the cycles of length equal or lower than j in G.*

We denote by ζ_j the $j-$hierarchical set of ζ. A generator ζ is hierarchical iff ζ_j is $j-$hierarchical for every j.

Lemma 1. *A minimum cycle basis of any graph is hierarchical.*

Proof. Let us consider a minimum cycle basis B. Assume that B is not hierarchical *i.e.* there is an integer j such B_j is not $j-$hierarchical.

Since B_j is not $j-$hierarchical, then there is a cycle c of length lower than or equal to j which cannot be generated with B_j. Therefore the cycle c doesn't belongs to B.

9

Since B_j is a cycle basis, there is a set of cycles $\{c_1, c_2, ..., c_\alpha\}$ in B with $c = c_1 \oplus c_2 \oplus$ ... $\oplus c_{\alpha-1} \oplus c_\alpha$. Let us assume that c_α is a cycle of maximum length in the set $\{c_1, c_2, ..., c_\alpha\}$. Since B_j doesn't generate c then the size of c_α is greater than j.

The binary operator $\oplus$ is commutative so $c_1 \oplus c_2 \oplus ... \oplus c_{\alpha-1} \oplus c = c_\alpha$. We denote by B' the set of cycles obtain by removing c_α and adding c in B (i.e. $B' = B \backslash \{c_\alpha\} \cup \{c\}$). As $\{c_1, c_2, ..., c_{\alpha-1}, c\} \subset B'$, $c_\alpha = c_1 \oplus c_2 \oplus ... \oplus c_{\alpha-1} \oplus c$ and B a cycle basis, so is B'. The weight of the cycle basis B' is $|B'| = |B| - |c_\alpha| + |c|$. The weight of B' is lower than the weight of B (a contradiction because B is a minimum cycle basis). Then B is hierarchical.

$\square$

Definition 6. *Let G be a molecular graph, an integer j and ζ_j be a j-hierarchical generator of cycles in G. The graph of cycles of G induced by ζ_j is denoted $G^{\zeta_j} = (V^{\zeta_j}, E^{\zeta_j}, \mu, \nu, \theta)$ with the edge-set $E^{\zeta_j} = E_1^{\zeta_j} \cup E_2^{\zeta_j}$.*

- *The vertex-set V^{ζ_j} is ζ_j.*

- *The edge-set E^{ζ_j} define the relationship between cycles of V^{ζ_j} according to to their proximity in G.*

    - *$[c_1, c_2] \in E_1^{\zeta_j}$ iff c_1 and c_2 belong to the same 2-connected components of G and they have at least one common vertex.*

    - *$[c_1, c_2] \in E_2^{\zeta_j}$ iff c_1 and c_2 belong to different 2-connected components and there is a path p from a vertex of c_1 to a vertex of c_2 in G such that all edges of p doesn't belongs to a cycle in V^{ζ_j}.*

- *For each vertex $c \in V^{\zeta_j}$, $\mu(c)$ is the weight of the cycle c;*

- *For each edge $e \in E_k^{\zeta_j}$, $\nu(e) = k$;*

- *For each edge $e = [c_1, c_2] \in E^{\zeta_j}$, $\theta(e)$ is the distance from c_1 to c_2 in G. If $e \in E_1^{\zeta_j}$ then $\theta(e)$ is the number of common edges between c_1 and c_2 in G. Otherwise $\theta(e)$ is the length of the shortest path between a vertex of c_1 and a vertex of c_2 in G.*

10

In the Example 3, we have $\mu(c_1) = 6$ as the length of the cycle c_1, $\nu([c_1, c_2]) = 1$, $\nu([c_2, c_3] = 2$ and $\theta([c_2, c_3]) = 2$ (the smallest path from a vertex of c_2 to a vertex of c_3 in the molecular graph). In the following section, we describe how to compute a generator of cycles for a molecular graph and the relationship between its cycles.

Cycles generator of a graph

A generator ζ of cycles as we define contains cycles such that each edge which belongs at least to a cycle is represented. It is computed by using a minimum cycle basis and adding additional cycles.

