Algorithmes de graphes pour l'analyse des conformations de dynamique moléculaire

Ylene Aboulfath

université PARIS-SACLAY

Algorithmes de graphes pour l'analyse de conformations de dynamique moléculaire

Graphs algorithms for molecular dynamic conformers analysis

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique
Référent : Université de Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans l'unité de recherche **DAVID** (Université Paris-Saclay, UVSQ), sous la direction de **Thierry Mautor**, Maître de Conférences, et le co-encadrement de **Marc-Antoine Weisser**, Maître de Conférences.

Thèse soutenue à Versailles, le 23 octobre 2024, par

Ylène ABOULFATH

Composition du jury

Membres du jury avec voix délibérative

Alain Denise Professeur, LISN, Université Paris-Saclay	Président
Guillaume Fertin Professeur, LS2N, Université de Nantes	Rapporteur & Examinateur
Yann Ponty Directeur de Recherche CNRS, LIX, École Polythechnique de Paris	Rapporteur & Examinateur
Anne-Claude Camproux Professeure, Université Paris 7	Examinatrice
Marie-Pierre Gaigeot Professeure, LAMBE, Université d'Evry Val d'Essonne	Examinatrice

THESE DE DOCTORAT

NNT : 2024UPASG063

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Algorithmes de Graphes pour l'analyse de conformations de dynamique moléculaire
Mots clés : Graphes, Algorithmes, Bases de cycles minimum, Dynamique moléculaire

Résumé : La thèse présente une nouvelle méthode pour l'analyse et la comparaison de trajectoires de dynamique moléculaire basée sur l'algorithmique de graphes. Nous considérons une trajectoire comme une suite de graphes représentant l'évolution des liaisons chimiques entre les atomes d'une molécule en mouvement.

Traditionnellement, l'analyse de dynamique moléculaire repose sur l'énergie potentielle, mais nous avons choisi de nous en abstraire et de proposer une méthode basée sur la topologie des graphes, en particulier celle des cycles. Au cours de l'évolution de la molécule, les cycles et leurs interactions représentent la structure de la molécule. Certains cycles différents peuvent cependant avoir un rôle similaire dans la structure, nous les qualifions alors de polymorphes. Partant de ces cycles polymorphes, nous définissons le polygraphe, un graphe représentatif de la dynamique d'une trajectoire, dans lequel les sommets sont des ensembles de cycles polymorphes.

La thèse présente à la fois la méthodologie permettant le calcul et l'utilisation de ce polygraphe, ainsi que l'étude de la complexité des problèmes sous-jacents à sa construction.

Dans le même temps, nous proposons plusieurs algorithmes pour répondre aux problèmes posés et obtenir ainsi un polygraphe. Par la suite, nos algorithmes sont évalués afin de définir un protocole de construction du polygraphe à partir d'une suite de graphes constituant une trajectoire.

Enfin, nous présentons nos résultats sur l'utilisation du polygraphe pour l'analyse des trajectoires. Le polygraphe permet d'avoir une vue globale, tandis que des sous-polygraphes permettent de représenter la structure en cycles polymorphes de chacun des graphes de la trajectoire. De cette façon, si deux graphes d'une trajectoire ont le même sous-polygraphe alors, leurs structures sont équivalentes et nous pouvons conclure qu'il n'y a pas de différences structurelles majeures entre ces graphes. Nous répétons cela sur tous les graphes de la trajectoire afin de définir des ensembles de graphes qui correspondent à la même structure en cycles polymorphes.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Graphs algorithms for molecular dynamic conformers analysis
Keywords : Graphs, algorithms, minimum cycle basis, molecular dynamic

Abstract : The thesis presents a new method for the analysis and comparison of molecular dynamics trajectories based on graph algorithms. We consider a trajectory as a sequence of graphs representing the evolution of chemical bonds between the atoms of a moving molecule.

Traditionally, molecular dynamics analysis relies on potential energy, but we chose to abstract from this and propose a method based on graph topology, particularly that of cycles. During the molecule's evolution, the cycles and their interactions represent the structure of the molecule. However, some different cycles can play a similar role in the structure, and we then qualify them as polymorphic. From these polymorphic cycles, we define the polygraph, a representative graph of the trajectory's dynamics, where the vertices are sets of polymorphic cycles.

The thesis presents both the methodology for calculating and using this polygraph, as well as the study of the complexity of the underlying problems in its construction.

At the same time, we propose several algorithms to address the posed problems and thus obtain a polygraph. Subsequently, our algorithms are evaluated to define a protocol for constructing the polygraph from a sequence of graphs constituting a trajectory.

Finally, we present our results on using the polygraph for trajectory analysis. The polygraph provides an overall view, while sub-polygraphs represent the polymorphic cycle structure of each graph in the trajectory. In this way, if two graphs in a trajectory have the same sub-polygraph, then their structures are equivalent, and we can conclude that there are no major structural differences between these graphs. We repeat this for all graphs in the trajectory to define sets of graphs that correspond to the same polymorphic cycle structure.

4

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à la réalisation de cette thèse.

Tout d'abord, je remercie mes encadrants, Thierry Mautor et Marc-Antoine Weisser, pour leurs conseils tout au long de ce travail. Je tiens également à remercier Dominique Barth et Dimitri Watel pour leurs précieuses contributions qui ont enrichi cette thèse.

Mes remerciements vont également à mes collaborateurs Marie-Pierre Gaigeot, Sana Bougueroua et Alvaro Cimas, membres du laboratoire de physique théorique LAMBE et initiateurs du sujet. Leur expertise dans la validation des résultats a été d'une grande valeur pour la réussite de ce travail.

Je souhaite également exprimer toute ma reconnaissance aux membres du laboratoire DAVID pour leur accueil chaleureux et leur soutien constant durant ces quatre années.

Enfin, je tiens à remercier mes amis, ma famille et mon compagnon, qui n'ont cessé de me soutenir tout au long de ce parcours. Je souhaite également avoir une pensée particulière pour mon père, qui a été le premier à m'encourager sur cette voie, mais qui n'aura pas pu voir ce projet aboutir.

À vous tous, merci. C'est grâce à vos encouragements et à votre soutien que j'ai pu persévérer et mener ce travail à son terme.

Table des matières

Introduction **9**

1 Graphes pour l'analyse de trajectoires **13**
1.1 Système moléculaire . . . 13
1.2 Dynamique d'un système moléculaire . . . 15
1.2.1 Évolution d'un système moléculaire au cours du temps . . . 15
1.2.2 Exploration de la surface d'énergie potentielle . . . 17
1.3 Graphes pour la caractérisation des structures moléculaires . . . 20
1.3.1 Transition d'une trajectoire de dynamique moléculaire à une suite de graphes moléculaires . . . 20
1.3.2 Caractérisation d'une trajectoire par la catégorisation de ses graphes moléculaires 22
1.3.3 Caractérisation d'une trajectoire à partir de motifs topologiques issus de ses graphes moléculaires . . . 27
1.4 Approche proposée pour l'analyse de trajectoires . . . 29

2 Caractérisation d'une trajectoire de dynamique moléculaire **33**
2.1 Trajectoire de dynamique moléculaire . . . 33
2.2 Modélisation de la structure d'une conformation . . . 35
2.2.1 Base de cycles minimum . . . 35
2.2.2 Caractérisation d'une conformation par ses cycles . . . 39
2.3 Caractérisation de la structure d'une dynamique moléculaire . . . 42
2.3.1 Polymorphisme de cycles . . . 43
2.3.2 Polygraphe d'une trajectoire . . . 46
2.4 Analyse de trajectoires de dynamique moléculaire . . . 49
2.5 Synthèse de la démarche suivie . . . 51

3 Sélection des cycles des conformations **53**
3.1 Calcul d'une base de cycles minimum . . . 54
3.2 Maximiser l'intersection de bases de cycles minimum . . . 56
3.2.1 Proximité entre intersection de bases de cycles et intersection de matroïdes . . 57
3.2.2 Complexité de MCBI dans le cas général . . . 65
3.2.3 Une méthode gloutonne pour le calcul des bases inspirée par max-MCBI . . . 74
3.3 Vers des méthodes de voisinages pour la sélection des bases . . . 76
3.3.1 Modéliser la répartition des cycles dans un ensemble de bases de cycles minimum 76
3.3.2 Mise en oeuvre . . . 79
3.4 Synthèse sur les méthodes de sélection des cycles . . . 86

4 Calcul du polygraphe d'une trajectoire **89**
4.1 Complexité du problème PCP . . . 90
4.1.1 NP-completude pour les trajectoires de dynamique moléculaire . . . 90
4.1.2 Inapproximabilité du problème min-PCP . . . 95
4.1.3 Paramètres de la complexité du problème PCP . . . 97
4.2 Méthodes proposées pour résoudre min-PCP . . . 98
4.2.1 Modélisation d'une partition en polycycles par un polygraphe . . . 98
4.2.2 Méthode exacte pour résoudre min-PCP . . . 105
4.2.3 Méthode heuristique pour tendre vers une solution à min-PCP . . . 107
4.3 Synthèse sur le calcul du polygraphe . . . 109

5 Comparaison des méthodes de calcul du polygraphe **111**
5.1 Calcul de la partition en polycycles . . . 112
5.1.1 Données de test : trajectoires simulées . . . 112
5.1.2 Évaluation des méthodes de calculs de la partition en polycycles . . . 114
5.2 Calcul des bases de cycles minimum . . . 119
5.2.1 Données de test : des trajectoires générées . . . 120
5.2.2 Comparaison des méthodes de sélection des bases de cycles minimum . . . 121
5.3 Procédure de calcul du polygraphe d'une trajectoire . . . 135

6 Analyse et comparaison de trajectoires par le polygraphe **137**
6.1 Données expérimentales . . . 137
6.2 Analyse de trajectoires de dynamique moléculaire . . . 138
6.2.1 Trajectoire semi-empirique du peptide ECCA* . . . 139
6.2.2 Trajectoire empirique du peptide Z-Ala$_6$ . . . 141
6.2.3 Trajectoire semi-empirique du polysaccharide Chondroitine disulfate . . . 145
6.3 Comparaison de trajectoires de dynamique moléculaire . . . 150
6.3.1 Correspondance entre polygraphes . . . 150
6.3.2 Analyse des trajectoires de Z-Ala$_6$ . . . 152
6.3.3 Analyse des trajectoires de Gramicidine . . . 160
6.4 Synthèse sur l'analyse et la comparaison de trajectoires . . . 162

Conclusion **163**

Bibliographie **165**

Introduction

L'analyse des trajectoires de dynamique moléculaire est essentielle pour comprendre les processus complexes au sein des molécules en mouvement. Les trajectoires obtenues par des simulations de dynamique moléculaire fournissent des données riches sur l'évolution des molécules et de leurs constituants au fil du temps. Ces informations sont cruciales pour révéler les mécanismes fondamentaux des réactions chimiques, les interactions entre molécules et leurs changements de conformation structurelle.

Traditionnellement, cette analyse repose sur l'énergie potentielle, une métrique qui quantifie l'énergie interne d'une molécule en fonction de la disposition de ses atomes dans l'espace. L'énergie potentielle permet d'identifier les états les plus stables de la molécule et les transitions entre ces états. Cependant, bien que cette méthode soit efficace pour déterminer les états stables et les chemins de transition, elle présente certaines limitations. Les variations topologiques qui ne se traduisent pas nécessairement par des changements significatifs d'énergie ne sont pas détectées. De plus, le calcul de cette métrique, nécessitant la résolution d'équations complexes, est coûteux. Enfin, l'étude de systèmes biomoléculaires complexes a montré que les motifs structuraux jouent un rôle crucial dans la fonction biologique, comme en témoigne l'exemple de l'ARN. Il est donc plausible que des interactions spécifiques et des conformations transitoires aient des implications fonctionnelles significatives.

C'est dans ce contexte que s'inscrit notre travail, proposant une méthode novatrice basée sur l'algorithmique de graphes et plus particulièrement sur la topologie des cycles pour l'analyse des trajectoires de dynamique moléculaire. En nous affranchissant de la métrique de l'énergie potentielle, nous pouvons explorer la dynamique moléculaire sous un angle nouveau, en mettant l'accent sur les structures topologiques et leur évolution. Cette approche nous permet de détecter et d'analyser des motifs récurrents et des structures polymorphes, offrant une perspective complémentaire à l'analyse traditionnelle.

Notre méthode repose sur la modélisation des trajectoires de dynamique moléculaire sous forme de suites de graphes. Dans cette représentation, chaque graphe correspond à un instantané de la structure moléculaire, où les nœuds représentent les atomes et les arêtes représentent les liaisons chimiques. En particulier, nous distinguons les liaisons covalentes, qui forment le squelette immuable de la molécule, des liaisons hydrogène, qui peuvent apparaître et disparaître au cours du temps. Cette distinction permet de capturer la dynamique des interactions faibles, souvent au cœur des processus biologiques et chimiques.

Nous proposons alors une méthode de caractérisation structurelle des graphes par les cycles, indépendamment de leur place dans la trajectoire. Néanmoins, dans un graphe les cycles peuvent être trop nombreux pour être tous considérés. Nous nous orientons donc vers la base de cycles minimum, un ensemble de cycles générateur de poids minimum qui permet le calcul de tous les cycles du graphe. En centrant notre analyse sur les cycles, nous identifions les contraintes structurales et les motifs répétitifs qui constituent la dynamique moléculaire. Les cycles correspondent ainsi à des motifs structuraux qui peuvent être fondamentaux pour la stabilité et la fonction de la molécule. En particulier,

nous observons que certains cycles peuvent jouer des rôles similaires malgré des différences dans leur composition et sont donc qualifiés de polymorphes. Pour les identifier il nous a fallu définir un ensemble de propriétés caractéristiques. Ces cycles polymorphes, qui évoluent en réponse aux changements dans la structure globale, sont les éléments de base de notre polygraphe, un graphe représentatif de la dynamique d'une trajectoire.

À partir du polygraphe, nous caractérisons finalement la structure des différents graphes de la trajectoire. Une fois cette caractérisation effectuée, les graphes sont replacés dans le temps de la trajectoire pour analyser la dynamique de celle-ci. L'utilisation des cycles polymorphes pour établir la structure des graphes est suffisamment flexible pour permettre la comparaison de différentes trajectoires représentant une même molécule. Cela permet de comparer des polygraphes et donc de comparer les structures observées dans différentes trajectoires de dynamique moléculaire. Ainsi, nous pouvons mieux comprendre l'impact des paramètres d'une dynamique moléculaire sur les structures parcourues.

Cette approche complémentaire enrichit les méthodes traditionnelles et offre de nouvelles opportunités pour l'analyse et la comparaison des trajectoires moléculaires, avec des applications potentielles dans divers domaines de la chimie et de la biologie.

Le manuscrit est organisé comme suit.

Le Chapitre 1 présente les résultats de la littérature concernant l'analyse des trajectoires de dynamique moléculaire. Nous y examinons la place de la théorie des graphes et des cycles dans ce domaine. Notre approche, qui utilise la topologie des cycles, se distingue par son caractère novateur, car peu de travaux se concentrent sur cette perspective spécifique. Cependant, les cycles sont des outils bien connus dans la littérature pour l'étude des structures moléculaires, et notre méthode s'inscrit dans cette suite tout en ouvrant de nouvelles possibilités pour l'analyse des dynamiques moléculaires.

Le Chapitre 2 propose une présentation globale de la méthode avant d'entrer dans les détails des problèmes sous-jacents aux différentes étapes de la méthode. Nous y introduisons les étapes nécessaires à la construction du polygraphe à partir d'une trajectoire ainsi que les notions de théorie des graphes nécessaires.

Les chapitres suivants sont plus techniques, car ils présentent non seulement nos algorithmes pour résoudre les problèmes auxquels nous nous sommes heurtés, mais ils traitent également de leur complexité.

Le premier de ces problèmes est celui de la sélection des cycles et est étudié dans le Chapitre 3. En effet, notre approche utilise une base de cycles minimum pour choisir les cycles représentatifs de la structure. Or, un graphe possède plusieurs bases de ce type. Nous devons donc en sélectionner une parmi toutes celles disponibles, en gardant à l'esprit que l'objectif est d'identifier par la suite des cycles polymorphes. Dans ce chapitre, nous présentons plusieurs méthodes pour faire cette sélection. L'une d'entre elles cherche à sélectionner des bases de cycles minimum qui ont la plus grande intersection possible. Nous montrons la complexité de ce problème ainsi que sa proximité avec le problème de l'intersection des matroïdes. Une seconde approche consiste en l'exploration d'un voisinage. Nous parcourons alors de nombreux ensembles de bases de cycles minimum à la recherche de celui qui sera le plus favorable à la résolution du second problème de cette thèse.

Ce second problème consiste à identifier le plus petit nombre de cycles polymorphes

différents lorsque l'on considère l'ensemble des cycles représentatifs de la trajectoire. Ce problème de partitionnement est l'objet du Chapitre 4. Un cycle polymorphe est un ensemble de cycles tous polymorphes deux à deux. Nous étudions alors le problème du partitionnement de l'ensemble des cycles de la trajectoire en un nombre minimal de cycles polymorphes. Là encore, nous présentons plusieurs approches, incluant une méthode exacte ainsi qu'une heuristique.

Tous nos algorithmes sont testés et évalués dans le Chapitre 5. Nos critères d'évaluation incluent non seulement l'identification d'un nombre minimal de cycles polymorphes mais aussi la rapidité d'exécution. Pour conclure ce chapitre, nous établissons une méthodologie pour le calcul du polygraphe à partir d'une suite de graphes constituant une trajectoire.

Enfin, dans le Chapitre 6, nous testons le polygraphe sur un ensemble de trajectoires obtenues par différentes méthodes de simulation. Nous présentons les outils annexes que nous avons mis en place pour l'analyse et la comparaison de ces trajectoires de dynamique moléculaire. Nous introduisons le chronogramme pour représenter les périodes d'apparitions des différents cycles polymorphes au cours de la trajectoire. La coexistence de plusieurs de ces cycles polymorphes approxime une structure de la molécule. Ainsi, nous pouvons estimer les différentes structures parcourues au cours d'une trajectoire de dynamique moléculaire. De plus, nous proposons une méthode de comparaison des polygraphes permettant ainsi de comparer les différentes structures observées dans chacune des trajectoires. Nous pouvons alors étudier les conséquences d'une modification d'un paramètre environemental sur les structures explorées par une trajectoire.

1 - Graphes pour l'analyse de trajectoires de dynamique moléculaire

Nous explorons dans cette thèse diverses méthodes d'analyse de trajectoires de dynamique moléculaire. Ces trajectoires retracent l'évolution d'une molécule à travers le temps dans un environnement défini (température, pression, ...). Dans notre étude, nous mettons particulièrement l'accent sur l'utilisation des graphes comme outil efficace dans ce domaine.

Nous débutons en considérant un modèle classique de molécule tel que décrit dans [29]. Ensuite, nous explorons les origines de la dynamique moléculaire [37, 39], en mettant en lumière les variations d'énergie qui induisent cette dynamique. En effet, une dynamique moléculaire est directement liée à l'idée de mouvement et donc à l'énergie qui permet ce mouvement. Les méthodes traditionnelles de chémo-informatique pour mener ce type d'analyse reposent sur des fonctions d'énergie et des techniques d'optimisation. Toutefois, notre objectif dans cette thèse est de sortir du cadre de ces méthodes conventionnelles pour proposer une alternative innovante basée sur la théorie des graphes.

À cet effet, nous présentons des approches de graphes qui ont prouvé leur efficacité dans des contextes moléculaires similaires, telles que la comparaison de molécules et la recherche de motifs dans les ARNs. Cela nous conduira à examiner comment la théorie des graphes peut être appliquée à l'analyse de trajectoires de dynamique moléculaire, à la fois de manière isolée et dans l'optique d'identifier des éléments topologiques caractéristiques pour les molécules.

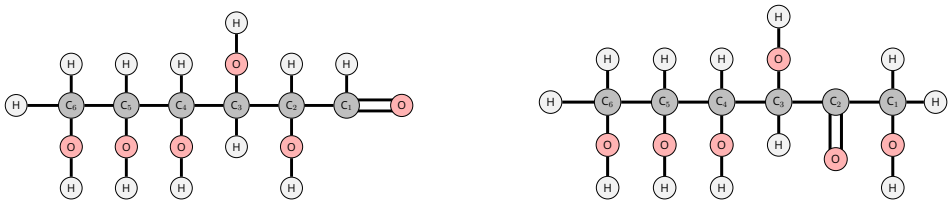
Les sections suivantes détaillent notre approche, mettant en évidence les principaux éléments qui constituent le fondement de notre contribution.

1.1 . Système moléculaire

Une **molécule**, ou système moléculaire, est un agrégat d'atomes maintenus ensemble par des forces liant les atomes, appelées liaisons chimiques.

Les liaisons entre les atomes d'une molécule déterminent sa forme générale, et la diversité de ces liaisons peut créer des variantes appelées isomères. Un isomère partage les mêmes atomes qu'une autre molécule, mais ses liaisons diffèrent, conduisant à une structure chimique distincte et à des propriétés spécifiques.

Prenons l'exemple du glucose et du fructose (voir Figure 1.1). Ces deux molécules sont des isomères, composées chacune de six atomes de carbone, douze d'hydrogène et six d'oxygène. Les différences chimiques entre ces molécules se manifestent par exemple, par des variations de l'indice glycémique, et par une métabolisation par des enzymes spécifiques (fructokinase pour le fructose et glucokinase pour le glucose). La Figure 1.1 illustre clairement que la distinction entre ces molécules réside dans les liaisons entre les atomes de carbone, communément appelées *liaisons covalentes*. Nous pouvons observer que dans le cas du glucose, le carbone C_1 est lié par deux liaisons à un oxygène. En revanche, dans le cas du fructose, cet atome est lié à un atome d'oxygène par une seule liaison et à un atome d'hydrogène supplémentaire. Notons que c'est à présent le carbone C_2 qui est lié à un oxygène par deux liaisons. Étant donné que les liaisons et l'agencement



(a) Molécule de Glucose (b) Molécule de Fructose

Figure 1.1 – Illustration de deux isomères dont la formule brute est $C_6H_{12}O_6$.

des atomes dans cette portion de la molécule diffèrent, la structure moléculaire diffère également.

Définition 1. *Une **liaison covalente** représente une interaction chimique puissante entre deux atomes, résultant du partage de une ou plusieurs paires d'électrons entre eux. Concrètement, les atomes impliqués dans la liaison contribuent à la formation de doublets d'électrons liants. Ce partage d'électrons crée une attraction mutuelle forte entre les deux atomes, les maintenant proches au sein de la molécule formée.*

Le nombre de liaisons covalentes auxquelles un atome peut participer dépend du nombre d'électrons libres dont il dispose. Ce nombre varie en fonction du type chimique de l'atome, déterminant ainsi le nombre maximal de liaisons covalentes qu'il peut former. À titre d'exemple, l'atome de carbone a la capacité de créer quatre liaisons covalentes, tandis que l'azote en forme trois, l'oxygène deux, et l'hydrogène une seule.

Les liaisons covalentes sont considérées comme *fortes* en raison de leur résistance à la rupture. La formation de ces liaisons implique une attraction mutuelle entre les atomes, déterminée par leur compatibilité électronique et leur proximité physique. La distance typique entre les atomes dans une liaison covalente varie de $0,1$ à $0,2$ nm[1]. Ces liaisons jouent un rôle essentiel dans la définition de la structure moléculaire car elles représentent des liens forts entre les atomes.

La structure d'une molécule est définie par la position spatiale de ses atomes et donc également par leur interaction. L'agencement des atomes dans l'espace et leur compatibilité électronique induit la formation de liaisons chimiques. Cette structure spatiale relative avec la position des atomes et les liens qui les unissent détermine les propriétés chimiques de la molécule.

Ainsi, les liaisons covalentes, bien que cruciales pour la structure d'une molécule, ne sont pas les seules forces d'intérêt. D'autres liaisons chimiques, dites secondaires, jouent également un rôle essentiel. Ces liaisons électroniques sont caractérisées par une intensité plus faible par rapport aux liaisons covalentes. Parmi ces liaisons secondaires, on retrouve les liaisons hydrogène, ionique ou halogène.

La température et la pression environnementales fournissent de l'énergie à la molécule, induisant des mouvements au sein des atomes dans l'espace, connus sous le nom d'agitation atomique. Bien que les liaisons covalentes restent inchangées, l'agitation atomique peut influencer les liaisons secondaires. Effectivement, ces liaisons secondaires ne produisent pas une attraction assez forte pour empêcher les atomes impliqués de s'éloigner. Par conséquent, elles peuvent se rompre si les atomes s'éloignent ne serait-ce que

1. 1 nanomètre (nm) équivaut à 10^{-9} mètres

de quelques nanomètres. De manière similaire, lorsque les atomes se rapprochent, de nouvelles liaisons secondaires peuvent se former, ce qui contribue à la dynamique globale de la molécule.

Nous observons la coexistence de plusieurs structures pour une molécule donnée en fonction des liaisons présentes. Chaque structure est supposée correspondre à un niveau d'énergie distinct. Le niveau d'énergie joue un rôle majeur dans la stabilité moléculaire et constitue un aspect fondamental de l'analyse de la dynamique moléculaire. On parle alors d'énergie potentielle, qui peut être vue comme une somme de l'énergie propre à chaque atome, elle-même dépendante de la position relative.

Les différentes structures d'une molécule ne partagent pas toutes le même niveau de stabilité. La stabilité globale d'une structure se manifeste classiquement par un niveau d'énergie bas, ce qui implique qu'une quantité significative d'énergie est nécessaire pour rompre cette structure et accéder à une autre.

Ainsi, les notions de stabilité et d'énergie potentielle se positionnent au cœur de toute analyse de dynamique moléculaire.

1.2 . Dynamique d'un système moléculaire

1.2.1 . Évolution d'un système moléculaire au cours du temps

Une **trajectoire de dynamique moléculaire**, ou plus simplement trajectoire, représente l'évolution d'un système moléculaire au fil du temps dans un environnement défini. Cet environnement est régi par plusieurs paramètres tels que la température, la pression et les forces d'interaction entre chaque atome. Les trajectoires sur lesquelles nous avons travaillées pendant cette thèse sont toutes issues de simulations numériques. Ainsi, la méthode utilisée pour définir les forces d'interaction fait partie des paramètres de la simulation. On distingue principalement la méthode empirique, où les forces sont dérivées d'un potentiel fixé empiriquement, et la méthode semi-empirique, ou *ab initio*, dans laquelle les forces sont calculées à partir des premiers principes de la mécanique quantique.

Comme nous l'avons déjà evoqué, les atomes se déplacent en fonction de l'énergie du système induite par la température et la pression et n'ont donc pas une position fixe. Leurs mouvements sont régis par les forces d'interaction considérées, déterminant ainsi la direction de leurs déplacements et induisant des changements dans la forme tridimensionnelle de la molécule, tels que la formation ou la rupture de liaisons secondaires.

Dans cette thèse, notre attention se porte sur l'évolution de la structure d'une molécule au cours d'une dynamique moléculaire. Les dynamiques qui nous intéressent impliquent des peptides ou de petites protéines en phase gazeuse, où seules les liaisons hydrogène évoluent au fil du temps.

Définition 2. *Une* **liaison hydrogène** *est une interaction spécifique entre un atome d'hydrogène et un atome électronégatif. L'atome d'hydrogène, lié à un atome donneur, interagit avec l'atome électronégatif, appelé accepteur, qui possède un doublet non liant, c'est-à-dire une paire d'électrons non impliqués dans une liaison covalente.*

La liaison hydrogène est une liaison secondaire qui ne peut impliquer que des héteroatomes, c'est-à-dire des atomes avec des propriétés électroniques particulières. Parmi

les hétéroatomes couramment impliqués dans les liaisons hydrogène, on retrouve : l'oxygène, l'azote ou le phosphore. En revanche, le carbone et l'hydrogène ne sont pas des hétéroatomes.

Une liaison hydrogène se forme lorsque les deux hétéroatomes, c'est-à-dire l'atome donneur et l'atome électronégatif, impliqués sont à une distance d'environ $0,25$ nm. La Figure 1.2 illustre une liaison hydrogène entre les atomes d'oxygène de deux molécules d'eau (H_2O). Dans cet exemple, l'atome O_1 joue le rôle de l'accepteur, tandis que l'atome O_2 tient le rôle du donneur. Nous observons également l'hydrogène intermédiaire entre les deux atomes d'oxygène impliqués dans la liaison.

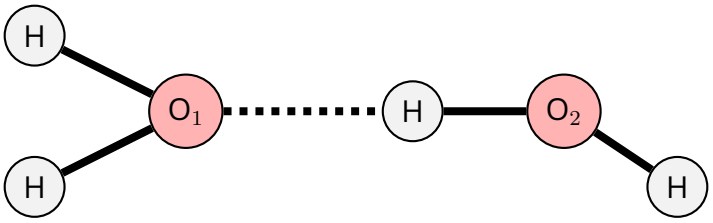


Figure 1.2 – Exemple d'une liaison hydrogène entre deux atomes d'oxygène.

Au cours d'une dynamique moléculaire, l'agitation atomique impacte les liaisons secondaires qui existent entre les atomes. Bien que ces liaisons ne soient pas directement observables, nous considérons qu'elles existent lorsque des atomes compatibles sont suffisamment proches. Ainsi, en se basant sur la position dans l'espace des atomes et sur leurs types chimiques respectifs, il est possible de déterminer les liaisons présentes.

Les trajectoires de dynamique moléculaire sont des simulations des mouvements des atomes dans l'espace au cours du temps. Ainsi, le résultat d'une trajectoire se présente sous la forme d'une séquence de positions prises par chaque atome pendant cette période. Des travaux menés par [10, 11] présentent une conversion de ce modèle tridimensionnel vers un modèle bidimensionnel, celui des conformations.

Une conformation correspond à un graphe dans lequel un ensemble de liaisons coexistent. Une conformation caractérise une structure moléculaire sans considération de la position des atomes dans l'espace. Rappelons que les liaisons covalentes ne sont pas impactées dans les dynamiques considérées ici. Par conséquent, nous avons l'assurance qu'une même molécule change de conformation en fonction des liaisons hydrogène présentes. Autrement dit, deux conformations d'une même trajectoire ont les mêmes atomes, les mêmes liaisons covalentes, mais un ensemble différent de liaisons hydrogène. Ces travaux nous permettent donc de considérer une suite de conformations plutôt que la séquence initiale des positions.

La Figure 1.3 illustre deux conformations d'une même molécule, Z-Ala_6. La formule brute de Z-Ala_6 est $C_{26}H_{39}N_7O_8$, mais les atomes d'hydrogène ne sont pas représentés dans les conformations de la Figure 1.3. Chaque atome d'hydrogène participe à une seule liaison covalente, et afin de ne pas surcharger la représentation, ils sont exclus de l'illustration. Ainsi, les liaisons hydrogène sont directement représentées entre l'atome donneur et l'atome accepteur de la liaison, sans inclure l'hydrogène intermédiaire.

Le niveau d'énergie d'un système représente sa capacité à se transformer. Cette énergie, appelée énergie potentielle, constitue la base de l'évaluation et de la comparaison des

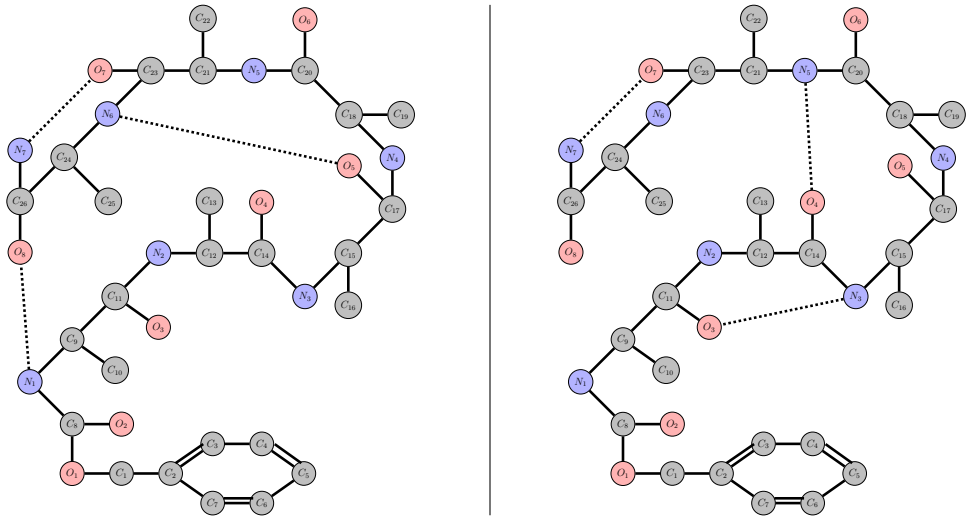


Figure 1.3 – Deux conformations du peptide Z-Ala$_6$. La couleur des sommets correspond au type chimique de l'atome : rouge pour oxygène, bleu pour azote et noir pour carbone. Les liaisons covalentes sont représentées par des traits pleins, et les liaisons hydrogène par des traits en pointillés.

conformations d'une molécule. Plus l'énergie potentielle est basse, plus il est difficile pour la molécule de changer de conformation, faisant de cette métrique un indicateur largement utilisé. Néanmoins, elle est très coûteuse à calculer car elle necessite la prise en compte de la position dans l'espace de chaque atome et l'évaluation des forces d'interaction entre eux.

1.2.2 . Exploration de la surface d'énergie potentielle

La surface d'énergie potentielle représente l'espace de toutes les structures possibles d'un système moléculaire et peut être visualisée comme une surface vallonnée en plusieurs dimensions. L'exploration structurelle consiste à énumérer les structures accessibles par une molécule dans un environnement défini. Il s'agit donc d'une exploration partielle de la surface d'énergie potentielle.

Une simulation de dynamique moléculaire constitue une exploration structurelle partielle et déterministe à partir d'une structure tri-dimensionnelle initiale, suivant un environnement fixé, produisant ainsi une trajectoire. Cette trajectoire représente un chemin sur la surface d'énergie potentielle et est souvent visualisée, comme sur la Figure 1.4, par une coupe 2D de la surface d'énergie potentielle. Cette représentation facilite la reconnaissance des différentes structures prises par la molécule au cours de la trajectoire. Pour conduire une exploration structurelle complète, les trajectoires pour une même molécule sont multipliées afin de cartographier l'ensemble de la surface d'énergie.

Lors de l'analyse d'une trajectoire de dynamique moléculaire, les structures explorées peuvent être catégorisées en deux groupes distincts : les structures stables et les structures instables.

- Les **structures stables** sont situées au fond de ce que l'on appelle *puits d'énergie* ou *bassins conformationnels*. Dans l'exemple de la Figure 1.4, on observe que la coupe présente trois bassins distincts. Il est facile de visualiser qu'une quantité d'énergie importante serait nécessaire pour passer du bassin B1 au bassin B2. Ainsi, quitter

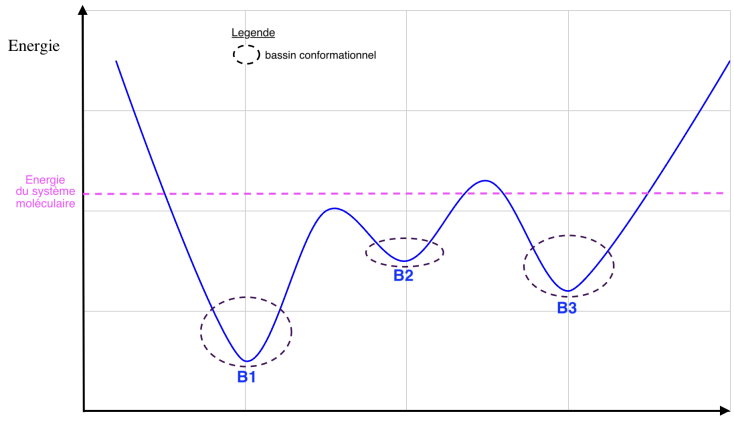


Figure 1.4 – Schéma d'une coupe de la surface d'énergie potentielle.

une structure stable donc sortir d'un bassin, demande une quantité significative d'énergie. La *barrière d'énergie* représente l'énergie nécessaire pour franchir cette transition d'un bassin à un autre. Malgré tout, au sein d'un bassin, quelques mouvements atomiques peuvent induire de légères variations, générant ainsi des structures différentes mais très proches les unes des autres.

- Les **structures instables**, en revanche, ne sont pas au fond des puits d'énergies mais plutôt à leurs jonctions. Ce sont des structures de transitions, également appelées états de transition, qui peuvent *tomber* dans plusieurs puits d'énergie. Elles représentent des structures très flexibles capables d'admettre facilement la création de différentes liaisons hydrogène. Dans l'exemple de la Figure 1.4, ces structures se situeraient au sommet des petites collines séparant les bassins.

La ligne horizontale sur la Figure 1.4 indique la limite fixée lors d'une simulation pour l'énergie du système, imposée par les paramètres de la simulation. Si cette limite est trop basse, certaines zones de la surface d'énergie potentielle deviennent inaccessibles. Il est important de noter que, même si une partie de la surface d'énergie est inaccessible en raison de la limitation d'énergie, les bassins conformationnels peuvent rester atteignables, car la surface est en réalité bien plus complexe qu'une coupe en deux dimensions. Cela est important car l'identification des structures stables, situées au fond des bassins conformationnels, est la principale problématique de l'analyse de la dynamique moléculaire.

Les paramètres de la simulation ont un impact majeur sur les résultats de l'analyse de la trajectoire. Si l'énergie du système est trop basse, cela peut limiter l'exploration des bassins conformationnels, tandis qu'une énergie trop élevée peut entraîner une agitation des atomes telle que la formation de liaisons devient difficile.

L'exploration de la surface d'énergie potentielle a pour objectif de découvrir des structures diverses. À partir de ces structures, les méthodes de minimisation d'énergie permettent d'identifier les structures stables. La Figure 1.5 schématise ce processus.

Les méthodes de minimisation ont pour point de départ une ou plusieurs structures déjà identifiées. Puis, à partir de ces dernières, elles recherchent des structures appartenant au puits d'énergie (ou bassin conformationnel) le plus proche.

Sans nécessairement cartographier la surface d'énergie potentielle, notre principale problématique réside dans l'exploration de cette surface. L'objectif est de catégoriser les

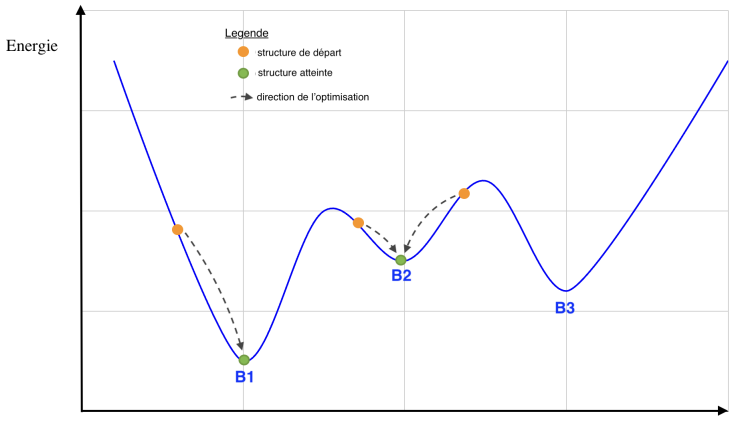


Figure 1.5 – Schéma d'une coupe de la surface d'énergie potentielle illustrant le principe de la minimisation d'énergie potentielle.

conformations représentatives des structures explorées sans avoir recours au calcul coûteux de l'énergie potentielle.

Pour atteindre cet objectif, nous introduisons le concept de **méta-structure**. Une **méta-structure** est un ensemble de conformations qui appartiennent à un même bassin conformationnels. Elle représente un état global de la molécule durant une période de la trajectoire. Ainsi, elle correspond à un agencement globalement similaire de ses atomes qui admet certaines variations impactant de manière mineure l'énergie potentielle de la molécule. Ainsi, une méta-structure moléculaire est une notion plus large que celle de conformation. Un ensemble de conformations peut correspondre à une même méta-structure et une conformation peut apparaître dans plusieurs méta-structures. Par conséquent, notre étude vise à reconnaître les méta-structures rencontrées au cours d'une trajectoire de dynamique moléculaire, proposant ainsi une alternative aux méthodes traditionnelles.

Une dynamique moléculaire peut représenter des systèmes lourds tels que des complexes moléculaires. Cependant, dans le cadre de cette thèse, nous nous concentrons uniquement sur des trajectoires qui correspondent à l'évolution d'une molécule de type peptide ou petite protéine, en phase gazeuse. Même dans ce contexte, qui pourrait sembler relativement simple, le calcul d'une trajectoire de dynamique moléculaire impliquant une seule petite molécule peut durer plusieurs jours. Il en va de même pour le coût de l'analyse d'une telle trajectoire qui est déjà très élevé.

La méthode que nous présentons et qui sera introduite dans le chapitre suivant (Chapitre 2), repose exclusivement sur la théorie des graphes. Par conséquent, elle ne nécessite pas le calcul de l'énergie potentielle. Elle utilise les graphes sous-jacents aux conformations pour les catégoriser et définir les méta-structures de la molécule. Dans certains travaux de post-analyse, nous pourrons discuter de l'utilisation de l'énergie potentielle pour l'étude de quelques structures sélectionnées en amont par la catégorisation en méta-structure, évitant ainsi les explosions de coût calculatoire. Cette thèse présente une nouvelle méthode d'analyse des trajectoires de dynamique moléculaire, basée sur l'évolution de la structure topologique de la molécule, et utilisant exclusivement la théorie des graphes.

1.3 . Théorie des graphes pour la caractérisation des structures moléculaires

Il est en effet naturel d'utiliser la théorie des graphes pour modéliser un système moléculaire. Dans ce modèle, les atomes sont représentés par les sommets du graphe, tandis que les liaisons chimiques sont représentées par les arêtes. On parle de graphes moléculaires. De nombreuses références ont souligné l'efficacité de cette approche pour la représentation des molécules [44, 45, 50]. Dans cette section, nous présentons les approches classiques basées sur la théorie des graphes qui sont éventuellement applicables aux trajectoires de dynamique moléculaire. Nous examinons ainsi les limites et les perspectives d'utilisation des graphes pour notre étude.

