UNIVERSITÉ DE PARIS-SACLAY

N° attribué par la bibliothèque

THÈSE

pour obtenir le grade de

Docteur de l'Université de Paris-Saclay

Spécialité : **Informatique**

préparée au laboratoire **DAVID**

dans le cadre de l'École Doctorale **EDSTIC**

présentée et soutenue publiquement
par

Stéfi NOULEHO ILEMO

le XX - mois - XXXX

Titre :

Algorithmique de graphes pour l'aide à la décision dans la construction moléculaire

Directeur de thèse : **Dominique BARTH**
Co-directeur de thèse : **Marc-Antoine WEISSER**

Jury

M. Z,	Président du jury
M. U,	Rapporteur
M. V,	Rapporteur
M. W,	Examinateur
M. Dominique BARTH,	Directeur de thèse
M. Marc-Antoine WEISSER,	Co-directeur de thèse

Résumé

Abstract

Table des matières

Introduction

Bla bla bla.

Chapitre 1

Présentation détaillée du sujet

1 Titre de la premiere section

Bla bla bla.

2 Titre de la seconde section

Re-bla bla bla. Une citation [**?**].

3 Conclusion

Conclusion du chapitre.

Chapitre 2

État de l'art sur la similarité moléculaire

La similarité moléculaire joue un rôle important dans de nombreux aspects de la chémo-informatique tels que la recherche de similarité, la classification de bases de données moléculaires, la découverte de médicaments et l'analyse de la diversité moléculaire. La similarité englobe les positions atomiques, les conformations, la forme et la disposition spatiale des propriétés moléculaires [6].

Cependant, la similarité entre deux molécules doit être optimisée par rapport aux applications examinées. C'est un domaine large et en constante évolution. Il existe de multiples mesures de similarité décomposés entre autre en ceux qui sont basés sur le graphe moléculaire, ceux qui basés de la conformation et ceux qui demandent un calcul de la fonction d'onde moléculaire. On distingue :

- Les **mesures unidimensionnelles** [7] sont basées sur les propriétés physico-chimiques des molécules telles que le volume, la polarisation et le poids. Puisqu'elles ne prennent pas en considération les informations géométriques de la molécules, ces mesures sont généralement utilisés pour le clustering des bases de données [8].
- Les **mesures** $2D$ sont basées sur le graphe moléculaire. La mesure la plus répandue pour comparer les structures chimiques représentées en utilisant les empreintes est le coefficient de Tanimoto [1]. La recherche de sous-structure via la sous-structure commune maximum (en anglais, Maximum Common Subgraph MCS [18] se classe aussi dans cette catégorie.
- Les **mesures 3D** tel que l'empreinte $3D$ utilisent l'alignement des molécules dans l'espace. Contrairement aux méthodes $1D$ et $2D$, elles tiennent compte de la taille et de la forme des structures moléculaires (propriétés conformationnelles) [16]. Cependant, elles restent moins utilisées car beaucoup plus complexes [2, 5].

La similarité que nous recherchons durant la thèse se situe sur la partie structurelle notamment les mesures $2D$ sur les graphes moléculaires. Dans la section 1, nous présenterons le quelques mesures de similarité basés sur le calcul d'empreintes dans le graphe moléculaire, plus précisement le coefficient de Tanimoto. Ensuite dans la Section 2, nous exposerons la mesure de similarité basée sur le problème Maximum Common Edge Subgraph.

Ce chapitre contient également dans la Section 3, les définitions préliminaires sur les notions de cycles, bases de cycles, union de bases de cycles dans les graphes.

Ces notions de la théorie de graphes seront utilisés dans les chapitres suivants.

1 Les mesures de similarité basés sur les empreintes

Les empreintes ou suites binaires sont calculées à partir de la présence ou de l'absence de caractéristiques moléculaires. Elles sont généralement comparées en utilisant un coefficient de similarité en tant que mesure de la similarité entre les structures.

Pour calculer la similarité il faut dans un premier temps associer à chaque molécule, une empreinte. Dans un second temps choisir le coefficient adapté pour obtenir une mesure.

1.1 Le calcul des empreintes

Les principales types d'empreintes sont les empreintes basées sur les notions de sous-structures, les empreintes topologiques ou basées sur le chemin, et les empreintes circulaires.

- Les empreintes basées sur des motifs (sous- structures) définissent les bits l'empreinte en fonction de la présence dans le composé de certaines sous-structures ou caractéristiques d'une liste donnée de motifs. Ces empreintes sont intéressantes lorsqu'elles sont utilisées avec des molécules susceptibles d'être recouvertes par la liste de motifs choisie. Elles le sont moins lorsque les molécules contiennent d'autres motifs, car leurs présence ne seraient pas représentés. Leur nombre de bits de l'empreinte est déterminé par le nombre de motifs, et chaque bit renseigne sur la présence ou à l'absence d'une caractéristique donnée dans la molécule.
- Les empreintes topologiques analysent tous les fragments de la molécule en parcourant un chemin linéaire jusqu'à un certain nombre de liaisons, puis en hachant chacun de ces chemins pour créer l'empreinte. Ainsi toute molécule peut produire une empreinte digitale cohérente et sa longueur peut être ajustée. Ce type d'empreintes est utilisé pour la recherche de sous-structures. L'empreinte la plus connu sous le nom de Daylight [] comporte jusqu'à 2048 bits et code toutes les motifs possibles à travers une molécule jusqu'à une longueur donnée.
- Les empreintes circulaires ressemblent aux empreintes topologiques à la différence qu'au lieu de rechercher des chemins dans le graphe moléculaire, ils stockent l'environnement de chaque atome jusqu'à un rayon fixé. On distingue parmis les empreintes cirulaires : Extended-Connectivity Fingerprint (ECFP diamètre 4 et 6) basé sur l'algorithme de Morgan, Functional-Class Fingerprints (FCFP 4 et 6) et Molprint2D.

1.2 Coefficients de similarité

Les mesures de similarité suivantes supposent que deux empreintes avec de nombreux bits en commun sont similaires. Ce sont des mesures brutes mais étonnamment efficaces pour une large gamme d'applications.

En effet, l'information de présence ou d'absence d'un motif est moins exhaustive que le nombre d'occurrences de celui-ci, ce qui, à son tour, ne donne aucune information sur la façon dont les motifs survenus sont répartis dans la molécule.

Le tableau 1.2 présente les différentes mesures et distances connues pour obtenir un coefficient de similarité. La plus utilisée est le coefficient de Tanimoto qui consiste à faire le ratio entre le nombre de bits communs à 1 et le nombbre total de bits à 1 dans les deux empreintes. Ce coefficient est compris dans $[0, 1]$ quelque soit la taille des empreintes.

TABLE 2.1 – **Coefficients de similarité et distance basés sur les empreintes**

Mesure de similarité	Intervalle	Formule
Similarité de Cosine	$[0, 1]$	$\frac{c}{\sqrt{a \times b}}$
Coefficient de Tanimoto	$[0, 1]$	$\frac{c}{\sqrt{a+b-c}}$
Coefficient de Dice	$[0, 1]$	$\frac{2 \times c}{\sqrt{a+b}}$
Distance euclidienne	$[0, N]$	$\sqrt{a + b - 2 \times c}$
Distance de Hamming	$[0, N]$	$a + b - 2 \times c$
Coefficient de Forbes	$[0, 1]$	$\frac{c \times m}{a \times b}$
Distance de Soergel	$[0, 1]$	$\frac{a+b-2 \times c}{\sqrt{a+b-c}}$

On considère deux empreintes moléculaires A et B. On note m, le nombre total de bits dans A et B ; a, b et c représente respectivement le nombre de bits à 1 dans A, le nombre de bits à 1 dans B et le nombre de bits à 1 à la fois dans A et B.

exemple ?