In this section, we present algorithms to compute a cycle generator ζ_j. Let us consider a molecular graph $G = (V, E, w_V, w_E)$ that may be non connected. We called the structural graph of a molecular graph to be the maximum subgraph of G without any vertex with a degree lower than 2 in the subgraph. We delete the bridges in the structural graph and we denote G_i with $i \in [1..K]$ the $2-$connected components (K is the number of components) computed with the bridgeless graph of the structural molecular graph.

Algorithm 1: Generator of a graph.

Data: A molecular graph G, an integer j = maximum length of cycles .

Result: Generator ζ_j of a molecular graph G

1 $\zeta_j = \emptyset$ set of cycles of the generator;

2 $T = \emptyset$ set of cycles;

3 Remove all the bridges and leaves in G ;

4 Extract the $2-$connected components ($G_1, G_2, ..., G_K$ components);

5 Compute for each $G_i, i \in [1..K]$, a minimum cycle basis $B^i = \{c_1^i, c_2^i, ..., c_k^i\}$;

6 foreach *Cycle basis B^i* **do**

7   **foreach** *Couple of cycles $c_a^i, c_b^i \in B^i$* **do**

8   Let a cycle $c = c_a^i \oplus c_b^i$;

9   **if** *c is an elementary cycle in G_i and $c \notin B^i$ and $|c| = max(|c_a^i|, |c_b^i|)$* **then**

10   Add c to T;

11   **end**

12   **end**

13 $B^i = B^i \cup T$;

14 $T = \emptyset$;

15 end

16 $\zeta_j = \cup B^i$;

17 Remove in ζ_j all the cycles with a length bigger than the parameter j;

We choose to compute a minimum cycle basis of a graph with the Horton algorithm.[21] The algorithm is described here :

Algorithm 2: Horton algorithm.

Data: A graph G

Result: A minimum cycle basis of G

1   $B = \emptyset$;

2   Find shortest chains between all pairs of vertices in each G;

3   **foreach** *vertex v and edge $[x, y]$ in each G* **do**

4   Create the circuit $C(v, x, y) = P(v, x) + P(v, y) + [x, y]$;

5   **if** $P(v, x) \bigcap P(v, y) = \{v\}$ **then**

6   Add $C(v, x, y)$ to B;

7   **end**

8   **end**

9   Order all the cycles of B by length;

10   Use a greedy algorithm (Gauss elimination) to find the minimum cycle basis B from its set of cycles;

Horton algorithm is polynomial $O(n * m^3)$.[21] The complexity of the algorithm 1 is lower than $O(n^2 * m^3)$:

- **Step 2:** is polynomial $O(m^2)$

- **Step 4:** Horton algorithm is called K times and each G_i have at least n vertex. An upper bound is $O(K * (n * m^3))$

- **Step 5 to 12:** number of operations is : $K * k_i^2$, with k_i the number of cycles in the computed minimum cycle basis of G_i.

Since similar molecules have similar structural parts, to compute the similarity between molecules we are going to do it on their graphs of cycles. In the next section we present a similarity calculation on graphs and we check if the results are coherent in terms of similarity of molecules.

Similarity evaluation

To measure similarity of two molecules on their corresponding graphs of cycles, we solve the Maximum Common Edge Subgraph (MCES) problem.[1] Considering two graphs $G = (V, E)$ and $G' = (V', E')$, a mapping of G on G' is a function $\pi : V \to \mathcal{P}(V')$, such that for any $v \in V$, $\pi(v) \neq \emptyset$. We say that G is π-isomorphic to G' iff there exist an isomorphism $\mathcal{I}$ between G and G' such that, for any $v \in V$, we have $\mathcal{I}(v) \in \pi(v)$. Solving problem MCES constrained by π consists in finding the maximum subgraph of G being π-isomorphic to a subgraph of G'. This problem has been shown to be NP-complete.[1]