1.3.1 . Transition d'une trajectoire de dynamique moléculaire à une suite de graphes moléculaires

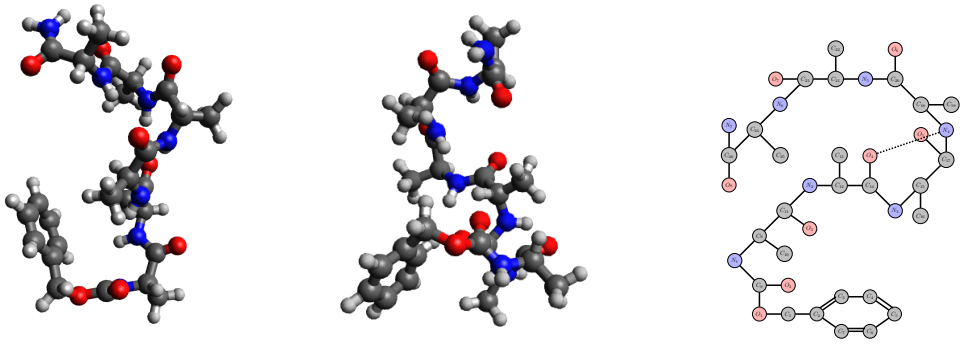
Avant d'étudier les graphes moléculaires, nous devons d'abord transformer une trajectoire, initialement représentée par une série de positions tri-dimensionnelles, en une série de graphes. En effet, à partir d'une position initiale pour chaque atome et d'un environnement donné, une trajectoire est visualisée sous la forme d'une série d'images tridimensionnelles. Chacune de ces images (ou *snapshots*) représente la position des atomes d'une molécule dans l'espace à un moment t de la simulation, induisant ainsi la présence de liaisons chimiques.

Ainsi, pour chaque image tri-dimensionnelle de la molécule, un graphe moléculaire est construit, comprenant uniquement les liaisons formées par les atomes à l'instant t de la simulation.

Dans ce graphe moléculaire, les sommets représentent les atomes étiquetés par leur type chimique (carbone, azote, oxygène, ...), et les arêtes représentent les liaisons chimiques, également étiquetées par leur type (covalente, hydrogène, ...). Ainsi, plusieurs images tridimensionnelles peuvent correspondre au même graphe moléculaire si les positions des atomes induisent les mêmes liaisons hydrogène. La Figure 1.6 illustre un exemple de deux images correspondant à un même graphe. Nous pouvons observer que les positions de nombreux atomes diffèrent considérablement d'une image à l'autre. Pourtant, les deux images correspondent au même graphe qui ne comporte qu'une seule liaison hydrogène. Ce graphe ne contient pas d'information sur la position des atomes non impliqués dans des liaisons. Le graphe moléculaire est la représentation d'une conformation de la molécule.

Les travaux préliminaires [10, 11] permettent la transformation d'une série d'images tridimensionnelles en une suite de conformations, comme illustré dans la Figure 1.7. Il est à noter que certaines conformations peuvent apparaitre à plusieurs moments de la trajectoire. Ces conformations apparaissent sur plusieurs images non consécutives et sont dites **récurrentes**. Dans l'exemple de la Figure 1.7, la conformation 2 est un exemple de conformation récurrente. À l'inverse, certaines conformations apparaissent trop peu de temps et sont dites **fugaces**. Dans [11], les conformations fugaces sont celles apparaissant au total pendant moins de 4% du temps de simulation. Cette limite existe pour ne pas risquer de considérer des artéfacts comme des structures importantes. Les conformations fugaces sont alors supprimées de la suite de conformations de la trajectoire.

Ces travaux nous permettent de considérer une suite de conformations étiquetées, c'est-à-dire une suite de graphes moléculaires.



(a) Image n°79. (b) Image n°2553. (c) Graphe correspondant.

Figure 1.6 – Exemples de deux images tridimensionnelles extraites d'une même trajectoire et du graphe moléculaire (bidimensionnel) leur correspondant. Les images des figures 1.6a et 1.6b sont extraites d'une trajectoire du peptide Z-Ala_6. Une vue des liaisons covalentes présentes entre les atomes a été ajoutée pour faciliter la comparaison entre les figures. Dans la Figure 1.6c, les liaisons covalentes sont représentées par des traits pleins, et les liaisons hydrogène par des traits en pointillés.

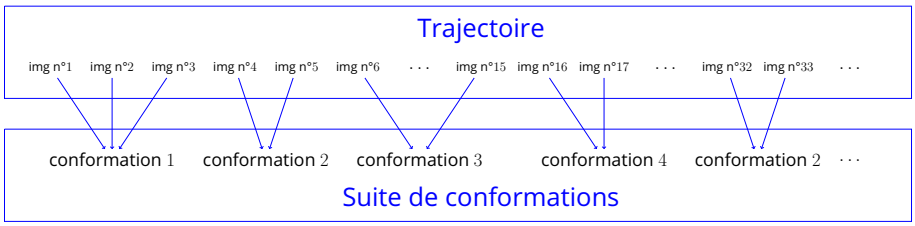


Figure 1.7 – Schéma de la transformation d'une trajectoire de dynamique moléculaire à une suite de conformations.

En étudiant les conformations dans l'ordre fourni par la trajectoire, nous constatons que les conformations successives diffèrent peu. En effet, l'intervalle de temps entre deux images est déterminé par l'utilisateur de l'outil de simulation. Ainsi, dans les trajectoires que nous considérons ici, l'intervalle entre deux images est très court (environ 0,4 fs [2]). Deux conformations consécutives ne peuvent différer que par une, voire exceptionnellement deux liaisons hydrogène.

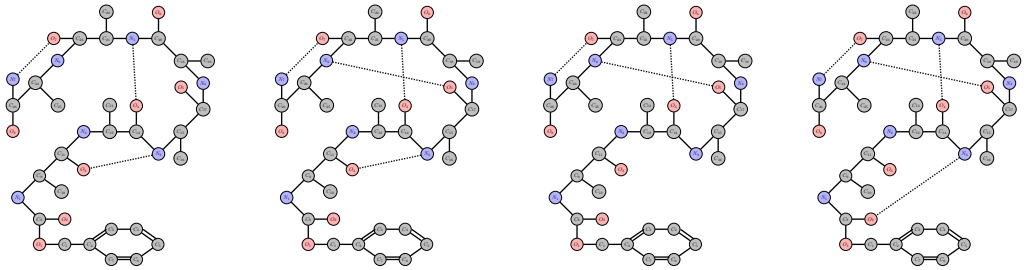
La Figure 1.8 illustre cette observation avec quatre conformations successives. D'une conformation à la suivante, nous pouvons observer qu'une seule liaison hydrogène apparaît ou disparaît.

Cette suite de graphes, représentant les différentes conformations de la molécule au cours de la trajectoire, constitue la base de notre analyse de la dynamique moléculaire. Caractériser cette suite de graphes sera donc la première étape de notre analyse.

Quelques références, telles que celles de [4, 24] explorent l'utilisation de concepts topologiques issues de la théorie des graphes pour caractériser la structure moléculaire.

En revanche, la plupart des ouvrages de référence en modélisation moléculaire, comme [38], abordent la modélisation d'une molécule pour ensuite explorer les méthodes d'analyses classiques comme celles évoquées précédemment (simulation, énergie potentielle,

2. 1 femtoseconde (fs) équivaut à 10^{-15} secondes.



(a) conformation 1 (b) conformation 2 (c) conformation 3 (d) conformation 4

Figure 1.8 – Illustration d'une suite de conformations. Les conformations contiennent chacune un ensemble de liaisons hydrogène différent : N_3-O_3, N_5-O_4, N_7-O_7 dans la Figure 1.8a, N_3-O_3, N_5-O_4, N_6-O_5, N_7-O_7 dans la Figure 1.8b, N_5-O_4, N_6-O_5, N_7-O_7 dans la Figure 1.8c, et N_3-O_2, N_5-O_4, N_6-O_5, N_7-O_7 dans la Figure 1.8d.

...), ou proposent la modélisation moléculaire pour des problématiques particulières telles que la modélisation de réaction chimique, ou la découverte de médicament.

Dans le domaine plus spécifique de l'analyse de trajectoire de dynamique moléculaire, des avancées techniques récentes facilitent l'étude des dynamiques moléculaires [26, 46]. Pour autant, l'outil de visualisation graphique VMD [28] demeure la référence depuis plusieurs décennies. Il permet d'observer les conformations de manière indépendante les unes des autres et est largement utilisé pour une analyse visuelle des trajectoires moléculaires.

Dans ce contexte, la question de la caractérisation d'un ensemble de conformations d'une même molécule ne semble pas être un problème majeur. Les méthodes d'analyse classiques sont bien établies et paraissent indispensables. En proportion, il existe peu de références offrant une approche globale pour l'analyse de la structure des molécules. Et, aucune étude topologique des graphes d'une trajectoire de dynamique moléculaire n'a été recensée jusqu'à présent.

Dans les prochaines sections, nous examinerons les méthodes disponibles dans la littérature afin de déterminer si nous pouvons les appliquer à notre problème.

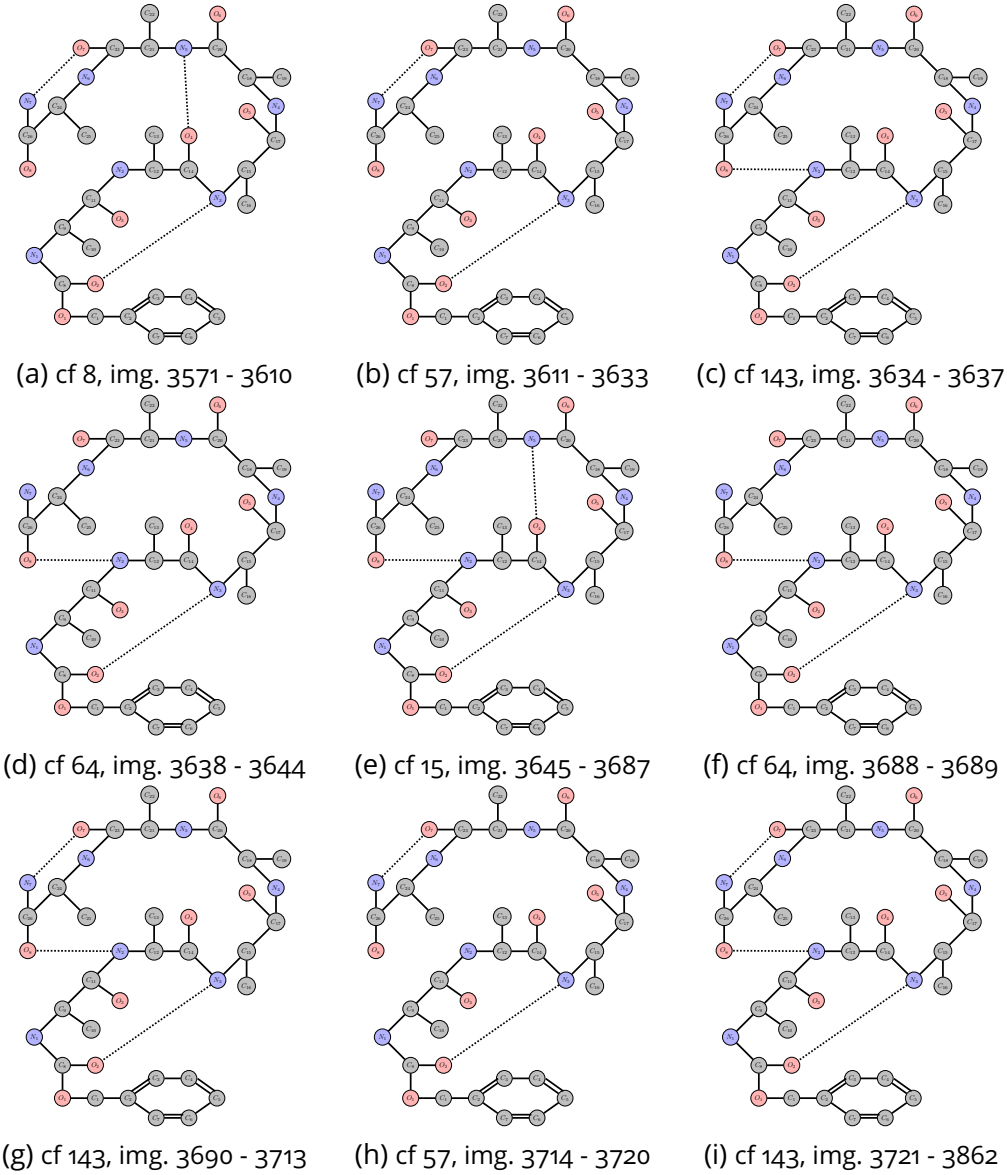
1.3.2 . Caractérisation d'une trajectoire par la catégorisation de ses graphes moléculaires

Dans cette section, nous proposons d'analyser une trajectoire de dynamique moléculaire par la catégorisation de ses graphes moléculaire.

Traditionnellement, l'analyse d'une trajectoire repose sur l'utilisation de l'énergie potentielle pour classifier les structures observées. Cependant, ici, nous explorons la possibilité d'obtenir des résultats similaires en se basant uniquement sur les graphes moléculaires et les événements liés aux liaisons hydrogène. Nous examinons plusieurs approches dans cette démarche, mais aucune ne se révèle concluante. Tout d'abord, nous tentons de catégoriser les transitions de liaisons hydrogène afin d'identifier celles conduisant nécessairement à des conformations représentant des structures stables, ou au contraire, à des conformations représentant des structures instables. Ensuite, une seconde approche consiste à associer à chaque conformation une estimation de son niveau énergétique, dans le but d'approximer les bassins conformationnels et d'en déduire la stabilité des structures liées à la conformation. Malheureusement, ces tentatives de catégorisation des graphes n'aboutissent pas aux résultats escomptés. En conséquence,

nous nous orientons vers une nouvelle méthode que nous détaillons dans la Section 1.3.3.

Nous présentons maintenant notre étude sur les possibilités de catégorisation des graphes, afin d'expliquer les limites de ces méthodes. Nous considérons alors la trajectoire à travers sa série de graphes, la Figure 1.9 illustre cette vue sur une portion de la trajectoire. Rappelons que les graphes de chaque conformation ne diffèrent que par leurs liaisons hydrogène.



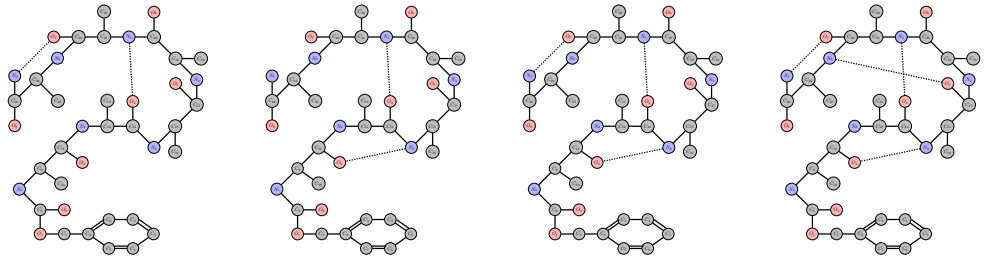
(a) cf 8, img. 3571 - 3610 (b) cf 57, img. 3611 - 3633 (c) cf 143, img. 3634 - 3637

(d) cf 64, img. 3638 - 3644 (e) cf 15, img. 3645 - 3687 (f) cf 64, img. 3688 - 3689

(g) cf 143, img. 3690 - 3713 (h) cf 57, img. 3714 - 3720 (i) cf 143, img. 3721 - 3862

Figure 1.9 – Illustration de la suite de graphes correspondant à environ 300 images d'une trajectoire de dynamique moléculaire du peptide Z-Ala$_6$. La trajectoire complète comprend un total de 12000 images. Rappelons que les liaisons covalentes sont en traits plein et les liaisons hydrogène en pointillés.

Nous débutons en examinant les transitions d'un graphe au suivant dans le but de reconnaître celles qui conduisent à des structures très stables, ou au contraire, instables.



(a) cf 17 (b) cf 78 (c) cf 86 (d) cf 87

⇢ cf 17 ⟶ cf 86 ⟶ cf 17 ⟶ cf 86 ⟶ cf 17 ⟶ cf 86 ⟶ cf 78 ⟶ cf 86 ⟶ cf 87 ⟶ cf 86 ⇢

(e) Séquence d'appariton des conformations extraites de la trajectoire.

Figure 1.10 – Exemple d'une séquence répétitive extraite d'une trajectoire de dynamique moléculaire du peptide Z-Ala_6. Les conformations indiquées dans la séquence de la Figure 1.10e sont illustrées sous la forme de graphes dans les figures 1.10a, 1.10b, 1.10c et 1.10d.

Chaque transition correspond à l'apparition ou à la disparition d'une, ou exceptionnellement, deux liaisons hydrogène. Notre objectif est alors de déterminer si l'apparition ou la disparition d'une certaine liaison est nécessairement associée à la formation d'une conformation stable, ou inversement. Cependant, cette hypothèse ne se vérifie pas dans le contexte des conformations. En effet, la stabilité d'une structure est induite par la présence d'un ensemble de liaisons hydrogène. Il est bien connu que chaque liaison hydrogène correspond à une certaine quantité d'énergie potentielle ; par conséquent, l'ajout de liaisons hydrogène à une structure entraîne une diminution de son énergie potentielle. Ainsi, la structure est définie par l'ensemble des liaisons qui coexistent et non par une liaison spécifique. De plus, étant donné que les liaisons hydrogènes sont liées à la présence d'électrons libres, certains couples de liaisons hydrogène sont incompatibles. Par conséquent, une structure est constituée d'un ensemble de liaisons hydrogène compatibles.

En fonction des liaisons déjà présentes, l'apparition ou la disparition d'une liaison aura un impact très différent. Les apparitions ou les disparitions de liaisons hydrogène qui constituent une transition ne sont pas quantifiables en termes de stabilité. Ainsi, nous ne pouvons pas nous baser sur l'apparition ou la disparition d'une seule liaison pour définir la catégorie du graphe de départ ou du graphe d'arrivée.

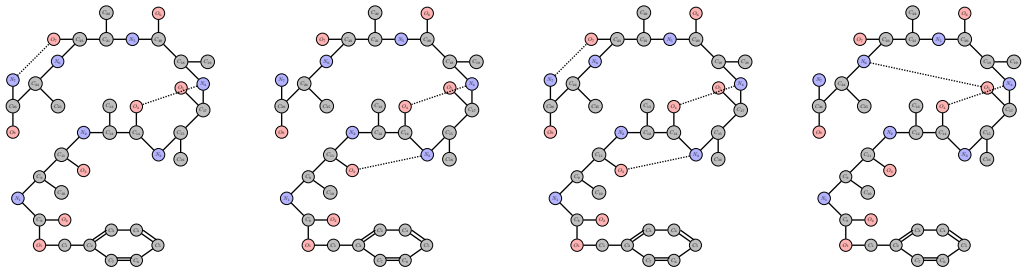
Même en supposant que nous disposions d'une structure stable, nous ne pourrions pas généraliser nos conclusions aux autres structures stables de la trajectoire. En effet, une structure stable représente le fond d'un bassin conformationnel. Les autres structures stables, correspondant à d'autres bassins conformationnels, auront des niveaux d'énergie différents, et par conséquent, un nombre et une composition en liaisons hydrogène totalement différents. En résumé, les événements d'apparition et de disparition des liaisons hydrogène ne nous permettent pas de catégoriser les conformations parcourues.

Pour aller plus loin, comme le montre la Figure 1.10, nous observons la présence de séquences au cours desquelles les mêmes conformations apparaissent en boucle dans la trajectoire. Ces séquences répétitives sont problématiques car elles peuvent indiquer soit une légère agitation au sein d'un bassin, soit une grande flexibilité caractéristique des états transitifs. Nous sommes donc incapables de catégoriser ces périodes. Cela nous laisse dans l'incertitude quant à la stabilité des structures traversées.

En étudiant la suite de graphes d'une trajectoire, comme illustré dans la Figure 1.9, nous remarquons que des changements structuraux se produisent tout au long de la trajectoire. Pour autant, tous ces changements ne mènent pas nécessairement à des structures stables ; certains conduisent à des structures instables et, certains ne modifient pas la méta-structure.

Notre objectif est donc de reconnaître les changements significatifs, c'est-à-dire ceux qui correspondent à une modification importante du niveau d'énergie potentielle.

Nous allons à présent examiner les conformations, et par extension, les graphes moléculaires qui les représentent. À partir des simulations faites sur le peptide Z-Ala$_6$, nous avons identifié plusieurs conformations distinctes dont les structures associées présentent des énergies potentielles très proches. Étant donné la proximité temporelle de ces structures, nous supposons qu'elles se situent *a priori* dans le même bassin conformationnel, et donc qu'elles correspondent à la même méta-structure. Ces conformations, bien que différentes par définition, présentent des changements mineurs de l'une à l'autre qui n'altérent pas significativement les relations entre les atomes. La Figure 1.11 illustre l'exemple de quatre conformations provenant du même bassin conformationnel dans une trajectoire de Z-Ala$_6$. En utilisant des méthodes d'analyse de trajectoire classique, nous avons pu identifier plusieurs bassins conformationnels au sein de cette trajectoire, dont l'un oscille entre des structures représentées par ces quatres conformations spécifiques. Ces conformations apparaissent et disparaissent au sein du bassin au cours d'une séquence de moins de 300 images (soit 150 fs.). Elles apparaissent non seulement au sein de ce bassin particulier, mais également à de nombreuses autres périodes de la trajectoire. Nous avons envisagé de regrouper les conformations en les étiquetant selon le bassin auquel elles appartiennent, mais cela s'avère impossible car une conformation peut appartenir à plusieurs bassins. En effet, comme nous l'avons observé expérimentalement dans plusieurs de nos trajectoires, certaines conformations sont caractérisées par un sous-ensemble de liaisons que l'on retrouve dans plusieurs bassins. Il n'est donc pas envisageable, dans l'état actuel, de catégoriser les graphes moléculaires, et donc les structures qu'ils représentent.



(a) conformation 100 (b) conformation 102 (c) conformation 103 (d) conformation 105

Figure 1.11 – Les graphes moléculaires de plusieurs conformations appartenant à un même bassin conformationnel.

Nous proposons maintenant de nous intéresser à l'énergie potentielle des conformations. Une approche possible serait d'estimer l'énergie potentielle des structures associées à une conformation en utilisant les paramètres du graphe. Cependant, en observant la Figure 1.12, nous pouvons rapidement rejeter une telle approche. La capacité d'une

conformation à apparaitre à différentes périodes de la trajectoire lui permet d'appartenir à différents bassins conformationnels, chacun pouvant varier en termes d'énergie. Il est même envisageable qu'une conformation appartienne à la fois à plusieurs puits d'énergie tout en étant également associée à des états transitifs.

Ainsi, étant donné une conformation, l'énergie potentielle a été calculée pour chacune des images correspondantes dans la trajectoire. Nous avons déterminé l'intervalle des énergies potentielles de chaque conformation parcourue.

Comme le montre clairement la Figure 1.12, ces intervalles sont si vastes qu'il est impossible de catégoriser les conformations à partir de ce paramètre. Ce résultat n'est pour autant pas surprenant, car l'énergie potentielle est calculée à partir des coordonnées tri-dimensionnelles. Cela indique que nous ne pourrons pas estimer ou reproduire l'énergie potentielle à partir du graphe moléculaire.

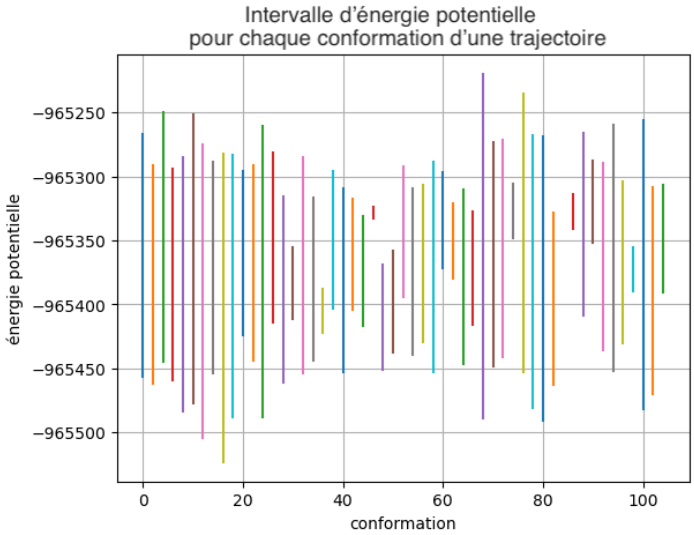
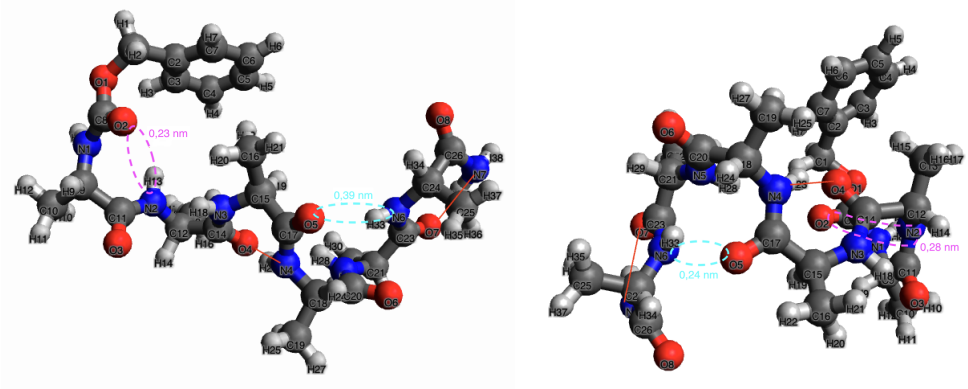


Figure 1.12 – Intervalles d'énergie de plusieurs conformations. Chaque trait vertical représente les énergies potentielles calculées pour une conformation.

En passant d'une image tri-dimensionnelle à un graphe moléculaire, plusieurs structures de la molécule sont regroupées en une seule conformation. Certains atomes peuvent occuper des positions très différentes. Même s'ils n'appartiennent pas aux liaisons hydrogène présentes dans la conformation, ces atomes peuvent être impliqués dans les conformations futures dans d'autres liaisons. Cependant, en raison de leurs coordonnées réelles, ils ne pourront pas former les mêmes liaisons dans les structures futures. La Figure 1.13 illustre cela avec l'exemple de deux images tri-dimensionnelles correspondant à la même conformation. Nous observons que dans la Figure 1.13a, les atomes de la liaison hydrogène N_6-O_5 sont extrémements éloignés ($0,3959$ nm). Tandis, que dans la Figure 1.13b ils ne sont plus qu'à seulement $0,246$ nm. Ils sont donc assez proches pour qu'une petite agitation suffise à faire apparaître la liaison hydrogène. Ces deux figures, 1.13a et 1.13b, sont à environ 2500 images d'écart. En raison de l'éloignement des atomes, cette liaison (N_6-O_5) ne pouvait en aucun cas se former à partir de la conformation de l'image 357. À l'image 3011 la liaison est à présent accessible. Il s'agit pourtant de la même conformation et donc du même graphe moléculaire. Sans contexte nous ne pouvons pas catégoriser les structures représentées par les conformations.



(a) Image de la conformation 100 avant l'apparition de la liaison N_2-O_2 (img. 357).

(b) Image de la conformation 100 avant l'apparition de la liaison N_6-O_5 (img. 3011).

Figure 1.13 – Deux images de la conformation 100 extraites de la même trajectoire mais qui précédent des conformations différentes. Les traits rouges indiquent les liaisons hydrogène présentes. Les pointillés indiquent les liaisons hydrogène absentes. Les lignes roses représentent la liaison N_2-O_2 et les lignes bleues représentent la liaison N_6-O_5.

Ainsi, c'est la dynamique que nous devons réussir à modéliser en reconnaissant les différentes structures de la molécule explorées lors d'une trajectoire. En considérant la suite des graphes, nous ne parvenons pas à catégoriser les structures parcourues. Ces graphes sont riches en informations structurelles, mais les spécificités atomiques ne nous permettent d'avoir une vision d'ensemble.

Nous envisageons donc une autre méthode de caractérisation de la structure d'une molécule. Celle-ci repose bien entendu, également, sur la vision d'une trajectoire comme une suite de graphes moléculaires. Néanmoins, nous ne nous intéresserons plus aux changements atomiques de ces graphes. Nous souhaitons, à présent, étudier la trajectoire à travers un prisme différent.

1.3.3 . Caractérisation d'une trajectoire à partir de motifs topologiques issus de ses graphes moléculaires

Une méthode de caractérisation classique repose sur l'utilisation de la topologie des graphes moléculaires [4, 15]. Elle consiste à caractériser la structure moléculaire en se basant sur les paramètres du graphe ou sur les motifs spécifiques qu'on y identifie. C'est cette approche que nous allons explorer dans cette section.

Définition 3. *Un **motif** est un ensemble d'éléments identifiables qui est spécifique ou répétitif dans la structure moléculaire. Dans un graphe moléculaire, un motif est un sous-graphe de celui-ci.*

Dans notre contexte, l'apport d'une telle méthode réside dans le fait que les motifs sont des éléments distincts et reconnaissables des graphes moléculaires. Chaque conformation contient un certain nombre de motifs, qui peuvent être partagés par plusieurs conformations ou qui peuvent être spécifiques à certaines. En se concentrant sur les motifs, nous changeons la granularité de notre analyse, pour sortir des spécificités atomiques et mieux appréhender les structures moléculaires dans leur ensemble.

Le choix du motif dépend de la problématique spécifique. Par exemple, dans le cas

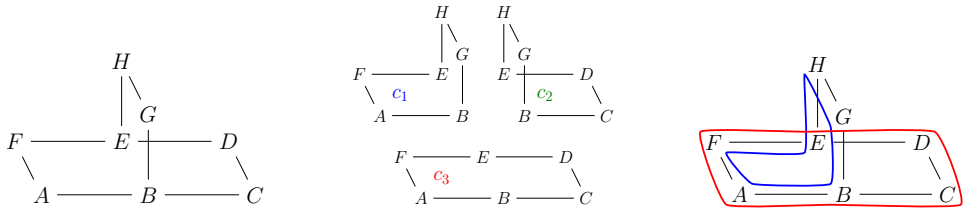
de l'ARN, les motifs sont recherchés en raison de leurs rôles dans la structure tertiaire [3] [22]. Dans le domaine du vivant, les motifs d'intérêts sont souvent ceux qui caractérisent les sites de fixation. C'est le cas, par exemple, dans les protéines, où le motif est alors une suite d'acides aminés qui permettent la fixation d'un agent à la protéine. Dans le contexte d'une étude structurelle, les cycles du graphe constituent souvent le premier choix pour identifier et caractériser une structure à partir d'un graphe moléculaire [8, 23]. Ils sont considérés comme des briques structurelles, surtout lorsqu'ils sont de petite taille. En effet, ils mettent en évidence la présence de multiples contraintes spatiales et permettent de localiser ces contraintes dans la molécule. La décomposition en cycles est très utilisée depuis plusieurs décennies [5, 20, 21, 40]. Cette décomposition en petits cycles est utile dans de nombreux problèmes bio-chimiques, comme les problèmes de comparaison structurelle de multiples molécules [42, 55].

Toutes les décompositions en cycles ne sont pas équivalentes, et il est important de choisir une décomposition en fonction du problème spécifique à résoudre. De manière générale, deux critères sont pris en compte lors de la sélection d'une décomposition en cycles : la taille des cycles et le nombre de cycles de l'ensemble. Le nombre de cycles doit être suffisant pour représenter de manière adéquate la structure, et les petits cycles sont privilégiés en raison de leur similitude avec les cycles couramment observés à l'oeil dans les molécules. Nous pouvons également considérer qu'un petit cycle représente plus de contraintes structurelles simultanées. En effet, dans le cas des petits cycles la géométrie des atomes impliqués combine de nombreuses contraintes pour permettre aux liaisons d'exister simultanément.

Cela a donné lieu au problème du *Smallest Set of Smallest Rings* (SSSR) au sein de la communauté de chémo-informatique. Comme son nom l'indique, ce problème consiste à chercher le plus petit ensemble de cycles contenant les plus petits cycles qui représentent l'entièreté de la structure moléculaire. Ce problème possède un équivalent dans la communauté informatique : le problème de la base de cycles minimum d'un graphe [27, 41]. Étant donné un graphe, le nombre de cycles nécessaires pour représenter tous les cycles du graphe est connu et fixé par les dimensions (en nombre de sommets et en nombre d'arêtes) du graphe en question. Dans un graphe, une base de cycles est un ensemble de cycles générateurs de tous les cycles du graphe. Le problème de la base de cycles minimum consiste donc à rechercher une base de cycles dont le nombre d'éléments est fixé par la dimension du graphe et dont la somme du poids de ses éléments est minimale. Les notions inhérentes aux cycles et aux bases de cycles seront définies formellement dans le Chapitre 2.

Un graphe peut admettre plusieurs bases de cycles minimum. Ainsi, dans la littérature, nous constatons que ce ne sont pas de simples bases de cycles minimum qui sont utilisées pour la caractérisation de structures moléculaires. En effet, elles sont souvent jugées insuffisantes pour les analyses topologiques, car elles ne sont pas déterministes et comme elles sont très restrictives, les résultats sont dépendants de la base choisie [8]. La base de cycles minimum ne fournit qu'un seul angle d'étude des relations entre les cycles d'un graphe moléculaire. Ainsi, un candidat fréquemment utilisé est l'union de toutes les bases de cycles minimum [9, 51]. Nous considérons ainsi tous les cycles qui appartiennent à au moins une base de cycles minimum du graphe. Ces cycles sont alors qualifiés de

3. La structure tertiaire correspond au repliement dans l'espace qui détermine la fonction d'une protéine.



(a) Un graphe (b) Les cycles *relevants* (c) Une base du graphe

Figure 1.14 – Illustration, étant donné un graphe, de l'ensemble de cycles relevants et d'une base de cycles minimum. Dans la Figure 1.14c, le cycle c_2 n'a pas été selectionné par il est obtenable par la combinaison du cycle c_1 et du cycle c_3.

cycles **relevants** ou pertinents en français. La Figure 1.14 présente un exemple simple où l'ensemble de cycles relevants est facilement énuméré et est illustré par la Figure 1.14b. Chaque paire de cycles relevants forme une base de cycles minimum, comme celle illustrée par la Figure 1.14c.

L'ensemble de cycles relevants est bien plus exhaustif qu'une base de cycles minimum classique, cependant son calcul est nettement plus lourd. En effet, un graphe peut contenir un nombre de cycles relevants exponentiel en fonction de son nombre d'arêtes.

Le calcul d'une seule base de cycles minimum présente l'avantage de pouvoir être effectué en temps polynomial [27]. Pour cela, il suffit de considérer un ensemble bien défini de cycles dans le graphe et d'en extraire une base au sens mathématique. Un compromis a été proposé dans le cadre des comparaisons topologiques [43] en utilisant une base de cycles minimum dite étendue. Les auteurs ont commencé par appliquer un algorithme classique, tel que celui mentionné précédemment, pour obtenir une base de cycles minimum. Ensuite, ils ont examiné les autres cycles générés par l'algorithme et les ont inclus dans l'ensemble s'ils étaient des cycles relevants. Ce faisant, toutes les bases de cycles minimum existantes au sein de l'ensemble de cycles générés initialement sont prises en compte. Ce compromis s'est montré efficace pour les comparaisons topologiques au sein d'une base de données de molécules, un contexte dans lequel le coût de calcul de l'union des bases de cycles minimum rendait ce modèle inenvisageable.

1.4 . Approche proposée pour l'analyse de trajectoires de dynamique moléculaire

Comme nous l'avons vu, l'approche traditionnelle pour l'analyse des trajectoires de dynamique moléculaire repose sur la mesure de l'énergie potentielle. Cependant, cette méthode nécessite des calculs complexes et coûteux que nous cherchons à éviter. En se concentrant sur la modélisation par les conformations, nous cherchons à nous affranchir de la nécessité de connaître précisément l'énergie potentielle, et à proposer une analyse à partir des graphes moléculaires.

Nous proposons donc d'adopter une approche basée sur la caractérisation par des motifs structuraux. En particulier, nous caractérisons chaque conformation par une base de cycles minimum. Contrairement aux méthodes de la littérature qui se concentrent sur une seule structure, notre approche doit prendre en compte une multitude de conformations. Ainsi, nous choisissons d'utiliser des bases de cycles minimum pour caractériser

les conformations, et ce, malgré les recommandations de la littérature en faveur de l'utilisation d'ensembles plus exhaustifs. En effet, chaque conformation que nous traitons correspond à un graphe moléculaire possédant potentiellement un nombre exponentiel de cycles relevants. Le coût de calcul d'un ensemble de cycles plus exhaustif serait prohibitif pour notre méthode. De plus, notre approche va au-delà de la simple constitution d'une base de cycles minimum. Une fois chaque conformation caractérisée par une base de cycles, nous pouvons étudier la dynamique en examinant l'évolution des cycles au fil de la trajectoire. Ainsi, même si chaque conformation est caractérisée par un ensemble restreint de cycles, la trajectoire dans son ensemble est caractérisée par un ensemble de bases de cycles minimum.

À partir de la caractérisation des conformations par une base de cycles minimum, nous proposons une caractérisation de l'ensemble de la trajectoire de dynamique moléculaire. Pour cela, nous considérons les cycles et les interactions qu'ils ont avec les autres cycles dans les différentes bases de cycles minimum utilisées pour la caractérisation des conformations. Nous envisageons maintenant la trajectoire non plus comme une séquence de conformations, mais plutôt comme les cycles qui apparaissent et disparaissent au fil du temps. Cette approche nous permet d'observer un phénomène d'équivalence structurelle entre les cycles, que nous avons défini comme le polymorphisme de cycles, qui représente la capacité d'un cycle à évoluer tout en conservant le même rôle dans la structure de la molécule. Nous regroupons alors les cycles qui ont un impact similaire sur cette dernière, ce qui nous permet d'obtenir des ensembles représentatifs des différentes contraintes qui peuvent s'appliquer au système moléculaire pour définir sa structure. De cette manière, nous définissons la ou les méta-structures principales du système. À partir de ces méta-structures, nous pouvons identifier les structures d'intérêt observées au cours de la trajectoire. Ces structures sont celles qui persistent suffisamment longtemps pour marquer une période dans la trajectoire, des périodes qui peuvent être assimilées à l'exploration d'un bassin conformationnel.

Les étapes de la méthode sont schématisées dans la Figure 1.15, qui se divise en plusieurs parties, chacune illustrée par une couleur différente.

Prélablement, nous effectuons la transition d'une suite de conformations issues des travaux [10, 11] vers un ensemble de graphes. En effet, notre objectif est d'établir une caractérisation pour chaque graphe moléculaire correspondant à une conformation. Les aspects tels que la récursivité des conformations n'ont pas d'impact sur la caractérisation de la trajectoire, cependant, ils seront d'intérêt pour l'identification des méta-structures de la trajectoire.

La première partie de notre analyse porte sur la caractérisation de la trajectoire par ses cycles. Les cycles sont identifiés à partir des bases de cycles des conformations, puis constituent le point de départ de la recherche et du regroupement de cycles équivalents. Cette partie, concentrée sur la caractérisation d'une trajectoire, explore plusieurs problèmes rattachés à la théorie des graphes. Cette caractérisation est découpée en deux problèmes qui seront abordés dans les chapitres suivants (Chapitre 3, Chapitre 4).

À partir de la caractérisation de la trajectoire, nous passons à la deuxième partie de notre méthode, qui concerne l'analyse proprement dite de la trajectoire. La modélisation de la structure moléculaire par les cycles offre de nombreuses possibilités, dont l'énumération des structures d'intérêt de la dynamique. Cette modélisation permet également

d'aller plus loin en pouvant être utilisée pour comparer des trajectoires entières. Ces aspects seront discutés dans le dernier chapitre de la thèse (Chapitre 6).

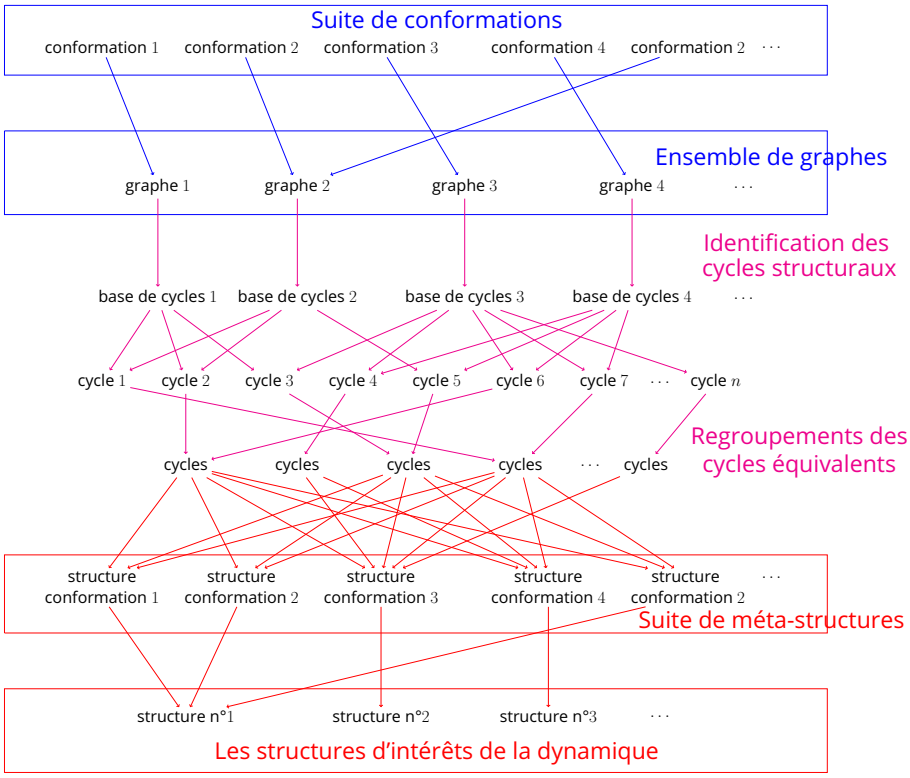


Figure 1.15 – Schéma représentant les étapes de la méthode que nous proposons.