2 Similarité basé sur le Maximum Common Edge Subgraph (MCES)

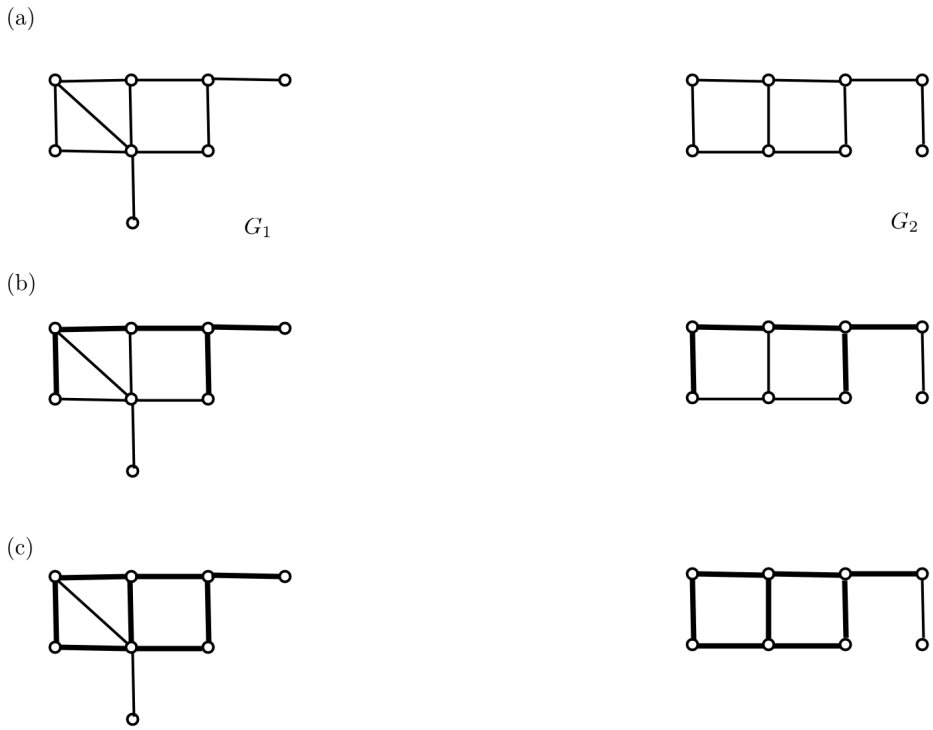
Dans cette section, nous allons présenter le calcul de similarité basé sur la recherche de sous-graphe commun ayant un nombre maximum d'arêtes. Toutefois nous nous limiterons à la définition de la mesure exacte et nous n'allons pas présenter les heuristiques[refs].

Soient deux graphes moléculaires $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$.

Définition 2.1. Un *sous-graphe commun induit* $G_{1,2}$ de G_1 et G_2 est un graphe isomorphe à des sous-graphes induits de G_1 et G_2. Un graphe $G_{1,2}$ est un sous-graphe commun induit maximum (MCIS) s'il contient un nombre maximum de sommets dans G_1 et G_2. Un graphe $G_{1,2}$ est un sous-graphe commun ayant un nombre d'arêtes maximum (MCES) s'il contient un nombre maximum d'arêtes communs dans G_1 et G_2.

Un MCIS et un MCES peuvent être des sous-graphes non connexes.

FIGURE 2.1 – (a) Deux graphes G_1 et G_2, (b) en gras un MCIS et (c) en gras un MCES.



Pour le calcul de similarité trouver un sous-graphe commun ayant un nombre d'arêtes maximum est pertinent [18]. Dans la suite, on note $G_{1,2} = (V_{1,2}, E_{1,2})$ un sous-graphe commun ayant un nombre maximum d'arêtes à la fois dans G_1 et G_2. La mesure de similarité $\mathrm{sim}(\mathrm{G_1}, \mathrm{G_2})$ est :

$$\mathrm{sim}(\mathrm{G_1}, \mathrm{G_2}) = \frac{(|V_{12}| + |E_{12}|)^2}{(|V_1| + |E_1|) \times (|V_2| + |E_2|)} \tag{2.1}$$

La recherche d'un sous-graphe commun ayant un nombre maximum d'arêtes se fait en trois étapes : le calcul du graphe produit, la recherche d'une clique maximum et l'extraction d'un sous-graphe commun à arêtes maximum entre G_1 et G_2.

2.1 Le calcul du graphe moléculaire produit

La première étape consiste à calculer le graphe moléculaire produit des linegraphes G_1 et de G_2 noté $L(G_1) \diamond (G_2)$. Le linegraphe $L(G_1)$ d'un graphe G_1 est un graphe dans lequel les arêtes de G_1 sont des sommets et deux sommets de $L(G_1)$ sont adjacents si et seulement si les arêtes correspondantes dans G_1 sont adjacentes. Ce graphe moléculaire produit est telle que : $V(G_1 \diamond G_2) = V(L(G_1)) \times V(L(G_2))$.

Soient deux sommets $(u_i, v_i), (u_j, v_j) \in V(L(G_1) \diamond L(G_2))$. L'arête $[(u_i, v_i), (u_j, v_j)] \in E(L(G_1) \diamond L(G_2))$ si et seulement si :

— $(u_i, u_j) \in E(L(G_1))$, $(v_i, v_j) \in E(L(G_2))$ et $w(u_i, u_j) = w(v_i, v_j)$ **OU**

— $(u_i, u_j) \notin E(L(G_1))$, $(v_i, v_j) \notin E(L(G_2))$.

La fonction $w(u_i, u_j) = w(v_i, v_j)$ vérifie la compatibilité entre les atomes et les liaisons. Elle associe les atomes et les liaisons du même type.

Exemple 2.1.1. On considère deux graphes moléculaires G_1 et G_2 de la Figure 2.1.1(a). Les Figures 2.1.1(b) et 2.1.1(c) montrent respectivement le calcul du linegraph des graphes et les sommets du graphe produit.

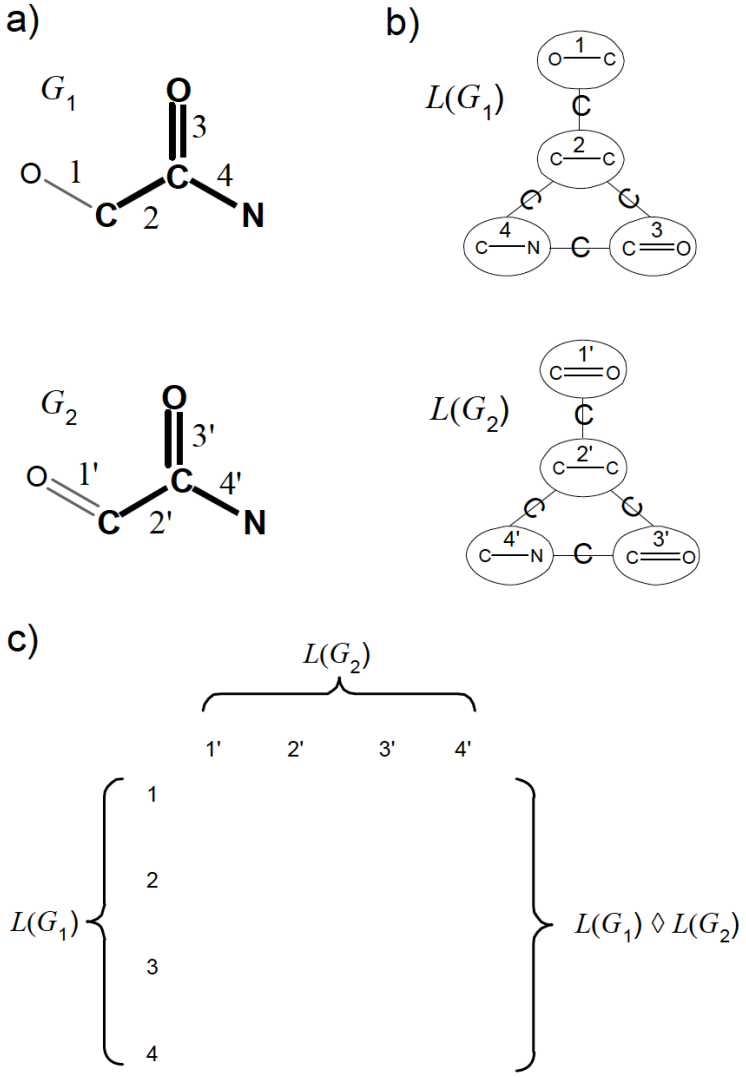


FIGURE 2.2 – Construction d'un graphe moléculaire produit

2.2 La recherche d'une clique maximum et l'extraction d'un sous-graphe commun à arêtes maximum

En utilisant le graphe produit des linegraphes de G_1 et G_2, on recherche une clique maximum à partir de laquelle on extrait un sous-graphe commun à arêtes maximum.

Exemple 2.2.1. En reprenant le graphe moléculaire produit de l'exemple précédent, on cherche une clique maximum.