Let us consider two graphs of cycles $G_1^{\zeta_j} = (V_1^{\zeta_j}, E_1^{\zeta_j})$ and $G_2^{\zeta_{j'}} = (V_2^{\zeta_{j'}}, E_2^{\zeta_{j'}})$ of two molecules M_1 et M_2. In our context, mapping π is defined such that for any $v \in V_1^{\zeta_j}, \pi(v) = \{v' | v' \in V_2^{\zeta_{j'}} \text{ and } ||v| - |v'|| \leq 0, 2 * min(|v|, |v'|)\}$. This function π in graph of cycles allow two cycles to match if they have similar length. The value $0, 2$ has been fixed experimentally. For molecules, the function π in MCES calculation[1] maps atoms of the same type. Here, we denote the graph $G_{\zeta_{12}} = (V_{12}^{\zeta_{j''}}, E_{12}^{\zeta_{j''}}, \mu, \nu, \theta)$ as a Maximum Common Edge Subgraph (MCES) of G_{ζ_1} and G_{ζ_2} constrained by π. The function μ associates to each cycle, it length; the function ν is for the type of relationship between cycles and the function θ for the label of edges.

Considering two graphs G_{ζ_1} and G_{ζ_2}, the similarity is the ratio between the sum of vertices and edges of the MCES $G_{\zeta_{12}}$ and the product of the sum of vertices and edges of G_{ζ_1} and the sum of vertices and edges of G_{ζ_2} :

$$sim(G_{\zeta_1}, G_{\zeta_2}) = \frac{(|V_{12}^{\zeta_{j''}}| + |E_{12}^{\zeta_{j''}}|)^2}{(|V_1^{\zeta_j}| + |E_1^{\zeta_j}|) \times (|V_2^{\zeta_j}| + |E_2^{\zeta_j}|)} \tag{1}$$

Note that finding a maximum common edge subgraph between G and G' is similar to find the maximum clique in the product graph of the linegraph induced by G and G'. In the next section, we experiment the similarity on graph of cycles.

Experimental results

We compare the computation of measures of similarity using the method given in[1] on the one hand by using molecular graphs (MG) and on the other hand by using graph of cycles (GC). The database of molecules used is a freely available dictionary of small molecular entities called Chemical Entities of Biological Interest ChEBI. This database contains 90130 molecules. According to their structural configuration, we choosed seven molecules in ChEBI. Considering the structural part of molecules from a chemical point of view, some of these molecules have many similar molecules in the database while others don't. For each one, we compute the similarities with all the molecules in the database. We then compare the distributions of the obtain similarities in the two contexts MG and CG, and we focus on the 20 most similar pairs of molecules in each context.

To make sure that the two methods computed the similarity on the structural part, we removed all the leaves and bridges in all molecular graphs.

The computation has been done on a cluster Intel(R) Xeon CPU E5-260 v3 @2.40GHz with 64G of RAM. To find a maximum clique in a graph to solve MCES, we did a linear program resolved by SCIP[1] (Solving Constraint Integer Programs). Because of the number of molecules in the database (90130 molecules, knowing that many other database are larger) and since the similarity calculation between two graphs may have an exponential runtime due to the NP-completeness of the problem MCES, we chosed to fix an upper bound of similarity computation of similarity for each pair of molecules. This time depend on the size of the considered graphs (MG or GC). For example, if the maximum time for each similarity is 20 seconds, then the whole computation requires ± 20 days on the cluster. As a consequence of the time limitation, some similarities are not computed for some pairs of molecules in the MG context.

We then compared the distributions of the obtained similarities in the two contexts MG and CG, and we focused on the 20 most similar pairs of molecules in each context. Our

---

[1]http://scip.zib.de/

goal is to evaluate and compare the performances of the two approaches MG and CG from three points of view : the execution time required to calculate the measure of similarity for each pair of molecules, the capacity of each approach to distinguish the pairs of real similar molecules (*i.e.*, the ones having similar core structures) and finally the capacity of discriminating real similar, meduim similar and not similar pairs of molecules.