2 - Caractérisation d'une trajectoire de dynamique moléculaire

Ce chapitre présente notre approche pour la caractérisation d'une trajectoire de dynamique moléculaire en utilisant la théorie des graphes. Cette méthode nous permet de distinguer les modifications structurelles significatives, des changements mineurs.

Notre méthode exclut l'utilisation de l'énergie potentielle et des graphes tridimensionnels, se concentrant uniquement sur les graphes moléculaires des conformations. Comme établi dans le Chapitre 1, une trajectoire est représentée par une suite de conformations prises par la molécule. Nous proposons donc de caractériser chaque conformation de manière indépendante à partir de ses cycles, et ainsi d'obtenir un graphe représentatif de sa structure, c'est-à-dire représentatif des intéractions existantes entre les atomes de la molécule. En définissant des ensembles d'équivalence entre les cycles, nous pouvons alors définir une méta-structure globale de la molécule, modélisée par un graphe que nous appelons "polygraphe". Chaque sommet de ce polygraphe correspond à un ensemble de cycles équivalents. Les méta-structures d'intérêt sont alors des sous-graphes au sein de ce polygraphe.

L'objectif de ce chapitre est de présenter la démarche globale suivie dans cette thèse. Ainsi, nous introduisons les différentes étapes qui nous conduisent au calcul du polygraphe dont les aspects plus spécifiques seront abordés dans les chapitres suivants (Chapitre 3, Chapitre 4). Les différentes méthodes possibles pour repondre à ces questions spécifiques seront comparées dans le Chapitre 5. Enfin, nous présenterons quelques applications de ce polygraphe à l'analyse des trajectoires de dynamique moléculaire, ce point sera ensuite élaboré dans le Chapitre 6.

2.1 . Trajectoire de dynamique moléculaire

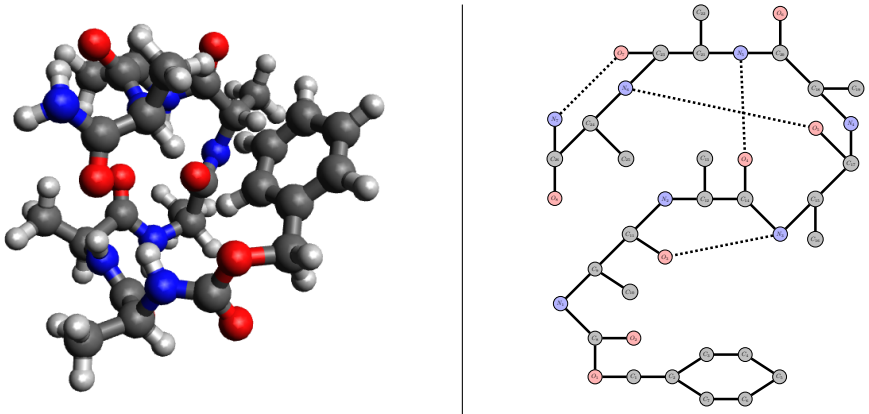
Dans cette section, nous présentons la trajectoire de dynamique moléculaire dans le contexte des graphes. Ces notions ont déjà été introduites dans le Chapitre 1 et nous les abordons maintenant de manière formelle en les situant dans le domaine de la théorie des graphes.

Comme nous l'avons expliqué dans le Chapitre 1, une trajectoire est à l'origine une suite d'images de la molécule. Ainsi, les travaux de [10, 11] transforment ce format pour nous fournir une suite de conformations à considérer en entrée. Cette suite de conformations est obtenue après l'identification des différentes conformations de la trajectoire et la suppression des conformations fugaces.

Il est à noter que dans le cadre de cette thèse, nous nous limitons aux trajectoires de dynamique moléculaire impliquant des liaisons covalentes et, pour liaisons faibles, uniquement des liaisons hydrogène.

Définition 4 (Conformation)**.** *Une conformation est un graphe moléculaire* $G = (V, E \cup H)$*, où* V *est l'ensemble des sommets correspondant aux atomes du système moléculaire, et les arêtes de l'ensemble* $\{E \cup H\}$ *représentent les liaisons chimiques entre les atomes. Deux types de liaisons chimiques sont à distinger : les liaisons fortes (correspondant ici aux liaisons*

covalentes) représentées par les arêtes de l'ensemble E, et les liaisons faibles (correspondant ici aux liaisons hydrogène) représentées par les arêtes de l'ensemble H.



(a) Une image tridimensionnelle de la molécule.

(b) La conformation obtenue à partir de l'image de la Figure 2.1a.

Figure 2.1 – Exemple du graphe moléculaire obtenu à partir d'une image tridimensionnelle extraite d'une trajectoire de dynamique moléculaire. Les couleurs utilisées dans les figures représentent le type chimique de l'atome en question : bleu pour azote, rouge pour oxygène, gris pour le carbone et blanc pour l'hydrogène. Le graphe de la Figure 2.1b possède deux types d'arêtes, les traits pleins représentent des liaisons covalentes tandis que les pointillés représentent les liaisons hydrogène.

Il est important de noter que le modèle que nous proposons pourrait être étendu pour inclure d'autres types de liaisons chimiques.

Nous considèrons, ici, des trajectoires de dynamique moléculaire présentant des liaisons fortes immuables et des liaisons faibles variables.

Une trajectoire de dynamique moléculaire est une séquence de conformations, chacune représentant un état spécifique du système moléculaire observé au cours de la simulation. Bien que les conformations consécutives soient différentes, certaines d'entre elles sont récurrentes. Aussi, pour un système moléculaire donné, toutes les conformations de toutes les trajectoires partagent le même squelette ou backbone. Pour clarifier ces concepts, nous introduisons les définitions suivantes.

Définition 5 (Trajectoire)**.** *Une trajectoire est une suite $G_1, G_2, G_3, \ldots, G_{M-1}, G_M$ avec $G_i = (V, E \cup H_i)$ avec $1 \leq i \leq M$ et H_i correspond ici aux liaisons hydrogène de la conformation G_i.*

Remarque 1. *De plus, pour tout $1 \leq i \leq M-1$ on a $H_i \neq H_{i+1}$. Notons que rien n'interdit pour $i, j \in [1, M]$ avec $|i-j| > 1$ d'avoir $H_i = H_j$ et donc $G_i = G_j$.*

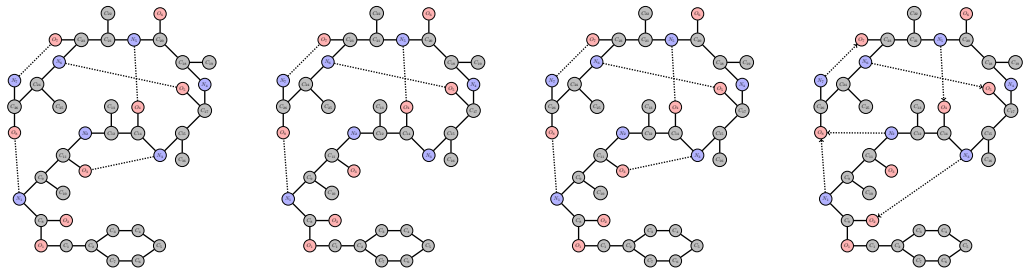
Définition 6 (Backbone)**.** *Le backbone désigne le sous-graphe commun à toutes les conformations d'une trajectoire. Il s'agit du sous-graphe (V, E) du graphe G de la Définition 4.*

Remarque 2. *Le backbone, tel que défini, correspond au sous-graphe des liaisons covalentes. Dans certains cas, des liaisons hydrogène peuvent être présentes dans toutes les conformations d'une trajectoire, ce qui implique leur appartenance au sous-graphe commun de cette*

trajectoire. Une alternative que nous n'avons pas explorée dans cette thèse consiste à envisager le backbone comme le sous-graphe commun à toutes les conformations, ce qui inclurait les liaisons hydrogène si nécessaire.

Définition 7. *Étant donnée une trajectoire $G_1, G_2, \dots G_M$, nous avons $\mathcal{G} = \{G_1, G_2, \dots, G_N\}$ avec $N \leq M$, l'ensemble des conformations différentes présentes au sein de la trajectoire.*

Afin de faciliter la compréhension des étapes de la méthode décrite dans ce chapitre, nous allons suivre le traitement d'un exemple simplifié. Cette trajectoire est illustrée dans la Figure 2.2 et se compose de quatre conformations, dont trois sont distinctes.



(a) Graphe G_1. (b) Graphe G_2. (c) Graphe G_1. (d) Graphe G_3.

Figure 2.2 – Un exemple simplifié d'une trajectoire G_1, G_2, G_1, G_3, ainsi nous avons $\mathcal{G} = \{G_1, G_2, G_3\}$

.

2.2 . Modélisation de la structure d'une conformation

Dans le Chapitre 1, nous avons établi que les cycles constituent d'excellents éléments de représentation de la structure moléculaire. Dans cette section, nous proposons une méthode de caractérisation des conformations basée sur l'analyse de leurs cycles.

Nous utilisons une approche basée sur une base de cycles minimum pour caractériser chaque conformation. Cette base de cycles permet de représenter l'ensemble des cycles du graphe en utilisant un nombre restreint d'entre eux, tout en privilégiant les cycles de petite taille. Ce problème est connu sous le nom de base de cycles de poids minimum, ou *minimum cycle basis* en anglais, et est référencé dans la littérature de chémo-informatique [8, 40] sous le nom *smallest set of rings*. Cette méthode de modélisation de la structure par les cycles est déjà connue dans la littérature [9, 23, 43, 51].

2.2.1 . Base de cycles minimum

Dans cette section, nous reprenons une définition générale des cycles afin d'introduire la notion de base de cycles minimum d'un graphe. Ces définitions ont été initialement proposées dans [7].

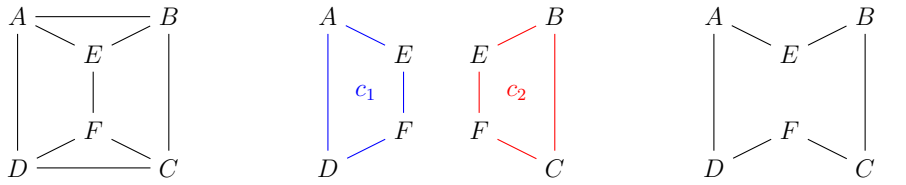
Définition 8 (Cycle)**.** *Étant donné un graphe $G = (V, E)$, un cycle est un sous-graphe de G où chaque sommet présente un degré pair.*

Définition 9 (Poids d'un cycle)**.** *Étant donné un cycle c, son poids est noté $\omega(c)$ et est égal à son nombre d'arêtes.*

Dans le graphe $G = (V, E)$, un cycle c de poids l est noté $c = \{e_1, e_2, \ldots, e_l\}$, où $e_i \in E$ pour $1 \leq i \leq l$.

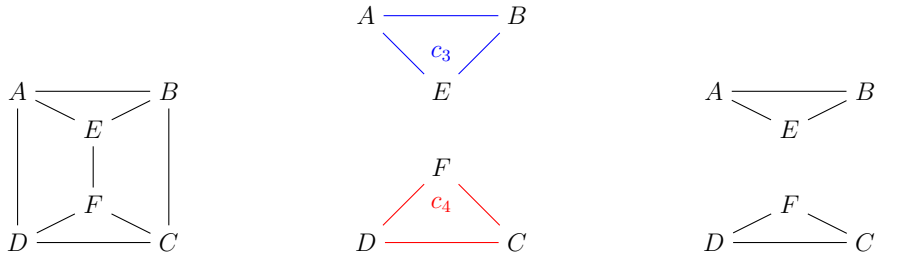
Définition 10 (Somme de deux cycles)**.** *La somme de deux cycles* ($\oplus$)*, également appelée composition, notée* $c_1 \oplus c_2$*, est le cycle dont l'ensemble d'arêtes correspond aux arêtes présentes dans* c_1 *ou dans* c_2*, mais pas dans les deux. Autrement dit, on a* $c = c_1 \oplus c_2 = \{c_1 \cup c_2\} \backslash \{c_1 \cap c_2\}$.

Les Figures 2.3 et 2.4 illustrent cette opération sur un exemple.



(a) Un graphe G (b) Deux cycles de G (c) Le cycle $c_1 \oplus c_2$.

Figure 2.3 – Illustration de la somme de deux cycles c_1 et c_2.



(a) Un graphe G (b) Deux cycles de G (c) Le cycle $c_3 \oplus c_4$.

Figure 2.4 – Illustration de la somme de deux cycles c_3 et c_4 avec $c_3 \cap c_4 = \emptyset$

Remarque 3. *La composition* ($\oplus$) *est une opération commutative et associative. Ainsi, étant donnés trois cycles* c_1*,* c_2 *et* c_3*,* $c_1 \oplus c_2 \oplus c_3 = (c_1 \oplus c_2) \oplus c_3 = c_1 \oplus (c_2 \oplus c_3) = c_2 \oplus (c_1 \oplus c_3)$.

En théorie des graphes, un cycle est souvent défini comme une chaîne fermée, c'est-à-dire une séquence d'arêtes partant d'un sommet et revenant à ce même sommet. On parle de cycle **élémentaire** lorsque cette chaîne ne repasse jamais deux fois par le même sommet. En repartant de la Définition 8, un cycle élémentaire est alors un sous-graphe connexe où chaque sommet a un degré de deux. Cette définition des cycles élementaires est par essence plus restrictive que la définition des cycles que nous avons introduite. Par exemple, la composition de deux cycles élementaires n'induit pas nécessairement un cycle élémentaire. Ainsi, nous adoptons la Définition 8 qui est plus large mais qui permet surtout de définir l'espace vectoriel des cycles.

L'ensemble des cycles du graphe G, désigné par $\mathcal{C}_G$, lorsqu'il est muni de l'opération $\oplus$, forme un espace vectoriel dans le corps $\mathbb{Z}/2\mathbb{Z}$, souvent désigné sous le terme d'espace des cycles [14]. En effet, intuitivement, l'ensemble $\mathcal{C}_G$ doté de l'opération $\oplus$ constitue un groupe commutatif $(\mathcal{C}_G, \oplus)$. Cette opération est à la fois commutative et associative. De plus, le cycle vide, c'est à dire celui formé d'aucune arête, est l'élément neutre de l'espace. Par conséquent, chaque cycle est son propre inverse.

Remarque 4. *Afin de simplifier les notations des conformations et lorsque cela ne crée pas d'ambiguïté, l'ensemble des cycles d'une conformation G_i est noté $\mathcal{C}_i$.*

Le nombre de cycles dans l'ensemble $\mathcal{C}_G$ peut être exponentiel par rapport au nombre d'arêtes. Par conséquent, cet ensemble ne peut pas être directement utilisé pour caractériser une conformation. Nous considérons donc un sous-ensemble de $\mathcal{C}_G$ qui reste représentatif de tous les éléments de l'ensemble.

Ainsi, $\mathcal{C}_G$ constitue l'espace des cycles, un espace vectoriel, ce qui implique que les familles génératrices de $\mathcal{C}_G$ sont des choix pertinents pour caractériser les conformations. En effet, une famille génératrice d'un ensemble permet, par composition, d'atteindre tous les éléments dudit ensemble. Les bases de cycles sont des familles génératrices comprenant un nombre restreint de cycles.

Définition 11 (Base de cycles)**.** *Étant donné un graphe G, une base de cycles de l'espace vectoriel $\mathcal{C}_G$ est un ensemble de cycles linéairement indépendants qui constitue à la fois une famille libre et génératrice de $\mathcal{C}_G$.*

La dimension de l'espace $\mathcal{C}_G$ correspond au nombre de vecteurs dans une base de cet espace, elle se note μ_G. Cette valeur, aussi appelée nombre cyclomatique, vaut $|E| - |V| + x$, où x représente le nombre de composantes connexes du graphe $G = (V, E)$. La Figure 2.5a ci-dessous illustre une base de cycles sur un exemple.

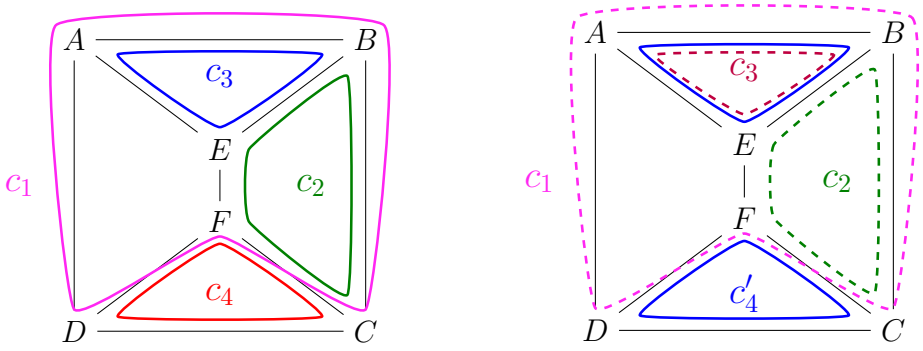
Comme discuté dans le Chapitre 1, les petits cycles sont souvent considérés comme plus aptes à représenter de manière précise la structure d'une molécule. Nous nous orientons donc vers les bases de cycles contenant de petits cycles pour mieux caractériser les conformations. Afin de donner la priorité à ces petits cycles, il est nécessaire de définir le poids d'une base de cycles en fonction du poids de ses cycles. Ceci nous amène à la définition suivante.

Définition 12 (Poids d'une base de cycles)**.** *Le poids d'une base de cycle $B = \{c_1, c_2, \ldots, c_{\mu_G}\}$, noté ω_B, est la somme des poids des cycles qui la composent : $\omega_B = \sum_{i=1}^{\mu_G} \omega(c_i)$.*

Définition 13 (Base de cycles minimum)**.** *Une base de cycles est qualifiée de minimum si son poids est minimal. Autrement dit, il n'existe pas de bases de cycles avec un poids plus faible.*

Les Figures 2.5 et 2.6 représentent des bases de cycles d'un même graphe, la Figure 2.5 illustre des bases de cycles quelconques, tandis que la Figure 2.6 illustre spécifiquement des bases de cycles minimum. Nous pouvons observer que toutes les bases de cycles de ces figures ont le même nombre de cycles car il s'agit d'un invariant de l'espace vectoriel associé. Ainsi, en minimisant la somme des poids des cycles qui composent la base, cela garantit l'obtention des cycles les plus légers, comme nous le verrons en détail au Chapitre 3. Étant donné la Définition 9 du poids d'un cycle dans le contexte non pondéré des conformations, cela revient donc à avoir les cycles les plus courts en nombre d'arêtes. Nous pouvons également observer avec la Figure 2.5 le poids supplémentaire induit par la présence d'un cycle non élementaire.

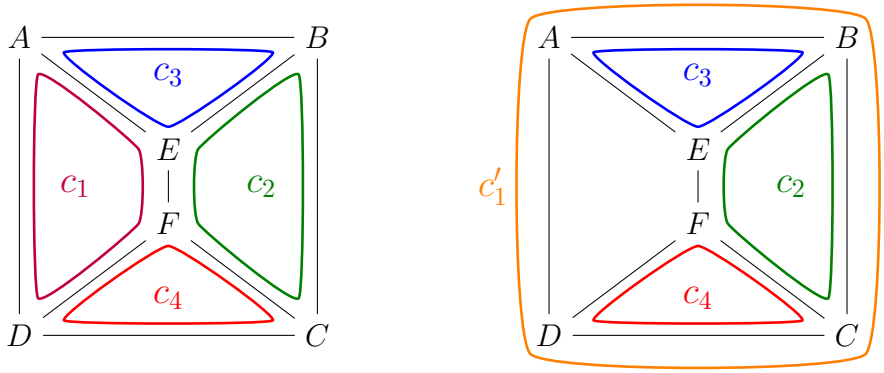
Bien que la Définition 8 propose une définition des cycles plus générale que celle des seuls cycles élementaires, les *bases de cycles de poids minimum* ne sont composées que de cycles élémentaires. Cela est dû au fait que les cycles élémentaires sont les cycles de poids les plus faibles, ce qui les rend préférables dans une base de cycles de poids minimum. Ainsi, nous exprimons ce résultat connu dans la Propriété 1.



(a) Une base b_1 de poids 15 contenant les cycles
c_1 $(A-B-C-F-D)$,
c_2 $(B-C-F-E)$,
c_3 $(A-B-E)$ et
c_4 $(C-D-F)$.

(b) Une base b_2 de poids 18 contenant les cycles
c_1 $(A-B-C-F-D)$,
c_2 $(B-C-F-E)$,
c_3 $(A-B-E)$ et
c'_4 $\big((A-B-E)\cup(C-F-D)\big)$.

Figure 2.5 – Étant donné un graphe G avec $\mu_G = 4$, les Figures 2.5a et 2.5b illustrent des bases de cycles de G quelconques. Les cycles c_1, c_2 et c_3 sont communs aux deux bases de cycles, ils sont indiqués en pointillés dans la Figure 2.5b pour que le nouveau cycle c'_4 soit bien visible. Contrairement aux autres cycles, le cycles c'_4 n'est pas un cycle élementaire, il est une union de cycles ainsi nous avons $c'_4 = c_3 \oplus c_4 = c_3 \cup c_4$.



(a) Une base b_3 contenant les cycles
c_1 (A-E-F-D), c_2 (B-C-F-E),
c_3 (A-B-E) et c_4 (C-D-F).

(b) Une base b_4 contenant les cycles
c'_1 (A-B-C-D), c_2 (B-C-F-E),
c_3 (A-B-E) et c_4 (C-D-F).

Figure 2.6 – Étant donné un graphe G avec $\mu_G = 4$, les Figures 2.6a et 2.6b illustrent des bases de cycles minimum de G; Ainsi, chacune contient deux cycles de poids 3 et deux cycles de poids 4 pour un poids total de 14.

Propriété 1. *Tous les cycles d'une base de cycles minimum sont des cycles élémentaires.*

Démonstration. Supposons qu'une base de cycles minimum B de l'ensemble $\mathcal{C}$ contienne un cycle non élémentaire, noté c. Alors, il existe $c_1 \in \mathcal{C}$ et $c_2 \in \mathcal{C}$, deux cycles non nuls et disjoints tels que $c = c_1 \oplus c_2 = c_1 \cup c_2$.

c_1 et c_2 ne peuvent pas tous les deux appartenir à B comme c, c_1 et c_2 ne sont pas linéairement indépendant. Supposons, sans perte de généralité, que $c_1 \notin B$.

Étant donné que B est une base il existe un sous ensemble de cycles $D \subset B$ tel que

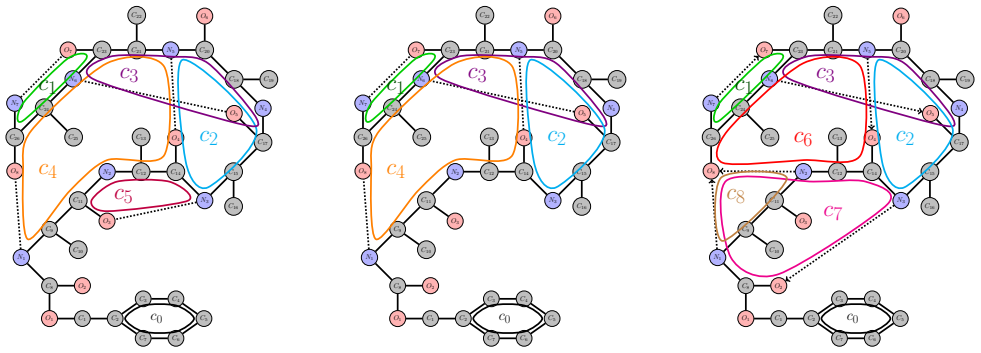
$c \oplus \bigoplus_{d \in D} d = c_1$. Puisque $(\mathcal{C}, \oplus)$ est un groupe commutatif, on a $c_1 \oplus \bigoplus_{d \in D} d = c$. Cela implique l'existence d'une base $B' = B \cup \{c_1\} \setminus \{c\}$ telle que $\omega_{B'} = \omega_B - \omega(c) + \omega(c_1)$. Par définition, puisque c n'est pas un cycle élémentaire, on a $\omega(c) > \omega(c_1)$ ce qui implique, $\omega_{B'} < \omega_B$.

Cependant, il ne peut exister une telle base B' car B est par définition une base minimum. Ainsi, les cycles d'une base de cycles minimum sont des cycles élémentaires. □

Comme illustré dans les Figures 2.6a et 2.6b, un graphe G peut avoir plusieurs bases de cycles minimum. L'ensemble de toutes ces bases de cycles minimum est noté $\mathcal{MCB}(G)$. Parmi toutes les bases minimum disponibles, une seule est choisie pour caractériser une conformation. Plusieurs paramètres peuvent orienter ce choix. Cet aspect, ainsi que l'impact sur la méthode, seront abordés dans le Chapitre 3.

Pour l'instant, pour chaque conformation G_i, une base de cycles minimum, notée B_i, est choisie arbitrairement parmi $\mathcal{MCB}(G_i)$.

Revenons à la trajectoire illustrée dans la Figure 2.2 qui est composée de trois conformations distinctes. La Figure 2.7 présente les bases de cycles minimums sélectionnées pour chacune de ces conformations.



(a) Graphe G_1. (b) Graphe G_2. (c) Graphe G_3.

Figure 2.7 – Illustration de la base de cycles minimum sélectionnée pour chacune des conformations de l'exemple simplifié de la Figure 2.2. Les cycles sont étiquetés par leurs liaisons hydrogène et sont représentés par la même couleur s'ils sont identiques. Les cycles sont numérotés afin d'alléger les figures les cycles sont numérotés. Les liaisons hydrogènes présentes dans ces cycles sont :
Le cycle c_0 (noir) ne contient aucune liaison.
Le cycle c_1 (vert) : N_7-O_7.
Le cycle c_2 (bleu) : N_5-O_4.
Le cycle c_3 (violet) : N_6-O_5.
Le cycle c_4 (orange) : N_1-O_8, N_5-O_4.
Le cycle c_5 (rose foncé) : N_3-O_3.
Le cycle c_6 (rouge) : N_2-O_8, N_5-O_4.
Le cycle c_7 (rose) : N_1-O_8, N_2-O_8, N_3-O_2.
Le cycle c_8 (marron) : N_1-O_8, N_2-O_8.

2.2.2 . Caractérisation d'une conformation par ses cycles

La base de cycles minimum d'une conformation est une caractérisation efficace de sa structure. Elle se compose d'un ensemble restreint de petits cycles à partir desquels

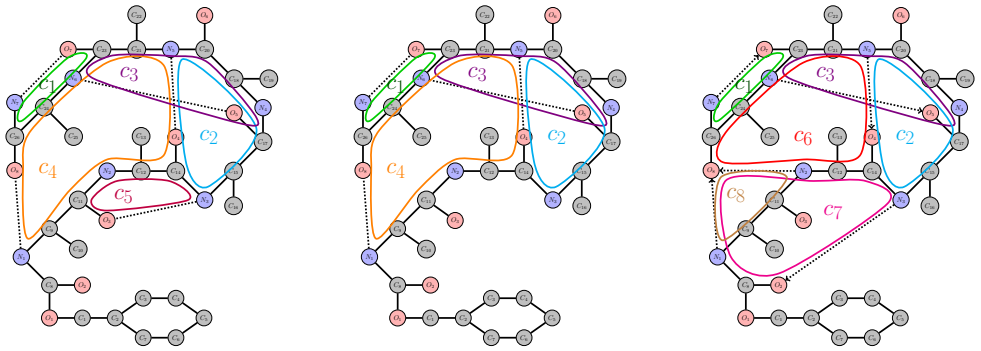
tous les cycles de la conformation peuvent être générés par composition. De plus, comme nous le verrons dans le Chapitre 3, son calcul est peu coûteux. Elle est donc facile à utiliser dans le cadre du traitement de trajectoires contenant plusieurs conformations.

Dans le contexte de l'analyse de la dynamique moléculaire, tous les éléments de la structure ne sont pas pertinents. En particulier, les cycles formés uniquement par des liaisons covalentes ne nous intéressent pas car ils sont immuables, et ne contribuent pas à la dynamique de la molécule. Bien qu'ils soient présents dans la structure moléculaire, ils ne sont pas représentatifs de sa dynamique. Ainsi, après avoir initialement pris en compte tous les cycles pour définir une base de cycles réprésentative de tous les cycles du graphe, une sélection est effectuée au sein de cette base pour ne conserver que ceux qui participent réellement à la dynamique structurelle de la conformation.

Notre attention se porte sur les cycles qui sont impliqués dans la dynamique structurelle de la trajectoire, ou qui la représentent de manière significative.

Définition 14 (Ensemble de cycles dynamiques d'une conformation)**.** *Étant donné une conformation G_i, son ensemble de cycles dynamiques est un ensemble de cycles B_i^* tel qu'il existe une base de cycles minimum $B_i \in \mathcal{MCB}(G_i)$ avec $B_i^* \subseteq B_i$ et $\forall c \in B_i^*, c \cap H_i \neq \emptyset$. En d'autres termes, il s'agit d'un sous-ensemble d'une base de cycles minimum dont tous les cycles contiennent au moins une arête représentant une liaison hydrogène.*

La Figure 2.8 présente l'ensemble des cycles sélectionnés pour chacune des conformations de l'exemple simplifié. Cette représentation prend en compte les bases de cycles choisies et illustrées précédemment dans la Figure 2.7. Nous observons que le cycle c_0, qui ne comprenait que des liaisons covalentes, n'est plus pris en compte à présent.



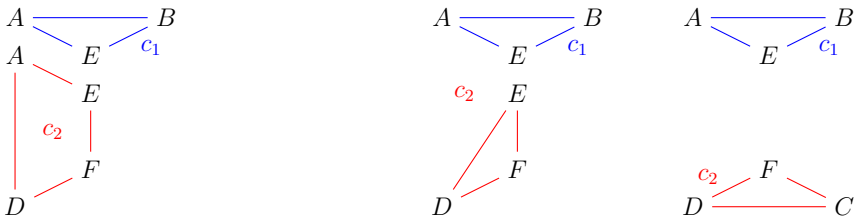
(a) Graphe G_1. (b) Graphe G_2. (c) Graphe G_3.

Figure 2.8 – Illustration de l'ensemble de cycles dynamiques pour chacune des conformations présentées dans la Figure 2.2. Les cycles sont étiquetés et colorés comme ils l'étaient dans la Figure 2.7.

Les cycles, bien que pertinents pour représenter la structure d'une conformation, ne sont pas utilisés seuls. En effet, les intéractions entre les cycles sont tout aussi importantes que les cycles eux-mêmes.

Définition 15 (Intéraction de deux cycles)**.** *Nous disons de deux cycles, c_1 et c_2, qu'ils intéragissent s'ils partagent au moins une arête, c'est-à-dire si $c_1 \cap c_2 \neq \emptyset$.*

La Figure 2.9 ci-dessous illustre cette définition de l'intéraction entre deux cycles à travers plusieurs exemples. Nous observons bien, dans la Figure 2.9b, que si deux cycles partagent uniquement des sommets et pas d'arêtes alors ils n'interagissent pas.

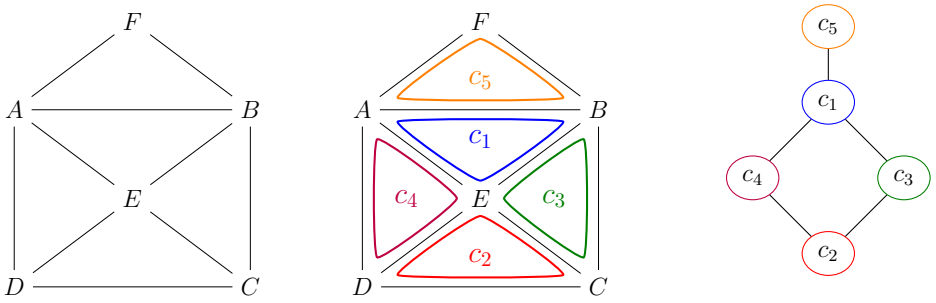


(a) Exemple où c_1 et c_2 intéragissent. (b) Deux exemples où c_1 et c_2 n'intéragissent pas.

Figure 2.9 – Illustrations de couples de cycles avec et sans intéractions.

Les intéractions entre les cycles modélisent des contraintes fortes sur la structure moléculaire. Afin de caractériser efficacement la structure et donc les liens entre les atomes au sein d'une conformation, il est nécessaire d'associer les cycles et leurs intéractions. Pour ce faire, nous représentons la structure d'une conformation par un graphe appelé le graphe des cycles. Dans ce graphe, les intéractions entre les cycles constituent ses arêtes.

Définition 16. *Un graphe de cycles est un graphe $GC = (C, E_C)$ où C est un ensemble de cycles, et dans lequel il existe une arête entre deux cycles de C si ces cycles interagissent. En d'autres termes, pour $c_1, c_2 \in C$, nous avons $[c_1, c_2] \in E_C$ si et seulement si $c_1 \cap c_2 \neq \emptyset$.*



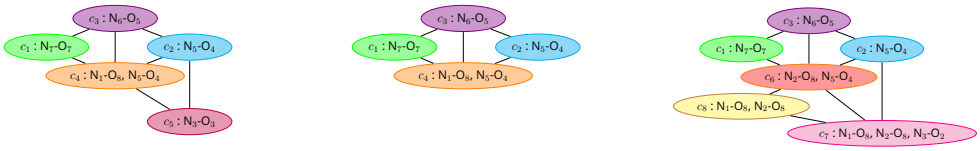
(a) Un graphe G (b) Un ensemble de cycles C (c) Le graphe de cycles GC

Figure 2.10 – Exemple d'un graphe de cycles construit à partir d'un ensemble de cycles. La Figure 2.10b illustre un ensemble de cycles C du graphe G de la Figure 2.10a. La Figure 2.10c illustre quant-à elle le graphe de cycles obtenu à partir de l'ensemble C. En accord avec la Définition 15, le graphe de cycles de la Figure 2.10c ne contient pas d'arête $[c_3, c_4]$ car c_3 et c_4 ont uniquement un sommet en commun et non une arête.

Prenons une conformation G_i avec son ensemble de cycles dynamiques B_i^*, nous définissons le graphe des cycles de la conformation, noté GC_i, où B_i^* constitue l'ensemble de sommets.

La Figure 2.11 présente les graphes de cycles construits pour chacune des conformations de l'exemple simplifié. Ces graphes ont été obtenus à partir des cycles sélectionnés et illustrés précédemment dans la Figure 2.8.

Le graphe de cycles représente une caractérisation de la structure d'une conformation. Ainsi, au lieu d'une suite de conformations, nous considérons la suite de graphes de cycles correspondants, notée $\mathcal{GC}$. Cette suite de graphes de cycles permet de mettre en évidence la dynamique de la molécule tout au long d'une trajectoire. La Figure 2.12 illustre, à travers un schéma, les étapes de cette section permettant la caractérisation de chacune des conformations au sein d'une trajectoire.



(a) Graphe de cycles GC_1. (b) Graphe de cycles GC_2. (c) Graphe de cycles GC_3.

Figure 2.11 – Illustration des graphes de cycles obtenus pour chaque conformation de l'exemple simplifié de la Figure 2.2. Les cycles sont étiquetés par leurs liaisons hydrogène et leur couleur correspond à celle de la Figure 2.8.

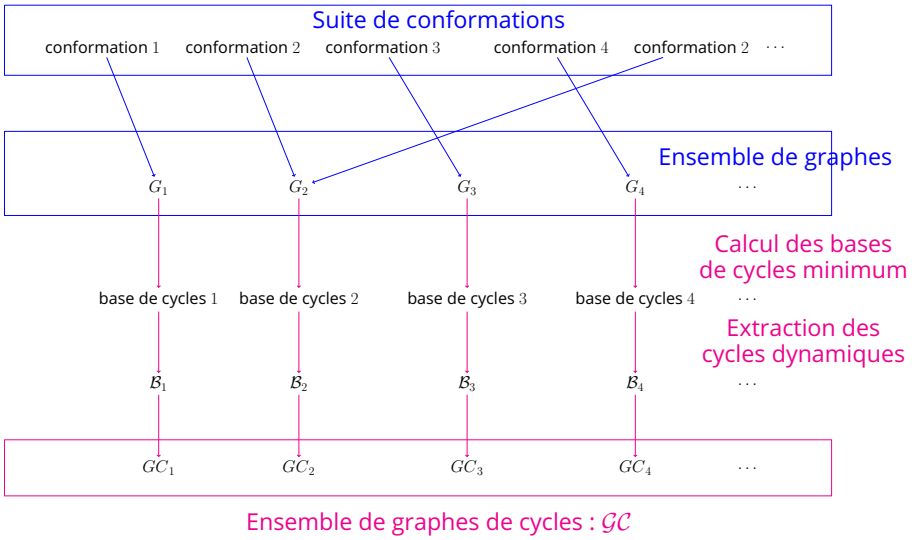


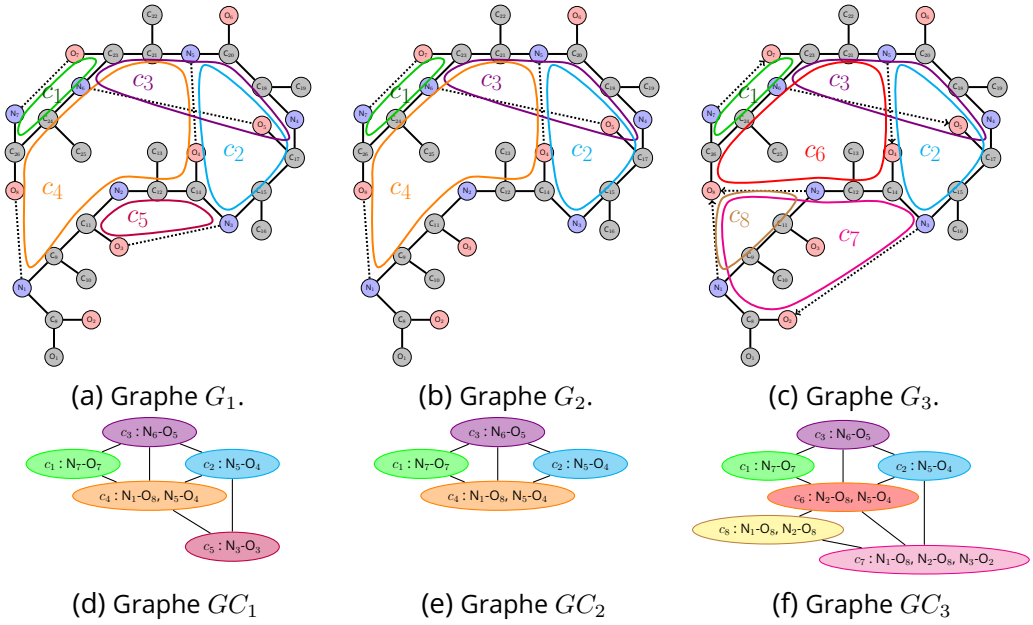
Figure 2.12 – Schéma du passage d'un ensemble de conformations à un ensemble de graphes de cycles.

2.3 . Caractérisation de la structure d'une dynamique moléculaire

Les cycles apparaissent et disparaissent au cours du temps, entrainant avec eux des contraintes structurelles qui se font et se défont. La Figure 2.13 reprend les cycles et les graphes de cycles qui constituent la caractérisation des conformations de l'exemple hypothétique G_1, G_2, G_1, G_3 que nous avons suivi. Nous pouvons y observer un autre type d'évolution des cycles. Les trois graphes présentent un cycle central, mais qui diffère par ses liaisons hydrogène. Le cycle c_4 (N_1-O_8, N_5-O_4 en orange) apparaît dans les Figures 2.13d et 2.13e, puis se transforme pour devenir le cycle c_6 (N_2-O_8, N_5-O_4 en rouge) dans la Figure 2.13f. Une interprétation de ce phénomène est que ce cycle central est primordial pour la structure de la molécule. Ainsi, les cycles c_4 et c_6 peuvent représenter la même contrainte sur la structure, mais leur évolution dépend des autres cycles présents et, par conséquent, des autres contraintes de la molécule.

Un exemple inverse est illustré par le couple de cycles c_4 et c_8. Bien qu'ils présentent des similitudes, c_4 est un cycle central avec de nombreux voisins, tandis que c_8 est un petit cycle très contraint avec peu de voisins. Ces deux cycles semblent donc représenter des contraintes différentes plutôt que la même contrainte qui aurait évolué.

Dans la suite, nous introduisons la notion de polymorphisme des cycles qui capture



(a) Graphe G_1. (b) Graphe G_2. (c) Graphe G_3.

(d) Graphe GC_1 (e) Graphe GC_2 (f) Graphe GC_3

Figure 2.13 – Illustration des informations connues après caractérisation des conformations de la trajectoire illustrée dans la Figure 2.2. Les graphes des Figures 2.13a, 2.13b et 2.13c ont été tronqués pour représenter uniquement les zones avec des liaisons hydrogène.

la capacité des cycles à changer de forme tout en maintenant leur rôle au sein de la structure de la molécule. À partir de ce modèle, nous définissons le polygraphe, un graphe représentatif des contraintes qui coexistent au cours de la trajectoire. Le polygraphe est un outil puissant qui caractérise les contraintes présentes au cours de la trajectoire de dynamique moléculaire. Le polygraphe n'est pas simple à construire et le problème associé à sa construction sera traité dans le Chapitre 4. Enfin, comme nous le verrons à la fin de cette section et plus en détail dans le Chapitre 6, le polygraphe permet d'analyser et de comparer différentes trajectoires d'un même système moléculaire.

2.3.1 . Polymorphisme de cycles

Dans cette Section, nous introduisons donc la notion d'évolution ou de *polymorphisme des cycles*. Cette notion caractérise les variations des cycles qui influent sur la structure. Un ensemble de *cycles polymorphes*, ou **polycycle**, est un regroupement de cycles représentant la même contrainte structurelle. Autrement dit, c'est un ensemble de cycles différents qui interagissent de la même façon dans les différents graphes de cycles.