	$(1,1')$	$(1,2')$	$(1,3')$	$(1,4')$	$(2,1')$	$(2,2')$	$(2,3')$	$(2,4')$	$(3,1')$	$(3,2')$	$(3,3')$	$(3,4')$	$(4,1')$	$(4,2')$	$(4,3')$	$(4,4')$
$(1,1')$	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
$(1,2')$	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$(1,3')$	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$(1,4')$	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
$(2,1')$	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
$(2,2')$	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
$(2,3')$	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
$(2,4')$	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
$(3,1')$	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
$(3,2')$	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
$(3,3')$	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
$(3,4')$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
$(4,1')$	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
$(4,2')$	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
$(4,3')$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
$(4,4')$	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1

On trouve une clique maximum de taille 3. La taille de la clique est égale au cardinal de $E(G_{12})$. On fait ensuite une projection des arêtes de la clique dans G_1(ou G_2) pour obtenir un sous-graphe commun ayant un nombre maximum d'arêtes à la fois dans G_1 et G_2. Dans cet exemple, on a :

- $|V(G_1)| = 5$ et $|E(G_1)| = 4$
- $|V(G_2)| = 5$ et $|E(G_2)| = 4$
- $|V(G_{12})| = 4$ et $|E(G_{12})| = 3$

$$\text{sim}(\text{G}_1, \text{G}_2) = \frac{(4+3)^2}{(5+4)\times(5+4)} = 0,6$$

La recherche de clique maximum est un problème NP-Complet. Il existe des heuristiques tels que les algorithmes de branch and bound [18].

3 Quelques préliminaires de la théorie de graphes

Dans cette section, nous allons définir quelques notions de la théorie de graphes (ref claude berges graph theory and application) que nous utiliserons dans les chapitres suivants. Les informations sur les cycles constituent une partie importante de la topologie structurelle utilisée pour identifier et caractériser les structures moléculaires. Il est donc d'une importance cruciale d'obtenir ces informations pour diverses tâches en chimie computationnelle.

On considère un graphe simple et non orienté $G = (V, E)$ avec n et m respectivement le nombre de sommets et le nombre d 'arêtes de G. On note $V = \{v_1, v_2, ..., v_n\}$ avec v_i le sommet d'indice i et l'ensemble $E = \{e_1, e_2, ..., e_m\}$ tel que $e_j = [v_i, v_i']$.

3.1 Chaînes et cycles

Définition 3.1. Une chaîne est une suite fini d'arêtes consécutives $e_1, e_2, ..., e_k$ de G. La longueur d'une chaîne est le nombre d'arêtes qui la composent.

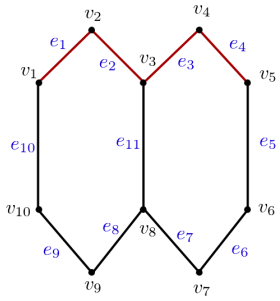
Une chaîne est élémentaire si et seulement si elle passe au maximum une fois par chaque sommet. Une chaîne simple est une chaîne ne passant pas deux fois par une même arête.

Définition 3.2. Un cycle est une chaîne simple dont les extrémités sont identiques. Un cycle élémentaire ne contient pas d'autres cycles.

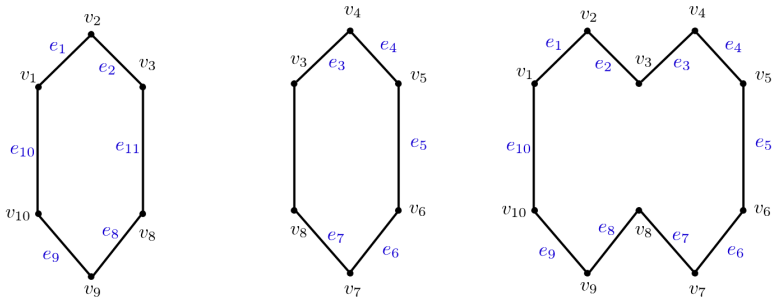
Nous allons représenter un cycle élémentaire c par un vecteur $v_c = (e_1^c, e_2^c, ..., e_m^c)$ de taille m avec $e_i^c = 1$ si et seulement si $e_i^c \in c$ et $e_i^c = 0$ sinon. La longueur d'un cycle notée $|c|$ vaut $\sum_{i=1}^{m} e_i^c$.

Exemple 3.1.1. Soit le graphe G de la Figure 3.1.1 avec $V = \{v_1, v_2, v_3, ..., v_9, v_{10}\}$ et $E = \{e_1, e_2, ..., e_{10}, e_{11}\}$. La chaîne (e_1, e_2, e_3, e_4) est une chaîne élémentaire de G. Les cycles $c_1 = (e_1, e_2, e_{11}, e_8, e_9, e_{10})$ et $c_2 = (e_3, e_4, e_5, e_6, e_7, e_{11})$ sont des cycles élémentaires dont les vecteurs sont respectivement :

$$v_{c_1} = (1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1)$$
$$v_{c_2} = (0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1)$$



G



c_1 c_2 c_3

FIGURE 2.3 – Chaînes et cycles élémentaires dans un graphe

On considère deux cycles élémentaires c_1 et c_2 dans un graphe G telle que : $v_{c_1} = (e_1^{c_1}, e_2^{c_1}, ..., e_m^{c_1})$ et $v_{c_2} = (e_1^{c_2}, e_2^{c_2}, ..., e_m^{c_2})$.

Définition 3.3. L'union disjointe des cycles c_1 et c_2 notée c_{12} est la différence symétrique des arêtes du graphe appartenant à c_1 et c_2. On note $c_{12} = c_1 \oplus c_2$ et le vecteur associé est $v_{c_{12}} = (e_1^{c_1} \oplus e_1^{c_2}, e_2^{c_1} \oplus e_2^{c_2}, ..., e_m^{c_1} \oplus e_m^{c_2})$.

L'opérateur $\oplus$ représente le XOR binaire sur les e_i^c. L'union de deux cycles disjoints donne un vecteur composé de deux cycles disjoints. Dans l'Exemple 3.1.1, le cycle c_3 est l'union disjointe des cycles c_1 et c_2.

Lorsque dans un graphe, deux sommets d'un cycle sont reliés par une arête qui n'appartient pas au cycle, cette arête est appelée *corde* du cycle. Un cycle de G dit induit lorsqu'il n'a pas de cordes.

Définition 3.4. Un cycle est dit **pertinent** s'il ne peut pas être obtenu par combinaison de cycles de longueur strictement plus petite.

3.2 Connexité et isthme

Définition 3.5. Un graphe G est **connexe** si toute paire de sommets $u, v \in V$ peut être reliée par une chaîne dans G.

Un sous-graphe connexe maximal de G est appelé composante connexe de G. Un graphe qui n'est connexe possède au moins 2 composantes connexes.

Soit un graphe connexe G et un ensemble d'arêtes E' de G. Si le graphe $G - E'$ n'est pas connexe, on dit que E' est un séparateur de G. Si dans $G - E'$, deux sommets u et v sont dans deux composantes connexes différentes, on dit que E' sépare u de v ou E' est un $(u, v)-$séparateur.

Définition 3.6. Pour $k \geq 1$, on dit que G est $k-$connexe s'il a au moins $k + 1$ sommets et aucun ensemble de $k - 1$ sommets ne le sépare. Un graphe est k-arête-connexe, s'il a au moins deux sommets et aucun ensemble de $k - 1$ arêtes ne le sépare.

La valeur maximale k pour laquelle une graphe G est $k - 1$connexe est appelée connexité de G et notée $\kappa(G)$. En particulier, lorsque $k = 2$ on parle de graphe $2-$connexe ou bi-connexe. De même, la valeur maximale k pour laquelle une graphe G est $k-$arête-connexe est appelée arête-connexité de G et notée $\kappa\,(G)$.

Définition 3.7. Une arête d'un graphe est un **isthme** si sa suppression augmente le nombre de composantes connexes du graphe.

Si le graphe est connexe, une arête est un isthme si et seulement si elle n'appartient à aucun cycle.

3.3 Base de cycles

Dans un graphe, on appele générateur de cycles ζ, un ensemble fini de cycles tel que tout cycle de G peut être obtenu par combinaison des cycles de ζ.

Définition 3.8. On appelle une base de cycles $\mathcal{B}$, un ensemble de cycles linéairement indépendants et générateur.