In this paper we are going to present the results of two molecules : Docetaxel Anhydrous and Amphotericin B. All the results will be available on : reference arkive a mettre ici

Docetaxel Anhydrous

Docetaxel Anhydrous has a generator of cycles with different lengths (4, 6 and 8). The graph of cycles has 6 vertices and we can see that it maximum connected subgraph with edges of type 1 is the kernel of this molecule (see Figure 6).
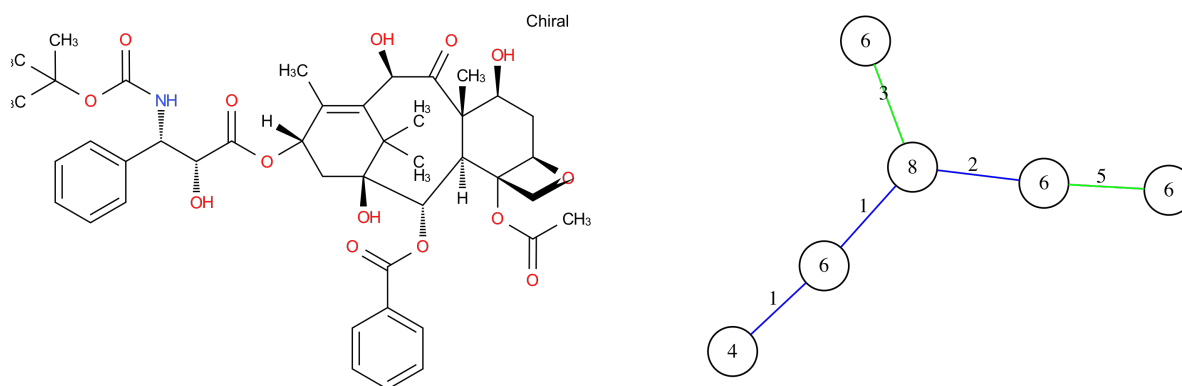
Figure 5: Molecular graph and graph of cycles of Docetaxel anhydrous.

Here are the distributions of similarity on MG and GC :
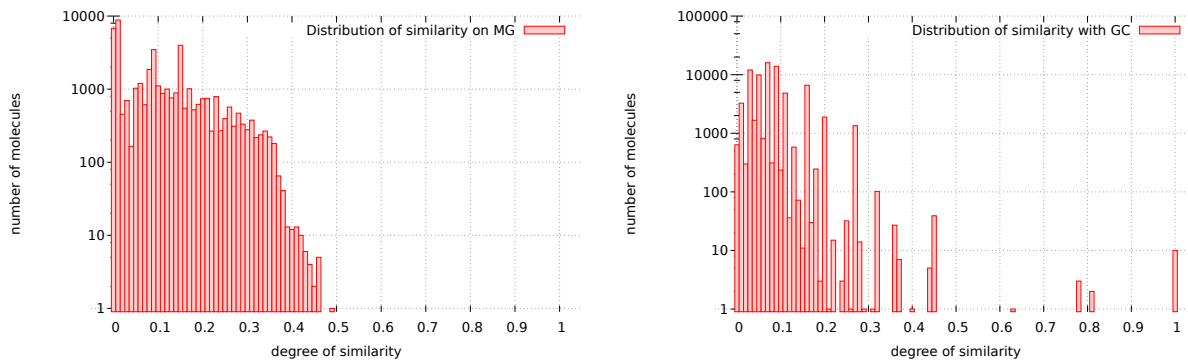
Figure 6: Distribution of similarity of Docetaxel anhydrous on molecular graphs (MG) and on graphs of cycles (GC).

According to the distribution of similarity on GC, 4 categories of similar molecules can be extracted :

- 9 molecules are totally similar to Docetaxel (they are isomers). In fact, they have exactly the same graph of cycles.

- 5 molecules are partially similar; 2 of them have a similarity degree equals to 0.81 differ from Docetaxel only on 1−connected part in MG. Their GC are subgraphs of the graph of cycles of Docetaxel, one cycle linked with an edge of type 2 is missing. The 3 other molecules (with a degree of similarity of 0.78) have the same structure as Docetaxel with more cycles. The GC of these molecules have GC of Docetaxel as subgraph of (they have one cycle more and two edges of type 2).

- 1 molecule is the kernel of Docetaxel. The degree of similarity is 0.63).