En connaissant les cycles présents dans au moins un graphe de cycles de la trajectoire, et nous allons rechercher un partitionnement en ensembles de cycles, chacun représentant une contrainte. Ainsi, nous avons l'ensemble des cycles de la trajectoire $\mathcal{C} = \bigcup_{i=0}^{N} B_i^*$, et nous considérons une partition $\mathcal{P}$ de $\mathcal{C}$. Dans la suite, nous présentons les trois propriétés qu'une de ces parties doit vérifier pour être identifiée comme un ensemble de cycles polymorphes au sein de ces familles. Pour illustrer nos explications, nous utiliserons la Figure 2.13.

Nous observons que les graphes partagent un certain nombre de cycles en commun, mais ils ont également des cycles qui leur sont propres. Comme nous l'avons déjà évoqué,

par exemple, le cycle c_4 est commun aux graphes G_1 et G_2 (Figures 2.13a et 2.13b), mais il n'est pas présent dans le graphe G_3 où il est "remplacé" par le cycle c_6. Bien que ces cycles soient différents, ils occupent la même position dans les graphes de cycles. Ainsi, ils ont un rôle similaire dans la structure moléculaire et semblent interchangeables. Cela les caractérise en tant que cycles polymorphes.

Nous avons défini, ci-dessous, les propriétés 2, 3, et 4, qui vérifient l'interchangeabilité potentielle des cycles, la similarité de leurs voisinages et la ressemblance de leurs compositions. Ces propriétés, que nous avons appelées propriétés du polymorphisme, permettent de reconnaitre les ensembles de cycles polymorphes. À la fin de cette section, nous présenterons une définition formelle d'un ensemble de cycles polymorphes au sein de cette partition appuyée sur ces propriétés.

Notation 1. *Étant donné un cycle c, $\mathcal{GC}_c$ représente l'ensemble des graphes de cycles qui incluent c parmi leurs sommets. Dans le cas d'un ensemble de cycles $D = \{d_1, d_2, \dots d_q\}$, $\mathcal{GC}_D$ désigne l'ensemble des graphes de cycles qui incluent l'un des cycles de D parmi leurs sommets, soit $\mathcal{GC}_D = \bigcup_{d \in D} \mathcal{GC}_d$.*

Propriété 2 (Absence de coexistence des cycles)**.** *Étant donné un ensemble de cycles $\{c_1, c_2, \dots, c_q\}$, pour tout $i, j \in [1, q]$ avec $i \neq j$, les cycles c_i et c_j ne coexistent dans aucun graphe de cycles de $\mathcal{GC}$. Autrement dit, $\mathcal{GC}_{c_i} \cap \mathcal{GC}_{c_j} = \emptyset$.*

Les cycles d'un ensemble de cycles polymorphes doivent être interchangeables pour la structure moléculaire. Cependant, s'ils apparaissent simultanément, cela n'est pas possible car un cycle ne peut pas en remplacer un autre s'il est déjà présent dans le graphe. Cela revient à considérer chaque cycle comme une contrainte structurelle. Ainsi, un polycycle représente une même contrainte structurelle à travers chacun de ses cycles. De plus, une même contrainte structurelle ne peut pas être représentée deux fois dans une conformation.

Prenons l'exemple des cycles c_2 et c_5, tous deux présents dans la Figure 2.13d. Étant donné que ces cycles coexistent dans le graphe de cycles GC_1, ils ne peuvent pas se remplacer. Considérons maintenant le cycle c_7 présent dans le graphe de cycles GC_3 de la Figure 2.13f, et absent des graphes de cycles GC_1 et GC_2. Dans le graphe GC_1 de la Figure 2.13d, nous avons le cycle c_4 qui par contre n'est pas présent dans GC_3. Ainsi, c_7 ne coexiste pas avec le cycle c_4. L'ensemble de cycles $\{c_4, c_7\}$ vérifie donc la Propriété 2. De manière similaire, nous pouvons conclure que $\{c_5, c_7\}$ vérifie également la Propriété 2, car c_5 n'appartient qu'à GC_1.

La Propriété 2 vérifie si les cycles ne coexistent pas dans un graphe de cycles, et que donc ils sont intérchangeable. Les prochaines propriétés vont établir si cette interchangeabilité est valide structurellement. En effet, la seconde propriété concerne la compatibilité des voisinages des cycles dans les différents graphes où ils apparaissent. Même si les cycles polymorphes ne coexistent pas ensemble, ils coexistent avec d'autres cycles. Or, il est nécessaire que les cycles polymorphes intéragissent de la même façon.

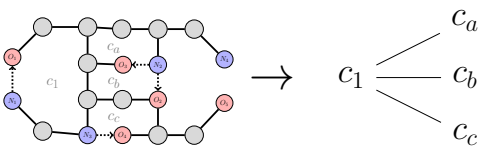
Propriété 3 (Intéractions identiques)**.** *Étant donnés un ensemble de cycles $\{c_1, c_2, \dots, c_q\}$ et une partition $\mathcal{P}$, pour tout graphe de cycles GC de $\mathcal{GC}_{\{c_1, c_2, \dots, c_q\}}$ dans lequel il existe au moins un sommet d n'appartenant pas à $\{c_1, c_2, \dots, c_q\}$ donc d appartient à une partie D de $\mathcal{P} \setminus \{c_1, c_2, \dots, c_q\}$, et notons c un cycle de $\{c_1, c_2, \dots, c_q\}$:*

- *si* $[c, d]$ *est une arête de* GC *alors pour tout graphe de cycles* $GC' \in \mathcal{GC}_D \cap \mathcal{GC}_{\{c_1,c_2,\ldots,c_q\}}$, *il existe une arête* $[c', d']$ *pour tout couple de cycles appartenant à* GC' *tels que* $c' \in \{c_1, c_2, \ldots, c_q\}$ *et* $d' \in D$.
- *si* $[c, d]$ *n'est pas une arête de* GC *alors pour tout graphe de cycles* $GC' \in \mathcal{GC}_D \cap \mathcal{GC}_{\{c_1,c_2,\ldots,c_q\}}$, *il n'existe aucune arête* $[c', d']$ *dans* GC' *telle que* $c' \in \{c_1, c_2, \ldots, c_q\}$ *et* $d' \in D$.

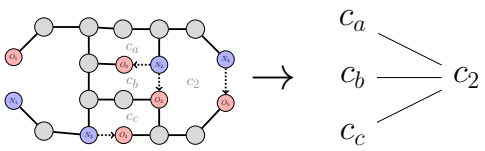
Remarque 5. *La formulation de cette propriété permet d'assurer que les voisinages de tous les cycles de* $\{c_1, c_2, \ldots, c_q\}$ *sont similaires. En pratique, puisque les trois propriétés du polymorphisme doivent être vérifiées par l'ensemble* $\{c_1, c_2, \ldots, c_q\}$, *la Propriété 2 nous assure qu'il n'existe qu'un seul cycle de* $\{c_1, c_2, \ldots, c_q\}$*dans chaque graphe de cycles de* $\mathcal{GC}_{\{c_1,c_2,\ldots,c_q\}}$.

Reprenons l'exemple de la Figure 2.13 et du cycle c_7 du graphe de cycles GC_3 pour illustrer cette propriété. À partir de la Propriété 2, nous avons identifié $\{c_4, c_7\}$ et $\{c_5, c_7\}$ comme des polycycles potentiels. Dans le graphe GC_3 de la Figure 2.13f, le cycle c_7 coexiste avec les cycles c_1, c_2, c_3, c_5, et c_8 mais il n'interagit pas avec les cycles c_1 et c_3. Dans les graphes GC_1 et GC_2, le cycle c_4 interagit avec c_1. Nous pouvons donc conclure que les voisinages des cycles c_4 et c_7 ne sont pas compatibles, et que l'ensemble $\{c_4, c_7\}$ ne vérifie pas la Propriété 3. En revanche, l'ensemble $\{c_5, c_7\}$ vérifie bien la propriété d'interactions identiques. Ils sont tous deux voisins de c_2 et n'interagissent pas avec c_1 et c_3 également présents dans GC_1 et dans GC_3.

Remarque 6. *Dans ce paragraphe, nous supposons que aucun polycycle n'a déjà été établi. Néanmoins, nous pouvons remarquer que nos conclusions sur le voisinage de l'ensemble* $\{c_5, c_7\}$ *sont valides même si comme nous l'avons dit en introduction de cette section que* c_4 *et* c_6 *sont polymorphes.*



(a) Un graphe de cycles GC obtenu à partir d'un graphe G. Les sommets du graphe GC sont les cycles : c_1, c_a, c_b et c_c.



(b) Un graphe de cycles GC' obtenu à partir d'un graphe G'. Les sommets du graphe GC' sont les cycles : c_2, c_a, c_b et c_c.

Figure 2.14 – Illustration d'un cas où $\{c_1, c_2\}$ vérifie les propriétés 2 et 3, sans pour autant que les cycles c_1 et c_2 n'aient d'atome en commun. Le graphe de cycles GC, respectivement GC', est obtenu en considérant pour sommets l'ensemble des faces internes de G, respectivement G'. Notons que dans ces exemples, l'ensemble des faces internes constitue bien un ensemble de cycles caractéristiques tel que nous l'avons défini.

Les propriétés 2 et 3 vérifient la capacité d'interchangeabilité d'un ensemble de cycles. Cependant, elles ne garantissent pas qu'une quelconque ressemblance atomique existe entre ces cycles. Même si la plupart des ensembles qui vérifient les propriétés ont des similitudes dans la composition atomique de leurs cycles, il peut exister des cas particuliers où cela n'est pas vérifié. La Figure 2.14 illustre un de ces cas particuliers. Dans cet exemple, les cycles c_1 et c_2 ont des positions symétriques dans les graphes de cycles, ce qui leur permet de vérifier les propriétés précédentes. Cependant, étant très éloignés dans les

graphes de départ, ils ne partagent aucun atome. Ces cycles ne sont donc pas localisés dans la même zone du graphe, et par la même de la molécule et ne représentent donc pas la même contrainte structurelle.

Nous présentons alors une troisième propriété pour vérifier que les cycles ne sont pas trop éloignés dans le graphe, et donc qu'ils peuvent constituer un même polycyle.

Notation 2. *Étant donné l'ensemble d'arêtes H correspondant ici aux liaisons hydrogène, notons $V(H)$ l'ensemble des sommets extrémités d'au moins une arête de H.*

Propriété 4 (Ancrage dans la structure dynamique)**.** *Étant donné un ensemble de cycles $\{c_1, c_2, \ldots, c_q\}$, tous ces cycles partagent au moins un sommet impliqué dans les liaisons variables. Ainsi, nous avons $V(c_1 \cap c_2 \cap \ldots \cap c_q) \cap V(H) \neq \emptyset$.*

Cette propriété vérifie qu'il existe au moins un point d'ancrage correspondant à un atome de $V(H)$ partagé par tous les cycles d'un ensemble. Nous pouvons à présent exprimer la définition d'un ensemble de cycles polymorphes à partir des propriétés que nous avons décrites.

Définition 17 (Ensemble de cycles polymorphes, polycycle)**.** *Étant donnés un ensemble graphes de cycles $\mathcal{GC}$, un ensemble d'arêtes H et une partition $\mathcal{P}$ de $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$ où B_i^* est l'ensemble des cycles présents dans $GC_i \in \mathcal{GC}$. Un ensemble de cycles $\{c_1, c_2, \ldots, c_q\}$ est un ensemble de cycles polymorphes s'il vérifie simultanément les trois propriétés :*

(Prop. 2) $\forall i, j \in [1, q]$ *avec* $i \neq j$, $\mathcal{GC}_{c_i} \cap \mathcal{GC}_{c_j} = \emptyset$.

(Prop. 3) Pour toute partie $D \in \mathcal{P} \backslash \{c_1, c_2, \ldots, c_q\}$, $\forall (GC_a, GC_b) \in (\mathcal{GC}_{\{c_1, c_2, \ldots c_q\}} \cap \mathcal{GC}_D)^2$, *notons* c_a, *respectivement* c_b, *le cycle de* $\{c_1, c_2, \ldots, c_q\}$ *présent dans* GC_a, *respectivement* GC_b, *nous avons* $\forall d \in D$ *:*

- *si* $[c_a, d] \in GC_a$ *alors* $[c_b, d'] \in GC_b$ *avec* $d' \in D$.
- *si* $[c_a, d] \notin GC_a$ *alors* $[c_b, d'] \notin GC_b$ *avec* $d' \in D$.

(Prop. 4) $V(\bigcap_{i=1}^{P} c_i) \cap V(H) \neq \emptyset$

Un polycycle est donc un ensemble de cycles particulier.

Définition 18 (Identité d'un polycycle)**.** *Étant donné un polycycle p, une* identité *de p est un cycle $c \in p$.*

2.3.2 . Polygraphe d'une trajectoire

Nous souhaitons établir les ensembles de cycles polymorphes dans la trajectoire afin qu'ils caractérisent la structure générale de la molécule. Ainsi, nous voulons définir une partition $\mathcal{P}$ de $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$ où B_i^* est l'ensemble des cycles présents dans le graphe de cycles $GC_i \in \mathcal{GC}$, dans laquelle chaque partie $p \in \mathcal{P}$ est un ensemble de cycles polymorphes tel que décrit par la Définition 17.

Inspiré par ce que nous avons fait pour représenter la structure d'une conformation, nous cherchons à minimiser le nombre d'ensembles de cycles polymorphes qui représentent la structure de la trajectoire. Nous avons formalisé ce problème comme le problème min-Partition de Cycles Polymorphes (min-PCP), et il constitue avec le problème de décision qui lui est associé l'objet du Chapitre 4.

La Figure 2.15 illustre ce que nous verrons dans cette section : l'obtention d'un polygraphe à partir d'un ensemble de graphes de cycles. Cette figure, est la suite de la Figure 2.12 qui aboutissait à l'obtention d'un ensemble de graphes de cycles. Ici, nous présentons comment à partir de cet ensemble nous obtenons le polygraphe d'une trajectoire de dynamique moléculaire.

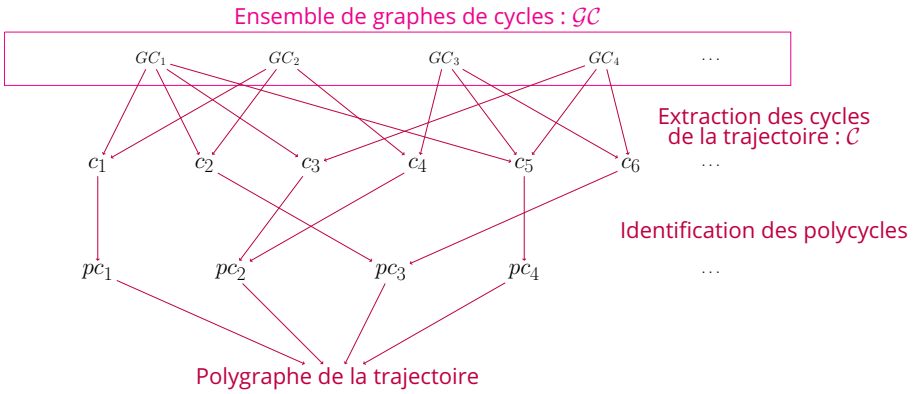


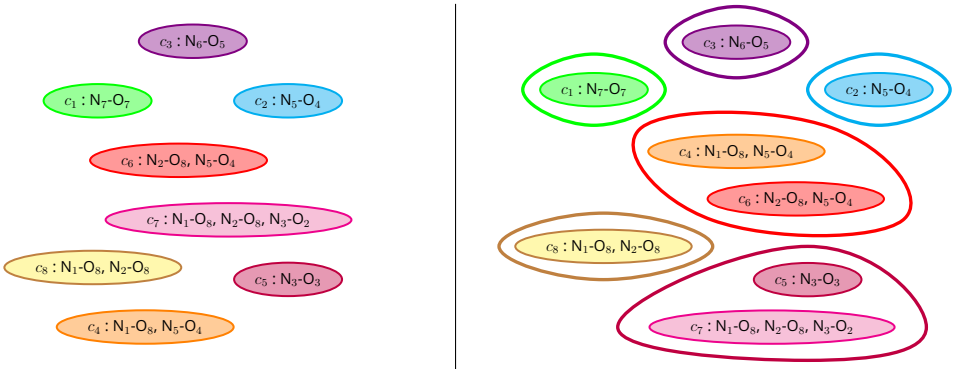
Figure 2.15 – Schéma de la construction du polygraphe à partir d'un ensemble de graphes de cycles.

Afin de caractériser la structure globale d'une molécule au cours d'une dynamique moléculaire, nous partons des conformations qui la composent. Nous savons que la structure de chacune de ces conformations est déjà caractérisée par un graphe de cycles et que les sommets de ces graphes de cycles forment l'ensemble des cycles de la trajectoire. En effet, si ces cycles sont des sommets dans un ou plusieurs graphes de cycles, c'est qu'ils jouent un rôle dans la structure. Nous disposons donc d'un ensemble de cycles caractéristique de la structure de la molécule. C'est à partir de cet ensemble, noté $\mathcal{C}$ que nous allons construire les polycycles de la trajectoire.

Reprenons notre exemple avec la Figure 2.16 qui illustre ce regroupement en polycycles à partir des cycles caractéristiques. Comme nous l'avons déjà évoqué, le cycle c_5 et le cycle c_7 sont identifiés comme polymorphes et ils constituent donc un polycycle, tout comme le cycle c_4 et le cycle c_6. En outre, la figure met en évidence plusieurs ensembles de cycles polymorphes, composés chacun d'un unique cycle. Ces ensembles, désignés comme des *singletons*, indiquent que, pour ces cycles particuliers, aucun équivalent n'a été identifié au cours de la trajectoire moléculaire. Cette observation souligne le rôle spécifique que jouent certains cycles dans la structure.

Suite à l'identification de ces ensembles de cycles équivalents, nous souhaitons obtenir une représentation concrète des liens entre les atomes de la molécule qui définissent sa structure. Nous allons donc construire un graphe des polymorphismes, ou polygraphe, directement inspiré des graphes de cycles. Dans ce polygraphe, les sommets sont les ensembles de cycles polymorphes, et une arête existe entre deux ensembles si, et seulement si, il existe au moins un graphe de cycles dans lequel des éléments des ensembles correspondants sont voisins. Le polygraphe est alors une synthèse des graphes de cycles possibles grâce au polymorphisme.

Définition 19 (Un graphe des polymorphismes ou polygraphe)**.** *Étant donnés un ensemble de graphes de cycles $\mathcal{GC}$ et une partition en polycycles $\mathcal{P}$, un polygraphe est un graphe $GP =$*



(a) L'ensemble des cycles de la trajectoire. (b) Les ensembles de cycles polymorphes.

Figure 2.16 – Illustration des regroupements de cycles polymorphes obtenus à partir des graphes de cycles présentés dans la Figure 2.13.

$(\mathcal{P}, I)$ dans lequel pour $p, p' \in \mathcal{P}$, il existe une arête $[p, p'] \in I$ si et seulement si il existe $GC \in \mathcal{GC}$ avec une arête $[c, c']$ tel que c est une identité de p et c' est une identité de p'.

La Figure 2.17 illustre concrètement cette idée, présentant le polygraphe généré à partir des ensembles de cycles polymorphes identifiés dans les graphes de cycles de l'exemple simplifié. Ce polygraphe propose une vue synthétique des variations structurales et des relations entre les cycles au cours de la trajectoire.

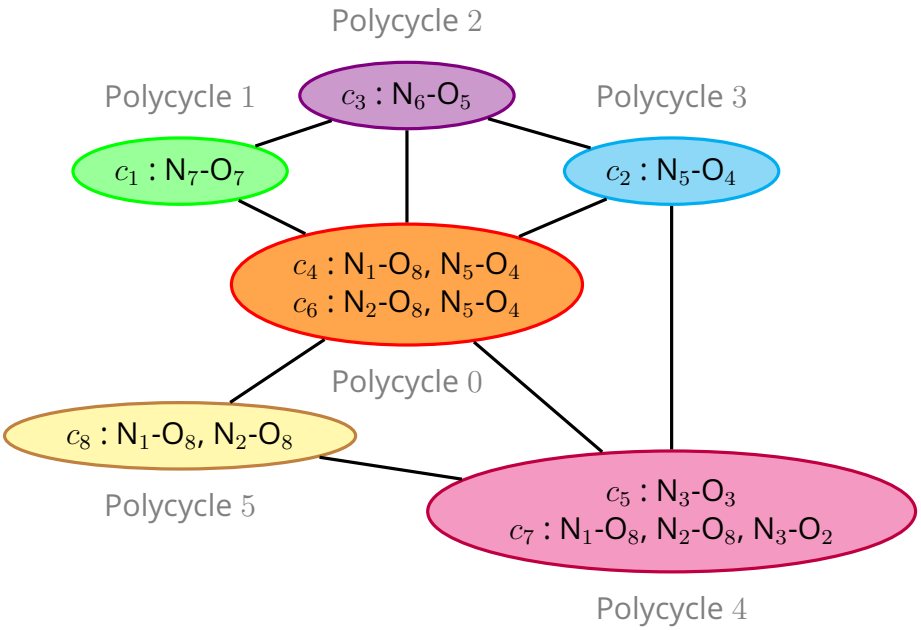


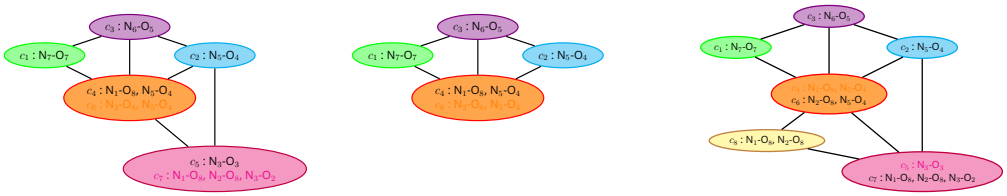
Figure 2.17 – Illustration du polygraphe obtenu à partir des graphes de cycles présentés dans la Figure 2.13. Les ensembles de cycles polymorphes correspondant aux sommets sont présentés dans la Figure 2.16.

Le polygraphe offre une caractérisation de l'évolution de la structure d'une molécule au cours d'une dynamique moléculaire.

2.4 . Analyse de trajectoires de dynamique moléculaire

Dans cette section, nous explorons un nouvel angle d'étude des trajectoires de dynamique moléculaire, accessible grâce au polygraphe.

Considérons le polygraphe P obtenu et les sous-polygraphes correspondants aux conformations de la trajectoire. Pour chaque conformation G_i de la trajectoire, nous construisons GP_i, un sous-graphe de P isomorphe au graphe de cycles GC_i. La Figure 2.17 illustre cette idée en présentant les sous-polygraphes correspondants aux graphes de cycles de l'exemple que nous avons suivi. Notons que le graphe GP_2 est un sous-graphe du graphe GP_1, lequel est lui-même un sous-graphe du graphe GP_3. Ces inclusions soulignent leur appartenance à une même structure sous-jacente. Ainsi, les conformations G_1, G_2 et G_3 de départ appartiennent à une seule méta-structure d'intérêt. Cette vision par le prisme du polygraphe permet d'identifier des motifs structuraux communs malgré des différences dans les graphes de cycles des conformations.



(a) Sous-polygraphe GP_1. (b) Sous-polygraphe GP_2. (c) Sous-polygraphe GP_3.

Figure 2.18 – Illustration des sous-graphes, du polygraphe de la Figure 2.17, isomorphes aux graphes de cycles de la Figure 2.8.

Le polygraphe ainsi que les sous-polygraphes permettent déjà de caractériser la trajectoire et ses conformations. Néanmoins, comme discuté dans le Chapitre 1, l'un des objectifs majeurs de l'analyse est la détermination de bassins conformationnels ou à défaut de méta-structures d'intérêts. Pour y parvenir, nous replaçons les conformations dans le temps de la trajectoire. Cela nous permet d'observer les apparitions et les disparitions des polycycles dans le temps. Ainsi, nous identifions des périodes au cours desquelles l'ensemble de polycycles présents reste le même, et chacun de ces ensembles peut éventuellement être associé à un bassin conformationnel.

La Figure 2.19 illustre cette vision avec un schéma. Dans cette représentation, nous partons de la trajectoire initialement présentée dans la l'exemple de la Figure 2.2, que nous étendons avec quelques conformations supplémentaires (G'_1, G''_1, G'_2 et G'_3), tout en supposant que le polygraphe de la trajectoire, illustré dans la Figure 2.17, reste inchangé. Sur ce schéma, nous notons deux périodes particulières, dénotées $S1$ et $S2$. La période $S1$ correspond à une période où tous les polycycles sont présents en majorité, ce qui pourrait éventuellement être associé à un bassin conformationnel. En revanche, la période $S2$ correspond à l'absence des polycycles 4 et 5, indiquant une structure plus flexible de la molécule car moins sujette aux contraintes.

Bien que le temps associé à chacune des conformations ne soit pas représenté sur le schéma de la Figure 2.19, il s'agit d'un élément essentiel pour déterminer les périodes d'intérêt. Nous aborderons cet aspect dans le Chapitre 6.

La combinaison du polygraphe et de cette vue des polycycles dans le temps offre une approche puissante pour l'analyse des trajectoires de dynamique moléculaire. En

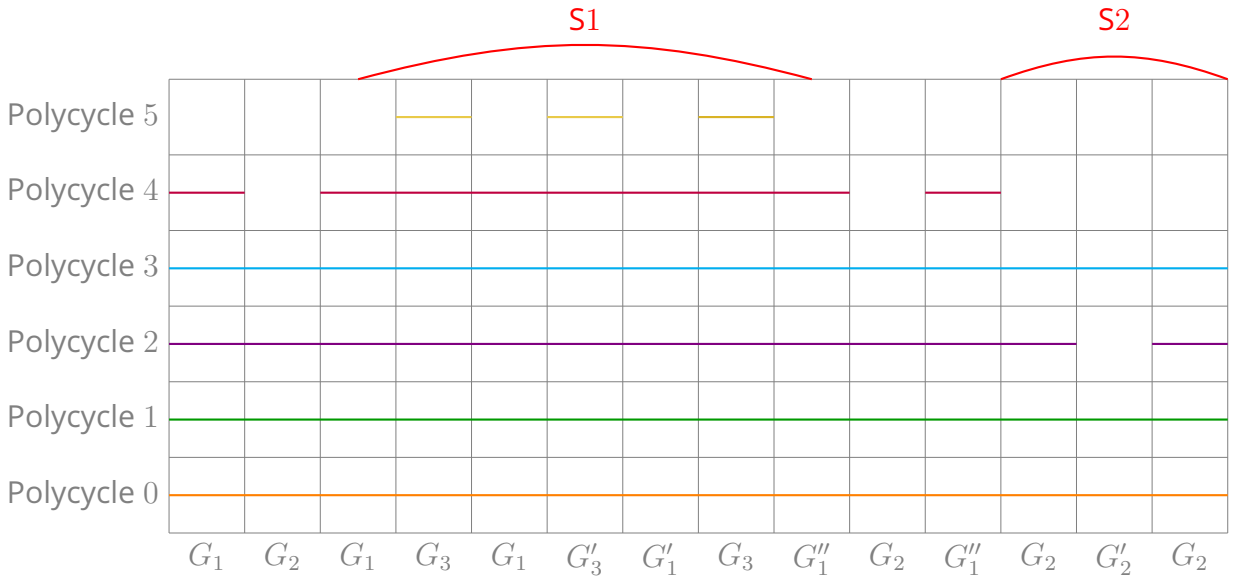
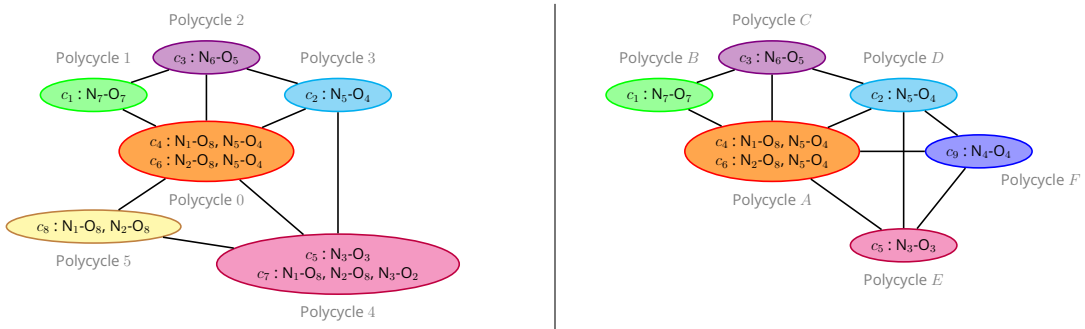


Figure 2.19 – Schéma des apparitions et disparitions des polycycles au cours d'une extension de la trajectoire de l'exemple simplifié que nous avons suivi jusque là. Les polycycles en ordonnée correspondent à ceux du polygraphe illustré dans la Figure 2.17. Les graphes sont identifiés en abscisse et suivent l'ordre temporel donné par la trajectoire. Une ligne est tracée dans une case si le polycycle correspondant à l'ordonnée apparaît dans la conformation correspondante en abscisse.

combinant ces informations, il est possible d'obtenir une compréhension approfondie de la dynamique et de l'évolution des structures moléculaires.

Le polygraphe offre une représentation forte qui caractérise une trajectoire de dynamique moléculaire rendant même possible la comparaison de différentes trajectoires. Bien que les polycycles soient définis spécifiquement pour chaque trajectoire, si les systèmes moléculaires sont les mêmes, les cycles sont comparables car ils sont représentés par les mêmes atomes. Ainsi, nous pouvons rechercher des correspondances entre les polycycles établissant ainsi les sommets similaires dans les polygraphes des trajectoires. Cette approche met en avant les similitudes et les différences dans les structures explorées par les différentes trajectoires. Cela permet, entre autres, de mesurer l'impact des conditions expérimentales, de l'environnement ou d'autres paramètres sur la dynamique moléculaire simulée. Par exemple, nous pouvons comparer des trajectoires à des températures différentes et en tirer des conclusions sur l'impact de ce paramètre sur la dynamique de la molécule.

La Figure 2.20 présente deux polygraphes représentant un même système moléculaire. Ces polygraphes sont donc comparables et les polycycles qui partagent un cycle identique sont considérés comme similaires. Ainsi, nous pouvons conclure que les polycycles 0 et A, 1 et B, 2 et C, 3 et D, ainsi que 4 et E sont similaires entre les deux polygraphes. Par contre, les polycycles 5 et F semblent spécifiques à chaque polygraphe. L'identification des polycycles spécifiques permet de tirer des conclusions sur les métastructures d'intérêt des trajectoires et donc sur l'exploration qui a été faite sur la surface



(a) Un polygraphe P.

(b) Un polygraphe P'.

Figure 2.20 – Illustration de deux polygraphes P et P' issus de deux trajectoires représentant un même système moléculaire. Les polycycles du polygraphe P de la Figure 2.20a sont identifiés par des chiffres, et les polycycles du polygraphe P' de la Figure 2.20b sont identifiés par des lettres.

d'énergie potentielle.

Le polygraphe offre une méthode d'analyse des trajectoires de dynamique moléculaire approfondie qui ne dépend pas de l'énergie potentielle. Grâce à cette approche, nous pouvons identifier des méta-structures d'intérêt qui peuvent être associées à des bassins conformationnels. De plus, en comparant les polygraphes obtenus à partir de différentes trajectoires décrivant le même système moléculaire, nous pouvons tirer des conclusions sur les spécificités de chaque trajectoire et sur l'impact de divers paramètres sur la dynamique de la molécule simulée.

2.5 . Synthèse de la démarche suivie

Dans la section précédente, nous avons mis en évidence que le polygraphe est un outil puissant pour l'analyse de la dynamique moléculaire. Il offre une perspective nouvelle sur l'étude de la dynamique et permet même la comparaison entre différentes trajectoires. Cependant, la pertinence du polygraphe dépend intrinséquement de la manière dont les cycles sont partitionnés en ensembles de cycles polymorphes, qui constituent les sommets du polygraphe. Comme nous le verrons dans le Chapitre 4, le calcul d'une partition minimale en ensembles de cycles polymorphes est un problème difficile. De plus, si l'on considère que la sélections des cycles formant la trajectoire peut influencer le résultat de cette partition la complexité de la tâche augmente encore. Plusieurs problèmes seront au cœur de cette thèse :

1. Comment choisir les bases de cycles de chaque conformation afin d'induire les ensembles de cycles dynamiques, et donc les cycles de la trajectoire ?
2. Comment partitionner l'ensemble des cycles de la trajectoire en le plus petit nombre possible d'ensembles de cycles polymorphes ?
3. Comment utiliser le polygraphe pour analyser et comparer des trajectoires ?

Ces problèmes constituent les différentes étapes de notre approche. Chaque étape doit

être traitée séquentiellement, mais il est important de noter que les résultats intermédiaires de chaque étape peuvent influencer de manière significative les étapes suivantes.

Plusieurs méthodes sont disponibles pour calculer une base de cycles d'un graphe en temps polynomial. Cependant, la question concernant le premier problème réside davantage dans la capacité d'une méthode de calcul de bases spécifiques à induire un ensemble de cycles garantissant un partitionnement optimal dans le cadre du second problème. Idéalement, nous souhaiterions résoudre ces deux problèmes simultanément, mais la complexité du second problème rend cette approche impraticable. Par conséquent, nous allons résoudre ces problèmes séquentiellement. Dans le Chapitre 5, nous étudierons la pertinence des solutions proposées pour le premier problème de sélection des bases dans le contexte du problème de partitionnement.

Le polygraphe offre de nouvelles perspectives pour l'étude des trajectoires de dynamique moléculaire. En regroupant l'information structurelle sous forme de polycycles, il rend les changements significatifs plus visibles. De plus, cette approche permet de regrouper des conformations différentes qui partagent la même structure sous-jacente.

Les méthodes traditionnelles d'analyse de structures sont souvent coûteuses en termes de calcul, ce qui rend le polygraphe particulièrement attrayant comme alternative. Dans le Chapitre 6, nous examinerons de manière détaillée les applications du polygraphe pour l'analyse et la comparaison de trajectoires.

3 - Sélection des cycles des conformations d'une dynamique moléculaire

Dans ce chapitre, nous nous penchons sur la méthode de sélection des bases de cycles pour caractériser les conformations. Comme évoqué dans le Chapitre 2, les bases de cycles sont un outil connu pour décrire les conformations au sein d'une trajectoire. En effet, en la représentant sous forme de graphe de cycles d'une conformation, la base de cycles minimale sous-jacente joue un rôle dans le calcul final du polygraphe de la trajectoire. Nous avons déjà établi qu'un graphe admet plusieurs bases de cycles minimales. À première vue, toutes semblent être de bonnes candidates pour caractériser les conformations moléculaires. Cependant, à l'échelle de l'ensemble de la trajectoire, certaines bases se révèlent plus avantageuses que d'autres. Nous les qualifions de *meilleures* bases de cycles, car elles favorisent un regroupement plus efficace au niveau des cycles ce qui permet d'aboutir à un polygraphe avec moins de sommets.

Le problème abordé dans ce chapitre consiste à rechercher un ensemble de bases de cycles minimum, avec une base par conformation de la trajectoire, de manière à ce que le polygraphe obtenu à partir de ces bases possède le moins de sommets possible.

Le chapitre est organisé comme suit. À la Section 3.1, nous présentons un algorithme classique pour le calcul des bases de cycles minimum d'un graphe, celui de Horton. Nous introduisons également une adaptation de cet algorithme dans le contexte des conformations de trajectoires de dynamique moléculaire.

Ensuite, dans la Section 3.2, nous abordons le problème de maximisation de l'intersection des bases de cycles. Notre proposition vise à maximiser l'intersection des bases de cycles minimales des conformations, dans le but d'améliorer la structure du polygraphe. Cette approche repose sur l'idée que favoriser l'intersection enrichit les informations structurelles contenues dans les voisinages, ce qui facilite davantage les regroupements. Nous analysons la complexité de ce problème, qui se rapproche du problème d'intersection des matroïdes, et démontrons qu'il est NP-complet lorsque trois graphes ou plus sont impliqués. Enfin, nous présentons une méthode de sélection des cycles inspirée de cette approche de maximisation.

Enfin, dans la Section 3.3, nous introduisons une méthode de descente pour la sélection des cycles. L'objectif de cette méthode n'est plus nécessairement de maximiser un paramètre tel que la dimension de l'intersection, mais plutôt de répartir les cycles entre les différentes bases de cycles minimum. En pratique, chaque solution sera associée à une valeur illustrant la diversité des bases sélectionnées, et nous chercherons à optimiser cette valeur au cours de la descente. L'objectif est d'atteindre un équilibre dans la représentation des cycles au sein des différentes bases de cycles minimales. Cette approche vise à ajouter de l'information structurelle de manière plus uniforme que ne le fait la méthode issue de l'intersection des bases.

Ce chapitre se conclut par une synthèse dans la Section 3.4.

3.1 . Calcul d'une base de cycles minimum

Dans cette section, nous présentons une méthode de calcul d'une base de cycles minimum pour un graphe. Cette méthode est applicable sur les conformations de manière indépendante les unes des autres.

Étant donné un graphe, un premier algorithme polynomial pour calculer une base de cycles minimum a été présenté par Horton en 1987 [27]. L'Algorithme 1 détaille le pseudo-code de cette procédure. Cet algorithme s'applique à un graphe simple sans boucles ni multi-arêtes. Sa complexité est en $O(|V| \times |E|^3)$ pour un graphe $G = (V, E)$. En effet, au plus $|V| \times |E|$ cycles sont théoriquement énumérés, même si en pratique nous atteignons rarement ce nombre l'étape la plus coûteuse se situe à la Ligne 4.

Algorithme 1 Calcul d'une base de cycles minimum, Horton [27]

Entrée un graphe $G = (V, E)$
Sortie une base de cycles minimum de G

1: Pour toute paire de sommets distincts $u, v \in V$, trouver $P(u, v)$ un chemin de poids minimum
2: Pour chaque sommet $v \in V$ et chaque arête $[x, y] \in E$ telle $v \neq x$ et $v \neq y$, créer le cycle $c_{v,x,y} = P(v, x) + P(v, y) + [x, y]$ ▷ Uniquement si $P(v, x) \cap P(v, y) = \emptyset$
3: Ordonner les cycles obtenus selon leur poids croissant.
4: Utiliser l'algorithme "Extraction d'une base" pour construire une base de cycles minimum B à partir de cet ensemble de cycles. ▷ Algorithme 2
5: **Renvoyer** B

Algorithme 2 Extraction d'une base

Entrée un ensemble de cycles C d'un graphe connexe G
Sortie une base génératrice de C

1: $B \leftarrow \emptyset$
2: **Tant que** $|B| < |E| - |V| + 1$ **faire** ▷ Si G n'est pas connexe, alors nous avons $|E| - |V| + x$ où x est le nombre de composantes connexe de G.
3: Supprimer de C son premier élément c
4: **Si** $B \cup \{c\}$ est une famille libre **alors**
5: $B \leftarrow B \cup \{c\}$
6: **Renvoyer** B

L'Algorithme 2 est une adaptation pour les bases de cycles de la procédure proposée en 1956 par Kruskal dans le contexte du problème du voyageur de commerce [35]. Dans sa forme générale, *the greedy algorithm* consiste à ajouter autant d'éléments que possible à un ensemble. Cette procédure est largement utilisée, notamment en algèbre linéaire [17]. Dans le cas présent, des cycles sont ajoutés à la base tant que celle-ci forme une famille libre, c'est-à-dire que les éléments qui la composent sont linéairement indépendants. Étant donné que l'Algorithme 2 conserve l'ordre des cycles donné en entrée, les cycles de poids les plus faibles sont sélectionnés en priorité et nous obtenons une

base de cycles minimum. La complexité de cet algorithme découle du coût associé à la vérification que l'union de la base et d'un nouvel élément reste une famille libre.

Horton suggère une représentation des cycles par des vecteurs binaires et l'utilisation du pivot de Gauss pour déterminer facilement l'indépendance des vecteurs. Cette approche permet d'appliquer facilement l'algorithme glouton de Kruskal pour extraire une base de cycles. L'étape d'élimination de Gauss, qui manipule une matrice de dimension $|V||E|^2$, a une complexité en $O(|V| \times |E|^3)$. C'est l'étape la plus coûteuse en termes de temps de calcul dans l'algorithme de Horton.

D'autres méthodes pour calculer les bases de cycles ont émergées pour le calcul ou l'approximation d'une base de cycles minimum [12, 32, 34, 41]. Ces méthodes sont devenues de plus en plus efficaces au fil des années et sont donc toutes polynomiales. Par exemple, [33] propose une méthode en $O(|E|^2 \times |V| + |E| \times |V|^2 \log |V|)$. Ces méthodes se distinguent entre autres par leur méthode d'extraction des bases de cycles. Contrairement à une approche incrémentale qui construit une base à la fois, ces méthodes construisent plusieurs bases simultanément, puis effectuent un tri pour sélectionner les bases de cycles les plus appropriées.

Dans le cadre de l'analyse des trajectoires, nous avons entrepris de modifier un algorithme classique pour y intégrer un paramètre spécifique à notre contexte, nous permettant de favoriser des bases de cycles minimum qui présentent des similarités. Notre choix s'est porté sur l'algorithme de Horton [27] pour sa facilité d'adaptation. Bien que des méthodes plus récentes et potentiellement un peu plus efficaces existent, celles-ci ne prennent pas en compte un ordre sur les cycles qui soit facilement manipulable, tout en garantissant que le résultat obtenu demeure une base de cycles minimum. De plus, la complexité de l'algorithme de Horton ne pose particulièrement problème que pour des graphes denses, ce qui n'est pas souvent le cas des graphes moléculaires.

Priorisation des cycles avec peu de liaisons variables Étant donné que l'Algorithme 2 conserve l'ordre des cycles donné en entrée, nous proposons d'ajuster l'ordonnancement des cycles afin de toujours prioriser les cycles les plus courts mais en ajoutant un ordre supplémentaire pour les cycles de même longueur. Nous souhaitons maintenant sélectionner les cycles avec le moins de liaisons variables afin de privilégier des cycles qui ont plus de chance d'être présent dans les bases de cycles minimum des autres graphes de la trajectoire. Nous proposons l'Algorithme 3, une modification de l'Algorithme 1, qui inclut les liaisons variables de nos graphes dans l'ordonnancement des cycles. Ainsi, lorsque deux cycles ont la même longueur, celui comportant le moins de liaisons variables est ordonné en premier. Les modifications sont indiquées en gras dans le pseudo-code.