Soit G un graphe de n sommets, m arêtes et p composantes connexes, la dimension d'une base de cycles est le nombre cyclomatique de $G : \nu(G) = m - n + p$. La longueur d'une base de cycles ζ est la somme des longueurs des cycles de la base. Il existe 2 principales types de bases de cycles : les base de cycles fondamentales et les bases de cycles de longueurs minimum.

Base de cycles fondamentale

Lorsque le graphe G est connexe, il est possible d'obtenir des bases de cycles en utilisant les arbres couvrants. Ces bases de cycles sont appelées **bases de cycles fondamentales** [**?**].

Soit T un abre couvrant arbitraire de G. Pour obtenir une base de cycles fondamentale, on rajoute à chaque fois à T, une arête de G qui n'appartient pas à T. A chaque ajout, on cree un cycle qui appartient à la base de cycle fondamentale. On repète le procédé $m - n + 1$ fois.

Cependant, trouver un abre couvrant tel que la base de cycles fondamentale soit minimum est un problème NP-Complet.

Base de cycles de longueur minimum

Une base de cycles de longueur minimum ou Mimimum Cycle Basis (MCB) est une base de cycles ayant une longueur minimale. Il existe plusieurs algorithmes pour trouver de telles bases. En chimie, les algorithmes developpés et applicables sur les structures moléculaires et graphes simples sont connus sous le nom de Smallest Set of Smallest Rings (SSSR) [12, 15, 19]. En informatique, les algorithmes MCB d'une grande robustesse, sont concus pour des graphes en général [4, 10, 11, 13].

Dans un graphe, il peut y avoir plusieurs bases de cycles de poids minimum. Lorsqu'il existe une unique base de cycles minimum $\mathcal{B}$ dans un graphe, cette base est aussi l'ensemble de cycles pertinents du graphe.

Union des bases de cycles de longueur minimum

Également appelé ensemble de cycles pertinents, l'union des bases de cycles de longueur minimum est défini comme le plus petit ensemble canonique de cycles qui décrit la structure cyclique d'un graphe [17]. Sa construction passe par l'obtention d'une représentation compacte servant de prototype pour les énumérer.

Chapitre 3

Modèle de graphe de cycles

Dans ce chapitre, nous allons définir formellement le modèle de graphes de cycles. Le graphe de cycles d'une molécule représente la partie structurelle de celle-ci et est construit en utilisant son graphe moléculaire. Nous construisons ce modèle pour calculer la similarité structurelle entre molécules. Ainsi, il est primordial que des molécules structurellement similaires aient des graphes de cycles similaires.

Cette représentation de la partie structurelle d'une molécule se fait à une échelle gros-grain. On ne se situe pas à l'échelle atomique comme dans les graphes moléculaires mais à l'échelle cyclique. L'élaboration de ce modèle gros-grain appelé par la suite graphe de cycles passe d'une part par la construction d'un ensemble de cycles pertinent pour décrire la structure moléculaire et d'autre part par l'interconnexion de ces cycles en fonction de leurs interactions dans la molécule. On souhaite garantir la notion de canonicité et la proximité sur les graphes de cycles. La canonicité signifie que deux graphes moléculaires isomorphes et ayant une numérotation de sommets différente aient des graphes de cycles isomorphes. Elle garantie également qu'à un graphe moléculaire, on associe un unique graphe de cycles quelque soit la numérotation des sommets. La proximité quand à elle signifie que si deux graphes moléculaires sont proches (elles diffèrent d'un cycle par exemple) alors les graphes de cycles le seront aussi. De plus, pour le chimiste, certains "grands" cycles dans les molécules ne définissent pas la structure de celle-ci. Il s'avère donc nécessaire de ne pas sauvegarder une certaine taille de cycles dans les graphes de cycles.

Dans la section 1, nous présentons et analysons une construction intuitive du graphe de cycles d'une molécule. Ensuite, nous définissons dans la section 2, le graphe de cycles qui capture la structure de cycles et l'interconnexion de ces cycles. Plus particulièrement, nous expliquons comment relier les cycles dans le graphe de cycles en respectant la structure du graphe moléculaire. Finalement dans la Section 3 nous présenterons cette section l'algorithme permettant d'avoir une base de cycles pertinente et suffisante pour capturer la partie structurelle des molécules. Nous verrons la notion d'hiérarchie pour limiter la taille des cycles dans le graphe de cycles.

1 Une représentation intuitive du graphe de cycles

On veut construire un graphe de cycles d'une molécule basé sur sa partie structurelle. Bien que le graphe moléculaire modélise globalement l'ensemble de l'informa-

tion structurelle, il n'encode pas explicitement l'information cyclique. L'hypothèse étant que des molécules similaires sont structurellement similaires et que la partie structurelle d'une molécule est défini par l'interconnexion des cycles pertinants présentes dans celle-ci. Un *cycle pertinent* est un cycle simple et élémentaire ne pouvant être obtenu par combinaison de cycles plus petits [17].

Pour s'assurer que l'on capture les cycles pertinents, on pourrait au premier abord, prendre tous les cycles de la molécule. Cependant, compter tous les cycles dans un graphe est NP-complet sachant qu'il peut y avoir un nombre exponentiel. Pour son usage dans la mesure de similarité, la représentation du graphe de cycles d'une molécule doit contenir un nombre minimal de cycles mais suffisant pour décrire au mieux sa structure ; il s'avère donc inutile de prendre tous les cycles de la molécule pour construire le graphe de cycles.

Une base de cycles d'un graphe quand elle contient un ensembe fini de cycles élémentaires (cf. chap2). Les cycles d'une base sont pertinents. Pour construire le graphe de cycles (Figure 1) d'une molécule $\mathcal{M}$, nous prenons intuitivement un hypergraphe $GC = (VC, EC, \mu)$ (défini par [3]) :

— L'ensemble de sommets VC est une base de cycles $\mathcal{B}$. Chaque sommet c correspond à un cycle pertinent et $\mu(c)$ correspond au sous-graphe moléculaire dans (M) contenant les atomes et liasons qui appartient au cycle c.
— Deux cycles c_1 et c_2 de VC sont reliés par une arête si et seulement si elles partagent au moins un atome dans le graphe moléculaire de $\mathcal{M}$ c'est à dire $\mu(c_1) \cap \mu(c_2) = \emptyset$.

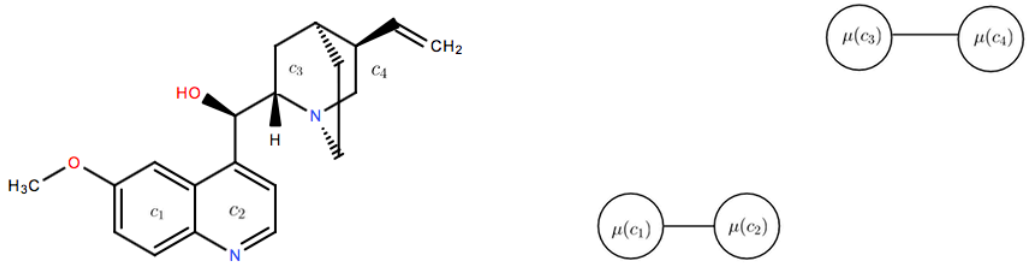


FIGURE 3.1 – Graphe moléculaire de la quinine et le graphe de cycles intuitif associé

En observant le graphe de cycles de la Figure 1, on ne peut pas savoir exactement si le graphe moléculaire possède 2 composantes connexes ou dans le contraire, sur quel cycle entre c_1 et c_2 les cycles c_3 et c_4 sont liées dans $\mathcal{M}$. Es-ce le cycle c_3 qui est relié au cycle c_1 ? ou au cycle c_2 ? ou à la fois à c_1 et c_2 ?

On pourrait tout à fait donner comme graphes moléculaires associés au graphe de cycles de la quinine, les graphes de la Figure 1.

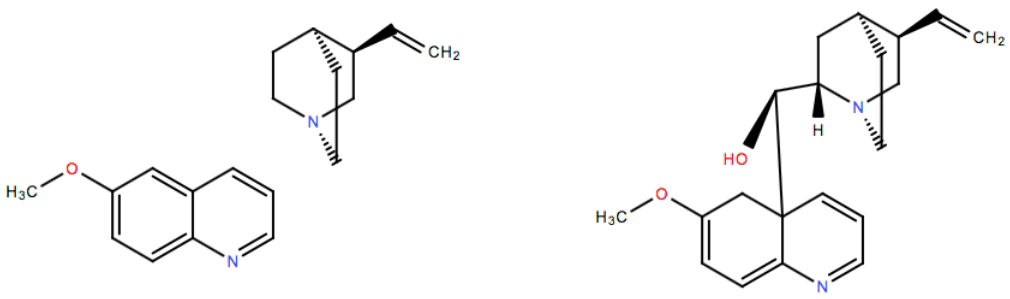


FIGURE 3.2 – Graphes moléculaires ayant le même graphe de cycles que la quinine.