- The rest of molecules with a degree lower than 0.45 are not similar to the target molecule.

In the distribution of similarity on MG, we fixed 30 seconds to compute the similarity of two molecules. Over 46846 of 90130 molecules where not computed (about 51.9%). None of the molecules in top 20 are chemically similar to to Docetaxel.

17

(a) **Baccatin III** with a similarity degree of $0,81$
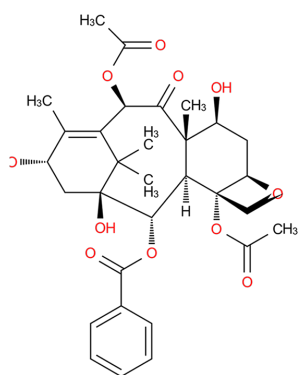
(b) **Paclitaxel** with a similarity degree of $0,78$
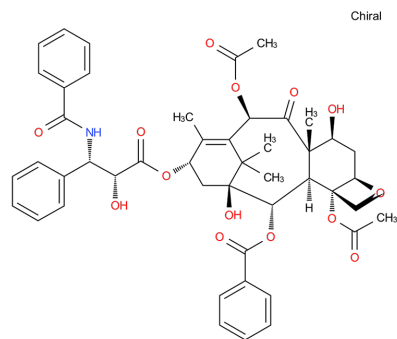
(c) **10-deacetyl-2-debenzoylbaccatin III** with a similarity degree of $0,63$
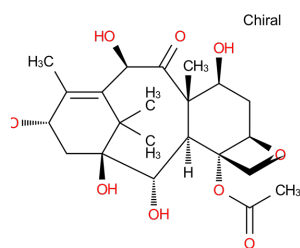
Figure 7: Three similar molecules to docetaxel with GC similarity

Amphotericin B

Amphotericin B has a particular cyclic structure so it mimimum cycle basis contains 3 cycles with a particular cycle of length 36. The corresponding graph of cycles thus contains 3 vertices (Figure 8).
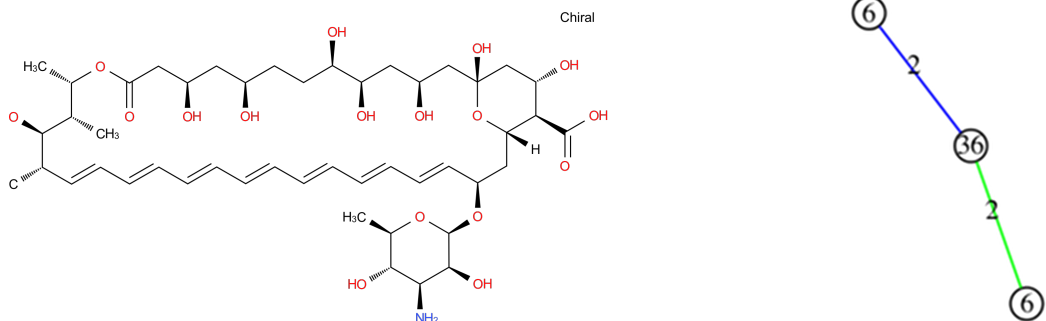
18

Figure 8: Amphotericin B molecular graph and it graph of cycles.