Concernant l'application de ces procédures sur une trajectoire moléculaire, notons que les algorithmes 1 et 3 sont applicables de manière indépendante à chacun des graphes. Bien que cela ne soit pas immédiatement évident, les bases obtenues par cet algorithme présentent toutefois des propriétés particulières. En effet, si les sommets et les arêtes des graphes sont toujours ordonnés de la même manière, alors les cycles construits le seront également. L'ordre de traitement des cycles a un impact direct sur la base extraite. Ainsi, nous obtenons des bases de cycles minimum qui sont déjà très similaires les unes aux autres. Cette propriété est un plus dans notre contexte moléculaire car elle signifie que dans la mesure du possible les mêmes cycles sont sélectionnés, ce qui va faciliter la

Algorithme 3 Modification de l'Algorithme 1

Entrée un graphe $G = (V, E)$
Sortie une base de cycles minimum de G

1: Pour toute paire de sommets distincts $u, v \in V$, trouver $P(u, v)$ un chemin de poids minimum
2: Pour chaque sommet $v \in V$ et chaque arête $[x, y] \in E$ telle que $v \neq x$ et $v \neq y$, créer le cycle $c_{v,x,y} = P(v, x) + P(v, y) + [x, y]$ ▷ Uniquement si $P(v, x) \cap P(v, y) = \emptyset$
3: **Ordonner les cycles obtenus selon leur poids croissant, et en cas d'égalité selon leur nombre de liaisons variables croissant**
4: Utiliser l'algorithme "Extraction d'une base" pour construire une base de cycles minimum B à partir de cet ensemble de cycles. ▷ Algorithme 2
5: **Renvoyer** B

comparaison et les étapes futures de l'analyse.

L'adaptation que nous proposons dans l'Algorithme 3 favorise d'autant plus la sélection de cycles identiques. De plus, la méthode vise à privilégier la sélection de cycles comportant un nombre minimal de liaisons variables. Ainsi, nous favorisons les cycles qui représentent le moins de contraintes structurelles simultanément, ce qui facilite l'identification des différentes contraintes présentes dans une conformation.

Dans ce chapitre, notre objectif est donc de construire des bases présentant certaines similarités et permettant le calcul d'un polygraphe avec peu de sommets. Bien que les méthodes proposées dans cette section s'appliquent indépendamment aux conformations, ce sont des méthodes déterministes qui favorisent la construction de bases présentant de fortes similitudes. Dans les prochaines sections, nous étudierons d'autres méthodes de sélection des bases de cycles basés sur d'autres critères de similarité, pour voir si celles-ci permettent de sélectionner des bases de cycles minimum plus favorable pour le calcul d'un polygraphe avec un nombre minimal de sommets.

3.2 . Maximiser l'intersection de bases de cycles minimum

Dans le cadre de l'optimisation du choix des bases pour le calcul du polygraphe, notre première hypothèse a été de maximiser l'intersection. En effet, l'idée sous-jacente est que maximiser l'intersection conduit à favoriser l'apparition de bases similaires ou à défaut de voisinages en partie similaires. Cette section présente nos travaux sur ce problème.

Rappelons qu'à partir d'une trajectoire, d'après les notations du Chapitre 2, nous notons $\mathcal{G}$ l'ensemble des graphes différents issus de cette trajectoire $\{G_1, G_2, \ldots, G_N\}$. Tous les graphes de $\mathcal{G}$ ont le même ensemble de sommets et partagent un sous-ensemble d'arêtes commun (le backbone).

Rappelons également que pour tout i tel que $1 \leq i \leq N$ B_i représente une base de cycles minimum de G_i, avec $B_i \in \mathcal{MCB}(G_i)$, où $\mathcal{MCB}(G_i)$ désigne l'ensemble des bases de cycles minimum du graphe G_i.

Considérons le problème suivant :

Problème 1 (Intersection de bases de cycles minimum, MCBI)**.** *Étant donné $\mathcal{G}$, un ensemble de N graphes $G_{1\leq i\leq N}$ partageant le même ensemble de sommets, et $\gamma \in \mathbb{N}$, existe-t-il un ensemble $\{B_1, \ldots, B_N\}$ avec $B_i \in \mathcal{MCB}(G_i)$ tel que $|\bigcap_{i=1}^{N} B_i| \geq \gamma$?*

Ainsi, étant donné un ensemble de graphes, le problème MCBI consiste à trouver un ensemble de bases de cycles minimum dont la cardinalité de l'intersection est supérieure à un entier γ. Le problème de maximisation associé à ce problème de décision est noté max-MCBI (Maximiser l'intersection de bases de cycles minimum). Il vise, quant à lui, à trouver un ensemble de bases de cycles minimum dont la cardinalité de l'intersection est maximale.

Dans la suite du chapitre, nous établissons la complexité de MCBI. Ce problème partage des similitudes avec le problème de l'intersection de matroïdes : ils sont tous deux NP-complets lorsqu'au moins trois graphes sont pris en considération mais peuvent être résolus en temps polynomial si il y a seulement deux graphes. Enfin, à la Section 3.2.3, nous proposons une méthode de sélection des cycles inspirée de ce problème max-MCBI.

3.2.1 . Proximité entre intersection de bases de cycles et intersection de matroïdes

Dans cette section, nous abordons le problème de l'intersection de matroïdes et établissons sa relation avec MCBI. Les matroïdes sont des structures généralisant le concept d'indépendance linéaire. Il existe donc un lien entre une base de cycles et un matroïde. Nous verrons dans la suite que l'on peut construire un matroïde à partir d'un graphe afin de résoudre MCBI. Cela fait du problème MCBI un cas particulier du problème de l'intersection de matroïdes.

Dans un premier temps, nous présentons le problème de l'intersection de matroïdes ainsi que sa complexité : ce problème est NP-complet à partir de trois matroïdes et est polynomial si seulement deux matroïdes sont impliqués. Dans un second temps, nous proposons une reduction du problème MCBI vers le problème de l'intersection de matroïdes, démontrant ainsi la polynomialité de MCBI lorsque seulement deux graphes sont impliqués.

Le problème de l'intersection de matroïdes

Un matroïde est un couple constitué d'un ensemble et d'une famille[1] de l'ensemble associé. La définition suivante a été initialement proposée dans [54].

Définition 20 (Matroïde)**.** *Un **matroïde** est un couple (E, I) où E est un ensemble fini non vide et I est une famille de parties de E tels que les deux axiomes suivants sont vérifiés :*

1. *Hérédité : $\forall X \in I$, si $X' \subset X$ alors $X' \in I$*
2. *Augmentation : Si $X \in I, Y \in I$ et $|X| < |Y|$ alors $\exists e \in Y, e \notin X$ tel que $\{X \cup \{e\}\} \in I$*

L'ensemble E est appelé *ground set* et les sous-ensembles de I sont appelés *indépendants*. Pour autant, il ne s'agit pas de familles libres au sens algébrique. Par exemple, il existe un matroïde appelé *matroïde libre* qui correpond au couple (E, I) où I est la famille de toutes les parties de E. Les matroïdes constituent un modèle bien plus général que

1. Le terme famille désigne un ensemble de sous ensembles.

celui des bases vectorielles que nous avons définies. Nous qualifions de linéaire un matroïde (E, I) dans lequel E est un ensemble de vecteurs et I est la famille composée de toutes les familles libres de E. Ce sont ces matroïdes qui nous serviront à généraliser le problème de base de cycles.

Problème 2 (Intersection de matroïdes)**.** *Étant donné un ensemble de matroïdes $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$ partageant le même ground set E, où $M_1 = (E, I_1)$, $M_2 = (E, I_2)$, ..., $M_N = (E, I_N)$ et un entier $\gamma \in \mathbb{N}$, existe-t-il un ensemble $X \in \{I_1 \cap I_2 \cap \ldots \cap I_N\}$ tel que $|X| \geq \gamma$?*

Ce problème recherche donc un sous-ensemble X de E qui est un indépendant dans chaque M_i pour tout i tel que $1 \leq i \leq N$. Notons que ce problème de l'intersection des matroïdes est le plus souvent décrit sous la forme du problème d'optimisation qui lui est associé. Ce problème d'optimisation consiste à rechercher un tel ensemble X de cardinalité maximale.

Ce problème est NP-complet lorsqu'au moins trois matroïdes sont impliqués [53]. Il est possible de connaître une borne supérieure à la solution du problème impliquant deux matroïdes [18]. Elle permet de montrer l'existence d'un algorithme polynomial de partionnement des indépendants [16] pour trouver une solution au problème de l'intersection de deux matroïdes. D'autres algorithmes se basant sur cette borne ont également été proposés [36, 19] pour résoudre ce problème ou certaines de ses variantes.

Réduction de MCBI vers le problème d'intersection de matroïdes

Nous consacrons cette section à la réduction de l'intersection de bases de cycles minimum à l'intersection de matroïdes. Nous commençons par démontrer le théorème suivant qui permet de transformer un graphe en matroïde.

Théorème 1. *Étant donné un graphe G, le couple $(\mathcal{C}, I)$ avec $\mathcal{C}$ l'ensemble des cycles de G et I la famille de parties X telles qu'il existe une base de cycles minimum B de G incluant X (autrement dit, $\exists B | X \subseteq B \in \mathcal{MCB}(G)$) définit un matroïde. Cette transformation est notée $M(G)$.*

Pour montrer ce théorème, il est nécessaire de prouver que les axiomes d'hérédité et d'augmentation sont bien vérifiés par le matroïde $(\mathcal{C}, I)$. Ce résultat est direct concernant l'axiome d'hérédité. Nous allons maintenant prouver plusieurs lemmes intermédiaires pour démontrer l'axiome d'augmentation.

Étant donné un graphe G, B est une base de cycles du graphe G et $\mathcal{C}$ est l'ensemble des cycles de ce graphe, la fonction $\lambda_B : B \times \mathcal{C} \to \{0, 1\}$ représente le rôle d'un cycle de B dans la génération d'un cycle de $\mathcal{C}$. Ainsi, $\lambda_B(c, d) = 1$ si et seulement si on a $c \in B$ et que c participe à la génération de $d \in \mathcal{C} \setminus B$. Autrement dit, il existe un sous-ensemble de cycles $B' \subset B$ qui permettent d'obtenir d par l'opération $\bigoplus B' = d$ avec $c \in B'$. Rappelons que la somme de deux cycles est décrite dans la Définition 10 (Chapitre 2, page 36). Ainsi, étant donné un ensemble de cycles $B = \{c_1, c_2, c_3\}$, $\bigoplus B = c_1 \oplus c_2 \oplus c_3$.

Lemme 1. *Étant donnée B une base de cycles du graphe G et deux cycles $c_1 \in B$ et $c_2 \notin B$, si $\lambda_B(c_1, c_2) = 1$ alors $(B \setminus \{c_1\}) \cup \{c_2\}$ est une base de cycles de G.*

Démonstration. Considérons un tel ensemble $B' = \{B \setminus \{c_1\}\} \cup c_2$. Alors, il existe un sous-ensemble de cycles $D \subset (B \setminus \{c_1\})$ tel que $c_2 = c_1 \oplus (\bigoplus D)$. De manière inverse, $c_1 = c_2 \oplus (\bigoplus D)$, démontrant ainsi que c_1 est généré par B'. Par conséquent, tout cycle couvert par B est également couvert par B'. De plus, étant donné que $|B| = |B'|$, nous pouvons conclure que B' est une base de cycles de G. □

Lemme 2. *Si $B \in \mathcal{MCB}(G)$ alors pour deux cycles c_1, c_2 avec $c_1 \in b$ et $c_2 \notin b$ et tels que $\lambda_B(c_1, c_2) = 1$ nous avons $\omega(c_1) \leq \omega(c_2)$.*

Démonstration. Considérons $B \in \mathcal{MCB}(G)$, $c_1 \in B$ et $c_2 \notin B$ tels que $\lambda_B(c_1, c_2) = 1$. Selon le Lemme 1, $B' = (B \setminus \{c_1\}) \cup \{c_2\}$ forme une base de cycles de G.

Notons $\omega(B)$, le poids de la base de cycles B, tel que $\omega(B) = \sum_{c \in B} \omega(c)$. Ainsi, on a $\omega(B') = \omega(B) - \omega(c_1) + \omega(c_2)$. Si $\omega(c_2) < \omega(c_1)$, alors on a $\omega(B') < \omega(B)$. Étant donné que B est, par définition, une base de cycles minimum, une telle base de cycles B' ne peut pas exister. □

Lemme 3. *Soient B_1, B_2 deux bases de cycles de G telles que si pour tout $B \in \{B_1, B_2\}$, pour chaque couple de cycles (c_1, c_2) où $c_1 \in B$, $c_2 \notin B$ avec $\lambda_B(c_1, c_2) = 1$, alors :*

- *$\omega(c_1) \leq \omega(c_2)$*
- *pour tout $c_2 \in \{B_2 \setminus B_1\}$, il existe $c_1 \in \{B_1 \setminus B_2\}$ tel que $(B_2 \setminus \{c_2\} \cup \{c_1\})$ est une base de cycles de G de même poids que B_2.*

Démonstration. Si nous pouvons démontrer qu'il existe $c_1 \in B_1$ avec $\omega(c_1) = \omega(c_2)$ et $\lambda_{B_2}(c_2, c_1) = 1$ alors le Lemme 1 permet de conclure. Soit $c_2 \in B_2 \setminus B_1$, considérons alors les ensembles disjoints suivants :

1. $D_1 = \{d \in B_1 \cap B_2 \mid \lambda_{B_1}(d, c_2) = 1\}$
2. $D_2 = \{d \in \{B_1 \setminus B_2\} \mid \lambda_{B_1}(d, c_2) = 1, \omega(d) < \omega(c_2)\}$
3. $D_3 = \{d \in \{B_1 \setminus B_2\} \mid \lambda_{B_1}(d, c_2) = 1, \omega(d) = \omega(c_2)\}$

Selon l'hypothèse sur B_1, aucun cycle de B_1 de poids plus élévé que c_2 ne peut générer c_2. Ainsi, on peut écrire la somme suivante $c_2 = \bigoplus D_1 \oplus \bigoplus D_2 \oplus \bigoplus D_3$.

Selon l'hypothèse sur B_2, pour tout cycle $d \in D_2$ et $e \in B_2$ tel que $\lambda_{B_2}(e, d) = 1$, nous avons $\omega(e) \leq \omega(d) < \omega(c_2)$, donc $e \neq c_2$. Ainsi, pour chaque cycle $d \in D_2$, on a $\lambda_{B_2}(c_2, d) = 0$. Nous pouvons donc écrire la somme des éléments de B_2 permettant d'obtenir c_2.

$$c_2 = \bigoplus D_1 \oplus \bigoplus_{d \in \{D_2 \cup D_3\}} \bigoplus_{e \in \{B_2 \setminus \{c_2\}\}} \lambda_{B_2}(e, d) \cdot d$$

Notons que $c_2 \notin D_1$ car $D_1 \subseteq B_1$. Ainsi, si pour tout $d \in D_3$ nous avons $\lambda_{B_2}(c_2, d) = 0$, alors c_2 n'apparaît pas dans la partie droite de l'équation ci-dessus. Cela suppose que c_2 est obtenable par composition des éléments de $\{B_2 \setminus \{c_2\}\}$ de poids inférieur à c_2. Étant donné que $c_2 \in B_2$, alors B_2 n'est pas linéairement indépendant.

Cette contradiction prouve l'existence d'un cycle $d \in D_3$ tel que $\lambda_{B_2}(c_2, d) = 1$. De plus, $d \in B_1$ et $\omega(d) = \omega(c_2)$, donc par le Lemme 1, nous pouvons conclure que $(B_2 \setminus \{c_2\}) \cup d)$ est une base de cycles de même poids que B_2. □

Lemme 4. *La base $B \in \mathcal{MCB}(G)$ si et seulement si B est une base de cycles de G et, pour tout couple c_1, c_2 avec $c_1 \in B$ et $c_2 \notin B$ tels que $\lambda_B(c_1, c_2) = 1$, nous avons $\omega(c_1) \leq \omega(c_2)$*

Démonstration. Afin de montrer cette équivalence, nous allons prouver les deux implications. La première implication correspond à la proposition du Lemme 2 : Si $B \in \mathcal{MCB}(G)$ alors pour c_1, c_2 avec $c_1 \in B$ et $c_2 \notin B$ tel que $\lambda_B(c_1, c_2) = 1$, nous avons $\omega(c_1) \leq \omega(c_2)$. Nous allons donc nous concentrer sur la seconde implication.

Considérons B_1 une base de cycles de G vérifiant la seconde proposition, et $B_2 \in \mathcal{MCB}(G)$. Le Lemme 3 précise que le poids de la base est conservé ainsi s'il est appliqué sur ces bases, B_1 et B_2 étant donné $c_2 \in \{B_2 \setminus B_1\}$, il existe $c_1 \in \{B_1 \setminus B_2\}$ tel que $(B_2 \setminus \{c_2\}) \cup \{c_1\} \in \mathcal{MCB}(G)$.

Nous pouvons alors appliquer ces échanges sur B_2 de façon repétée jusqu'à ce que $B_1 = B_2$. Le Lemme 3 indique bien que le résultat de cet échange conserve le poids de la base de cycles. Par conséquent, B_1 est une base de cycles minimum. □

Lemme 5. *Si $i_1 \in I$, $i_2 \in I$ et $|i_1| < |i_2|$ alors $\exists e \in \{i_2 \setminus i_1\}$ tel que $\{i_1 \cup \{e\}\} \in I$.*

Démonstration. Rappelons que $\mathcal{C}$ désigne l'ensemble de cycles du graphe G. Notons $\text{span}(i) \subseteq \mathcal{C}$ le sous-ensemble de $\mathcal{C}$ couvert par l'ensemble de cycles i. Ainsi, si $i \in I$ et $\text{span}(i) = \mathcal{C}$, alors $i \in \mathcal{MCB}(G)$.

Soit $i_1 \in I$ et $i_2 \in I$, si $|i_1| < |i_2|$, alors il existe $c \in \mathcal{C}$ tel que $c \notin \text{span}(i_1)$ et $c \in \text{span}(i_2)$. Autrement dit, il existe un élement c non couvert par i_1 mais couvert par i_2. Ainsi, il existe un cycle $e \in \{i_2 \setminus i_1\}$ tel que $\lambda_{i_2}(e, c) = 1$. Notons que, comme nous l'avons établi dans le Lemme 2, e ne peut pas être généré par un ensemble de cycles de poids inférieur.

Montrons qu'il existe une base $B \in \mathcal{MCB}(G)$ avec $\{i_1 \cup e\} \subseteq B$. Étant donné que e génère un cycle c qui n'est pas couvert par i_1, alors $\{i_1 \cup e\}$ est une famille libre. De plus, aucun des cycles de $\{i_1 \cup e\}$ ne peut être généré par un ensemble de cycles de poids inférieur. Nous avons donc que $\forall c_1, c_2$ avec $c_1 \in \{i_1 \cup e\}$ et $c_2 \notin \{i_1 \cup e\}$ tel que $\lambda_{\{i_1 \cup e\}}(c_1, c_2) = 1$, alors $\omega(c_1) \leq \omega(c_2)$. L'ensemble $\{i_1 \cup e\}$ respecte les propriétés d'un sous-ensemble d'une base de cycles établies par le Lemme 4. Il suffit alors de compléter $\{i_1 \cup e\}$ pour former B une base de cycles de G respectant la seconde propriété du Lemme 4 et obtenir notre résultat : $\{i_1 \cup e\} \subseteq B \in \mathcal{MCB}(G)$.

□

Nous disposons maintenant des éléments nécessaires pour prouver le Théorème 1.

Démonstration du Théorème 1. La transformation $M(G)$ définit bien un matroïde car le couple $(\mathcal{C}, I)$ où $\mathcal{C}$ est l'ensemble des cycles de G et I est la famille de parties X tel que $X \subseteq B \in \mathcal{MCB}(G)$ vérifie bien les deux axiomes des matroïdes.

1. Le premier axiome est évidemment vérifié car si $i \subset B \in \mathcal{MCB}(G)$ et $i' \subset i$ alors $i' \subset B \in \mathcal{MCB}(G)$
2. Le second axiome est vérifié par le Lemme 5.

□

À partir d'un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, nous construisons N matroïdes. Soit $\mathcal{C} = \bigcap_{j=1}^{N} \mathcal{C}_j$ où $\mathcal{C}_j$ désigne l'ensemble des cycles de G_j. Nous avons alors N les matroïdes $M(G_j)$ tels que $\forall j$ avec $1 \leq j \leq N$, $M(G_j) = (\mathcal{C}, I_j)$ où $\mathcal{C} = \bigcap_{k=1}^{N} \mathcal{C}_k$ et $\forall i \in I_j$, $i \subseteq B_j \in \mathcal{MCB}(G_j)$. Résoudre le problème de maximisation de l'intersection de $\{M(G_1), M(G_2), \ldots, M(G_N)\}$ revient alors à déterminer l'ensemble de cycles de cardinalité maximum qui est un sous-ensemble d'une base de cycles minimum de G_i pour tout

$1 \leq i \leq N$. Il ne reste alors plus qu'à compléter cet ensemble pour constituer les bases de cycles minimum des graphes G_i pour tout $1 \leq i \leq N$.

Nous avons donc bien une réduction du problème MCBI vers le problème de l'intersection des matroïdes. Pour deux matroïdes le problème de l'intersection est polynomial. Les algorithmes connus pour résoudre le problème de l'intersection de deux matroïdes, tels que celui d'Edmonds [16], sont polynomiaux en fonction de la dimension du ground-set. Or, ici, le ground-set est l'intersection des ensembles de cycles de plusieurs graphes et peut donc potentiellement être exponentiel en le nombre d'arêtes de ces graphes.

Nous allons maintenant montrer comment nous pouvons limiter la taille du ground-set tout en assurant la présence d'une solution optimale à max-MCBI. Cela nous montre que le problème MCBI est bien polynomial lorsque seulement deux graphes sont considérés. Pour ce faire, nous proposons l'Algorithme 4 qui va permettre de définir des ensembles de cycles candidats pour MCBI.

L'algorithme d'extraction des cycles candidats à MCBI : Algorithme 4. Nous présentons ici une version générale qui prend en argument un ensemble des graphes et retourne un ensemble de cycles candidats pour chacun des graphes. Les ensembles de cycles candidats ainsi construits vérifient qu'une solution optimale de MCBI est accessible depuis ces ensembles.

Remarque 7. *Le fait que l'on ne choisisse qu'un seul plus court chemin entre deux sommets dans la boucle de la ligne 3 peut amener à penser que des cycles peuvent être omis. C'est pour cela que nous avons également le traitement des cycles pairs à la ligne 7. En effet, les cycles qui peuvent être omis par la ligne 3 sont nécessairement des cycles pairs.*

L'Algorithme 4 retourne un ensemble de cycles de taille polynomiale pour chacun des graphes donnés à partir desquels nous pouvons définir un ensemble de bases de cycles minimum $B_i \subseteq S_i$, pour $1 \leq i \leq N$ telles que $\bigcap_{i=1}^{N} B_i$ est de cardinalité maximum.

Pour chaque graphe G_i avec $1 \leq i \leq N$, nous avons $M(G_i) = (L, I_i)$ où $L = \bigcap_{i=1}^{N} S_i$ et $X \in I_i$ s'il existe $B_i \in \mathcal{MCB}(G_i)$ tel que $X \subset B_i$.

Nous considérons par $M(G_i)$ la restriction à L du matroïde du Théorème 1. Ainsi, le Lemme 6 prouve qu'une solution optimale à max-MCBI peut être extraite depuis les ensembles candidats renvoyés par l'Algorithme 4. De cette façon, résoudre le problème de l'intersection maximale sur les matroïdes que nous avons définis, résoud bien le problème max-MCBI sur les graphes donnés.

Lemme 6. *Étant donné un ensemble de graphes* $\{G_1, \ldots, G_N\}$ *instance du problème* max-MCBI *et* $S_1, \ldots, S_N$ *les ensembles renvoyés par l'Algorithme 4, il existe une solution optimale* $\{B_1^*, \ldots, B_N^*\}$ *telle que* $\forall i$ *avec* $1 \leq i \leq N$, $B_i^* \subseteq S_i$.

Démonstration. Notons $B^* = \{B_1^*, B_2^*, \ldots, B_N^*\}$ une solution optimale du problème max-MCBI. Supposons qu'il existe un cycle c appartenant à la solution optimale B^* mais absent de l'ensemble candidat, c'est-à-dire que $c \in \bigcap_{i=1}^{N} B_i^*$ et $c \notin L$ où L est l'ensemble des cycles commun à tous les graphes énumérés par l'algorithme 4. Pour tout cycle c ainsi défini, il existe trois cycles c', c_1 et c_2 tels que $c' \in \bigcap_{i=1}^{N} S_i$ et $c = c_1 \oplus c' \oplus c_2$ avec $\omega(c_1) < \omega(c)$, $\omega(c') = \omega(c)$ et $\omega(c_2) < \omega(c)$. La Figure 3.1 illustre cette combinaison de cycles sur des exemples. Ainsi, tout cycle c peut-être remplacé par c' dans toutes les bases optimales telles que $(B_i^* \setminus \{c\}) \cup \{c'\} \in \mathcal{MCB}(G_i)$.

Algorithme 4 Ensembles candidats à MCBI

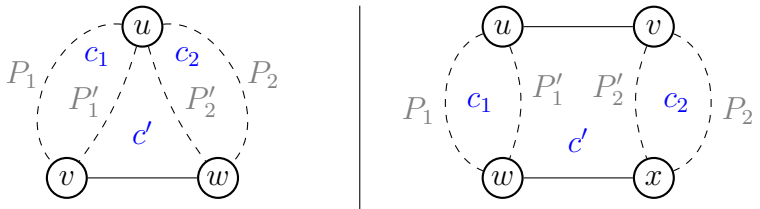
Entrée un ensemble de graphes $\{G_1, G_2, \ldots, G_N\}$
Sortie un ensemble de cycles de taille polynomiale pour chaque graphe G_i avec $1 \leq i \leq N$

1: $G = (V, E) \leftarrow (V, \bigcap_{i=1}^{N} E_i)$ avec $G_i = (V, E_i)$
2: $L \leftarrow \emptyset$

Gestion des cycles impairs
3: **Pour** $u \in V$, $[v, w] \in E$, $v \neq u$, $w \neq u$ **faire**
4: $c_{u,v,w} = P(w, u) + P(u, v) + [v, w]$ où $P(u, v)$ désigne un plus court chemin de u à v.
5: **Si** $P(w, u) \cap P(u, v) = \emptyset$ **alors**
6: ajouter le cycle $c_{u,v,w}$ à L

Gestion des cycles pairs
7: **Pour** $[u, v] \in E$, $[w, x] \in E$, $w \neq u$, $w \neq v$, $x \neq u$, $x \neq v$ **faire**
8: $c_{w,u,v,x} = P(w, u) + [u, v] + P(v, x) + [w, x]$
9: **Si** $P(w, u) \cap P(v, x) = \emptyset$ **alors**
10: ajouter le cycle $c_{w,u,v,x}$ à L
11: $c_{w,v,u,x} = P(w, v) + [v, u] + P(u, x) + [x, w]$
12: **Si** $P(w, v) \cap P(u, x) = \emptyset$ **alors**
13: ajouter le cycle $c_{w,v,u,x}$ à L

14: $S_1, S_2, \ldots, S_N \leftarrow \emptyset$
15: **Pour** i allant de 1 à N **faire**
16: calculer une base $B_i \in \mathcal{MCB}(G_i)$
17: $S_i \leftarrow L \cup B_i$
18: **Renvoyer** $S_1, S_2, \ldots, S_N$



(a) Le poids de c est impair (b) Le poids de c est pair

Figure 3.1 – Exemples des combinaisons $c_1 \oplus c' \oplus c_2 = c$ avec $\omega(c') = \omega(c)$. Si $|P_1| < |P_2|$, cela signifie qu'il existe un sommet $x \in P_2$ tel que le cycle passant par x, v et w correspond à l'un des deux exemples proposés.

Prouvons maintenant qu'un tel cycle c' existe. Considérons que c est un cycle impair, alors c n'a pas été énuméré lors de la boucle à la Ligne 3 de l'Algorithme 4. Donc, il existe des plus courts chemins P_1, P_2, P'_1 et P'_2 tous de même longueur. Comme c appartient à toutes les bases de cycles minimum, alors $\omega(c') = \omega(c)$, sinon les bases ne seraient pas minimum (en conséquence du Lemme 2). Ainsi, nous avons $|P_1| = |P'_1| = |P_2| =$

$|P'_2|$ comme illustré dans la Figure 3.1a. Le cycle c' existe et a été ajouté à $\bigcap_{i=1}^{N} S_i$ par l'Algorithme 4 à la place de c. Le même principe est applicable pour définir c' si c est un cycle pair, en considérant cette fois la boucle à la Ligne 7 de l'Algorithme 4. Notons que dans tous les cas, si $P_1 = P'_1$, alors c_1 est vide, et si $P_2 = P'_2$, alors c_2 est vide.

Pour conclure, si le cycle c tel que nous l'avons décrit existe, alors le cycle c' existe également. Ainsi, une nouvelle solution optimale à max-MCBI peut être construite en remplaçant c par c' dans toutes les bases B_i^* pour $1 \leq i \leq N$. Le Lemme 1 montre bien que de tels échanges mènent bien à des bases de cycles minimum. Ces bases ont une intersection de même cardinalité que les bases optimales B_i^* pour $1 \leq i \leq N$. Nous pouvons alors repéter ces échanges jusqu'à ce que $\bigcap_{i=1}^{N} B_i^* \subseteq \bigcap_{i=1}^{N} S_i$.

□

À présent, pour résoudre notre instance du problème de l'intersection de matroïdes nous devons définir un oracle polynomial qui vérifie si un ensemble X est un indépendant des matroïdes données. En effet, les algorithmes classiques pour résoudre le problème d'intersection de matroïdes, tels que [16], font appel à cet oracle de façon répétitive sur tous les ensembles X considérés.

Un oracle polynomial pour déterminer les sous-ensembles d'une base de cycles minimum. Nous commençons par introduire l'Algorithme 5 qui sera directement employé pour créer un oracle polynomial, permettant ainsi l'utilisation des algorithmes connus pour l'intersection de matroïdes.

L'Algorithme 5 prend en entrée un graphe G et un ensemble de cycles $\mathcal{C}$ et renvoie une base de cycles minimale B dans G de manière à maximiser la cardinalité de $|B \cap \mathcal{C}|$. Le Lemme 7 démontre que cet algorithme s'exécute en temps polynomial. Cet algorithme est très utile pour construire des bases de cycles à partir d'un ensemble donné, comme nous l'avons fait dans la preuve du Lemme 6.

Algorithme 5 Calculer une base de cycles minimum maximisant son intersection avec un ensemble.

Entrée un graphe $G = (V, E)$ et un ensemble de cycles $\mathcal{C}$ de G.
Sortie une base de cycles minimum de G

1: **Fonction** MCBset($G = (V, E)$, $\mathcal{C}$) : $B \in \mathcal{MCB}(G)$
2: $B' \leftarrow$ une base de cycles minimum de G
3: $\mathcal{S} \leftarrow$ les cycles $c \in \{\mathcal{C} \cup B'\}$ ordonnés par poids croissant, et en cas d'égalité les cycles de $\mathcal{C}$ sont placés en premier.
4: Utiliser l'algorithme d'extraction d'une base de cycles pour construire une base de cycles minimum B à partir de cet ensemble de cycles $\mathcal{S}$. ▷ Algorithme 2
5: **Renvoyer** B

Lemme 7. *Trouver une base de cycles minimum maximisant son intersection avec un ensemble de cycles donné se fait en temps polynomial avec l'Algorithme 5.*

Démonstration. L'Algorithme 5 est manifestement polynomial. Il se compose d'une série d'opérations, chacune étant polynomiale, suivie d'une boucle avec un nombre d'itérations

polynomial. Dans cette boucle, chaque itération vérifie si un élément est linéairement indépendant d'un ensemble, une opération également polynomiale. De plus, le calcul d'une base de cycles minimum B' à la ligne 2 s'effectue en temps polynomial [27, 33].

Considérons $\mathcal{C}$ l'ensemble de cycles donné en argument de la fonction MCBset, et B la base de cycles minimum renvoyée. Soit $B^* = (e_1, e_2, \ldots, e_{\mu(G)})$, une base de cycles qui maximise $|B^* \cap \mathcal{C}|$ et $|B \cap B^*|$. Montrons que $|B \cap \mathcal{C}| = |B^* \cap \mathcal{C}|$. Nous considérons alors que B^* est ordonée selon l'ordre stipulé à la Ligne 3 de l'Algorithme 5. Les cycles sont donc ordonnés par poids et en cas d'égalité les cycles de $\mathcal{C}$ sont placés en premier. Supposons que B ne maximise par l'intersection avec $\mathcal{C}$, c'est-à-dire que $|B \cap \mathcal{C}| < |B^* \cap \mathcal{C}|$. Alors, nous avons $e_i \in B^*$ le premier cycle de $B^* \cap \mathcal{C}$ n'appartenant pas à B.

La base de cycles B est bien minimum car il s'agit des $\mu(G)$ plus petits cycles de $\mathcal{S}$ constituant une famille libre. Il existe donc un sous-ensemble $D \subseteq B$ permettant de générer e_i, soit $e_i = \bigoplus_{d \in D} \lambda_B(d, e_i) \cdot d$. Par définition, $D \not\subseteq B^*$, car sinon B^* ne constitue pas une famille libre. À partir du Lemme 2, nous avons $\forall d \in D, \omega(d) \leq \omega(e_i)$. Donc, pour $d \in D$, nous obtenons $(B^* \setminus e_i) \cup \{d\} \in \mathcal{MCB}(G)$. Si $\omega(d) < \omega(e_i)$, alors la base issue de cet échange a un poids inférieur à B^* ce qui entre en contradiction avec la propriété de minimalité de B^*. Ainsi, il existe $\hat{d} \in D$ tel que $\omega(\hat{d}) = \omega(e_i)$. Les cycles e_i et $\hat{d}$ ayant le même poids, comme $\hat{d} \in b$ et $e_i \notin B$, alors nous pouvons déduire que l'ordre de la Ligne 3 a placé $\hat{d}$ avant e_i dans $\mathcal{S}$. Sachant que $e_i \in \mathcal{C}$, alors c'est également le cas pour $\hat{d}$. Soit $B' = (B^* \setminus \{e_i\}) \cup \{\hat{d}\}$ la base de cycles minimum issue de l'échange de e_i par $\hat{d}$. La base B' a le même poids que B^*, la même cardinalité d'intersection avec $\mathcal{C}$, mais présente une plus grande intersection avec B. Une telle base B' ne peut pas exister sans entrer en contradiction avec la définition de B^*.

Nous pouvons conclure que l'Algorithme 5 renvoie une base de cycles minimum B de G qui maximise $|B \cap \mathcal{C}|$ en un temps polynomial. □

Nous présentons maintenant, l'Algorithme 6, un oracle polynomial qui vérifie si un ensemble X est un indépendant. L'algorithme utilise la fonction MCBset décrite par l'Algorithme 5. Ainsi, l'Algorithme 6 prend en entrée un ensemble de cycles X et un graphe G. Il renvoie vrai si X est un indépendant du matroïde $M(G)$, et faux sinon.

Algorithme 6 Un ensemble est un sous-ensemble d'une base de cycles minimum

Entrée Un graphe $G = (V, E)$ et un ensemble de cycles e
Sortie Vrai, si e est un sous-ensemble d'une base de cycles minimum ; Faux sinon
1: $B =$ MCBset(G, e)
2: **Si** $e \subseteq B$ **alors**
3: **Renvoyer** Vrai
4: **Sinon**
5: **Renvoyer** Faux

Théorème 2. *Le problème* Intersection de bases de cycles minimum *est polynomial pour* $N = 2$.

Démonstration. Étant donnés deux graphes G_1 et G_2, les matroïdes $M(G_1)$ et $M(G_2)$ du Théorème 1 sont obtenus en utilisant l'Algorithme 4. L'Algorithme 6 est un oracle polynomial qui vérifie si un ensemble X est un indépendant dans $M(G_1)$ et dans $M(G_2)$. Le

problème MCBI pour $N = 2$ est donc résolu en temps polynomial, par l'application de cet oracle à tous les indépendants des matroïdes $M(G_1)$ et $M(G_2)$. □

3.2.2 . Complexité de MCBI dans le cas général

Dans cette section, nous examinons la complexité du problème MCBI dans un contexte général. Dans la section précédente, nous avons établi une réduction polynomiale vers le problème de l'intersection de matroïdes. Ainsi, la complexité du problème MCBI appartient à NP. Nous montrons à présent que le problème MCBI n'est pas plus simple que le problème d'intersection des matroïdes, ce qui implique que MCBI est lui aussi NP-complet lorsqu'il concerne au moins trois graphes.

Pour démontrer cela, nous modélisons les choix induits par une base de cycles minimum à l'aide d'un outil décrit dans la Section 3.2.2. Cet outil, ou gadget, permet de sélectionner un élément parmi p à l'aide d'une base de cycles minimum.

Ensuite, nous démontrons dans la Section 3.2.2 que le problème MCBI est bien NP-complet lorsqu'il implique au moins trois graphes. Pour cela, nous proposons une réduction polynomiale à partir du problème du chemin Hamiltonien. Cette preuve suit la même structure que la preuve de NP-complexité du problème d'intersection des matroïdes établie dans [53].

Choisir un cycle parmi p à l'aide d'une base de cycles minimum

Nous considèrons un ensemble de p cycles de taille 4 (carrés) disjoints, $C = c_1, c_2, \ldots, c_p$. À partir de cet ensemble C, nous construisons le graphe GAD$(c_1, c_2, \ldots, c_p)$ qui est planaire et connexe. La procédure consiste à entourer les cycles donnés en entrée par un ensemble de cycles plus légers. L'objectif étant que toute base de cycles minimum B de GAD$(c_1, c_2, \ldots, c_p)$ contienne exactement $p - 1$ cycles parmi les p cycles de C donnés en entrée. Ainsi, le calcul de n'importe quelle base de cycles minimum de GAD implique de choisir un seul cycle parmi p. Ce choix correspond au cycle qui n'appartient pas à la base.

Remarque 8. *Il est possible de généraliser cet outil* GAD *pour considérer des cycles de tout poids, tant que tous les cycles de l'ensemble donné ont le même poids. Cependant, il est nécessaire d'adapter la procédure en fonction du poids des cycles. Les cycles ajoutés par la procédure doivent être plus légers mais ne doivent pas suffire à générer l'ensemble des cycles du graphe final.*

Plusieurs algorithmes sont envisageables pour décrire cette procédure qui entoure nos cycles de même poids par des cycles plus légers. L'Algorithme 7 décrit une de ces procédures en détail. Ainsi, l'objectif de cet algorithme est d'illustrer la construction d'un graphe GAD à partir d'un ensemble de cycles donné. La figure associée, Figure 3.2, illustre cet algorithme avec un exemple. Il est important de noter que le plan du graphe résultant est primordial pour la suite. Nous considérons donc que le graphe résultant respecte le même plan que celui illustré dans la Figure 3.2. Dans ce graphe planaire, les cycles sont représentés alignés côte à côte, et les sommets ainsi que les arêtes sont ajoutés étape par étape sur la face extérieure du graphe.

Nous allons maintenant montrer que l'Algorithme 7 construit un graphe vérifiant la propriété suivante.

Algorithme 7 GAD

Entrée un ensemble de cycles $\{c_1, c_2, \ldots, c_p\}$ de poids 4.
Sortie un graphe G tel que $\forall B \in \mathcal{MCB}(G)$, $\exists c \in \{c_1, c_2, \ldots, c_p\}, \{\{c_1, c_2, \ldots, c_p\} \setminus \{c\}\} \subset B$.

1: Le graphe G est construit à partir de l'ensemble de cycles disjoints donné. Chaque cycle $c_{1 \leq i \leq p}$ est un carré composé des sommets A_i, B_i, C_i et D_i (dans cet ordre).
▷ Deux cycles c_i et c_j sont positionnés côte à côte dans le plan comme illustré dans la Figure 3.2

2: **Pour tout** $i \in [1, p-1]$ **faire**
3: Ajouter à G un sommet v_i entre c_i et c_{i+1}.
4: Ajouter à G les arêtes :$[B_i, v_i]$, $[C_i, v_i]$, $[A_{i+1}, v_i]$, $[D_{i+1}, v_i]$, $[B_i, A_{i+1}]$ et $[C_i, D_{i+1}]$
5: $E \leftarrow$ la face externe de G.
6: **Pour tout** $e_i = [x, y] \in E$ **faire**
7: Ajouter à G un sommet w_i et les arêtes $[w_i, y]$ et $[w_i, x]$
8: **Pour tout** i allant de 2 à $|E|$ **faire**
9: Ajouter une arête $[w_{i-1}, w_i]$ à G.
10: Si elle n'existe pas déjà : ajouter une arête $[w_1, w_i]$ à G
11: **Renvoyer** G

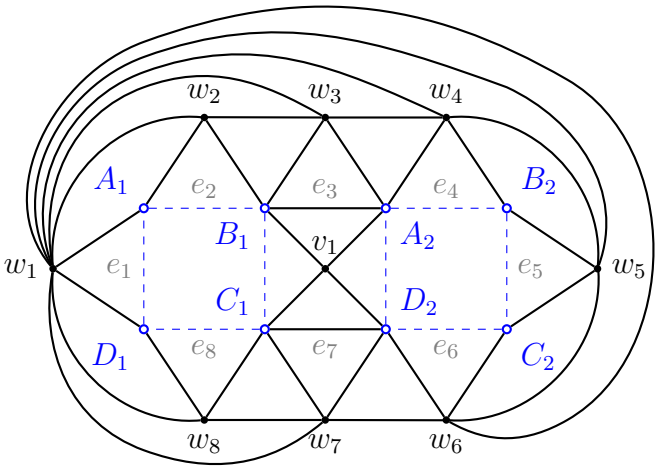


Figure 3.2 – Illustration détaillée d'un graphe GAD(c_1, c_2)

Propriété 5. $\forall B \in \mathcal{MCB}(\text{GAD}(c_1, \ldots, c_p))$, $\exists c \in C = \{c_1, \ldots, c_p\}$ *tel que* $\{C \setminus \{c\}\} \subset B$.