On constate dès lors que ce modèle n'est pas suffisant notamment :
— Au niveau de l'interconnexion des cycles : il semble logique que deux cycles ayant des atomes en commun soient liés dans GC mais il faudrait aussi rajouter des liens supplementaires entre cycles indirectement liés pour réduire l'ambigüité et mieux décrire la structure des molécules.
— Au niveau de la base de cycles : Elle n'est pas nécessairement unique ! Ce qui entrainerait dans la construction du graphe de cycles que l'on puisse avoir un graphe de cycles différent en fonction de la base choisi. En terme de similarité, cela reviendrait à dire qu'une molécule n'est pas totalement similaire à elle-même.

2 Interconnexion des cycles

Nous avons vu dans la section précédente que le graphe de cycles devrait avoir suffisamment d'arêtes pour décrire la partie structurelle de la molécule. Dans cette section nous allons définir les types d'intersections que nous utiliserons danns le graphe de cycles.

Définition 2.1. Étant donné un graphe moléculaire $G = (V, E)$ d'une molécule $\mathcal{M}$, son graphe de cycles GC est défini par un graphe simple, non orienté et etiqueté $GC = (VC, EC, \phi, \varphi, \chi)$ tel que :
— L'ensemble de sommets VC est constitué d'un nombre fini de cycles pertinents décrivant la structure de $\mathcal{M}$,
— La fonction $\phi : VC \mapsto \mathbf{N}$ pour étiqueter les sommets de GC,
— L'ensemble d'arêtes EC modélise les différentes interconnexions entre les sommets du graphe de cycles,
— Les fonctions $\varphi : EC \mapsto \{1, 2\}$ et $\chi : EC \mapsto \mathbf{N}$ permettent respectivement de distinguer les types d'arêtes et leurs étiquettes.

Nous distinguons 2 principaux types d'interconnexions entre cycles : les cycles directement liés et les cycles indirectement liés.

2.1 Le type 1 : Cycles directement liés

Deux cycles sont directement liés s'ils ont une relation directe dans la molécule. C'est à dire qu'ils partagent au moins un atome dans le graphe moléculaire de $\mathcal{M}$.

Chimiquement cela signifie qu'une interaction sur l'un des cycles affecte rapidement l'autre. Si deux cycles sont reliés par une arête de ce type alors ils appartiennent à une même composante 2−connexe.

Soit un graphe de cycles GC et deux cycles $c_1, c_2 \in VC$. Les fonctions d'étiquettages des arêtes sont telles que :

— $\varphi([c_1, c_2]) = 1$ si l'arête est de type 1,
— $\chi([c_1, c_2])$ pour une arête de type 1 est égale au nombre de liaisons chimiques en commun entre les cycles c_1 et c_2.

Si c_1 et c_2 partagent un seul atome alors $\chi([c_1, c_2]) = 0$.

Exemple 2.1.1. Le cycle élémentaire c_2 (situé au milieu du graphe moléculaire) est lié de part et d'autre par des arêtes de type 1. On constate aussi que le cycle c_1 et c_3 ne sont pas liés. Cette connexion n'est pas nécessaire car le graphe de cycle fournit l'information du voisin commun c_2 entre les cycles c_1 et c_3.

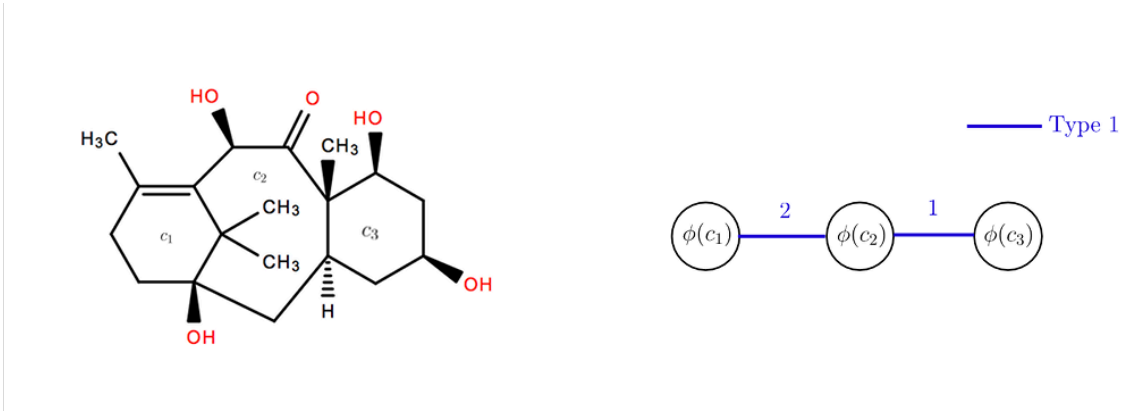


FIGURE 3.3 – Graphe moléculaire et graphe de cycles

2.2 Le type 2 : Cycles indirectement liés

Deux cycles sont indirectement liés lorsqu'ils sont reliés dans le graphe moléculaire par des chaînes. Dans la Figure 1, on a constaté qu'il fallait des informations supplémentaires pour décrire en la partie strcutrelle de la molécule.

Bien que les cycles c_2 et c_3 de la Figure 2.2.1 ne partagent pas d'atomes, une arête entre ces deux cycles dans GC permettrait de lever l'ambiguïté.

Soit un graphe de cycles GC et deux cycles $c_1, c_2 \in VC$. Les fonctions d'étiquettages des arêtes sont telles que :

— $\varphi([c_1, c_2]) = 2$ si l'arête est de type 2. C'est à dire qu'il existe dans G une chaîne d'atomes ne passant par aucun cycle de VC reliant les cycles c_1 et c_2,
— $\chi([c_1, c_2])$ pour une arête de type 2 est égale à la longueur de la plus petite chaîne reliant c_1 et c_2 dans G.

Exemple 2.2.1. Les cycles élémentaires c_2 et c_3 appartiennent à deux composantes 2−connexes différentes. La plus petite chaîne relaint ces cycles est de longueur 2.

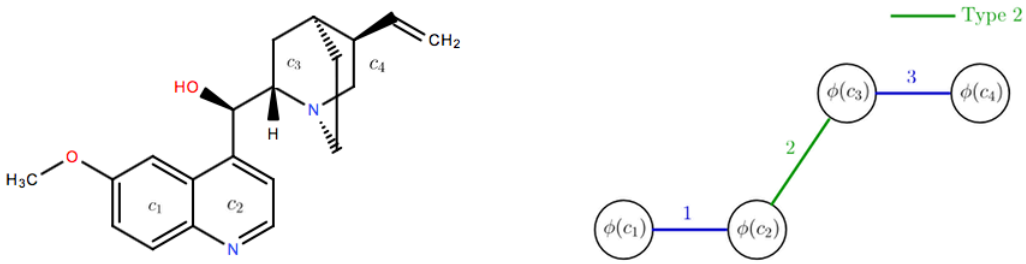


FIGURE 3.4 – Graphe moléculaire et graphe de cycles

Maintenant que nous avons défini les types d'interconnexions, nous allons procédér au choix de la base de cycles pour construire le graphe de cycles.

3 Le générateur pour le graphe de cycles

On rappelle que l'on souhaite avoir une représentation canonique de la partie structurelle de la molécule. Dans cette section, nous allons construire l'ensemble de sommets VC et définir la fonction $\phi(VC)$.

3.1 Une base de cycles comme générateur ?

Une base de cycles de longueur minimum est un candidat potentiel pour construire le graphe de cycles. Il contient un nombre fini de cycles et toutes les bases de cycles d'un graphe ont le même nombre de cycles.

En reprenant le graphe moléculaire de la quinine, on constate dans la Figure 3.1 que ce graphe possède 3 bases de cycles différentes. Deux cycles de l'ensemble $\{c_3, c_4, c_5\}$ suffisent pour obtenir une base.