The GC distribution of similarities concerns all the molecules of the database. This distribution given in Figure 9 shows 11 molecules fully similar to the target one (degrees of similarity equal to 1), and another distinguished set of molecules being partially similar to it (degrees of similarity equal to 0.7 or to 0.6). The other molecules can be considered as different from the target molecule (similarity lower than 0.5). Thus, the calculation using cycle graphs clearly discriminates the molecules into three classes, which the molecular graph approach does not do. Moreover, MG approach does not succeed in calculating similarity degrees for several molecules classified as very similar by the GC approach (50932 over 90130 molecules where not computed; that is 56.5%). This is due to a too important running time needed; the computation is stop because of the upper bound (20 seconds). Indeed, the required computation time is far exceeding the imposed limit.
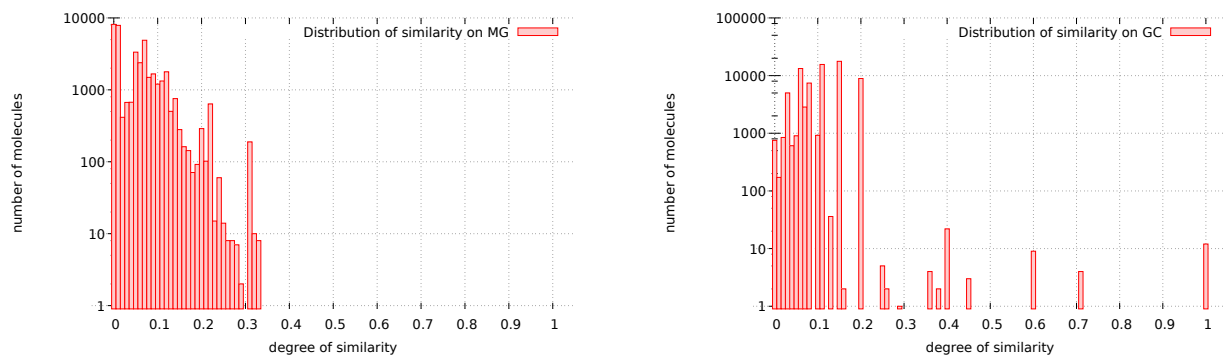

Figure 9: Distribution of similarity of Amphotericin B on molecular graphs (MG) and on graphs of cycles (GC).

Most of the totally similar molecules provided by the GC approach are either isomers of amphotericin B (amphotericin B methyl ester) or member of the same family (nystatin A1). Amphotericin belongs to the family of antifungal. The other fully similar molecules are not intuitively similar to amphotericin B considering their molecular graphs but the similarity in terms of cycle structure are chemically relevant (Figure 10). The molecules with degree of similarity equal to 0.7 in the GC distribution are the ones such that their graph of cycles have the one of Amphotericin B as subgraph, and the molecules with degree of similarity 0.6 are the ones which graph of cycles is the subgraph of the one of Amphotericin B. Note that these molecules are not discriminated in the MG approach.
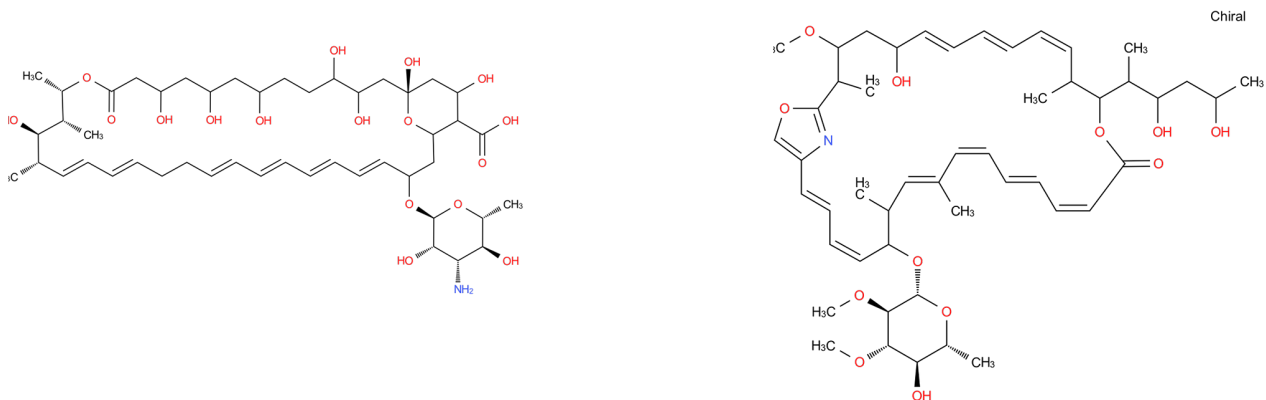

Figure 10: nystatin A1 (ChEBI id 473992) and Chivosazole A (ChEBI id 80057)

Conclusion

The objective of this article was to propose a quickly computable measure of molecule pairwise similarity that can distinguish between pairs of molecules whose core structures (i.e., the interconnection of elementary cycles) are similar. The experiments carried out lead to several conclusions:

- First, the graph of cycles approach allows the similarity of all the pairs of molecules to be calculated in a reasonable time, while the molecular graph approach can require

unrealistic execution times, especially for pairs of similar molecules.