Lemme 8. *Pour tout graphe G planaire, l'ensemble des faces intérieures est une base de cycles du graphe.*

Démonstration. D'après le théorème d'Euler pour les graphes planaires, l'ensemble des faces du graphe $G = (V, E)$ est de taille $|E| - |V| + 2$ en incluant la face extérieure. Il

y a donc $|E| - |V| + 1$ faces intérieures. Dans un graphe, deux faces ont, par définition, des ensembles d'arêtes différents et sont donc indépendantes. Une base de cycles d'un graphe connexe G est un ensemble de $\mu_G = |E| - |V| + 1$ cycles, tous linéairement indépendants. L'ensemble des faces intérieures est donc bien une base de cycles de G. □

Nous pouvons même aller plus loin car tout ensemble de μ_G faces de G constitue également une base de cycles du graphe. En effet, la face extérieure est générée par toutes les autres faces réunies.

Nous allons maintenant montrer que, étant donné le graphe GAD construit, ses faces sont soit des cycles de C, soit des triangles construits par la procédure, et que nous pouvons construire une base de cycles minimum à partir de celles-ci. On note $T = \{t_1, t_2, \ldots, t_q\}$ cet ensemble de triangles.

Lemme 9. *Chaque cycle de T utilise au moins un sommet ajouté par la procédure* GAD

Démonstration. Supposons qu'il existe un triangle $t \in T$ qui n'utilise aucun nouveau sommet. Les cycles donnés sont des carrés disjoints qui ne contiennent aucun triangle. Ainsi, pour qu'un tel triangle t existe, la procédure doit ajouter deux arêtes partageant la même extremité et ne devant impliquer aucun des nouveaux sommets. Considérons les arêtes ajoutées par la procédure qui n'impliquent pas de nouveaux sommets. La seule ligne de l'Algorithme 7 proposant ce type d'arêtes est la Ligne 4. Cependant, les arêtes ajoutées ne présentent pas d'extremité commune. Nous pouvons donc conclure. □

Les cycles de T sont, par définition, à la fois des triangles et des faces. Notons que ces triangles sont linéairement indépendants des cycles de C, ce qui signifie que les cycles de T ne peuvent pas être construits uniquement avec les cycles de C. En effet, comme le montre le Lemme 9 chaque cycle de T utilise au moins un sommet ajouté par la procédure GAD.

Notons que la procédure ajoute chaque arête dans le contexte de la construction d'un triangle. Ainsi, chaque arête ajoutée appartient à un triangle. De plus, tout cycle c avec $\omega(c) = 3$ est nécessairement une face du graphe GAD. En effet, en reprenant l'Algorithme 7, nous avons :

- les arêtes ajoutées par la Ligne 4 participent à plusieurs cycles de taille 3, tous passant par v_i ;
- les arêtes $[w_i, y]$ et $[w_i, x]$ ajoutées par la Ligne 7 participent au moins au triangle w_i, y, x ;
- les arêtes $[w_{i-1}, w_i]$ et $[w_1, w_i]$ ajoutées par les lignes 9 et 10 participent au moins au triangle w_1, w_{i-1}, w_i.

De plus, ces arêtes sont ajoutées sur la face extérieure du graphe, délimitant ainsi une nouvelle face pour chacune d'elles. En conséquence, chaque triangle construit par GAD correspond à une face du graphe résultant.

Le lemme suivant établit que tous les cycles de poids 4 n'appartenant pas à C ne peuvent pas appartenir à une base de cycles minimum.

Lemme 10. *Tout cycle c avec $\omega(c) = 4$ est soit un cycle de C, soit généré par des cycles de T.*

Démonstration. Tout cycle c avec $\omega(c) = 4$, qui n'appartient pas à l'ensemble C, est nécessairement composé d'au moins deux arêtes ajoutées par la procédure GAD. En effet, un

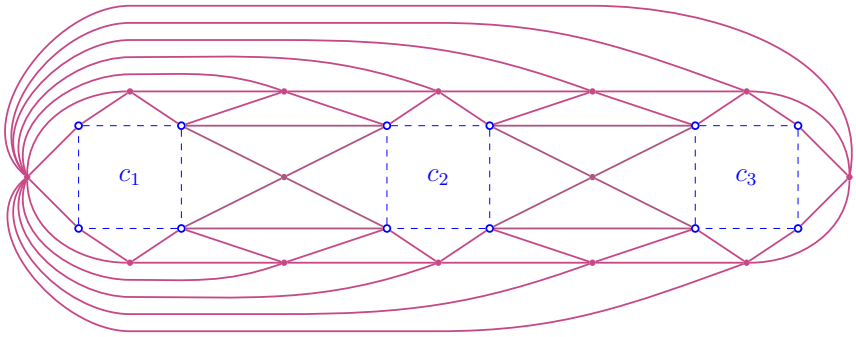


Figure 3.3 – Illustration du sous-graphe connexe obtenu par l'application d'une procédure GAD(c_1, c_2, c_3) sur un ensemble de trois cycles.

cycle composé de trois arêtes non ajoutées par la procédure est au moins de poids 5. Cela est dû au fait que les arêtes adjacentes aux carrés de départ participent à des triangles composés de deux arêtes ajoutées par GAD. Pour revenir à un cycle de poids 4, si celui-ci est composé de 3 ou 4 arêtes ajoutées par la procédure GAD, il est facile de voir que étant donné que chaque arête ajoutée par la procédure appartient à un triangle et que ces triangles sont des faces, alors c est généré par un sous-ensemble de T. Les cycles de poids 4 composés de deux arêtes ajoutées par la procédure GAD sont très spécifiques car ils sont construits lors de la première boucle de l'Algorithme 7. Ainsi, ils sont obtenables par la combinaison des quatres triangles construits à la même étape. Par conséquent, tout cycle c avec $\omega(c) = 4$ qui n'appartient pas à l'ensemble C est généré par un sous-ensemble de T. □

Comme défini précédemment, tout cycle de T représente une face (intérieure ou extérieure) du graphe GAD$(c_1, \ldots, c_p)$. Cependant, toute face n'est pas nécessairement un cycle de T. Les cycles de C sont également des faces du graphe GAD$(c_1, \ldots, c_p)$.

Nous considérons, $f \in T$, le triangle correspondant à la face extérieure de GAD. À partir de cette face extérieure, nous construisons la base de cycles $B = \{T \setminus \{f\}\} \cup C$, qui représente l'ensemble des faces intérieures. En utilisant cette base de cycles, nous pouvons définir la base de cycles minimum $B' \in \mathcal{MCB}(\text{GAD})$. Pour ce faire, il suffit de supprimer l'un des cycles de C, de le remplacer par f, et ainsi obtenir la base B'.

Il n'existe pas d'autres triangles que ceux de T, donc $T \subset B'$ est vide. Ainsi, aucun autre cycle de poids 3 ne peut être ajouté à la base. De même, aucun cycle de C dans B' ne peut être remplacé par un autre cycle n'appartenant pas à C. Comme le montre le Lemme 10, seuls les cycles de C peuvent être à la fois de longueur 4 et linéairement indépendants à l'ensemble T.

En conclusion, le graphe résultant de la procédure GAD satisfait la Propriété 5.

La Figure 3.3 illustre l'application d'une procédure GAD sur un ensemble de trois cycles. Dans cet exemple, toute base de cycles minimum du graphe contient deux carrés parmi c_1, c_2 et c_3.

Dans la suite, nous nous intéressons au graphe final obtenu par l'application de plusieurs procédures GAD sur différents sous-ensembles de cycles. Si les différentes procédures GAD sont disjointes, le graphe final contient alors plusieurs composantes connexes.

La base de cycles minimum d'un tel graphe est l'union des bases de cycles minimum des composantes connexes du graphe. Autrement dit, elle correspond à l'union des bases de cycles minimum des graphes GAD, qui respectent chacune la propriété 5.

En revanche, si plusieurs procédures GAD sont appliquées sur des ensembles de cycles non disjoints, le résultat est différent. En effet, lorsque deux procédures GAD partagent des cycles communs, cela réduit le nombre de carrés présents dans une base de cycles minimum de la composante connexe qui les contient. La Figure 3.4 illustre un tel exemple avec un ensemble de trois cycles c_1, c_2, et c_3, sur lequel nous avons appliqué GAD(c_1, c_2) et GAD(c_2, c_3). Puisque les deux procédures GAD partagent le cycle c_2, elles ne sont pas disjointes.

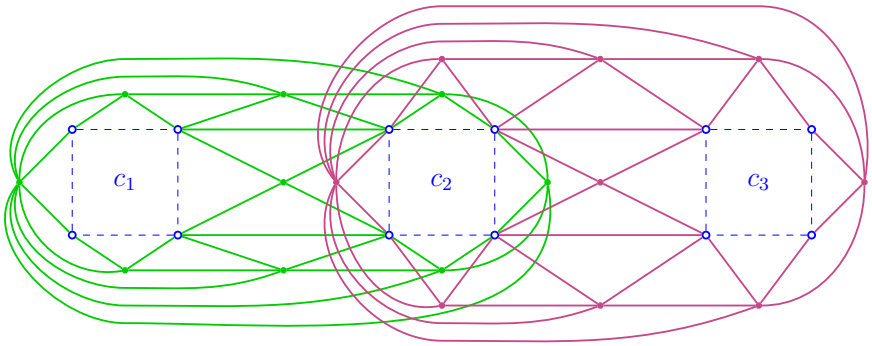


Figure 3.4 – Illustration du sous-graphe connexe obtenu par l'application de deux procédures GAD(c_1, c_2) et GAD(c_2, c_3) sur des ensembles de cycles liés. Les arêtes et sommets ajoutés par GAD(c_1, c_2) sont en vert, tandis que ceux ajoutés par GAD(c_2, c_3) sont en rose.

Dans le graphe final, une composante connexe contiendra cette union de GAD non disjoints, et le nombre de carrés dans la base de cycles minimum de cette composante connexe en sera réduit.

Lemme 11. *Étant donné un ensemble de carrés* $\{c_1, c_2, c_3\}$*, si* $G =$ GAD$(c_1, c_2) \cup$ GAD(c_2, c_3) *alors* $\forall B \in \mathcal{MCB}(G)$*,* B *contient exactement un et un seul carré parmi* c_1, c_2 *et* c_3*.*

Démonstration. Soit B une base de cycles minimum de $G =$ GAD$(c_1, c_2) \cup$ GAD(c_2, c_3). Notons T l'ensemble des triangles de GAD(c_1, c_2) et T' l'ensemble des triangles de GAD(c_2, c_3).

Comme démontré dans le Lemme 9, tous les triangles d'une procédure utilisent les arêtes ajoutées par celle-ci. Ainsi, nous avons $T \cap T' = \emptyset$. De plus, comme nous l'avons déjà évoqué dans la preuve de la Propriété 5, il n'existe pas de triangles qui n'appartiennent pas à la base du graphe GAD. Ainsi, nous avons $T \cup T' \subset B$. Or, $\forall i$ tel que $1 \leq i \leq 3, c_i \notin$ span$(T \cup T')$, donc B contient au moins un carré.

Sans perte de généralité, supposons $c_1 \in B$, soit $B = \{c_1\} \cup T \cup T'$. Alors $c_2 \in$ span(B) car $c_2 = \bigoplus T \oplus c_1$. De même, $c_3 = \bigoplus T' \oplus c_2 = \bigoplus T' \bigoplus T \oplus c_1$, donc $c_3 \in$ span(B).

Enfin, d'après le Lemme 10, nous savons que dans un graphe GAD tout carré linéairement indépendant de l'ensemble des triangles est l'un des carrés de départ. Ainsi, nous pouvons conclure. □

MCBI est NP-complet pour $N \geq 3$

Dans cette section, nous présentons une réduction polynomiale du problème du chemin hamiltonien vers le problème Intersection de bases de cycles minimum (MCBI).

Étant donné un graphe orienté $G = (V, A)$, le problème du chemin hamiltonien est de rechercher si un chemin hamiltonien existe entre s et t, deux sommets de G. Rappelons que dans un graphe, un chemin hamiltonien est un chemin passant une et une seule fois par tous les sommets du graphe. Dans le problème présent, le sommet s est considéré sans arc entrant, et le sommet t est considéré sans arc sortant. Ce problème est connu pour être NP-complet [31].

Remarque 9. *La réduction proposée s'inspire de celle utilisée dans [53] pour établir que le problème d'intersection d'au moins trois matroïdes est NP-complet.*

Considérons une instance $G = (V, A)$ du problème du chemin hamiltonien. Nous définissons trois graphes, G_1, G_2, et G_3, de telle sorte que chacun contienne un carré c_a pour chaque arc $a \in A$. La procédure GAD est ensuite appliquée sur différents ensembles de carrés afin de représenter les liens entre les arcs qu'ils représentent. Notons que, étant donné un sommet v, $\Gamma^+(v)$ désigne ses successeurs et $\Gamma^-(v)$ désigne ses prédécesseurs.

La Figure 3.5 illustre l'application des procédures GAD pour construire les graphes G_1, G_2 et G_3 sur un exemple. À chaque fois la procédure est appliquée autour d'un ensemble de cycles, par exemple étant donné l'ensemble de cycles {c_{a_1} et c_{a_2}}, GAD(c_{a_1}, c_{a_2}) remplace alors le sous-graphe correspondant aux cycles c_{a_1} et c_{a_2}.

- Dans G_1, $\forall v \in V$ et $\forall w_1, w_2 \in \Gamma^+(v)$, la procédure GAD($c_{(v,w_1)}$, $c_{(v,w_2)}$) est appliquée autour de $c_{(v,w_1)}$, $c_{(v,w_2)}$.
- Dans G_2, $\forall v \in V$ et $\forall w_1, w_2 \in \Gamma^-(v)$, la procédure GAD($c_{(w_1,v)}$, $c_{(w_2,v)}$) est appliquée autour de $c_{(w_1,v)}$, $c_{(w_2,v)}$.
- Dans G_3, considérons une base de cycles $\{d_1, d_2, \ldots, d_p\}$ du multigraphe non orienté sous-jacent à G. Pour tout $i \in [1, p]$, la procédure GAD($\{\bigcup_{a \in d_i} c_a\}$) est appliquée autour de l'ensemble de cycles $\{\bigcup_{a \in d_i} c_a\}$.

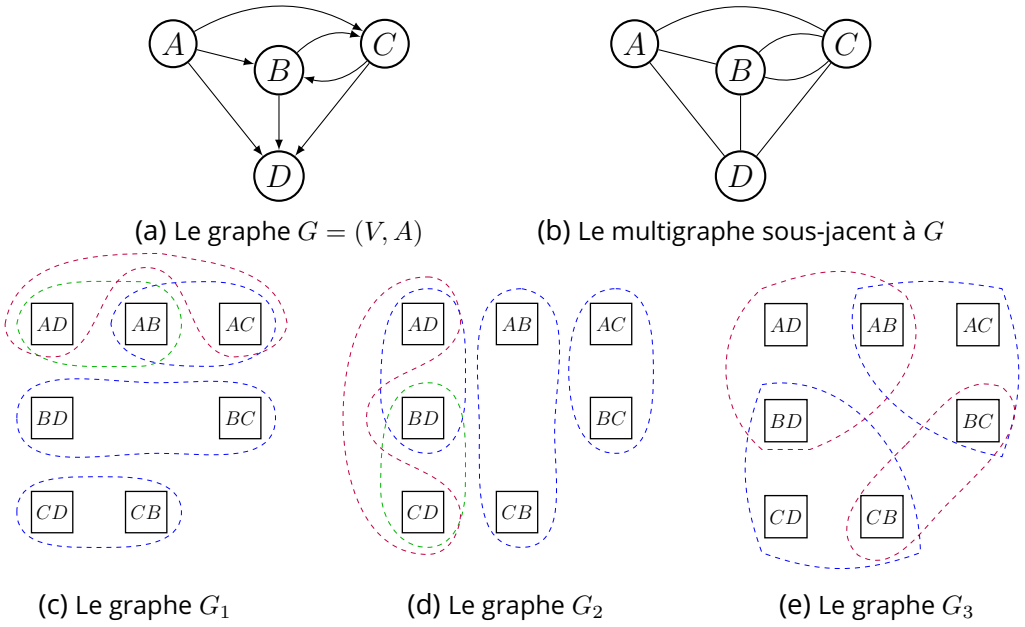
Les graphes G_1, G_2 et G_3 doivent former un ensemble de graphes instance du problème MCBI, ce qui signifie qu'ils doivent partager le même ensemble de sommets. Pour garantir cela, chaque nœud ajouté par une procédure GAD dans un graphe est également ajouté de manière déconnectée du reste des noeuds dans les deux autres graphes. Ainsi, les trois graphes ont le même ensemble de sommets.

Nous allons maintenant démontrer plusieurs lemmes concernant les carrés contenus par les bases de cycles minimum des graphes que nous avons construits.

Lemme 12. *Une base de cycles minimum de G_1 contient exactement un carré $c_{(v,w)}$ pour chaque sommet $v \in V$ tel que $|\Gamma^+(v)| \geq 1$.*

Démonstration. Le graphe G_1 est constitué d'une composante connexe pour chaque sommet v dans l'ensemble des sommets V. Chaque composante est formée par les procédures GAD correspondantes aux arcs sortants de v.

- Si $\Gamma^+(v) = \{w_1\}$, cela signifie qu'aucune procédure GAD n'a été appliquée. Dans ce cas, la composante connexe de v se résume au carré c_{v,w_1}.



(a) Le graphe $G = (V, A)$ (b) Le multigraphe sous-jacent à G

(c) Le graphe G_1 (d) Le graphe G_2 (e) Le graphe G_3

Figure 3.5 – Illustration de la construction des graphes G_1, G_2, et G_3 à partir d'un graphe orienté $G = (V, A)$. La base de cycles considérée dans G_3 est l'ensemble des faces intérieures, selon la représentation du multigraphe de la Figure 3.5b. Dans les Figures 3.5c, 3.5d et 3.5e, les lignes en pointillés représentent les ensembles de carrés sur lesquels la procédure est appliquée. Pour plus de lisibilité, des couleurs différentes sont utilisées pour illustrer les procédures GAD lorsqu'elles ont des carrés en commun.

- Si $\Gamma^+(v) = \{w_1, w_2\}$, alors une seule procédure GAD a été appliquée. Ainsi, le sous-graphe GAD(c_{v,w_1}, c_{v,w_2}) est la composante connexe correspondant à v. Or, la Propriété 5 précise bien qu'un des carrés de départ, donc c_{v,w_1} ou c_{v,w_2}, n'appartient pas à la base de cycles minimum du graphe résultant. Autrement dit, un seul des deux carrés de départ appartient à la base de cycles minimum.
- Si $|\Gamma^+(v)| > 2$, alors plusieurs procédures GAD constituent la composante de v. Pour chaque couple de sommets $(w_1, w_2) \in \Gamma^+(v)$, une procédure GAD a été appliquée. Cependant, ces procédures GAD sont liées, donc un seul carré appartiendra à la base de cycles minimum de cette composante connexe. Le Lemme 11 présente la preuve de cela dans le cas de deux procédures liées. Les mêmes observations nous permettent de constater que dans le cas présent, où toutes les procédures GAD sont liées entre elles, la conclusion est la même. Ainsi, si un seul carré est sélectionné, tous les autres sont accessibles par composition.

Une base de cycles minimum de G_1 est l'union des bases de cycles minimum de ses composantes connexes. Par conséquent, une base de cycles minimum de G_1 contient exactement un carré $c_{(v,w)}$ pour chaque sommet $v \in V$ tel que $|\Gamma^+(v)| \geq 1$. □

Lemme 13. *Une base de cycles minimum de G_2 contient exactement un carré $c_{(w,v)}$ pour chaque sommet $v \in V$ tel que $|\Gamma^-(v)| \geq 1$.*

Démonstration. Dans G_2, les procédures GAD ont été appliquées autour de couples de cycles correspondant à un arc entrant d'un même sommet v. Ainsi, de manière comparable à G_1, une base de cycles de G_2 ne contient pas plusieurs carrés correspondant à un arc entrant d'un même sommet v. □

Le Lemme 16 démontre que les carrés appartenant à une base de cycles minimum de G_3 représentent un arbre couvrant du multigraphe sous-jacent G. Nous introduisons préalablement les lemmes intermédiaires suivants qui seront nécessaires à la preuve.

Lemme 14. *$\forall B_3 \in \mathcal{MCB}(G_3)$, si c est un cycle de G alors $\{\bigcup_{a\in c} c_a\} \not\subseteq B_3$.*

Démonstration. Considérons c un cycle de G. Nous allons montrer que l'union de l'ensemble des triangles de G_3, notée T, et de l'ensemble des carrés $C_a = \bigcup_{a\in c} c_a$ n'est pas une famille libre.

Soient $B_3 = \{e_1, e_2, \ldots, e_l\}$ une base de cycles minimum de G_3 et $B = \{d_1, d_2, \ldots, d_p\}$ la base de cycles de G utilisée lors de la construction de G_3. Nous pouvons exprimer $c = \bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot d_i$. Rappelons que $\lambda_B(d_i, c)$ vaut 1 si d_i est nécessaire pour générer c à partir de B et 0 sinon.

La somme S, ci-dessous, décrit les ensembles de cycles obtenus à partir des applications des procédures $\mathsf{GAD}_i = \mathsf{GAD}(c_a, a \in d_i)$. L'objectif est de montrer par cette somme que si c est un cycle, alors $\{\bigcup_{a\in c} c_a\}$ est un ensemble lié.

$$
\begin{aligned}
S &= \bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot (c_a, a \in d_i) \\
&= \bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot \Big(\bigoplus_{t\in\{\mathsf{GAD}_i\cap T\}} t + \bigoplus_{c_a\in\{\mathsf{GAD}_i\cap C_a\}} c_a \Big) \\
&= \bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot \Big(\bigoplus_{t\in\{\mathsf{GAD}_i\cap T\}} t + \bigoplus_{a\in\{d_i\cap C\}} c_a \Big)
\end{aligned}
$$

L'ensemble des faces intérieures de GAD_i correspond à l'union de l'ensemble des triangles $\{\mathsf{GAD}_i \cap T\}$ et des carrés $\{c_a | a \in d_i\}$, alors

$$
\bigoplus_{t\in\{\mathsf{GAD}_i\cap T\}} t + \bigoplus_{a\in\{d_i\cap c\}} c_a = \bigoplus_{a\in\{d_i\setminus c\}} c_a
$$

$$
S = \bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot \bigoplus_{a\in\{d_i\setminus c\}} c_a
$$

Considérons chaque arc a de l'ensemble $\{\bigcup_{i=1}^{p} d_i\}$. Si $a \in c$, alors a apparaît un nombre impair de fois dans la somme $\bigoplus_{i=1}^{p} \lambda_B(d_i, c) \cdot d_i$, sinon, a y apparaît un nombre pair de fois.

Dans la somme S, ci-dessus, pour chaque arc $a \notin d_i$, le cycle c_a apparaît un nombre pair de fois. Ainsi, $S = 0$, et nous pouvons conclure que $\{C_a \cup T\}$ n'est pas une famille libre.

□

Lemme 15. *$\forall B_3 \in \mathcal{MCB}(G_3)$, un ensemble d'arcs e_A contient un cycle de G si et seulement si $\{\bigcup_{a\in e_A} c_a\} \not\subseteq B_3$.*

Démonstration. Considérons un ensemble d'arcs $e_A \subseteq A$ tel que l'union de l'ensemble des triangles de G_3, notée T, et de l'ensemble de carrés $C_a = \{c_a \mid a \in e_A\}$ ne forme pas une famille libre. Supposons que e_A est minimum, c'est-à-dire que $\forall C'_a$ obtenu par la suppression d'un arc dans e_A, l'ensemble $\{T \cup C'_a\}$ est une famille libre.

Nous avons nécessairement $|e_A| > 0$ car T étant une famille libre, si $|e_A| = 0$, alors $\{T \cup C_a\}$ est une famille libre. Notons $e_A = \{a_1, a_2, \ldots, a_{|e_A|}\}$; il existe un sous-ensemble $T' \subset T$ tel que $\bigoplus_{t \in T'} t + \bigoplus_{j=1}^{|e_A|} c_{a_j} = 0$.

Soit $B = d_1, d_2, \ldots, d_p$ la base de cycles de G utilisée lors de la construction de G_3. Prouvons maintenant que si T' contient un triangle de GAD_i, alors T' contient tous les triangles de GAD_i. Supposons qu'il existe au moins une arête $e \in \mathsf{GAD}_i$ partagée par deux triangles t et t' tels que $t \in T'$ et $t' \notin T'$. Notons que t et t' sont les seuls cycles de $\{T \cup C_a\}$ contenant e. Ainsi, la somme suivante ne peut pas valoir zéro car aucun des autres cycles ne peut annuler l'arête e ajoutée par t :

$$\bigoplus_{t \in T'} t + \bigoplus_{j=1}^{|e_A|} c_{a_j} \neq 0$$

Sans perte de généralité, considérons $T' = \bigcup_{i=1}^{q} \mathsf{GAD}_i$ pour un $q \leq p$. Montrons à présent que $\bigoplus_{i=1}^{q} d_i = e_A$. Pour tout $i \leq q$, $\bigoplus_{t \in \{T \cap \mathsf{GAD}_i\}} t + \bigoplus_{a \in d_i} c_a = 0$. Alors,

$$\bigoplus_{t \in T'} t + \bigoplus_{j=1}^{|e_A|} c_{a_j} + \bigoplus_{i=1}^{q} \Big(\bigoplus_{t \in \{T \cap \mathsf{GAD}_i\}} t + \bigoplus_{a \in d_i} c_a \Big) = 0$$

Sachant que chaque triangle de T' appartient à $\mathsf{GAD}_{1 \leq i \leq q}$, nous avons :

$$\bigoplus_{j=1}^{|e_A|} c_{a_j} + \bigoplus_{i=1}^{q} \bigoplus_{a \in d_i} c_a = 0 \iff \bigoplus_{j=1}^{|e_A|} c_{a_j} = \bigoplus_{i=1}^{q} \bigoplus_{a \in d_i} c_a$$

Les carrés de G_3 étant disjoints, cette égalité implique que $\bigoplus_{i=1}^{q} d_i = e_A$. Autrement dit, la base de cycles B génère e_A, et donc e_A est un cycle.

Notons que si e_A n'est pas *minimum*, comme nous l'avons supposé au début, alors e_A contient un cycle.

Pour conclure, e_A est un ensemble d'arcs contenant un cycle de G si et seulement si $\{T \cup C_a\}$ n'est pas une famille libre. □

Lemme 16. *$\forall B_3 \in \mathcal{MCB}(G_3)$, un ensemble d'arcs e_A contient un cycle de G si et seulement si $\{\bigcup_{a \in e_A} c_a\} \not\subseteq B_3$.*

Démonstration. Le Lemme 14 démontre que si e_A est un cycle de G, alors l'ensemble des carrés qui le représentent ne peut pas être un sous-ensemble de $B_3 \in \mathcal{MCB}(G_3)$. Le Lemme 15 montre la proposition inverse : si un sous-ensemble de carrés ne peut pas appartenir à une base $B_3 \in \mathcal{MCB}(G_3)$, alors les arcs correspondants décrivent un cycle de G.

Par conséquent, toute base $B_3 \in \mathcal{MCB}(G_3)$ représente un arbre couvrant du multi-graphe non-orienté sous-jacent à G.

□

Théorème 3. *Le problème* Intersection de bases de cycles minimum *est NP-Complet pour $N \geq 3$.*

Démonstration. Le Lemme 12 montre que les bases de cycles minimum de G_1 représentent un seul arc sortant pour chaque sommet $v \in V$ ayant au moins un successeur. Ensuite,

le Lemme 13 montre que les bases de cycles minimum de G_2 représentent un seul arc entrant pour chaque sommet $v \in V$ ayant au moins un prédécesseur. Enfin, le Lemme 16 montre que les bases de cycles minimum de G_3 représentent des chemins sans cycles de G.

L'intersection des bases de cycles de G_1, G_2 et G_3 représente un chemin élémentaire qui passe une seule fois par chaque sommet. Cette intersection a une cardinalité de $|V|-1$ si et seulement si G est un graphe hamiltonien. □

Cela conclut notre étude de la complexité du problème MCBI. Nous allons maintenant proposer des méthodes inspirées par ce problème pour sélectionner des bases de cycles minimum dans le cadre de la construction du polygraphe.

3.2.3 . Une méthode gloutonne pour le calcul des bases inspirée par max-MCBI

Dans cette section, nous présentons une méthode gloutonne pour le calcul des bases, directement inspirée par max-MCBI. La section précédente a montré que le problème max-MCBI est NP-complet lorsqu'il implique au moins trois graphes. Ainsi, la résolution exacte de ce problème sera toujours relativement coûteuse. Étant donné que la pertinence des bases de cycles minimum obtenues par ce problème pour le polygraphe n'est pas encore établie, nous proposons dans un premier temps une approche heuristique afin d'observer l'impact sur le polygraphe. Cette approche considère les graphes de la trajectoire un par un, et définit pour chacun une seule base de cycles minimum. Chaque base de cycles minimum est calculée en considérant l'intersection des bases de cycles minimum déjà établies. L'Algorithme 5, présenté dans la Section 3.2.1, décrit la fonction MCBset qui étant donné un graphe G et un ensemble de cycles C, calcule une base de cycles minimum de G dont l'intersection avec C est maximum.

Rappelons que $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ désigne l'ensemble des graphes de la trajectoire. La base de cycles minimum d'un premier graphe, supposons-le être G_1, est calculée à l'aide de l'algorithme de Horton (voir Section 3.1). Ensuite, pour chaque autre graphe de la trajectoire, soit pour tout i tel que $2 \leq i \leq N$, la base de cycles minimum de G_i est obtenue par l'appel de la fonction MCBset(G_i, $\bigcap_{j=1}^{i-1} B_j \cap G_i$), où B_j désigne la base de cycles minimum de G_j. Les détails de cette méthode sont présentés dans l'Algorithme 8.

Algorithme 8 Calcul d'un ensemble de bases de cycles minimum qui tendent à maximiser l'intersection globale.

Entrée Un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$
Sortie Un ensemble de bases de cycles minimum $\{B_1, B_2, \ldots, B_N\}$ tel que $\forall i, 1 \leq i \leq N, B_i \in \mathcal{MCB}(G_i)$

1: $\mathcal{B} \leftarrow \emptyset$
2: $B_1 \leftarrow$Horton(G_1)
3: Ajouter B_1 à $\mathcal{B}$
4: **Pour** i allant de 2 à N **faire**
5: $\quad B_i \leftarrow$MCBset(G_i, $\bigcap_{B \in \mathcal{B}} B \cap G_i$)
6: $\quad$ Ajouter B_i à $\mathcal{B}$
7: **Renvoyer** $\mathcal{B}$

Cependant, dans le contexte des trajectoires, résoudre max-MCBI pour l'ensemble des graphes de la trajectoire semble perdre de son intérêt. En effet, plus la dynamique représentée par une trajectoire est forte, moins les graphes se ressemblent, rendant ainsi l'intersection de toutes les bases vide. C'est pourquoi nous proposons de considérer les intersections locales entre les graphes que l'on sait proches. Il est important de rappeler que les conformations successives au sein d'une trajectoire ont peu de liaisons hydrogène distinctes, souvent une ou deux au plus. Ainsi, les graphes correspondants à ces conformations successives présentent, dans la plupart des cas, une relation d'inclusion, où le premier est un sous-graphe du second, ou inversement le second est un sous-graphe du premier.

La méthode que nous proposons maintenant considère les graphes dans leur ordre d'apparition au sein de la trajectoire. Ainsi, $T[x]$ désigne l'indice du $x^{\text{ème}}$ graphe observé dans la trajectoire.Inversement, nous index(G) désigne la première étape de la trajectoire qui correspond à l'observation de G.

Considérons deux graphes G_i et G_j tels que G_j apparaît pour la première fois dans la trajectoire en succédant à G_i. Autrement dit, pour $i \neq j$ et $1 \leq (i,j) \leq N$, index$(G_i) <$ index(G_j) et $T[\text{index}(G_j) - 1] = i$. Alors, nous calculons $B_j \in \mathcal{MCB}(G_j)$ par l'appel MCBset$(G_j, B_i \cap G_j)$ où B_i est la base de cycles minimum calculée pour G_i. Remarquons que si G_i est un sous-graphe de G_j, alors $B_i \cap G_j = B_i$. Soulignons également que le premier graphe de la trajectoire n'a aucun prédécesseur, sa base de cycles minimum est donc calculée à partir de l'algorithme de Horton (voir Section 3.1), comme dans l'approche précédente. Cette approche, maximisant localement les intersections de bases de cycles, est formalisée dans l'Algorithme 9.

Pour mieux comprendre la différence entre les deux méthodes, la Figure 3.6 présente un exemple.

$\forall i$ tel que $1 \leq i \leq 4$, $B_i \in \mathcal{MCB}(G_i)$	Algorithme. 8	Algorithme. 9
$B_1 =$	Horton(G_1)	Horton(G_1)
$B_2 =$	MCBset$(G_2, \bigcap_{i=1}^{1} B_i \cap G_2)$	MCBset$(G_2, B_1 \cap G_2)$
$B_3 =$	MCBset$(G_3, \bigcap_{i=1}^{2} B_i \cap G_3)$	MCBset$(G_3, B_1 \cap G_3)$
$B_4 =$	MCBset$(G_4, \bigcap_{i=1}^{3} B_i \cap G_2)$	MCBset$(G_4, B_2 \cap G_4)$

Figure 3.6 – Soit, la suite de graphes : $G_1, G_2, G_1, G_3, G_2, G_4, G_3$. Le tableau, ci-dessus, indique l'appel permettant le calcul de la base de cycles minimum en fonction de l'algorithme utilisé.

Les deux algorithmes que nous proposons offrent une alternative à l'utilisation de l'algorithme de Horton pour le calcul de chacune des bases de cycles minimum. Le second, présenté dans l'Algorithme 9, a été particulièrement orienté pour être appliqué aux trajectoires de forte dynamique moléculaire. Autrement dit, des trajectoires comportant de nombreux graphes très distincts les uns des autres.

Algorithme 9 Calcul d'un ensemble de bases de cycles minimum qui tendent à maximiser les intersections locales.

Entrée Un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, une liste d'entiers T qui représente l'ordre des graphes dans la trajectoire.

Sortie Un ensemble de bases de cycles minimum $\{B_1, B_2, \ldots, B_N\}$ tel que $\forall 1 \leq i \leq N, B_i \in \mathcal{MCB}(G_i)$

1: $\mathcal{B} \leftarrow \{B_1, B_2, \ldots B_N\}$ où $\forall i, 1 \leq i \leq N, B_i = Null$
2: $B_{T[1]} \leftarrow$ Horton$(G_{T[1]})$
▷ L'indice 1 désigne la première étape de la trajectoire
3: $x \leftarrow 2$
4: **Tant que** $\exists B \in \mathcal{B}, B = Null$ **faire**
5: **Si** $B_{T[x]} = Null$ **alors**
6: $B_{T[x]} =$ MCBset$(G_{T[x]}, B_{T[x-1]} \cap G_{T[x]})$
7: $x \leftarrow x + 1$
8: **Renvoyer** $\mathcal{B}$

3.3 . Vers des méthodes de voisinages pour la sélection des bases

Dans cette section, nous introduisons une nouvelle modélisation de la relation entre les cycles d'une trajectoire et les bases de cycles des conformations. Ce modèle propose une vision globale de l'ensemble des cycles impliqués dans les bases. Chaque solution représente la relation entre l'ensemble des cycles de la trajectoire et les différentes bases de cycles sélectionnées pour chacun des graphes. Les différentes solutions que nous allons explorer se distinguent par les bases de cycles sélectionnées pour chaque graphe. Ainsi, en explorant l'espace des solutions, nous devrons étudier divers critères pour déterminer ceux qui permettent de détecter les ensembles de bases de cycles qui permettant d'améliorer le polygraphe.

3.3.1 . Modéliser la répartition des cycles dans un ensemble de bases de cycles minimum

Nous présentons maintenant notre modélisation de la répartition des cycles dans des bases de cycles minimum par l'intermédiaire d'un graphe biparti. Nous introduisons ensuite une opération d'échange entre deux cycles, l'un étant remplacé par l'autre dans un ensemble de bases de cycles minimum. Cela afin de modifier l'ensemble des bases de cycles minimum et d'explorer un espace de solution. Enfin, nous présentons une trame générale pour notre méthode de voisinage.

Graphe biparti REPART$(\mathcal{G}, \mathcal{C}_{ext})$ La méthode va s'appuyer sur un graphe biparti noté REPART$(\mathcal{G}, \mathcal{C}_{ext})$ où $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ est l'ensemble des graphes d'une trajectoire, et où $\mathcal{C}_{ext}$ est un ensemble de cycles tel que pour tout cycle $c \in \mathcal{C}_{ext}$, avec $1 \leq i \leq N$ il existe $G_i \in \mathcal{G}$ et $B_i \in \mathcal{MCB}(G_i)$ tel que $c \in B_i$.

Remarque 10. *Rappelons que, $\mathcal{C}$, l'ensemble des cycles de la trajectoire est l'union des bases de cycles minimum sélectionnées pour chacun des graphes. Nous avons alors $\mathcal{C} \subseteq \mathcal{C}_{ext}$.*

Dans ce graphe REPART$(\mathcal{G}, \mathcal{C}_{ext})$ il existe une arête $[G_i, c_j]$ avec $1 \leq i \leq N$, et $1 \leq j \leq |\mathcal{C}_{ext}|$ si et seulement si $\exists B_i \in \mathcal{MCB}(G_i)$ et $B_i \subseteq \mathcal{C}_{ext}$ et $c_j \in B_i$. La Figure 3.7 illustre un tel

graphe biparti. Dans cette figure, on note l'association du graphe à un ensemble de bases de cycles minimum $\mathcal{B}$. En effet, étant donné un ensemble de bases $\mathcal{B} = \{B_1, B_2, \dots B_N\}$ dans lequel $\forall i$ avec $1 \leq i \leq N, B_i \subseteq \mathcal{C}_{ext}$, nous définissons une fonction de couverture $Cover_{\mathcal{B}}$. Notons que dans la Figure 3.7, les arêtes "couvertes" sont celles en pointillés rouges.

Dans notre méthode de voisinage, une solution courante est un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, \dots, B_N\}$. REPART$(\mathcal{G}, \mathcal{C}_{ext})$ et $Cover_{\mathcal{B}}$ sont des outils qui vont nous permettre de passer d'une solution courante à une solution voisine. Ainsi, étant donné l'ensemble $\mathcal{B}$, $Cover_{\mathcal{B}}$ indique les arêtes reliant un cycle au graphe pour lequel le cycle appartient bien à la base de cycles minimum. Nous avons donc, $\forall i$ avec $1 \leq i \leq N$ et $\forall j$ avec $1 \leq j \leq |\mathcal{C}_{ext}|$, $Cover_{\mathcal{B}}([G_i, c_j]) = 1$ si et seulement si $c_j \in B_i$ tel que $B_i \in \mathcal{MCB}(G_i)$ et $B_i \in \mathcal{B}$, sinon $Cover_{\mathcal{B}}([G_i, c_j]) = 0$.

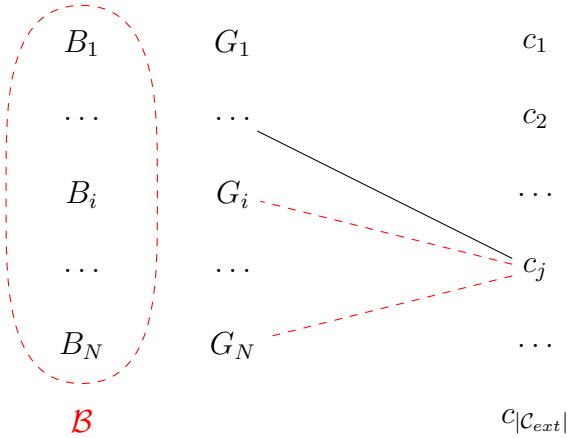


Figure 3.7 – Schéma d'un graphe biparti REPART$(\mathcal{G} \cup \mathcal{C}_{ext})$ avec $\mathcal{G} = \{G_1, \dots, G_i, \dots, G_N\}$ et $\mathcal{C}_{ext} = \{c_1, c_2, \dots, c_j, \dots, c_{|\mathcal{C}_{ext}|}\}$ associé à un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, \dots, B_i, \dots, B_N\}$. Les arêtes du graphe sont représentées en traits pleins ou en pointillés rouge, celles en pointillés représentent les arêtes couvertes par $\mathcal{B}$.

Solutions voisines Dans la méthode de voisinage que nous proposons, nous explorons des solutions de proche en proche. Ainsi, nous définissons deux solutions $\mathcal{B}$ et $\mathcal{B}'$ comme voisines si elles correspondent à un échange entre deux cycles c et c' dans toutes les bases de $\mathcal{B}$ dans lesquelles c' peut remplacer c. Celles-ci constituent alors un ensemble de bases de cycles minimum $Swap(c, c') \subset \mathcal{B}$ avec $c, c' \in \mathcal{C}_{ext}$, et la solution $\mathcal{B}' = \{B'_1, B'_2, \dots, B'_N\}$ est construite comme suit :

- $\forall i$ tel que $1 \leq i \leq N$, si $B_i \in \mathcal{B} \setminus Swap(c, c')$ alors $B'_i = B_i$.
- $\forall i$ tel que $1 \leq i \leq N$, si $B_i \in Swap(c, c')$ alors $B'_i = (B_i \setminus \{c\}) \cup \{c'\}$.

La Figure 3.8 illustre la création d'un ensemble $\mathcal{B}'$ donnant lieu à une couverture $Cover_{\mathcal{B}'}$, à partir d'un ensemble $\mathcal{B}$ et d'un ensemble $Swap(c, c')$.