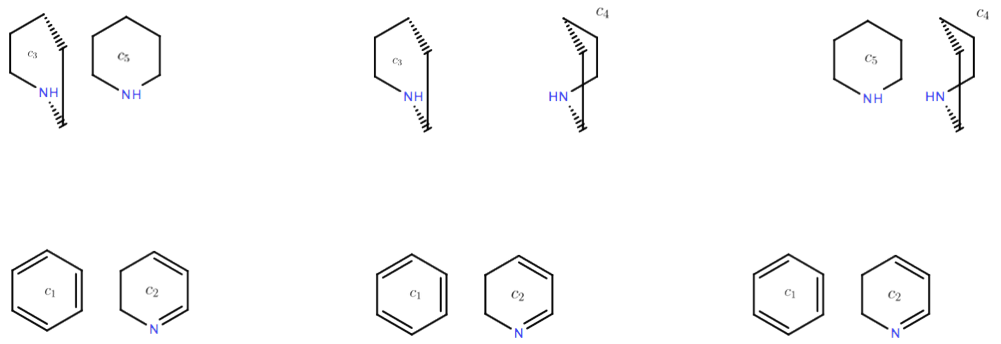


FIGURE 3.5 – Les bases de cycles pour le graphe moléculaire de la quinine

En construisant les graphes de cycles avec chacune de ces bases on obtient respectivement GC_1,GC_2 et GC_3. Les graphes GC_2 et GC_3 sont isomorphes. Le choix

d'une base de cycle pourrait donner un graphe de cycles différent pour une même molécule.

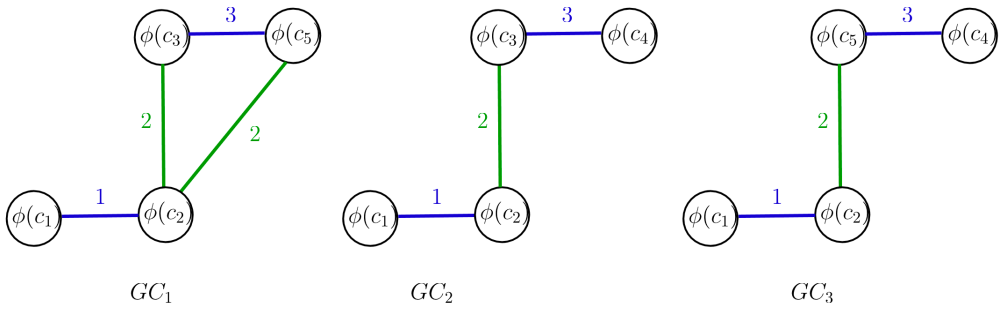


FIGURE 3.6 – Les graphes de cycles pour les différentes bases de cycles.

Dans ce cas, si on prend deux fois la même molécule et deux bases de cycles différentes, la mesure de similarité conclurait que les graphes de cycles ne sont pas totalement similaires.

Cela deviendrait encore plus ambigüe s'il existe des cycles raccrochés au composé xxxx (voir Figure 3.1). Si on continue d'étendre ce graphe moléculaire, les graphes de cycles d'une molécule pourraient prêter à confusion pour la similarité.

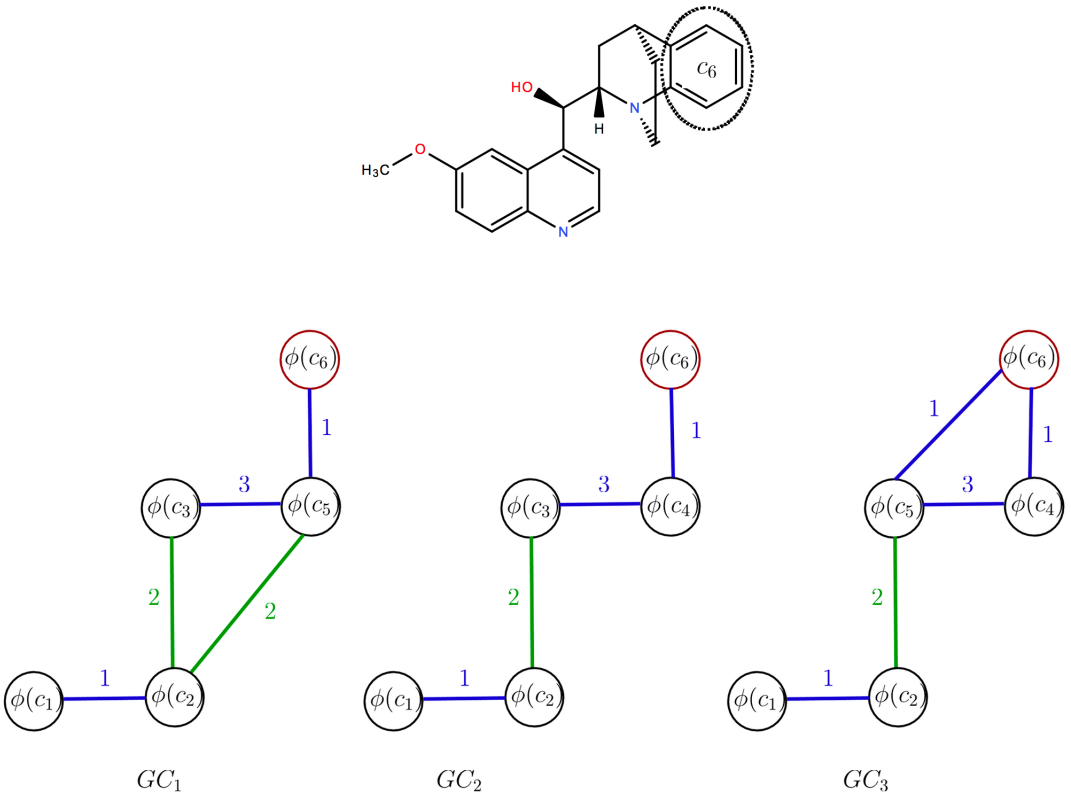


FIGURE 3.7 – L'ajout d'un cycle dans un graphe moléculaire augmente potentiellemnt le nombre de graphes de cycles possibles.

En observant l'algorithme de Horton, précisement au niveau de la selection des cycles indépendants, deux des cycles c_3, c_4 et c_5 seront selectionnés en fonction de l'ordre de classement qui n'est pas prédefini.

Pour construire un générateur canonique pour le graphe de cycles, nous utilisons l'algorithme de Horton et l'algorithme de McKay. La première étape consiste à numéroter canoniquement les sommets du graphe moléculaire en utilisant l'algorithme d'isomorphisme de McKay. Ensuite, cette numérotation de sommets est utilisée pour obtenir une base de cycle de longueur minimum $\mathcal{B}$ avec l'algorithme de Horton. Et finalement, on rajoute eventuellement des cycles pour garantir la canonicité.

3.2 Algorithme de McKay pour la numérotation canonique

Cet algorithme a été conçu pour résoudre le problème d'isomorphisme de graphes. Décider de l'existence d'un isomorphisme entre deux graphes G_1 et G_2 est dans NP. L'algorithme de Brendan McKay [9] cherche un étiquetage canonique des sommets d'un graphe permettant leur ordonnancement. Si deux graphes possèdent le même étiquetage alors ils sont isomorphes [14].

Cependant le problème peut être résolu en temps polynomial pour certaines classes de graphes, par exemple les graphes planaires ou les graphes de degré borné et en temps quasi-polynomial pour le cas général. Les graphes moléculaires se retrouvent dans cette catégorie car le degré des sommets est borné par le nombre de liaisons covalentes des atomes.

Définition 3.1. Une partition d'un ensemble E est une collection d'ensembles disjoints deux à deux et non vides (appelées blocs dans la suite) dont l'union est E. Une partition ordonnée de E est une partition dont on a ordonné les blocs. Lorsque chaque bloc d'une partition est de taille 1, la partition est dite discrète.

L'algorithme de McKay commence par établir un partitionnement ordoné et équitable des sommets en fonction du degré. Une partition dans lequel les sommets du même bloc sont deux à deux adjacents à un nombre identique de sommets dans un autre bloc est une partition **équitable**. Avec une partition équitable, l'algorithme introduit des distinctions supplementaires entre les sommets. A chaque étape, il examine tous les choix pertinents en explorant systématiquement l'espace des partitions ordonnées équitables à l'aide d'un arbre de recherche. L'algorithme s'arrête lorsque chaque partition est discrète.