- Secondly, the similarity measures obtained by the cycle graph approach distinguish the similar pairs of molecules, and not just the isomeric molecules of the target molecules.

- Finally, the proposed approach discriminates well the molecules very, little or not similar to a target molecule, while the approach by molecular graphs, when it can calculate similarity measures, is less discriminating.

These experiments therefore show that the proposed approach allows decision support for the determination of chemicals reactions that can be used to synthesize this target molecule from available compounds.

An extension of the proposed approach would be to be able to set the size of the cycles (parameter j) according to the characteristics of the molecular graph. Indeed in many cases, taking into account cycles of too large size can distort the similarity measurement because these cycles do not reflect the core structure of the molecules, while in some other cases, taking into account of such cycles is necessary to take all the core structures into account. It seems that the differentiation between these two cases depends, at least in part, on topological properties of the molecular graph, which requires further studies. Finally, the use of other metrics of similarity than the resolution of the MCES problem, for example the use of an editing distance between the cycle graphs, could also be considered.

References

(1) John W. Raymond, J.; Gardiner, E. J.; Peter, W. *The Computer Journal* **2002**, *45*.

(2) Reaxys. `https://www.elsevier.com/solutions/reaxys`, Accessed Nov, 2017.

(3) Chemical Entities of Biological Interest (ChEBI). `https://www.ebi.ac.uk/chebi/`, Accessed April 1, 2015.

(4) Zager, L. A.; Verghese, G. C. *Applied Mathematics Letters* **2008**, *45*, 86–94.

(5) Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; The American Chemical Society, 1988.

(6) Eckert, H.; Bajorath, J. *Drug Discovery Today* **2007**, *12*, 225–233.

(7) Michel, N.; Horst, B. *Bridging the Gap Between Graph Edit Distance and Kernel Machines*; World Scientific Publishing Co., Inc., 2007.

(8) Roger, S.; John, M.; Boyle, N. O.; Andrew J., G.; Stefan, S.; Darren V.S., G. *7th Joint Sheffield Conference on Chemoinformatics* **2015**,

(9) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. *Methods* **2015**, *71*, 58 – 63.

(10) Andreas, B.; Jeremy L., J.; Josef, S.; Sai C. K., S.; Meir, G.; John W., D. *J. Chem. Inf. Model.* **2009**, *49*.

(11) K., d. G.; F., C. *J. Chem. Inf. Model.* **2010**, *50*, 1660–1668.

(12) Benoit, G.; L., B.; D., V. *Quantitative Graph Theory Mathematical Foundations and Applications, CRC Press* **2015**, 425–470.

(13) Abu-khzam, F.; Samatova, N.; A. Rizk, M.; Langston, M. *IEEE/ACS International Conference on Computer Systems and Applications* **2007**, 367–373.

(14) Tatsuya, A.; Hiroshi, N. *Comput Struct Biotechnol J.* **2013**, *5*.

(15) Benoit, G.; L., B.; D., V. *9th IAPR-TC-15 Graph-Based Representations in Pattern Recognition* **2013**, 111.

(16) T., H.; T., G.; S., W. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2004**, 158–167.

(17) Michael F., L.; John D., H. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 930–936.

(18) Valerie J., G.; Geoffrey M., D.; John D., H.; Michael F., L.; Winfried, D. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260–270.

(19) Berge, C. *Dunod* **1963**,

(20) Gasteiger, J. *Handbook of Chemoinformatics : From Data to Knowledge (Representation of Molecular Structures)*, 1st ed.; Wiley, 2003.

(21) Joseph D., H. *SIAM Journal on Computing* **1987**, *16*, 358–366.

(22) Philippe, V. *Electr. J. Comb.* **1997**, *4*, 73–87.