Detecting molecular cages from graph of cycles

February 2024

Purpose of the approach: automatically detect if a molecule contains a cage, by studying the interconnections of cycles representing as graph of cycles.

1 Definitions

1.1 Molecular graph and graph of cycles

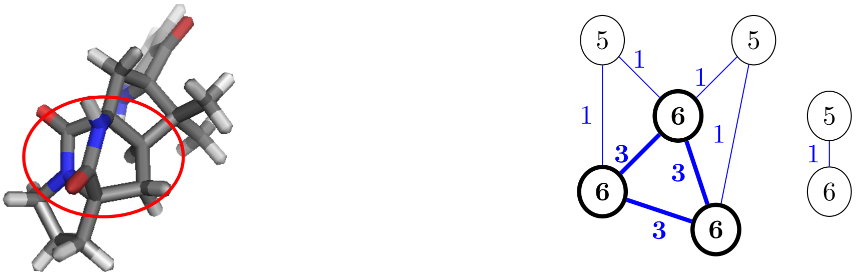
We refer to definitions of molecular graphs and graphs of cycles in [1].

A molecular graph is an undirected labeled graph $G = (V, E)$ encoding the structural information of a molecule. The set of vertices V encodes the atoms of the molecule and the set of edges E encodes the covalent bonds between atoms. Vertices are labeled by the corresponding chemical element and edges are labeled by the type of covalent bonds.

A graph of cycles is defined in [1] to represent the interconnections of cycles in a molecule. It is an undirected weighted graph $G_C = (V_C, E_C)$ constructed from a molecular graph G. The set of vertices V_C represents a set of cycles in the graph, and the set of edges E_C represents the connections between the cycles. Any pair of vertices x, y will belong to E_C if the two corresponding cycles share at least one atom in common in the molecular graph.

In our study, we will consider as the set of vertices V_C the union of all minimum cycles bases in G, as defined in [2]. A minimum cycle basis is a minimal set of cycles of the molecular graph G such as every cycle of G can be obtained from a linear combination of cycles of this set. However, we can have more than one minimum cycle basis for a molecular graph. The union of all minimum cycles bases is then the smallest set of cycles that computes the cyclic structure of a graph [2].

The vertices are weighted by the number of atoms in the cycles and the edges are weighted by the number of common atoms between the two cycles.



(a) Molecule in 3D (artificial corner circled in red)

(b) Graph of cycles of the molecule (artifical corner in bold)

Figure 1: Molecule containing an artificial corner (brevianamide A)

1.2 Corners

Définition Let G_C be the graph of cycles of a molecular graph G. A *corner* in G_C is a clique of size 3, *i.e.* a complete subgraph containing 3 vertices of G_C.

It corresponds to three cycles in the molecular graph with covalent bonds in common two by two, and can thus be seen as a part of a potential cage in the molecule. However, not every clique of size 3 in the graph of cycles belongs to a cage. This is why we define several types of corners that cannot belong to a cage:

- *the artificial corner*:

  Let c_1, c_2 and c_3 be a set of 3 cycles in the molecular graph G, containing respectively $|c_1|$, $|c_2|$ and $|c_3|$ atoms.

  This set of 3 cycles is called *artificial corner* if the corresponding vertices in the graph of cycles form a clique of size 3 and if:

  - any edge of one cycle of the clique also belongs to one of the other two cycles and
  - at least two cycles share the same number of atoms equal to $max(|c_1|, |c_2|, |c_3|)$.

  An example of such a corner is presented in Figure 1, with three cycles containing 6 atoms each.

  In this case, pairs of cycles in the clique belong to the same minimum cycle basis, but all 3 cycles do not. They all appear in the graph of cycles because we consider the union of all minimum cycle bases. However, they cannot be seen as a cage corner.

- *the almost artificial corner*:

  This type of corners is a clique of size 3 in the graph of cycles, in which 2 out of 3 cycles belong to an artificial corner and the last one does not.

- *the unopened corner*:

  We define this type of corners as a clique of size 3 in the graph of cycles, in which there is an edge common to all three cycles. This arrangement of cycles in space does not enable the formation of a cage (example in Figure 2).

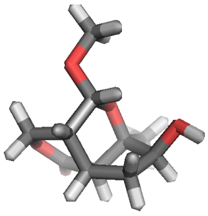


Figure 2: Molecule containing an unopened corner (jatamanin H). There are 3 cycles with common bonds but they do not form a cage.

- *the corner with double bonds at the junction of cycles (double bonded corners)*:

  This type of corners is a clique of size 3 in the graph of cycles, in which at least one double bond is present at the junction of two cycles. Double bonds make structures plane and thus cannot contribute to the formation of a cage. Aromatic rings are one example of such a case.

All the other corners are considered here as potentially cage corners.

1.3 Graph of corners

To represent the interconnections between corners in a graph of cycles, we define the notion of graph of corners. An example is presented in Figure 3.

Definition **Graph of corners**

Let G_c be the graph of cycles, as defined above from a molecular graph G.

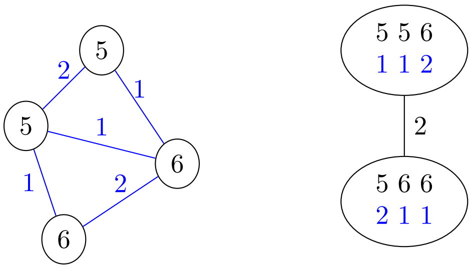
Let $\mathscr{K}_3$ be the set of all cliques of size 3 in G_c, that are considered as *cage corners*.

The graph of corners of G_c induced by $\mathscr{K}_3$ is denoted $G^{\mathscr{K}_3} = (V^{\mathscr{K}_3}, E^{\mathscr{K}_3}, \mu, \nu)$.