Ensuite à chaque noeud final de l'arbre est une partition ordonée et discrète. Chacune de ces partitions définit une nouvelle numérotation des sommets de V. La numeérotation se fait dans l'ordre dans lequel les sommets apparaisssent.

Chaque isomorphisme peut-être vu comme un mot en concatenant les ensembles de la partition. La numerotation canonique est l'isomorphisme ayant le mot le plus petit.pas tout a fait ! c'est un mot binaire, j'ecrirais la version exacte plus tard... DOIS JE RAJOUTER UN PETIT EXEMPLE POUR MONTRER COMMENT CA FONCTIONNNE ?

Cet algorithme est implémenté dans le package *nauty* [14] disponible sous differents systèmes d'exploitations (Windows, Mac OS, GNU/Linux).

3.3 Algorithme de Horton pour une base de cycles

Cet algorithme calcule une base de cycles de longueur minimum dans un graphe. Soit un graphe $G = (V, E)$, l'algorithme de Horton [11] est polynomial en $\mathcal{O}(|E|^3 \times |V|)$.

Il se base sur le théorème suivant :

Théorème 3.1. Soit un sommet v appartenant à un cycle c dans une base de cycles de longueur minimum $\mathcal{B}$. Alors il existe une arête (x, y) dans c tel que : $c = P(v, x) + P(v, y) + (x, y)$ avec $P(v, x)$ un plus court chemin entre v et x dans G.

La première étape de l'algorithme consiste à trouver s'il existe, un plus court chemin entre chaque paire de sommet. Ensuite, on verifie s'il existe des cycles élémentaires entre chaque sommet et chaque arête du graphe. Cela permet d'obtenir au maximum $|E| \times |V|$ cycles élémentaires. À l'étape 3, les cycles sont classés par taille croissante de manière à les selectionner dans cet ordre durant la dernière étape.

Algorithme 1 : Algorithme de Horton

Données : Un graphe simple $G = (V, E)$
Résultat : Une base de cycles de longueur minimum $\mathcal{B}$
1 Trouver un plus court chemin entre chaque paire de sommets de V ;
2 Pour chaque sommet v et chaque arête (x, y) du graphe G, créer un cycle $c(v, x, y) = P(v, x) + P(v, y) + (x, y)$;
3 Ordonner les cycles par taille croissant ;
4 Utiliser un algorithme glouton pour extraire une base de cycles de longueur minimum à partir des cycles obtenus ;

L'étape 4 consiste former une matrice binaire dans laquelle chaque ligne est un cycle (en respectant l'ordre effectué à l'étape 3). Chaque colonne représente une arête du graphe. Si une arête appartient à un cycle la case associé aura la valeur 1 et la valeur 0 sinon. L'élimination de Gauss peut alors être appliquée pour obtenir une base de cycle de longueur minimum.

Un exemple ?

3.4 Générateur canonique

Pour obtenir le générateur canonique ζ d'un graphe moléculaire, on calcule une base de cycles de longueur minimum $\mathcal{B}$ en ayant au préalable effectué une numérotation canonique des sommets. Ensuite pour résoudre le prolème engendré par la chiralité, on rajoute éventuellement à $\mathcal{B}$ l'intégralité des cycles respectant la règle suivante :

$\forall c, c' \in \mathcal{B}$ on calcule $c" = c \oplus c'$ en combinant c et c'. Si $c"$ est un cycle élementaire de taille plus petite ou égale à ceux de c et c' alors il pertinent pour le graphe de cycles.

La fonction d'etiquettage des sommets ϕ associe à chaque cycle, sa taille c'est dire le nombre d'atomes qu'il contient dans le graphe moléculaire.

Dans la figure 3.4, le graphe de cycles de la quinine conserve alors les trois cycles du composé chiral. Ces cycles partagent alors des arêtes de type 1 avec $\phi(c_i)$, la taille du cycle c_i.

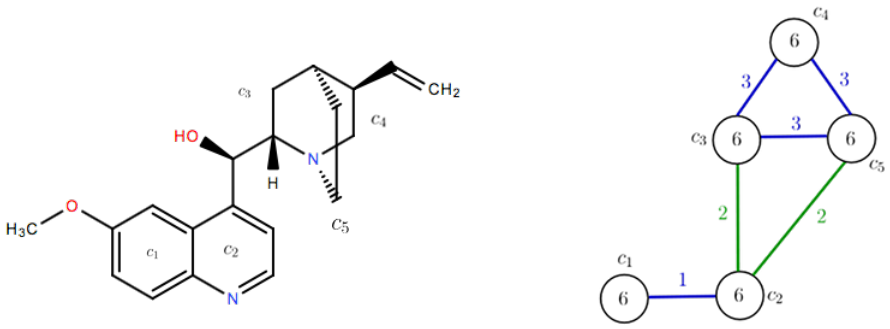


FIGURE 3.8 – Le graphe de cycles GC de la quinine.

La complexité de l'algorithme qui calcule le graphe de cycles d'un graphe G est $\mathcal{O}(n^2.m^3)$.

3.5 Générateur j-hierarchique

Nous allons introduire la notion de $j-$hierarchique avec l'exemple suivant :

Exemple 3.5.1. On considère deux molécules struturellement similaires : la Strychnine et la Vomicine. Sur la figure 3.5.1, si on prend en compte tous les cycles du générateur canonique, alors on trouverait que les graphes de cycles associés à ces molécules ne sont pas assez similaires avec en mesurant la similarité.

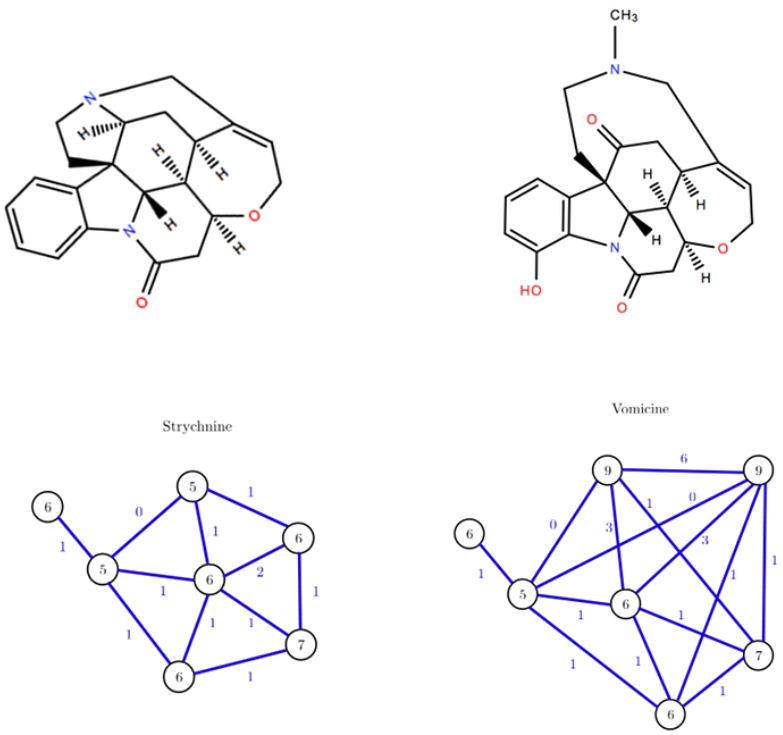


FIGURE 3.9 – Les molécules : Strychnine et Vomicine avec leur graphe de cycles.

En effet, deux cycles de la Strychnine (de taille respectivent 5 et 6) ont fusionnés entraînant l'apparition de deux cycles de taille 9 dans le graphe de cycles de la Vomicine. De plus, ces cycles ne sont pas chimiquement pertinents pour la Vomicine.

Si on reprend le générateur canonique de la Vomicine et on supprime les cycles de taille 9, le nouveau graphe de cycles obtenu (dans la figure 3.5.1) capte mieux sa similarité avec la Strychnine.

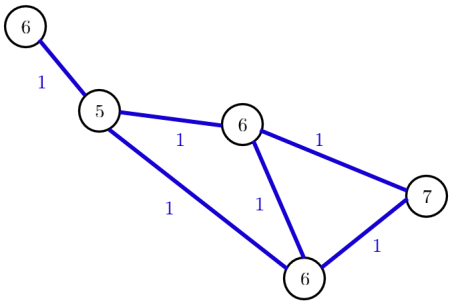


FIGURE 3.10 – Le graphe de cycles de la Vomicine en supprimant les cycles de taille 9.