L'ensemble $Swap(c, c')$ représente une transition de la solution courante $\mathcal{B}$ vers une solution voisine $\mathcal{B}'$. Ainsi, $\forall i$ avec $1 \leq i \leq N$, $B_i \in Swap(c, c')$ si et seulement si les deux propositions suivantes sont vérifiées :

1. $[G_i, c] \in$ REPART$(\mathcal{G}, \mathcal{C}_{ext})$, et $[G_i, c'] \in$ REPART$(\mathcal{G}, \mathcal{C}_{ext})$, $Cover_{\mathcal{B}}([G_i, c]) = 1$ et $Cover_{\mathcal{B}}([G_i, c']) = 0$. Autrement dit, c et c' peuvent être utilisés dans une base de cycles minimum de G_i mais seul c appartient à $B_i \in \mathcal{B}$.

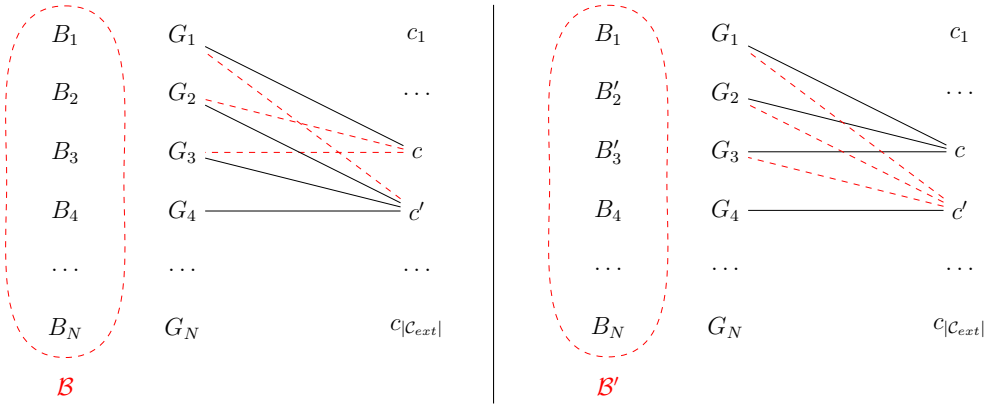


Figure 3.8 – Schéma d'une transition d'un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, \dots B_N\}$ vers un ensemble de bases de cycles minimum $\mathcal{B}' = \{B'_1, \dots B'_N\}$, à partir de $Swap(c, c') = \{B_2, B_3\}$. Nous avons donc $\mathcal{B}' = \{B'_1, B'_2, B'_3, B'_4, \dots, B'_N\} = \{B_1, B'_2, B'_3, B_4, \dots, B_N\}$.

2. $B'_i = (B_i - \{c\}) \cup \{c'\}$ est une base de cycles minimum de G_i. Autrement dit, nous avons $\lambda_{B_i}(c, c') = 1$ et $\omega(c) = \omega(c')$.

Trame générale de notre méthode de voisinage : Étant donné un paramètre X sur l'association de REPART$(\mathcal{G}, \mathcal{C}_{ext})$ et de l'ensemble $\mathcal{B}$, nous avons $Eval_X$(REPART$(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B})$ le résultat de l'évaluation de ce paramètre sur la solution $\mathcal{B}$. Notons que $Eval_X$(REPART$(\mathcal{G}, \mathcal{C}_{ext})$, $\mathcal{B}) \in \mathbb{R}$. Ainsi, nous exprimons la trame générale d'une méthode de voisinage sur REPART$(\mathcal{G}, \mathcal{C}_{ext})$ dans l'Algorithme 10.

La procédure proposée est très permissive et doit être adaptée au paramètre considéré. Supposons ici que nous souhaitions maximiser un paramètre X. À chaque itération, nous recherchons une solution $\mathcal{B}'$ meilleure que la solution $\mathcal{B}$, c'est-à-dire que

$$Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) < Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}')$$

.

Algorithme 10 Trame générale d'une méthode de voisinage sur REPART$(\mathcal{G}, \mathcal{C}_{ext})$ en fonction d'un paramètre X que l'on souhaite maximiser.

Entrée un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, et un nombre d'itérations K maximum.
Sortie un ensemble de bases de cycles minimum $\mathcal{B}$

1: Construction de REPART$(\mathcal{G}, \mathcal{C}_{ext})$ à partir de $\mathcal{G}$
2: $\mathcal{B}$ est l'ensemble de bases de cycles minimum calculé lors de la construction de REPART$(\mathcal{G}, \mathcal{C}_{ext})$.
3: **Pour** k allant de 1 à K **faire**
4: $V_s \leftarrow \emptyset$
5: **Pour tout** couple de cycles avec $c \in \mathcal{C}_{ext}$ et $c' \in \mathcal{C}_{ext}$ **faire**
6: calculer l'ensemble $Swap(c, c')$
7: **Si** $Swap(c, c') \neq \emptyset$ **alors** ajouter l'élement $\Big(c, c', Swap(c, c')\Big)$ a V_s
8: **Si** $V_s = \emptyset$ **alors**
9: **Renvoyer** $\mathcal{B}$
10: $\mathcal{B}' \leftarrow \mathcal{B}$
11: $next \leftarrow Vrai$
12: **Tant que** $|V_s| > 0$ et $next = Vrai$ **faire**
13: Sélection d'un élément $\Big(c, c', Swap(c, c')\Big)$ de V_s,
14: Construction de $\mathcal{B}'$ à partir de $\mathcal{B}$ et de $Swap_(c, c')$.
15: **Si** $Eval_X$(REPART$(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) < Eval_X$(REPART$(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}')$ **alors**
16: $next \leftarrow Faux$
17: **Pour** $B \in Swap(c, c')$ **faire**
18: G le graphe de $\mathcal{G}$ tel que $B \in \mathcal{MCB}(G)$.
19: mise à jour de la table de composition de G en fonction de B'.
20: $\mathcal{B} \leftarrow \mathcal{B}'$
21: Suppression de $c, c', Swap(c, c')$ dans V_s.
22: **Renvoyer** $\mathcal{B}$

3.3.2 . Mise en oeuvre

Dans cette section, nous décrivons les différentes étapes nécessaires à la mise en oeuvre de la méthode de voisinage telle qu'esquissée dans la Section 3.3.1. Nous abordons ainsi la construction du graphe biparti REPART$(\mathcal{G}, \mathcal{C}_{ext})$, le calcul des ensembles $Swap$, ainsi que les possibilités de métriques.

Construire le graphe biparti REPART$(\mathcal{G}, \mathcal{C}_{ext})$ à partir de $\mathcal{G}$

À partir de l'ensemble de graphes $\mathcal{G}$, nous définissons un ensemble $\mathcal{C}_{ext}$ qui constitue, avec $\mathcal{G}$, l'ensemble des sommets de REPART$(\mathcal{G}, \mathcal{C}_{ext})$. Dans la suite, nous présentons la construction de $\mathcal{C}_{ext}$ ainsi que celle des arêtes de REPART$(\mathcal{G}, \mathcal{C}_{ext})$.

$\mathcal{C}_{ext}$ est un ensemble de cycles où tous les cycles doivent appartenir à au moins une base de cycles minimum d'un des graphes de $\mathcal{G}$. Nous souhaitons également vérifier que pour tout graphe G_i de $\mathcal{G}$, il existe une base de cycles minimum tel que $B_i \subseteq \mathcal{C}_{ext}$. En

effet, les cycles de $\mathcal{C}_{ext}$ sont les seuls que nous possédons dans ce modèle ; aucun cycle supplémentaire ne sera calculé. Il est donc nécessaire que les bases de cycles minimum que nous recherchons soient accessibles uniquement depuis $\mathcal{C}_{ext}$.

Pour tout i tel que $1 \leq i \leq N$, étant donné le graphe G_i, nous définissons l'ensemble de cycles $\mathcal{C}_{ext}(G_i)$ tel que $\forall c \in \mathcal{C}_{ext}(G_i), \exists B_i \in \mathcal{MCB}(G_i)$ avec $c \in B_i$ et $B_i \subseteq \mathcal{C}_{ext}(G_i)$. Ainsi, étant donné un ensemble de graphes $\mathcal{G} = \{G_1, \ldots, G_N\}$, nous avons $\mathcal{C}_{ext} = \bigcup_{i=1}^{N} \mathcal{C}_{ext}(G_i)$.

Ainsi, l'Algorithme 11 décrit la procédure à appliquer sur chaque graphe G_i pour $1 \leq i \leq N$. Cette méthode suit le même principe que celle de Horton (Section 3.1). Nous commençons par énumérer des cycles élémentaires puis nous les ordonnons par paliers de poids croissants. La procédure se distingue alors car en même temps que des cycles sont ajoutés à $\mathcal{C}_{ext}(G_i)$, une base de cycles minimum $B_i \in \mathcal{MCB}(G_i)$ est construite.

Algorithme 11 Calcul de $\mathcal{C}_{ext}(G)$ pour un graphe G

Entrée un graphe $G = (V, E)$
Sortie l'ensemble de cycles $\mathcal{C}_{ext}(G)$ et $B \in \mathcal{MCB}(G)$ telle que $B \subseteq \mathcal{C}_{ext}(G)$.

1: Algorithme 1 Ligne 1 : Pour chaque paire de sommets, trouver un plus court chemin.
2: Algorithme 1 Ligne 2 : Pour chaque triplet (u, v, x) où $[u, v] \in E$ avec $x \in V$, $u \neq x$ et $v \neq x$, construire le cycle élementaire, si il existe, correspondant à la concatenation des plus courts chemins de u à x et de x à v, ainsi que de l'arête $[u, v]$.
3: $C = \{c_1, c_2, \ldots, c_q\}$ l'ensemble de cycles élémentaires ainsi construits et ordonnés par ordre de poids croissant.
4: $\mathcal{C}_{ext}(G) \leftarrow \emptyset$, $B \leftarrow \emptyset$
5: **Pour** j allant de 3 à $\omega(c_q)$ **faire**
6: $\quad C_j \leftarrow$ les cycles de C de poids j.
7: $\quad E \leftarrow \emptyset$
8: $\quad$ **Pour tout** $c \in C_j$ **faire**
9: $\quad\quad$ **Si** c est linéairement indépendant de B **alors** ajouter c à E.
10: $\quad$ **Pour tout** $c \in E$ **faire**
11: $\quad\quad$ **Si** c est linéairement indépendant de B **alors** ajouter c à B.
12: $\quad \mathcal{C}_{ext}(G) \leftarrow \mathcal{C}_{ext}(G) \cup E$
Renvoyer $\mathcal{C}_{ext}(G)$ et B

Pour chacun de ces paliers tous les cycles linéairement indépendants des cycles qui lui sont strictement plus légers peuvent faire partie d'une base de cycles minimum. En effet, chacun d'entre eux aurait pu être ajouté à la base de cycles minimum en étant classé premier de son palier. Ils sont donc tous ajoutés à $\mathcal{C}_{ext}$ avant que la base B_i ne soit complètée.

Remarque 11. *Si nous voulions être exhaustifs, alors il serait nécessaire de considérer tous les cycles élémentaires possibles et de ne pas nous restreindre à ceux énumérés par l'algorithme de Horton. Néanmoins, cela augmenterait dans certains cas la taille de $\mathcal{C}_{ext}$ au-delà d'une praticité raisonnable.*

La majeure partie des arêtes de REPART$(\mathcal{G}, \mathcal{C}_{ext})$ sont ajoutées en même temps que les cycles de $\mathcal{C}_{ext}$ sont calculés. Ainsi, $\forall i$ tel que $1 \leq i \leq N, \forall c \in \mathcal{C}_{ext}(G_i)$, il existe $[G_i, c] \in$ REPART$(\mathcal{G}, \mathcal{C}_{ext})$. Néanmoins, même si $c \notin \mathcal{C}_{ext}(G_i)$ alors l'arête $[G_i, c]$ peut exister s'il existe $B_i \in \mathcal{MCB}(G_i)$ tel que $c \in B_i$ et $B_i \in \mathcal{C}_{ext}$. Cela suppose que c n'ait pas été énuméré par l'Algorithme 11 lorsqu'il a été appliqué sur G_i. Il faut à présent vérifier, étant donné $B_i \in \mathcal{MCB}(G_i)$ et $B_i \subseteq \mathcal{C}_{ext}(G_i)$, s'il existe un cycle $c' \in B$ tel que $\lambda_B(c', c) = 1$ et $\omega(c') = \omega(c)$. Ainsi, par le lemme d'échange (Lemme 1), il existe $B' = (B \setminus \{c'\}) \cup \{c\}$ avec $B' \in \mathcal{MCB}(G_i)$ et $B' \subseteq \mathcal{C}_{ext}$. Donc, $[G_i, c]$ appartient à REPART$(\mathcal{G}, \mathcal{C}_{ext})$.

En pratique, nous utilisons des tables de composition pour identifier les cycles qui permettent la composition d'un cycle donné. Nous définissons une table de composition pour chacun des graphes de $\mathcal{G}$. Ainsi, une table est définie pour un graphe G_i et étant donnée une base de cycles B_i de celui-ci. Il s'agit d'un tableau à deux dimensions représentant le résultat de l'application $\lambda_{B_i}(c, d)$ entre chacun des cycles $c \in B_i$ et chacun des cycles $d \in \mathcal{C}_{ext}$. La table de composition d'un graphe doit donc être recalculée à chaque fois que la base de cycles minimum est modifiée. Néanmoins, elle est nécessaire à plusieurs étapes de la méthode de voisinage.

Pour calculer ces tables, nous proposons d'utiliser la méthode du pivot de Gauss à partir de B_i. Rappelons que les cycles correspondent à des vecteurs binaires sur les arêtes du graphe. La complexité de l'élimination de Gauss dépend donc du nombre maximal d'arêtes d'un graphe de $\mathcal{G}$.

Construire le graphe REPART$(\mathcal{G}, \mathcal{C}_{ext})$ à partir de $\mathcal{G}$ se fait en temps polynomial
Étant donné $\mathcal{G}$, la définition de $\mathcal{C}_{ext}$ se fait en temps polynomial par plusieurs appels à l'Algorithme 11. Rappelons que $\mathcal{C}_{ext} = \bigcup_{G \in \mathcal{G}} \mathcal{C}_{ext}(G)$. L'Algorithme 11 est polynomial tout comme l'algorithme de Horton dont il suit la méthodologie. De plus, cet algorithme renvoie, pour tout $1 \leq i \leq N$, $G_i \in \mathcal{G}$, non seulement $\mathcal{C}_{ext}(G_i)$ mais également $B_i \in \mathcal{MCB}(G_i)$ telle que $B_i \subseteq \mathcal{C}_{ext}(G)$.

Pour définir les arêtes de REPART$(\mathcal{G}, \mathcal{C}_{ext})$, nous pouvons déjà en définir la majeure partie par la construction sequentielle de $\mathcal{C}_{ext}(G_i)$, pour tout $1 \leq i \leq N$. Pour le reste, le parcours des tables de compositions permet d'ajouter celles manquantes.

Pour tout i tel que $1 \leq i \leq N$, la table de composition de G_i est calculée étant donné B_i calculée par l'Algorithme 11.

Définition des solutions voisines

Dans la Section 3.3.1, nous avons présenté comment, à partir d'un ensemble de bases de cycles minimum B, il est possible de construire un nouvel ensemble de bases de cycles minimum B' en effectuant une opération de *swap* entre deux cycles dans certaines bases de cycles de B. Nous allons maintenant montrer comment, à partir d'un ensemble de bases de cycles minimum B, nous pouvons définir les ensembles $Swap(c, c')$ nécessaires pour réaliser ces opérations.

Étant donné un cycle $c \in \mathcal{C}_{ext}$, et une solution $\mathcal{B}$, nous notons $Value_{\mathcal{B}}(c)$ le nombre de bases de $\mathcal{B}$ qui utilisent c. Ainsi, si $Value_{\mathcal{B}}(c) > 0$, alors le cycle c appartient à au moins une base de cycles minimum de $\mathcal{B}$. Nous avons

$$Value_{\mathcal{B}}(c) = \sum_{i=1}^{N} Cover_{\mathcal{B}}([G_i, c])$$

.

Considérons un couple de cycles c, c' avec $c \in \mathcal{C}_{ext}, c' \in\in \mathcal{C}_{ext}$ et $c \neq c'$ tel que $\omega(c) = \omega(c')$ et $Value_{\mathcal{B}}(c) > 0$. Rappelons que, $\forall i$ tel que $1 \leq i \leq N$, $B_i \in Swap(c, c')$ si et seulement si les deux propositions suivantes sont vérifiées :

1. $[G_i, c] \in$ REPART$(\mathcal{G}, \mathcal{C}_{ext})$, et $[G_i, c'] \in$ REPART$(\mathcal{G}, \mathcal{C}_{ext})$, $Cover_{\mathcal{B}}([G_i, c]) = 1$ et $Cover_{\mathcal{B}}([G_i, c']) = 0$. Autrement dit, c et c' peuvent être utilisés dans une base de cycles minimum de G_i mais seul c appartient à $B_i \in \mathcal{B}$.
2. $B'_i = (B_i - \{c\}) \cup \{c'\}$ est une base de cycles minimum de G_i. Autrement dit, nous avons $\lambda_{B_i}(c, c') = 1$ et $\omega(c) = \omega(c')$.

Pour établir l'ensemble des solutions voisines accessibles depuis la solution courante, nous considérons donc tous les couples de cycles tels que $c, c' \in \mathcal{C}_{ext}$. Pour chacun de ces couples, nous allons définir l'ensemble $Swap(c, c')$ en parcourant une à une les bases de cycles minimum de $\mathcal{B}$ pour vérifier si elles respectent les deux propriétés énoncées et, le cas échéant, les ajouter à $Swap(c, c')$. Ces propriétés sont relativement simples à vérifier. La première porte sur les arêtes de REPART$(\mathcal{G}, \mathcal{C}_{ext})$ et sur la couverture associée à ces cycles pour le graphe G. Tandis, que la seconde porte sur la relation linéaire entre c et c', nécessitant de vérifier $\lambda_B(c, c') = 1$. Cette information est directement accessible dans la table de composition de G en fonction de B.

Ainsi, étant donnés deux cycles c et c' de $\mathcal{C}_{ext}$ et un ensemble de bases de cycles minimum $\mathcal{B}$, établir l'ensemble $Swap(c, c') \subseteq \mathcal{B}$ se fait en temps polynomial à l'aide de la procédure décrite ci-dessus.

Nous avons déjà décrit comment, à partir de $Swap(c, c')$ et $\mathcal{B}$, nous construisons $\mathcal{B}'$. Cette étape consiste en un simple échange de l'élément c par l'élément c' dans les bases concernées et est donc polynomial.

Remarque 12. *Lorsqu'une solution $\mathcal{B}'$ est calculée, les tables de compositions associées aux bases modifiées doivent être recalculées. Donc, $\forall i$ tel que $1 \leq i \leq N$, si $B_i \neq B'_i$, alors la table de G_i doit être recalculée en fonction de B'_i. Pour cette mise à jour, nous utilisons toujours la méthode du pivot de Gauss, ce qui garantit que la procédure reste polynomiale.*

Algorithmes de voisinage

Maintenant que les mouvements (swap) permettant le remplacement d'un cycle c par un cycle c' dans un certain nombre de bases de cycles minimum ont été définis, il reste à préciser lequel de ces mouvements est choisi à chaque itération. L'Algorithme 10 propose une trame générale à une méthode de voisinage basée sur notre modèle, dans le cadre de l'optimisation d'un paramètre X.

Une itération de cet algorithme a une complexité en $O(|\mathcal{C}_{ext}|^2 \times \sum_{i=0}^{|\mathcal{G}|} |V|^2 \times (|E_i| + V \log V))$ avec pour tout i, $1 \leq i \leq |\mathcal{G}|, G_i$ est le graphe (V, E_i).

1ère Métrique proprosée : cardinalité de l'union des cycles Nous proposons de considérer le paramètre $X = |\mathcal{C}_{\mathcal{B}}|$, où $\mathcal{C}_{\mathcal{B}}$ est l'union des bases de cycles minimum de $\mathcal{B}$ que nous cherchons ici à minimiser. Selon notre modèle, nous avons $\mathcal{C}_{\mathcal{B}} = \bigcup_{c \in \mathcal{C}_{ext}}$ Cover$_{\mathcal{B}}(c) > 0$. Néanmoins, calculer la cardinalité de cet ensemble revient à :

$$Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \sum_{c \in \mathcal{C}_{ext} \mid Value_{\mathcal{B}}(c) > 0} 1.$$

Ce paramètre est en lien avec l'idée de MCBI. En effet, plus la cardinalité de $\mathcal{C}_\mathcal{B}$ est grande, plus les cycles sélectionnés dans les bases de $\mathcal{B}$ sont diversifiés. À l'inverse, plus la cardinalité de $\mathcal{C}_\mathcal{B}$ est petite, plus les cycles sélectionnés dans les bases de $\mathcal{B}$ sont similaires.

Remarque 13. *À défaut d'améliorer la solution, une faible cardinalité de $|\mathcal{C}_\mathcal{B}|$ signifie que moins de cycles sont à considérer pour le calcul du polygraphe.*

Proximité des solutions voisines Nous avons observé que les solutions voisines ne se demarquaient pas fortement, cela nous a amené à considérer, en plus des meilleures solutions, les solutions équivalentes. Cela revient à remplacer la condition du **Si** de la ligne 15 de l'algorithme par $Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) \geq Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}')$.

Le problème dans ce contexte est le risque de cyclage sur les mêmes solutions courantes. En effet, lorsqu'il y a équivalence entre les solutions, après la construction d'une solution $\mathcal{B}'$ à partir de l'ensemble $Swap(c, c')$ sur $\mathcal{B}$, le couple (c', c) peut être sélectionné à l'itération suivante. Pour éviter ces allers-retours entre solutions voisines, nous allons limiter les cycles pouvant être sélectionnés en ajoutant un statut d'échange.

Ainsi, seuls les cycles ayant un statut d'échange positif peuvent être sélectionnés lors d'une itération. Au départ, tous les cycles ont un statut d'échange positif. Après avoir été sélectionnés et les échanges appliqués, les cycles c et c' voient leur statut d'échange devenir négatif. Ce statut d'échange s'apparente à un statut "Tabou". Lorsqu'un cycle a un statut d'échange négatif, si l'une des bases de cycles qui l'utilise est modifiée, alors son statut d'échange redevient positif.

Premier algorithme Nous proposons une méthode de voisinage qui vise à minimiser $|\mathcal{C}_\mathcal{B}|$ dans l'Algorithme 12. Cet algorithme s'arrête après au plus K itérations, ou moins si aucune meilleure solution n'est trouvée par l'instruction de la Ligne 13.

À chaque itération, nous avons la solution $\mathcal{B}'$ construite à partir de $Swap(c, c')$ qui vérifie :

- Si $Value_\mathcal{B}(c) = |Swap(c, c')|$ et $Value_\mathcal{B}(c') > 0$ alors $Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}') + 1$. Cela revient à supprimer c de $\mathcal{C}$ et à le remplacer par c' qui appartenait déjà à $\mathcal{C}$, la cardinalité de $\mathcal{C}$ diminue donc de 1.
- Si $Value_\mathcal{B}(c) = |Swap(c, c')|$ et $Value_\mathcal{B}(c') = 0$ alors $Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}')$. Cela revient à supprimer c de $\mathcal{C}$ et à le remplacer par c' qui n'appartenait pas déjà à $\mathcal{C}$, la cardinalité de $\mathcal{C}$ est inchangée.

La solution $\mathcal{B}'$ a donc au pire la même valeur que la solution $\mathcal{B}$. De plus, dans cet algorithme un tableau Exc est utilisé pour conserver le statut d'échange d'un cycle.

Remarque 14. *Dans l'Algorithme 12 à la Ligne , nous sélectionnons le couple (c, c') de cardinalité maximum qui vérifie $Swap(c, c') \neq \emptyset$ et $Cover_\mathcal{B}(c) = |Swap(c, c')|$. Tout couple qui vérifie ces deux conditions permet de construire une solution voisine $\mathcal{B}'$ au pire de même poids que la solution courante $\mathcal{B}$. Pour ne pas avoir à faire un choix aléatoire, nous sélectionnons le couple dont la cardinalité est la plus grande.*

Variantes Pour nous assurer qu'une solution $\mathcal{B}'$ strictement meilleure est trouvée à chaque itération, il faut modifier l'instruction de la Ligne 10. Le couple (c, c') est sélectionné parmi les éléments de V_s, et cette ligne précise les propriétés de $Swap(c, c')$ pour qu'il soit ajouté à V_s. Pour qu'une solution strictement meilleure soit construite, il faudra vérifier

Algorithme 12 Méthode heuristique de voisinage pour minimiser $|\mathcal{C}_{\mathcal{B}}|$

Entrée un ensemble de graphes $\mathcal{G}$ et un nombre d'itérations maximum K.
Sortie un ensemble de bases de cycles minimum $\mathcal{B}$

1: Construction de REPART$(\mathcal{G}, \mathcal{C}_{ext})$ à partir de $\mathcal{G}$
2: $\mathcal{B}$ est l'ensemble de bases de cycles minimum calculé lors de la construction de REPART$(\mathcal{G}, \mathcal{C}_{ext})$.
3: $Exc \leftarrow$ un tableau de taille $|\mathcal{C}_{ext}|$ où chaque case est initialisée à 1.
4: ▷ Si $Exc[c] = 1$ alors le statut d'échange de c est positif
5: **Pour** k allant de 1 à K **faire**
6: $V_s \leftarrow \emptyset$
7: **Pour tout** couple de cycles avec $c \in \mathcal{C}_{ext}$ et $c' \in \mathcal{C}_{ext}$ **faire**
8: **Si** $Exc[c] = 1$ et $Exc[c'] = 1$ **alors**
9: Calculer $Swap(c, c')$
10: **Si** $Swap(c, c') \neq \emptyset$ et $Cover_{\mathcal{B}}(c) = |Swap(c, c')|$ **alors**
11: ajouter l'élement $\Big(c, c', Swap(c, c')\Big)$ a V_s
12: **Si** $V_s = \emptyset$ **alors**
13: **Renvoyer** $\mathcal{B}$
14: Soit, $\Big(c, c', Swap(c, c')\Big)$ l'élément de V_s dont $|Swap(c, c')|$ est maximum.
15: $\mathcal{B}' \leftarrow \emptyset$
16: **Pour** $B \in \mathcal{B}$ **faire**
17: **Si** $B \in Swap(c, c')$ **alors**
18: **Pour** $c \in B$ **faire**
19: $Exc[c] \leftarrow 1$
20: $B' \leftarrow (B \setminus \{c\} \cup \{c'\})$
21: G le graphe de $\mathcal{G}$ tel que $B \in \mathcal{MCB}(G)$.
22: mise à jour de la table de composition de G en fonction de B'
23: **Sinon**
24: $B' \leftarrow B$
25: Ajouter B' à $\mathcal{B}'$
26: $Exc[c] \leftarrow 0$
27: $Exc[c'] \leftarrow 0$
28: $\mathcal{B} \leftarrow \mathcal{B}'$
29: **Renvoyer** $\mathcal{B}$

que $Swap(c, c') \neq \emptyset$, que $Value_{\mathcal{B}}(c') > 0$, et que $Value_{\mathcal{B}}(c) = |Swap(c, c')|$. Ainsi, nous nous assurons de supprimer c en le remplaçant par un cycle déjà présent dans $\mathcal{C}$. De cette façon, nous aurons $Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}') < Eval_X(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B})$.

Une autre variante est de sélectionner le couple (c, c') aléatoirement parmi les éléments de V_s, au lieu de choisir l'élément de cardinalité maximale comme indiqué à la Ligne 14. Ensuite, nous considérons $\mathcal{B}'$ comme la prochaine solution uniquement si son évaluation est au moins équivalente à celle de $\mathcal{B}$.

2^{nde} **Métrique proposée : taux de couverture des cycles** Nous proposons maintenant de considérer également un paramètre Y connecté à l'ensemble de cycles $\mathcal{C}_\mathcal{B}$: le taux de couverture des cycles. Nous définissons le taux de couverture d'un cycle c comme $\tau(c) = \frac{Value_\mathcal{B}(c)}{\delta(c)}$ où $\delta(c)$ est le degré de c dans le graphe REPART$(\mathcal{G}, \mathcal{C}_{ext})$. Le taux de couverture de $\mathcal{C}_\mathcal{B}$ est alors $\tau(\mathcal{C}_\mathcal{B}) = \sum_{c \in \mathcal{C}_\mathcal{B}} \tau(c)$. Ainsi, nous avons :

$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \tau(\mathcal{C}_\mathcal{B}) = \sum_{c \in \mathcal{C}_{ext} \mid Value_\mathcal{B}(c) > 0} \frac{Value_\mathcal{B}(c)}{\delta(c)}.$$

A priori, nous cherchons à maximiser le taux de couverture d'un cycle c car cela suggère que le cycle c est utilisé dans de nombreuses bases où il peut être sélectionné. Néanmoins, ce paramètre du taux n'est pas si simple d'utilisation. Pour illustrer cela nous reprenons la Figure 3.8 présentée à la Section 3.3.1. Rappelons que les arêtes en pointillés colorés sont celles couvertes par l'ensemble des bases de cycles minimum associées. Considérons l'ensemble des bases de cycles minimum $\mathcal{B}$. Nous observons que $\tau(c) = \frac{2}{3}$ et $\tau(c') = \frac{1}{4}$. Ainsi,

$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \tau(c) + \tau(c') + \tau(\mathcal{C}_\mathcal{B} \setminus \{c, c'\})$$
$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \frac{11}{12} + \tau(\mathcal{C}_\mathcal{B} \setminus \{c, c'\})$$

Considérons maintenant l'ensemble des bases de cycles minimum $\mathcal{B}'$. Nous avons alors $\tau(c) = \frac{0}{3}$ et $\tau(c') = \frac{3}{4}$. Comme $c \notin \mathcal{C}_{\mathcal{B}'}$ et $c' \in \mathcal{C}_{\mathcal{B}'}$, nous avons :

$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}') = \tau(c') + \tau(\mathcal{C}'_\mathcal{B} \setminus \{c'\})$$
$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}') = \frac{3}{4} + \tau(\mathcal{C}'_\mathcal{B} \setminus \{c'\})$$

Puisque l'ensemble des cycles présents dans $\mathcal{C}'_\mathcal{B}$ est inchangé hormis pour c et c', notons $y = \tau(\mathcal{C}'_\mathcal{B} \setminus \{c'\}) = \tau(\mathcal{C}_\mathcal{B} \setminus \{c, c'\})$. Nous avons donc :

$$Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \frac{11}{12} + y > \frac{3}{4} + y = Eval_Y(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}')$$

Il est important de mentionner que le taux d'un cycle est directement lié à son degré dans REPART$(\mathcal{G}, \mathcal{C}_{ext})$. Ainsi, un cycle présent dans seulement quelques conformations aura facilement un taux de couverture très élevé comparé à d'autres cycles présents dans plus de conformations. Ce paramètre du taux n'est donc pas utilisable seul.

Nous proposons donc de distinguer les cycles obligatoires de $\mathcal{C}_\mathcal{B}$ des cycles optionnels. En effet, nous avons observé que certains cycles appartiennent toujours à $\mathcal{C}_\mathcal{B}$ peu importe les bases de cycles minimum sélectionnées. Ce sont des cycles qui font toujours partie d'au moins une base de cycles minimum. Ce phénomène est courant pour les petits cycles qui ne peuvent généralement pas être obtenus par la combinaison d'autres cycles plus petits.

Notons alors $\mathcal{C}^O_\mathcal{G}$ l'ensemble des cycles obligatoire de $\mathcal{G}$. Étant donné un graphe $G_i \in \mathcal{G}$, si $\forall B_i \in \mathcal{MCB}(G_i)$, $c \in B_i$, alors nous considérons que c est obligatoire dans G_i. Ainsi, nous avons $c \in \mathcal{C}^O_\mathcal{G}$. Le paramètre du taux ne semble intéressant que pour les cycles obligatoires, nous définissons donc

$$Eval_{Y'}(\text{REPART}(\mathcal{G}, \mathcal{C}_{ext}), \mathcal{B}) = \tau(\mathcal{C}^O_\mathcal{G}) = \sum_{c \in \mathcal{C}^O_\mathcal{G} \mid Value_\mathcal{B}(c) > 0} \frac{Value_\mathcal{B}(c)}{\delta(c)}.$$

Néanmoins, nous ne traitons là qu'une partie des cycles sélectionnés.

Pour que le paramètre Y soit le plus pertinent possible, nous proposons de la combiner au paramètre X pour considérer l'ensemble des cycles sélectionnés dans des bases de cycles lors de notre évaluation. En effet, un cycle obligatoire sera toujours au moins partiellement couvert, donc nous cherchons à maximiser la couverture de ce cycle. En revanche, un cycle optionnel peut éventuellement ne pas être utilisé, donc nous allons essayer de minimiser leur utilisation. Ce deuxième point revient à minimiser $|\mathcal{C}_\mathcal{B}|$. En effet, par définition, un cycle obligatoire ne peut pas être supprimé de $\mathcal{C}_\mathcal{B}$, donc seuls des cycles optionnels sont supprimés pour minimiser $|\mathcal{C}_\mathcal{B}|$.

Second algorithme L'Algorithme 12 propose déjà une heuristique de voisinage qui cherche à minimiser $|\mathcal{C}_\mathcal{B}|$, nous allons donc adapter cet algorithme pour inclure le paramètre du taux Y.

Définir le statut obligatoire d'un cycle est simple à partir de la table de composition d'un graphe. Notons $O(c) = 1$ si c est obligatoire et $O(c) = 0$ sinon. Considérons un cycle $c \in \mathcal{C}_{ext}$ et un graphe $G_i \in \mathcal{G}$ pour $1 \leq i \leq N$. Si pour tout $c' \in \mathcal{C}_{ext}$ tel que $c' \in G_i$ et $\omega(c') = \omega(c)$, nous avons $\lambda_{B_i}(c, c') = 0$ pour tout $B_i \in \mathcal{MCB}(G_i)$, alors $c \in \mathcal{C}^O_\mathcal{G}$. Cela revient à vérifier qu'il n'existe pas de cycles dans $\mathcal{C}_{ext}$ par lesquels c peut être remplacé dans une base de cycles minimum de G_i.

Nous proposons maintenant de ne vérifier qu'à la Ligne 10 que $Swap(c, c') \neq \emptyset$. Ensuite, lors de l'ordonnancement des couples disponibles à la Ligne 14, les couples sont priorisés selon l'ordre suivant :

1. $O(c) = 0$ et $O(c') = 1$.
2. $O(c) = 1$, $O(c') = 1$ et $\tau(c) > \tau(c')$
3. $O(c) = 0$, $O(c') = 0$ et $Value_\mathcal{B}(c) = |Swap(c, c')|$

En cas d'égalité, le couple (c, c') maximisant $|Swap(c, c')|$ est choisi. Ainsi, nous privilégions dans l'ordre : (1) d'utiliser le moins possible de cycles optionnels ou, à défaut, de limiter leur présence à un nombre réduit de bases de cycles ; (2) de maximiser le taux de couverture des cycles obligatoires ; et (3) de supprimer les cycles optionnels, même s'ils ne peuvent être remplacés par des cycles obligatoires.

Au final, cette méthode se positionne comme une amélioration de celle décrite par l'Algorithme 12.

3.4 . Synthèse sur les méthodes de sélection des cycles

Dans ce chapitre, nous avons présenté plusieurs méthodes de sélection des bases de cycles minimum à partir d'un ensemble de graphes. Dans notre contexte de dynamique moléculaire, l'objectif est de choisir des bases de cycles minimum pertinentes pour la construction du polygraphe puisque les cycles sélectionnés sont utilisés pour définir les polycycles de la trajectoire.

Tout d'abord, étant donné un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, il est possible de calculer un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$, où pour $1 \leq i \leq N$, $B_i \in \mathcal{MCB}(G_i)$, en temps polynomial avec l'algorithme de Horton. Nous avons introduit une variante de cet algorithme, que nous avons appelée "Horton modifié", et qui

a la même complexité. Cette variante favorise l'utilisation des arêtes correspondant à des liaisons variables.

Nous avons, ensuite, introduit le problème Maximiser l'intersection de bases de cycles minimum (max-MCBI), qui, étant donné un ensemble de graphes $\mathcal{G}$, consiste à trouver un ensemble de bases de cycles minimum $\mathcal{B}$ tel que l'intersection de ces bases soit maximale. Cette approche découle de notre hypothèse selon laquelle maximiser l'intersection des bases de cycles minimum pourrait conduire à des bases plus similaires, favorisant ainsi le calcul d'un polygraphe de meilleure qualité.

Nous avons démontré que le problème de décision MCBI est NP-complet lorsque trois graphes ou plus sont considérés, mais qu'il est polynomial dans le cas de seulement deux graphes. Ces résultats de complexité du problème MCBI font partie d'une publication à venir dans le cadre de la conférence IWOCA [52].

Nous avons élaboré une heuristique polynomiale dans le but de maximiser l'intersection entre certaines paires de bases de cycles minimum. Cette méthode, baptisée "MCBI ", est directement inspirée du problème max-MCBI, bien qu'elle ne le résolve pas de manière explicite.

Les méthodes "Horton" et la méthode "MCBI " calculent une seule base de cycles minimum pour chacun des graphes.

Pour élargir cette perspective et explorer un espace plus large de bases de cycles minimum, nous avons introduit un nouveau modèle basé sur un graphe biparti REPART$(\mathcal{G}, \mathcal{C}_{ext})$ qui représente la relation entre un cycle de $\mathcal{C}_{ext}$ et les bases de cycles minimum des graphes de $\mathcal{G}$. Ce graphe nous offre une vue plus globale du lien entre les cycles et les bases de cycles minimum que nous recherchons.

À partir de ce graphe biparti, nous avons défini une méthode de voisinage qui explore différents ensembles de bases de cycles minimum. À chaque étape, plusieurs bases de cycles minimum sont modifiées pour créer un nouvel ensemble de bases. Ainsi, contrairement aux méthodes précédentes, nous calculons de nombreux ensembles de bases de cycles minimum, ce qui nous permet d'avoir une vision plus étendue des possibilités.

Pour guider notre choix parmi tous les ensembles possibles, nous avons proposé deux métriques, toutes deux sont finalement liées à la cardinalité de $\mathcal{C}_{\mathcal{B}}$, qui permettent l'évaluation de la qualité des solutions obtenues.

Nous avons présenté les outils nécessaires à la mise en œuvre de notre heuristique de voisinage, ainsi qu'une première version de la méthode que nous avons nommée méthode de "Voisinage", accompagnée d'une variante améliorée.

Dans le Chapitre 5, nous discuterons de la pertinence des méthodes issues de ces paramètres pour le calcul du polygraphe d'une trajectoire. Nous évaluerons et comparerons les méthodes "Horton modifié", "MCBI " et "Voisinage" sur différents jeux de données pour établir leur contribution à la qualité du polygraphe résultant de leurs bases de cycles minimum.

En fonction des résultats de cette étude, nous pourrons alors proposer des perspectives d'amélioration pour ces méthodes.

4 - Calcul du polygraphe étant donné un ensemble de cycles d'une trajectoire de dynamique moléculaire

Dans le Chapitre 3, nous avons abordé le problème de la sélection des bases de cycles minimum pour chaque conformation dans une dynamique moléculaire. Nous avons en particulier proposé plusieurs méthodes pour sélectionner ces bases. À partir des algorithmes que nous avons présentés, nous sommes en mesure de choisir une base de cycles pour chaque conformation, et par conséquent, de constituer des ensembles de cycles caractéristiques.

Rappelons que l'ensemble de cycles caractéristiques d'une conformation est un sous-ensemble d'une base de cycles minimum, ne contenant que des cycles possédant au moins une liaison hydrogène.

Étant donné l'ensemble des graphes de la trajectoire $\mathcal{G} = \{G_1, G_2, \dots G_N\}$, pour chaque conformation G_i avec $1 \leq i \leq N$, nous avons donc un ensemble de cycles caractéristiques B_i^*. Ensuite, pour chaque conformation G_i avec $1 \leq i \leq N$, le graphe de cycles GC_i est construit tel que $GC_i = (B_i^*, F)$ où une arête $[c_1, c_2] \in F$ existe si les cycles c_1 et c_2 interagissent, c'est-à-dire partagent au moins une arête. Alors, nous avons l'ensemble des graphes de cycles de la trajectoire $\mathcal{GC} = \{GC_1, GC_2, \dots, GC_N\}$, et l'ensemble de cycles $\mathcal{C}$ de la trajectoire qui est induit par $\mathcal{GC}$.

Cet ensemble de cycles de la trajectoire est l'union des ensembles de cycles caractéristiques des conformations, soit $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$.

Le problème qui nous intéresse dans ce chapitre consiste à diviser cet ensemble de cycles $\mathcal{C}$ en des sous-ensembles de cycles polymorphes tels que décrit dans la Définition 17 (page 46). Nous formalisons le problème suivant.

Problème 3 (Partition de Cycles Polymorphes, PCP)**.** *Étant donnés un ensemble de graphes $\mathcal{G}$ ayant le même ensemble de sommets, un ensemble de graphes de cycles $\mathcal{GC}$, un sous-ensemble d'arêtes H de $\mathcal{G}$ et un entier $m \in \mathbb{N}$, existe-t-il une partition $\mathcal{P}$ de $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$ où B_i^* est l'ensemble de sommets de GC_i, telle que chaque partie est un ensemble de cycles polymorphes et $|\mathcal{P}| \leq m$?*

Cependant, notre objectif ne se limite pas uniquement à cette partition, nous cherchons une partition qui minimise le nombre de parties et donc le nombre d'ensemble de cycles polymorphes différents. Ainsi, le problème de minimisation associé est appelé min-Partition de Cycles Polymorphes (min-PCP). Ce problème de minimisation consiste à rechercher une partition de cycles polymorphes avec le plus petit nombre de parties possible.

Notation 3. *Étant donnés un ensemble de graphes $\mathcal{G}$, un ensemble de graphes de cycles $\mathcal{GC}$ obtenus à partir des ensembles de cycles séléctionnés pour chacun des graphes, et un ensemble de liaisons variables H, l'ensemble des partitions de cycles polymorphes est noté $< \mathcal{G}, \mathcal{GC}, H >$.*

Finalement, le polygraphe caractéristique de la trajectoire est construit à partir d'une partition $\mathcal{P}$ minimale telle que $\mathcal{P} \in < \mathcal{G}, \mathcal{GC}, H >$. La Définition 19 (page 47) formalise le polygraphe d'une trajectoire.