- The vertex-set $V^{\mathscr{K}_3}$ is the set of all cliques of size 3 in G_c, that are considered as cage corner.
- The edge-set $E^{\mathscr{K}_3}$ defines the relationship between the cliques of $V^{\mathscr{K}_3}$ according to their proximity in G_c

  > $[k_1, k_2] \in E^{\mathscr{K}_3}$ if and only if the cliques k_1 and k_2 share at least one common vertex in G_c (*i.e.* one common cycle in G). In the example of graph of cycles in the Figure 3, the cycles respectively labeled 5 and 6 in the middle are common to both cliques.

- For each vertex $k \in V^{\mathscr{K}_3}$, $\mu(k)$ is a sextuplet indicating the weight of each vertex and the weight of each edge of the clique k in G_c
- For each edge $e = [k_1, k_2] \in E^{\mathscr{K}_3}$, $\nu(e)$ is the number of common vertices between the cliques k_1 and k_2 in G_c.



(a) Graph of cycles with two cage corners

(b) Graph of corners

Figure 3: Example of a graph of corners (molecule arboduridine)

2 Experimental results

We consider a dataset of 81 molecules that have been classified by experts in three categories:

- Cage : the molecules that actually contain one cage (6 molecules)
- Precage : the molecules that are about to form a cage (26 molecules)
- Non cage : molecules that may resemble caged molecules but do not contain any (48 molecules)

We construct the graph of cycles for each molecule as described before, and we search for the different types of corners we previously defined.

2.1 Discriminate non cages from cages and precages

The number of corners we obtain in each molecule will be used to discriminate cages and precages from non cages. To do so, we define classifications according to our graph-based approach, and then aim to determine how consistent our classifications are with the classification made by experts.

We define the following classifications: A molecule will be considered as Cage or Precage (resp. Non cage) if its graph of cycles contains (resp. does not contain) at least one corner:

1. of any type (classification 1)
2. of any type except the artificial and almost artificial corners (classification 2)
3. of any type except the artificial , almost artificial and unopened corners (classification 3)
4. of any type except the artificial , almost artificial , unopened and double bonded corners (classification 4 for cage corners).

		Experts approach	
		Cage/Precage	Non cage
Graph-based approach	Cage/Precage	32	19
	Non cage	0	29

(a) All types of corners

		Experts approach	
		Cage/Precage	Non cage
Graph-based approach	Cage/Precage	31	12
	Non cage	1	36

(b) All types of corners expect artificial and almost artificial corners

		Experts approach	
		Cage/Precage	Non cage
Graph-based approach	Cage/Precage	31	8
	Non cage	1	40

(c) All types of corners expect artificial , almost artificial and unopened corners

		Experts approach	
		Cage/Precage	Non cage
Graph-based approach	Cage/Precage	31	4
	Non cage	1	44

(d) All types of corners except artificial , almost artificial , unopened corners and double bonded corners

Table 1: Comparison between experts and graph-based approaches using the number of corners. The number of molecules entering each case is indicated.

The results of consistency with the experts approach are presented in Table 1. These results confirm that removing the corners we identify as not cage corners allows us to increase the consistency with the experts classification, with only 5 misclassified molecules in the classification 4, against 19 misclassified molecules in the classification 1.

2.2 Discriminate precages from cages

For each molecule, we also calculate the corresponding graph of corners.

By studying molecules with cages and molecules with precages as classified by experts, we observed that the graph of corners for a molecule with a cage always contains at least one clique of size 3, i.e. at least 3 corners with one or two common cycles. It is not the case for most molecules with precages.

Based on this observation, we discriminate molecules with cages from molecules with precages as follows: A molecule is considered as a Cage (resp. a Precage) if its graph of corners contains at least (resp. does not contain) one clique of size 3.

With this criterium, we classify all 6 molecules with cages as Cage and 23 out of 26 molecules with precages as Precage. Two molecules with precages are classified as Cage and one as Non cage as seen in the previous subsection.

3 Summary : First algorithm to determine if a molecule contains a cage or a precage

The different steps to determine if a molecule contains a cage or a precage are summarized below.

```
Classifying a molecule M
Input: A molecular graph $G$ obtained from a molecule $M$
Output: The class of the molecule $M$ among Cage, Precage or Non cage

- Construct the graph of cycles $G_C$ from $G$
- Search for corners in $G_C$ and classify them by type (artificial , almost artificial , unopened , double bonded, or cage corners)
- Construct the graph of corners $G^{\mathscr{K}_3}$ only with cage corners
- Choose a class for $M$ as follows :
  if $G_C$ contains at least one cage corner then
      if $G^{\mathscr{K}_3}$ contains at least one clique of size 3 then
          $M$ is a Cage
      else
          $M$ is a Precage
  else
      $M$ is a Non cage
```

In Table 2 are presented the results of consistency between our graph-based approach using this algorithm and the experts approach.

		Experts approach		
		Cage	Precage	Non cage
Graph-based approach	Cage	6	2	0
	Precage	0	23	4
	Non cage	0	1	44

Table 2: Comparison between experts and graph-based approaches to discriminate Cages from Precages (with the presence of a clique of size 3 in the graph of corners) and Precages from Non cages (with at least one cage corner in the graph of cycles). The number of molecules entering each case is indicated.

References

[1] Stefi Nouleho Ilemo, Dominique Barth, Olivier David, Franck Quessette, Marc-Antoine Weisser, and Dimitri Watel. Improving graphs of cycles approach to structural similarity of molecules. *PLOS ONE*, 14(12):e0226680, December 2019. Publisher: Public Library of Science.

[2] Philippe Vismara. Union of all the Minimum Cycle Bases of a Graph. *The Electronic Journal of Combinatorics*, pages R9–R9, January 1997.