Définition 3.2. Soit un générateur canonique ζ et un entier j. Le générateur ζ est $j-$hierarchique si le sous-ensemble contenant tous les cycles de ζ ayant une taille inférieure ou égale à j génère tous les cycles de G ayant une taille inférieure ou égale à j.

Un générateur est *hierarchique* si et seulement si il est $j-$hierarchique pour tout j. On note ζ_j le générateur $j-$hierarchique de ζ. Le générateur canonique utilisé pour construire le graphe de cycles est hierarchique car la base de cycles de longueur minimum l'es.

Lemme 3.1. Toute base de cycles de longueur minimum d'un graphe est hierarchique.

Preuve 3.1. Soit G un graphe et $\mathcal{B}$ une base de cycles de longueur minimum de G. $\forall j, \mathcal{B}_j$ est une base de cycles de longueur minimum pour les cycles de longueur inférieur ou égal j.

Supposons que $\mathcal{B}$ n'est pas hierarchique ; Cela signifie qu'il existe un entier j tel que $\mathcal{B}_j$ n'est pas $j-$hierarchique.

Soit un entier j tel que $\mathcal{B}_j$ n'est pas $j-$hierarchique. Il existe par définition un cycle $c \notin \mathcal{B}$ tel que $|c| \leq j$ et c ne peut-être généré en utilisant les cycles de $\mathcal{B}_j$.

Prenons dans $\mathcal{B}$ un ensemble de cycles $\{c_1, c_2, ..., c_\alpha\}$ tel que $c = c_1 \oplus c_2 \oplus ... \oplus c_{\alpha-1} \oplus c_\alpha$. Supposons que c_α est un cycle de longueur maximum dans $\{c_1, c_2, ..., c_\alpha\}$. Puisque c ne peut-être généré par $\mathcal{B}_j$ alors $|c_\alpha| > j$.

De plus, $\oplus$ est associatif et commutatif, on a $c_\alpha = c_1 \oplus c_2 \oplus ... \oplus c_{\alpha-1} \oplus c$. Soit $\mathcal{B}' = \mathcal{B} \backslash \{c_\alpha\} \cup \{c\}$, $\mathcal{B}'$ est aussi une base de cycle.

La longueur de la base $\mathcal{B}'$ est $|\mathcal{B}'| = |\mathcal{B}| - |c_\alpha| + |c|$, donc $|\mathcal{B}'| < |\mathcal{B}|$. Il y'a contradiction car $\mathcal{B}$ est une base de cycle de longueur minimum.

De facon générale, les cycles chimiquement pertinents pour la structure d'une molécule sont de taille inférieur ou égale à 8. Mais il arrive que des cycles très grands le sont dans certains graphes moléculaires. Dans la figure 3.5, l'Amphotéricin B possède un cycle de taille 36 qui fait sa particularité.

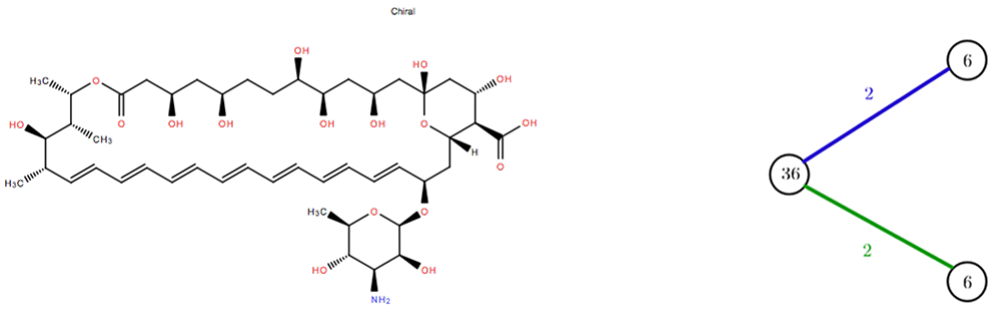


FIGURE 3.11 – Le graphe moleculaire et le graphe de cycles de l'Amphotéricin B.

Dans ce chapitre, nous avons définit formellement le graphe de cycles. Pour le calculer, nous utilisons l'algorithme de McKay pour la numérotation canonique des sommets du graphe moléculaire et ensuite l'algorithme de Horton pour construire un générateur canonique. Par la suite, nous rajoutons éventuellement des cycles supplémentaires pour assurer la canonicité du graphe de cycles. Le choix du paramètre j sera important pour mesurer la similarité dans le chapitre suivant et permet de valider le modèle de graphe de cycles.

Chapitre 4

Évaluation de la similarité moléculaire

1 Titre de la premiere section

Bla bla bla

2 Titre de la seconde section

Re-bla bla bla

3 Conclusion

Conclusion du chapitre.

Chapitre 5

Similarité pour la catégorisation des réactions

1 Titre de la premiere section

Bla bla bla

2 Titre de la seconde section

Re-bla bla bla

3 Conclusion

Conclusion du chapitre.

Conclusion

Bla bla bla.

Bibliographie

[1] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations ? *Journal of cheminformatics*, 7(1) :20, 2015.

[2] Frédérique Barbosa and Dragos Horvath. Molecular similarity and property similarity. *Current topics in medicinal chemistry*, 4 :589–600, 02 2004.

[3] Didier Villemin BeHartkenoit Gaüzère, Luc Brun. Représentation des cycles d'une molécule sous forme d'hypergraphe. *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*, 06 2014.

[4] Franziska Berger, Peter Gritzmann, and Sven de Vries. Minimum cycle bases for network graphs. *Algorithmica*, 40(1) :51–62, 2004.

[5] SA Bero, AK Muda, YH Choo, NA Muda, and SF Pratama. Similarity measure for molecular structure : a brief review. In *Journal of Physics : Conference Series*, volume 892, page 012015. IOP Publishing, 2017.

[6] Philip Michael Dean. *Molecular similarity in drug design*. Springer Science & Business Media, 2012.

[7] Steven L Dixon and Kenneth M Merz. One-dimensional molecular representations and similarity calculations : methodology and validation. *Journal of Medicinal Chemistry*, 44(23) :3795–3809, 2001.

[8] Geoffrey M Downs, Peter Willett, and William Fisanick. Similarity searching and clustering of chemical-structure databases using molecular property data. *Journal of Chemical Information and Computer Sciences*, 34(5) :1094–1102, 1994.

[9] Stephen G. Hartke and Andrew Radcliffe. Mckays canonical graph labeling algorithm. *Contemporary Mathematics book series*, 479, 02 2013.

[10] Alexander Golynski and Joseph D Horton. A polynomial time algorithm to find the minimum cycle basis of a regular matroid. In *Scandinavian Workshop on Algorithm Theory*, pages 200–209. Springer, 2002.

[11] Joseph Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.*, 16 :358–366, 04 1987.

[12] Chang Joon Lee, Young-Mook Kang, Kwang-Hwi Cho, and Kyoung Tai No. A robust method for searching the smallest set of smallest rings with a path-included distance matrix. *Proceedings of the National Academy of Sciences of the United States of America*, 106 :17355–8, 09 2009.

[13] Telikepalli Kavitha, Kurt Mehlhorn, Dimitrios Michail, and Katarzyna Paluch. A faster algorithm for minimum cycle basis of graphs. In *International Collo-*

quium on Automata, Languages, and Programming, pages 846–857. Springer, 2004.

[14] Brendan D. McKay and Adolfo Piperno. Practical graph isomorphism, {II}. *Journal of Symbolic Computation*, 60(0) :94 – 112, 2014.

[15] Cheng Qian, William Fisanick, Dale E. Hartzler, and Steven W. Chapman. Enhanced algorithm for finding the smallest set of smallest rings. *Journal of Chemical Information and Computer Sciences*, 30 :105–110, 05 1990.

[16] Woong-Hee Shin, Xiaolei Zhu, Mark Bures, and Daisuke Kihara. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules*, 20(7) :12841–12862, 2015.

[17] Philippe Vismara. Union of all the minimum cycle bases of a graph. *Electr. J. Comb.*, 4, 01 1997.

[18] John W. Raymond, Eleanor Gardiner, and Peter Willett. Rascal : Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.*, 45 :631–644, 04 2002.

[19] Antonio Zamora. An algorithm for finding the smallest set of smallest rings. *Journal of Chemical Information and Computer Sciences*, 16 :40–43, 02 1976.