Dans la suite, nous examinons la complexité du problème de partitionnement dans la Section 4.1, et exposons les solutions que nous envisageons dans la Section 4.2. Ce chapitre se conclut par une synthèse dans la Section 4.1.3.

4.1 . Complexité du problème PCP

Cette section aborde la complexité du problème PCP dans le contexte des trajectoires de dynamique moléculaire. En effet, si nous considérons des graphes quelconques, sans propriétés particulières, la complexité de PCP est facile à établir en raison de sa proximité avec le problème d'isomorphisme de sous-graphes.

Dans le contexte de cette thèse et des trajectoires de dynamique moléculaire, les graphes que nous considérons sont particuliers. Ce sont des graphes moléculaires qui possèdent des propriétés spécifiques, bien que toutes ne soient pas pleinement établies dans la littérature. Une propriété connue est que le nombre d'arêtes d'un graphe moléculaire est limité par le nombre de liaisons possibles entre les atomes. Comme discuté dans le Chapitre 1, les liaisons covalentes ou hydrogène impliquent des électrons. Ainsi, le degré d'un sommet dans un graphe moléculaire est borné par la valence [1] de l'atome qu'il représente. Par exemple, un atome de carbone peut former jusqu'à quatre liaisons avec d'autres atomes, donc un sommet représentant un atome de carbone aura un degré maximal de quatre. Ces limites rendent les graphes moléculaires généralement planaires ou presque, à quelques arêtes près.

Dans la Section 4.1.1, nous établissons la complexité de PCP pour des instances proches des graphes moléculaires. Nous montrons que le problème reste NP-complet même lorsque les instances considérées proviennent d'un ensemble de graphes planaires. Ensuite, dans la Section 4.1.2, nous présentons la preuve d'inapproximabilité du problème de minimisation, min-PCP. Nous concluons en présentant nos remarques sur les bornes restantes à explorer dans la Section 4.1.3.

4.1.1 . NP-completude pour les trajectoires de dynamique moléculaire

Dans cette section, nous proposons une réduction polynomiale depuis une instance du problème d'isomorphisme de sous-graphes induits (ISI). Ce problème prend en entrée deux graphes G_a et G_b, et vérifie si G_a est isomorphe à un sous-graphe induit de G_b. Le problème ISI est NP-complet, même dans le cas où G_a et G_b sont des graphes planaires extérieurs (outerplanar) [49].

Définition 21 (Les graphes planaires extérieurs, ou outerplanar)**.** *Un graphe* outerplanar *peut être dessiné de telle sorte que tous ses sommets appartiennent à la face extérieure.*

Ainsi, à partir d'une instance (G_a, G_b) du problème ISI, nous construisons les graphes planaires G_1 et G_2 accompagnés de leurs ensembles de cycles respectifs C_1 et C_2 induisant le graphe de cycles $\mathcal{GC}_1$, respectivement $\mathcal{GC}_2$. Nous obtenons une instance du problème $(\mathcal{G} = \{G_1, G_2\}, \mathcal{GC} = \{\mathcal{GC}_1 \cup \mathcal{GC}_2\}, m)$ du problème PCP en trois étapes. La

1. La valence d'un élement chimique est le nombre maximum de liaisons qu'il peut former de par sa configuration électronique.

construction des graphes G_1 et G_2 à partir des graphes G_a et G_b est illustrée dans les Figures 4.2 et 4.3. La Figure 4.2 représente toutes les étapes de la construction de G_1 à partir de G_a. Tandis que la Figure 4.3, quant à elle, représente seulement les étapes majeures de la construction de G_2 à partir de G_b.

Soit (G_a, G_b) deux graphes *outerplanar* formant une instance du problème ISI.

Notation 4. *Le nombre de sommets de G_a est noté n_a et le nombre de sommets de G_b est noté n_b. Nous considérons que $n_a \leq n_b$.*

La réduction suivante présente la construction de G_1 depuis G_a. La même réduction est appliquée, respectivement, depuis G_b pour obtenir G_2. L'ensemble de graphes $\{G_1, G_2\}$ obtenu est un ensemble de graphes planaires. Nous introduisons la définition suivante qui nous sera utile dans cette section.

Définition 22 (Une arête *libre*)**.** *Une arête est dite* libre *si ses deux extremités n'appartiennent qu'à un seul cycle et qu'elle est sur la face extérieure.*

La réduction est la suivante :

1. Construction d'un graphe $G_1 = (V, E \cup H)$ à partir de G_a.

   (a) Pour chaque sommet u de G_a, nous construisons un cycle c_u. De plus, ces cycles sont construits de telle sorte que pour chaque arête $[u, v]$ de G_a, il existe une arête e dans G_1 telle que $e \in c_u$ et $e \in c_v$, c_u contient $\delta_{G_a}(u)+3$ arêtes où $\delta_{G_a}(u)$ désigne le degré de u dans le graphe G_a et au moins une arête de c_u est *libre*.
   Les Figures 4.2a et 4.2b illustrent cette étape sur un exemple. Pour plus de clarté, nous présentons maintenant une procédure pour construire de tels cycles à partir du plongement du graphe G_a dans le plan :
   - Pour chaque nœud u, les arêtes adjacentes sont numérotées de 1 à $\delta_{G_a}(u)$ dans le sens trigonométrique. Les arêtes numérotées 1 et $\delta_{G_a}(u)$ sont celles qui touchent la face extérieure.
   - Pour chaque nœud u, deux nouveaux nœuds u_1 et u_2 sont placés sur la face extérieure de sorte que les droites suivant les arêtes 1 et $\delta_{G_a}(u)$ séparent ces nouveaux nœuds des autres arêtes de u. Une arête $[u_1, u_2]$ est ajoutée entre ces deux nouveaux nœuds.
   - On crée un sommet e_1 entre le sommet u_2 et l'arête 1 tel que l'arête $[u_2, e_1]$ ne crée pas de croisement dans le plan. L'arête $[u_2, e_1]$ est ajoutée à G_a.
   - Ensuite, pour chaque arête numérotée i, le sommet e_{i+1} est crée entre les arêtes i et $i+1$. L'arête $[e_i, e_{i+1}]$ est ajoutée à G_a.
   - Enfin, l'arête $[e_{\delta_{G_a}+1}, u_1]$ est ajoutée à G_a.
   - Ainsi, pour chaque nœud u, nous avons dessiné un cycle c_u, composé de $\delta(u)+3$ sommets et d'une arête libre $[u_1, u_2]$. La Figure 4.1 illustre la construction d'un tel cycle.

   (b) Nous ajoutons $|G_2| - |G_1|$ sommets non connectés au graphe G_1 de telle sorte que G_1 et G_2 aient le même ensemble de sommets.

   La Figure 4.2a représente le graphe G_a de départ. La Figure 4.2b représente G_1 après la construction de n_1 cycles. Enfin, la Figure 4.2c représente G_1 à la fin de cette première étape, donc après l'ajout des sommets déconnectés. Pour bien comprendre

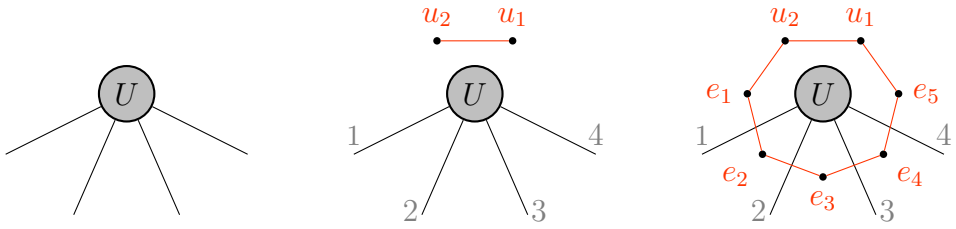


Figure 4.1 – Illustration des étapes de la construction d'un cycle c_U à partir du plongement du noeud U dans la plan.

cette étape, il est nécessaire de considérer également le graphe G_b qui donne lieu à G_2 et qui est illustré par la Figure 4.3a.

Toutes les arêtes créées lors de cette première étape appartiennent à l'ensemble des arêtes variables H.

2. Nous ajoutons une chaîne x, y, z dans E, ainsi qu'une arête $[x, z] \in H$ dans G_1. Pour chaque cycle c_u, une arête libre $[a, b]$ est remplacée par une chaîne a, x, b dans E. Cette nouvelle étape est illustrée dans la Figure 4.2d.
   Notons que chaque cycle c_u appartient à l'ensemble de cycles de G_1 (resp. G_2) noté C_1 (resp. C_2). Le cycle formé par les sommets x, y, z appartient également à C_1 (resp. C_2) et est noté c_1^+ (resp. c_2^+). Les sommets x, y et z sont les mêmes dans les deux graphes.
3. Pour chaque sommet v de G_1, si $[v, x]$ n'appartient pas à G_1 alors nous ajoutons $[v, x] \in E$. Cette dernière étape est illustrée dans la Figure 4.2e.
   Notons que les arêtes ajoutées dans cette dernière étape n'appartiennent à aucun cycle de C_1 (resp. C_2).

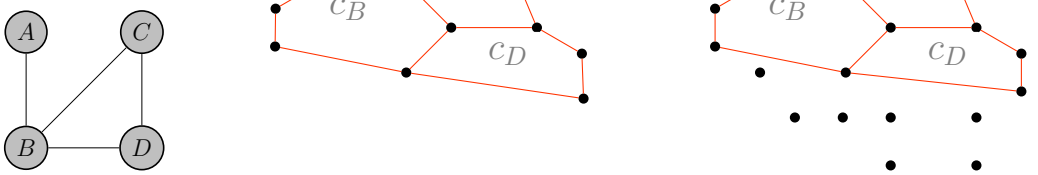
Les arêtes de E forment un sous-graphe connexe commun aux deux graphes qui est représenté par les arêtes noires dans les Figures 4.2 et 4.3. Les arêtes de H appartiennent toutes à au moins un cycle de C_1 ou de C_2. Notons que cette seconde propriété nous rapproche du modèle des trajectoires. Dans le contexte des graphes moléculaires, le backbone de la molécule est connexe ainsi, l'ajout d'une liaison variable dans ce graphe induit nécessairement la formation d'au moins un cycle.

Lemme 17. *La réduction proposée se fait en temps polynomial.*

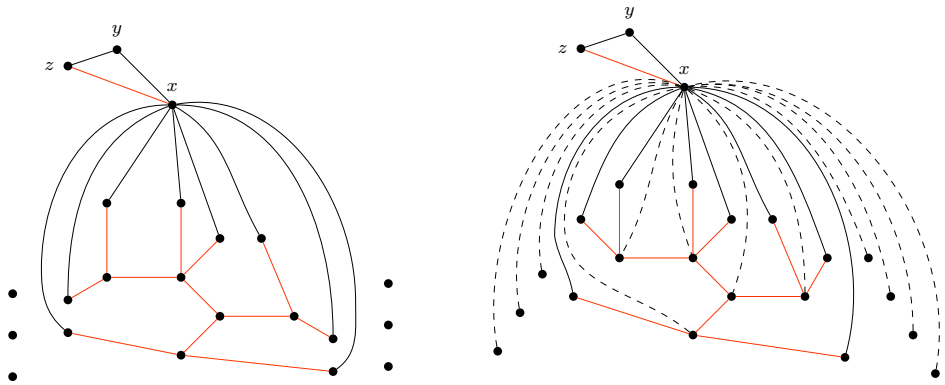
Démonstration. La réduction est une suite d'opérations simples que l'on peut diviser en trois étapes consécutives. Considérons le graphe $G_1 = (V, E \cup H)$ obtenu à partir de G_a avec n_a le nombre de sommets de G_a et m_a le nombre d'arêtes de G_a.

1. n_a cycles sont crées à l'étape 1, si on considère la procédure que nous décrivons alors cette étape est en $O(n_a \times m_a)$.
2. $m_a + 3 \times n_a$ sommets sont créés dans G_1, il faut alors ajouter les sommets supplémentaires créés dans G_2.
3. L'étape de remplacement des arêtes libres $[a, b]$ en un chaîne a, x, b est en $O(n_a)$.
4. La dernière étape consiste en l'ajout de $m_a + n_a$ arêtes $[v, x]$.

Le nombre d'opérations par étape est donc borné par le nombre de sommets et d'arêtes du graphe de départ. Cette réduction est polynomiale. □

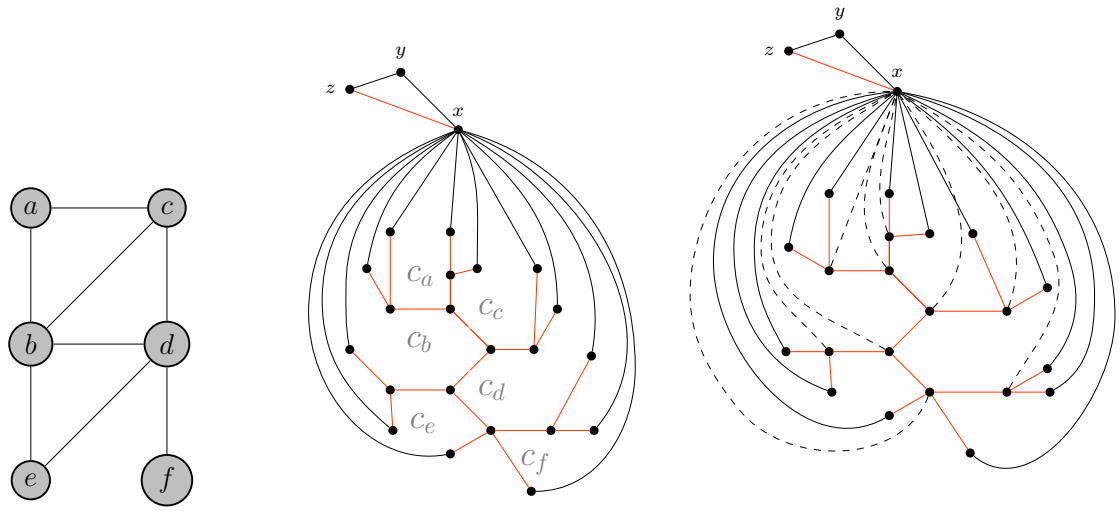


(a) Le graphe G_a (b) Le graphe G_1 après l'étape $1(a)$ (c) Le graphe G_1 après l'étape $1(b)$



(d) Le graphe G_1 après l'étape 2 (e) Le graphe G_1 final obtenu après l'étape 3

Figure 4.2 – Illustration des étapes de la construction du graphe G_1 à partir du graphe *outerplanar* G_a. Les arêtes colorées sont celles appartenant à H tandis que les arêtes noires sont celles appartenant à E. Dans la Figure 4.2e, les lignes en pointillés représentent les arêtes ajoutées lors de l'étape 3.



(a) Le graphe G_b (b) Le graphe G_2 après l'étape 2 (c) Le graphe G_2 final

Figure 4.3 – Illustration des étapes majeures de la construction du graphe G_2 à partir du graphe *outerplanar* G_b. Les arêtes colorées sont celles appartenant à H tandis que les arêtes noires sont celles appartenant à E. Dans la Figure 4.3c, les lignes en pointillés représentent les arêtes ajoutées lors de l'étape 3.

Lemme 18. *Le graphe G_1 (ou G_2) obtenu à partir de G_a (respectivement G_b) par la réduction proposée est planaire.*

Démonstration. Retraçons l'évolution de la planarité du graphe au fil des étapes de la réduction.

1. Le graphe G_1 est construit à partir du plongement planaire extérieur de G_a (voir Figures 4.2a et 4.2c). Chaque cycle est défini à partir d'un sommet de la face extérieure de G_a. Rappelons que G_a est un graphe *outerplanar* donc chacun de ses sommets est sur sa face extérieure. Lorsque deux cycles partagent une arête dans G_1, cela signifie que les sommets correspondants forment une arête dans G_a. Or, les arêtes de G_a ne peuvent pas se croiser, donc il ne peut pas y avoir d'arêtes qui se croisent dans G_1.
2. Dans chaque cycle c_u, il existe une arête *libre* qui est donc localisée sur la face extérieure de G_1. Le remplaçement d'une arête *libre* $[a, b]$ par une chaîne a, C, b, ne peut pas induire de croisement dans le graphe.
3. En conservant le plan dans lequel est dessiné le graphe G_1, les arêtes lors de la dernière étape peuvent être dessinées depuis l'intérieur du cycle c_u. La Figure 4.2e illustre cette troisième étape.

Pour conclure, le graphe obtenu par la réduction proposée est planaire. □

Théorème 4. *Le problème* Partition de Cycles Polymorphes *(*PCP*) est* NP-complet *et ce même si chaque graphe G_i de $\mathcal{G}$ est planaire et si chaque arête de H appartient à au moins un cycle de $\mathcal{C}$.*

Démonstration. Considérons la réduction polynomiale (voir Lemme 17) qui, à partir d'une instance (G_a, G_b) du problème ISI, construit une instance PCP $(\{G_1, G_2\}, \{GC_1, GC_2\}, H, m)$. Chaque graphe de $\{G_1, G_2\}$ est planaire (voir Lemme 18). Aussi, les sommets de G_1 et de G_2 sont identiques. Enfin, par construction, chaque arête de H appartient à au moins un cycle de $\mathcal{C} = \{C_1 \cup C_2\}$ avec C_1 l'ensemble de sommets de GC_1 et C_2 l'ensemble de sommets de GC_2.

Il est à noter que parmi les propriétés du polymorphisme vérifiées par chaque partie d'une partition appartenant à $< \{G_1, G_2\}, \{GC_1, GC_2\}, H >$, la propriété d'ancrage (Propriété 4, page 46) nécessite que les cycles d'une partie aient en commun au moins un sommet impliqué dans une liaison de H. Le sommet x vérifie cette propriété et est commun à tous les cycles de C_1 et C_2. Ainsi, dans le contexte de cette réduction depuis (G_a, G_b), tous les ensembles de cycles vérifient cette propriété.

Supposons que le graphe G_a soit isomorphe à G'_b, un sous-graphe induit de G_b. Alors, il existe une fonction f qui associe les nœuds de G'_b de sorte que $[u, v] \in G_a$ si et seulement si $[f(u), f(v)] \in G'_b$. Nous créons ainsi une partition $\mathcal{P}$ composée de :

- Pour chaque nœud $u \in G_a$, une partie $\{c_u, c_{f(u)}\}$.
- Pour chaque nœud $w \in G_b \setminus G'_b$, une partie $\{c_w\}$.
- La partie $\{c_1^+, c_2^+\}$.

Nous avons là une partition $\mathcal{P}$ de cardinalité $m = n_b + 1$.

Montrons à présent que toutes les parties de $\mathcal{P}$ sont des polycycles. La propriété d'ancrage n'est pas un problème ici, donc nous nous concentrons sur les deux autres propriétés du polycycle.

Pour une partie $\{c_u, c_{f(u)}\}$, c_u désigne le cycle obtenu à partir du nœud u du graphe G_a donc $c_u \in G_a$, et $c_{f(u)}$ désigne le cycle obtenu à partir du nœud $f(u)$ du graphe G_b donc $c_{f(u)} \in G_b$. De plus, par définition, nous avons $c_1^+ \in C_1$ et $c_2^+ \in C_2$. Ainsi, toutes les parties de cardinalité deux sont bien composées de cycles venant de graphes différents et vérifient alors la propriété de non-coexistence (Propriété 2, page 44).

Une arête $[u, v] \in G_a$ est isomorphe à l'arête $[f(u), f(v)] \in G'_b$. Donc, nous avons $c_u \cap c_v \neq \emptyset$ si et seulement si $c_{f(u)} \cap c_{f(v)} \neq \emptyset$. Ainsi, nous avons l'assurance que la propriété de voisinage (Propriété 3, page 44) est vérifiée pour toutes les parties de $\mathcal{P}$.

Nous avons donc $\mathcal{P} \in \langle \{G_1, G_2\}, \{GC_1, GC_2\}, H \rangle$ et $|\mathcal{P}| = n_b + 1$. Par définition, nous avons $n_a \leq n_b$, $n_a + 1 = |C_1|$ et $n_b + 1 = |C_2|$. Ainsi, $|\mathcal{P}| = |C_2|$ et il ne peut donc pas exister de partition $\mathcal{P}' \in \langle \{G_1, G_2\}, \{GC_1, GC_2\}, H \rangle$ avec $|\mathcal{P}'| < |\mathcal{P}|$.

Considérons, à présent, une partition $\mathcal{P} = \langle \{G_1, G_2\}, \{GC_1, GC_2\}, H \rangle$ de cardinalité $m = n_b + 1$ où chaque partie est un ensemble de cycles polymorphes.

La partition $\mathcal{P}$ est obtenue à partir d'un ensemble de deux graphes, donc chaque partie contient un ou deux cycles. La partition est de cardinalité $n_b + 1$, soit $|\mathcal{P}| = |C_2|$. Ainsi, il ne peut pas exister de partie de cardinalité 1 qui soit un singleton d'un cycle de C_1. Nous avons donc $n_a + 1$ parties de cardinalité 2, incluant la partie $\{c_1^+, c_2^+\}$.

Nous définissons une fonction f telle que si $\{c_u, c_v\}$ est une partie de $\mathcal{P}$, alors $f(u) = v$ avec $u \in G_a$ et $v \in G_b$.

Toute partie de $\mathcal{P}$ doit respecter la propriété de voisinage. Considérons deux cycles c_u, c'_u de C_1. S'il existe une arête $[c_u, c'_u] \in GC_1$, alors les parties $\{c_u, c_v\}$ et $\{c_{u'}, c_{v'}\}$ existent uniquement si $[c_v, c'_v] \in GC_2$. Réciproquement, s'il n'existe pas d'arête $[c_u, c'_u]$ dans GC_1, alors les parties $\{c_u, c_v\}$ et $\{c_{u'}, c_{v'}\}$ existent uniquement si $c_v \cap c'_v = \emptyset$. Ainsi, il existe une arête $[v, v']$ dans G'_b si et seulement si il existe une arête $[u, u']$ dans G_a.

Nous avons donc f qui désigne une fonction d'isomorphisme entre G_a et G'_b.

Le problème ISI est NP-complet [3] ainsi nous pouvons conclure que le problème Partition de Cycles Polymorphes (PCP) est également NP-complet et ce même si chaque graphe $G \in \mathcal{G}$ est planaire et si chaque arête de H appartient à au moins un cycle de $\mathcal{C}$. □

4.1.2 . Inapproximabilité du problème min-PCP

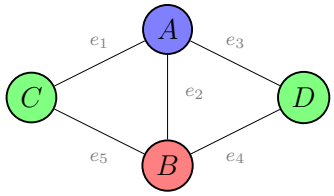
Cette section présente l'étude de l'inapproximabilité du problème min-PCP par l'intermédiaire du problème de coloration de graphe. Le problème de coloration est inaproximable, ainsi en établissant la proximité entre ce problème et min-PCP nous pouvons conclure sur l'inapproximabilité de notre problème. Notons que nous ne nous réduisons plus ici aux instances dont les graphes sont exclusivement planaires.

Nous proposons une réduction polynomiale depuis une instance du problème de coloration de graphe. Ce problème prend en entrée un graphe $G = (V_G, E_G)$ et un entier k, et recherche si une coloration du graphe en au plus k couleurs existe. Le problème de coloration est NP-complet pour $k > 2$ [30]. Nous considérons le problème de minimisation qui lui est associé, noté MCG, et qui consiste à trouver le nombre chromatique d'un graphe G.

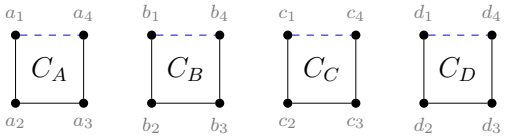
Nous construisons une instance du problème min-PCP à partir d'une instance $G = (V_G, E_G)$ du problème du nombre chromatique d'un graphe. En premier lieu le graphe

P_G est établi depuis G et nous définissons $\{E \cup H\}$ son ensemble d'arêtes. La Figure 4.4 illustre les étapes de construction de P_G.

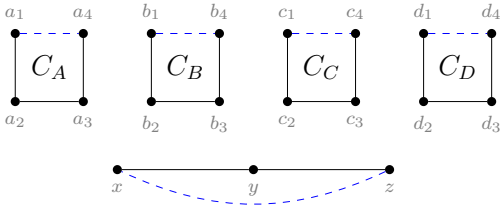
1. Pour chaque sommet $u \in V_G$, un cycle c_u de longueur 4 est construit. Ce cycle c_u est défini par u_1, u_2, u_3, u_4 avec $[u_1, u_4]$ sa seule arête appartenant à H.
2. Une chaîne appartenant à E est ajoutée avec ses trois nouveaux sommets x, y, z. Une arête $[x, z] \in H$ est ajoutée. Le cycle formé par les sommets x, y et z est noté d.
3. Dans chaque cycle c_u, l'arête $[u_2, u_3]$ est remplacée par les deux arêtes de E : $[u_2, x]$ et $[u_3, z]$. Le cycle ainsi modifié u_1, u_2, x, z, u_3, u_4 est toujours noté c_u.



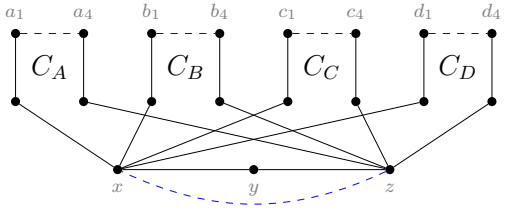
(a) Un graphe 3-coloriable G



(b) Le graphe P_G après l'étape 1.



(c) Le graphe P_G après l'étape 2.



(d) Le graphe P_G après l'étape 3.

Figure 4.4 – Illustration des étapes de la construction du graphe P_G à partir du graphe G. Dans les Figures 4.4b, 4.4c et 4.4d, on distingue les arêtes de E en trait noir plein et les arêtes de H en trait pointillés bleu.

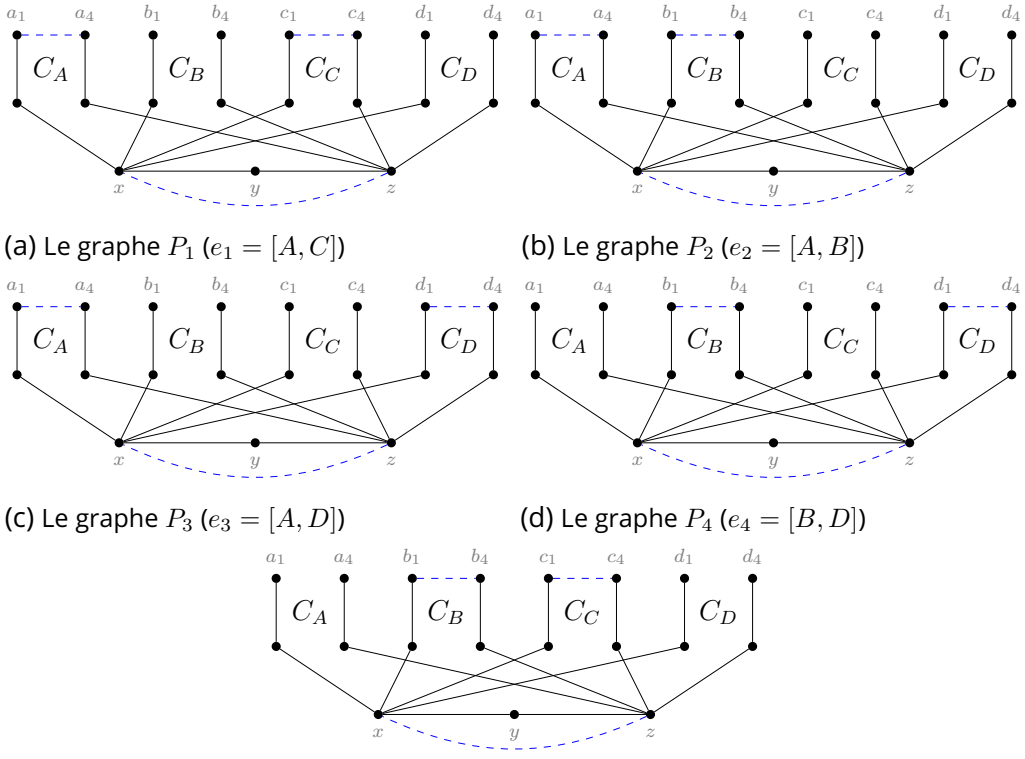
À partir du graphe P_G, nous constituons un ensemble de graphes $\mathcal{G} = \{P_1, \ldots, P_{|E_G|}\}$. Pour chaque arête $e_i = [v, w] \in E_G$, nous construisons le graphe P_i (un sous-graphe de P_G) dans lequel, pour chaque sommet $w \in V_G \setminus \{v, w\}$, l'arête $[u_1, u_4]$ de c_w a été supprimée. Nous avons $C_i = \{c_v, c_w, d\}$, l'ensemble de cycles associé à P_i, et GC_i le graphe de cycles associé avec C_i son ensemble de sommets.

Nous avons l'ensemble de cycles $\mathcal{C} = \bigcup_{i=1}^{|E_G|} C_i$ où C_i est l'ensemble de sommets de GC_i et nous pouvons noter que $|\mathcal{C}| = |V_G|+1$. À partir d'une instance $G = (V_G, E_G)$ du problème du nombre chromatique, nous avons ainsi construit une instance $(\{G_1, G_2\}, \{GC_1, GC_2\}, H)$ de min-PCP.

Théorème 5. *Le problème min-*Partition de Cycles Polymorphes *ne peut pas être approximé à un facteur* $|\mathcal{C}|^{1/5-\epsilon}$ *pour tout* ϵ.

Démonstration. Un graphe P_G est construit depuis G avec un ensemble d'arêtes H, selon la procédure proposée.

Cette réduction d'une instance $G = (V_G, E_G)$ du problème du nombre chromatique vers une instance $< \{G_1, G_2\}, \{GC_1, GC_2\}, H >$ de PolyPart est clairement polynomiale. En effet, la procédure est itérative et chaque étape est limitée soit par $|V_G|$, soit par $|E_G|$.



(a) Le graphe P_1 ($e_1 = [A, C]$)

(b) Le graphe P_2 ($e_2 = [A, B]$)

(c) Le graphe P_3 ($e_3 = [A, D]$)

(d) Le graphe P_4 ($e_4 = [B, D]$)

(e) Le graphe P_5 ($e_5 = [B, C]$)

Figure 4.5 – Illustration des graphes de l'ensemble $\mathcal{G} = \{P_1, \ldots, P_{|E_G|}\}$.

Soit, l'ensemble de cycles $\mathcal{C} = \bigcup_{i=1}^{|E_G|} C_i$ où C_i est l'ensemble de sommets de GC_i et tous les cycles de $\mathcal{C}$ ont le sommet x, extremité d'une arête de H, en commun. La propriété d'ancrage (Propriété 4, page 46) est donc toujours vérifiée. De plus, les cycles ont tous l'arête $[x, z]$ en commun, et partagent donc le même voisinage. La propriété de voisinage (Propriété 3, page 44) est donc également toujours vérifiée. Ainsi, seule la propriété de coexistence (Propriété 2, page 44) qui vérifie l'absence de coexistence dans les graphes de cycles est discriminante pour établir les ensembles polymorphes dans le cas présent.

Une partition $\mathcal{P} \in < \{G_1, G_2\}, \{GC_1, GC_2\}, H >$ de cardinalité m correspond à une répartition en $m - 1$ stables de V_G. En effet, le cycle d appartient à l'ensemble de cycles de chacun des graphes, il forme donc une partie composée d'un seul cycle. Hormis ce singleton, chaque autre ensemble de cycles polymorphes correspond à un stable de V_G. Nous avons alors un $m - 1$-coloriage de G.

Dans [6], il a été établi que trouver une approximation du nombre chromatique de G dans l'intervalle d'erreur $|V_G|^{\frac{1}{5}-\epsilon}$, est difficile et ce pour tout $\epsilon > 0$. Cela montre l'inapproximabilité du problème MCG et donc par conséquent l'inapproximabilité de min-PCP. □

4.1.3 . Paramètres de la complexité du problème PCP

Nous avons établi que le problème PCP est NP-complet dans le Théorème 4. Cette preuve considère des instances de PCP dans lesquelles tous les graphes sont planaires et toutes les arêtes de H appartiennent à au moins un cycle. Ces restrictions ont pour objectif de nous rapprocher d'instances correspondant à des trajectoires de dynamique moléculaire.

Plusieurs pistes semblent maintenant prometteuses pour préciser la complexité de PCP et min-PCP, ainsi que pour améliorer la modélisation de ces problèmes pour une application pratique comme celle de l'analyse de trajectoires de dynamique moléculaire.

Ajouter la borne maximum du degré des sommets dans les graphes moléculaires pourrait potentiellement restreindre l'espace de recherche et conduire à des résultats plus précis. En effet, comme nous l'avons évoqué dans la Section 4.1.1, les atomes ne peuvent former qu'un nombre restreint de liaisons chimiques, ainsi le degré d'un sommet est borné. C'est pourquoi, ce paramètre sera d'autant plus d'intérêt dans le contexte de l'analyse de trajectoires.

De même, prendre en compte la minimalité des bases de cycles dans la preuve de la complexité pourrait améliorer la pertinence de la modélisation pour l'application que nous en avons. En effet, dans nos preuves nous utilisons un ensemble de cycles qui définit une base de cycles sans vérifier si celle-ci est minimale, alors qu'en pratique nous utilisons spécifiquement des bases de cycles minimum. Les graphes étant planaires et les cycles construits définissant des faces, il faudrait donc vérifier que la face extérieure a un poids plus grand que les faces internes pour considérer des bases de cycles minimum.

En ce qui concerne l'inapproximabilité de min-PCP, étendre la preuve à des graphes moléculaires permettrait une compréhension plus approfondie des limites de l'approximation dans ce contexte spécifique. Pour l'instant nous n'établissons aucune propriété sur les graphes utilisés dans la preuve du Théorème 5.

Ces pistes de recherche offrent des opportunités pour explorer davantage les aspects théoriques des problèmes PCP et min-PCP.

4.2 . Méthodes proposées pour résoudre min-PCP

Cette section propose une représentation d'une partition en polycycles par un polygraphe. L'objectif est d'utiliser ce polygraphe pour accéder aux solutions améliorantes, c'est-à-dire les partitions en polycycles possèdant une partie de moins. À partir de cette représentation, nous présentons deux méthodes pour calculer une solution au problème de minimisation qui nous intéresse.

La Section 4.2.2 présente une approche exacte qui consiste à énumérer tous les ensembles de cycles polymorphes possibles afin d'aboutir à une partition minimisant le nombre de polycycles. Bien que cette méthode soit garantie de fournir une solution optimale, elle peut être longue et coûteuse en termes de temps de calcul. Ainsi, dans la Section 4.2.3, nous proposons une heuristique gloutonne dont la rapidité en fait une alternative intéressante.

4.2.1 . Modélisation d'une partition en polycycles par un polygraphe

Dans cette section, nous presentons le polygraphe qui représente une partition $< \mathcal{G}, \mathcal{GC}, H >$ quelconque. Rappelons qu'un polygraphe est un graphe dans lequel les sommets sont des polycycles, et il existe une arête entre deux polycycles si ces derniers partagent une arête dans un graphe de cycles de $\mathcal{GC}$. Cette définition est celle introduite dans le Chapitre 2 et correspond au polygraphe attendu pour caractériser une trajectoire de dynamique moléculaire.

Nous optons pour une représentation par polygraphe pour modéliser les partitions en polycycles, ce qui offre une approche flexible et pratique. Cette représentation nous permet d'ajouter des arêtes pour exprimer les contraintes spécifiques entre les éléments de la partition. En intégrant ces contraintes dans le polygraphe, nous facilitons la recherche de solutions améliorantes à partir d'une partition donnée, dans les cas où de telles solutions existent.

Considérons une partition en polycycles $P = \{p_1, p_2, \dots p_k\}$ de cardinalité k. Une solution améliorante est une partition en polycycles P' de cardinalité $k-1$ telle qu'il existe $p_i, p_j \in P$ avec $i, j \in [1, k]$ et $i \neq j$, et il existe $p' \in P'$ tel que $p' = p_i \cup p_j$, alors $P' = (P \setminus \{p_i, p_j\}) \cup \{p'\}$. En d'autres termes, une solution améliorante peut être obtenue à partir de P en fusionnant deux polycycles pour en former un seul. Cette approche de construction d'un polygraphe, étant donnés deux polycycles compatibles, est décrite par l'Algorithme 13.

Algorithme 13 Méthode de construction d'une solution améliorante à partir d'un polygraphe

Entrée Un polygraphe $GP = (P, I)$ où P est son ensemble de sommets et I est son ensemble d'arêtes, et deux polycycles p et q tels que $p \in P$, $q \in P$ et $p \cup q$ est un ensemble de cycles polymorphes.

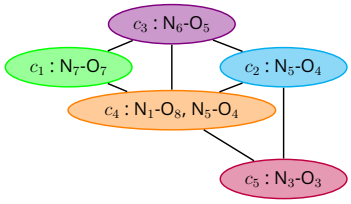
1: **Fonction** Ameliore_Polygraphe(p, q, GP) : un polygraphe
2: $\mathcal{C} \leftarrow P \setminus \{p, q\}$
3: $GP^* \leftarrow GP[\mathcal{C}]$
▷ Le sous graphe de GP contenant uniquement les sommets présents dans $\mathcal{C}$
4: Ajouter le sommet $p' = p \cup q$ à GP^*
5: **Pour tout** $r \in \mathcal{C}$ **faire**
6: **Si** $[p, r] \in I$ ou $[q, r] \in I$ **alors**
7: Ajouter l'arête $[p', r]$ à GP^*
8: **Renvoyer** GP^*

Dans la suite de cette section, nous présentons comment reconnaître une solution améliorante étant donnée une partition en polycycles, et comment construire le polygraphe correspondant à cette solution améliorante. Nos méthodes nécessitent une partition de départ, nous considérons alors la partition en singletons. Dans cette partititon où chaque polycycle est composé d'un seul cycle, toutes les propriétés d'une partition en cycles polymorphes sont respectées.

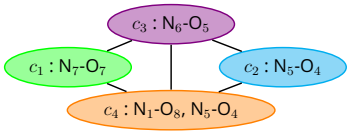
Un polygraphe efficace. Afin de facilement reconnaître une solution améliorante, nous devons identifier les couples de polycycles qui peuvent être fusionnés. Il s'agit des couples qui, une fois unis, vérifient les trois propriétés du polymorphisme. Notons que les couples qui partagent une arête dans le polygraphe classique ne vérifient pas la propriété de non-coexistence. À partir de cette observation, nous avons cherché à ajouter d'autres arêtes qui seraient indicatives de l'absence de compatibilité entre ces deux polycycles. Ainsi, nous avons défini les arêtes de contraintes, un second type d'arête, utilisée uniquement lors de la représentation de partition et qui specifie que deux polycycles ne peuvent pas être fusionnés. L'objectif est que ces arêtes ne nécessitent pas de traitement spécifique lors de la transititon vers une solution améliorante. Ainsi, elles sont conservées

avec la fusion des polycycles dans le polygraphe de la solution améliorante. Nous ajoutons donc une arête de contraintes entre deux polycycles p et p', si il existe une graphe $GC \in \mathcal{GC}$ tel que $p, p' \in GC$ et $[p, p']$ n'est pas une arête de GC.

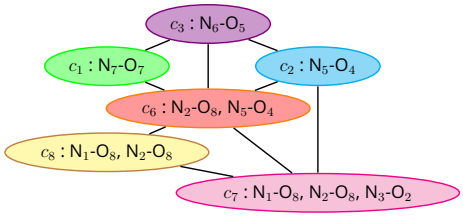
Dans la Figure 4.6, nous observons trois graphes de cycles à partir desquels nous construisons la partition en singletons représentée par le polygraphe illustré dans la Figure 4.7.



(a) Graphe de cycles GC_1.



(b) Graphe de cycles GC_2.



(c) Graphe de cycles GC_3.

Figure 4.6 – Trois graphes de cycles obtenus pour une trajectoire hypothétique.
Figure initialement présentée dans le Chapitre 2.

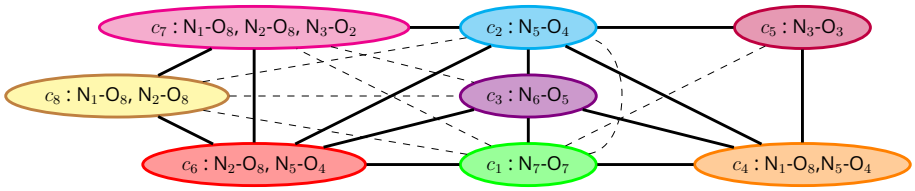


Figure 4.7 – Polygraphe correspondant à la partition en singletons issue des graphes de cycles de la Figure 4.6. Les arêtes en trait plein correspondent aux arêtes classiques, et les arêtes en pointillés aux arêtes de contraintes.

Ajouter des arêtes de contraintes réduit le nombre de paires de sommets à considérer dans le polygraphe. Cependant, les deux autres propriétés du polymorphisme ne peuvent pas être facilement représentées par des arêtes de contraintes, contrairement à ce qui était possible avec la propriété de non-coexistence (Propriété 2, page 44).

La propriété d'ancrage (Propriété 4, page 46), concerne l'existence d'au moins un sommet commun à tous les cycles de l'ensemble et qui soit impliqué dans certaines liaisons spécifiques. Lorsque deux polycycles sont fusionnés pour former un seul polycycle, il est possible que cette propriété ne soit plus vérifiée. Prenons l'exemple de trois polycycles p, p_1 et p_2 tels que $V(p) \cap V(p_1) \cap V(H) = \{X\}$ et $V(p) \cap V(p_2) \cap V(H) = \{Y\}$. Supposons que p_1 et p_2 soient fusionnés en un seul polycycle p_{12}. Alors, $V(p) \cap V(p_{12}) \cap V(H) = \emptyset$. Ainsi, le polycycle p et le nouveau polycycle p_{12} ne valident pas la propriété d'ancrage, même si les couples (p, p_1) et (p, p_2) la vérifiaient auparavant. Il est donc nécessaire de vérifier la propriété pour chaque nouveau couple si la propriété est vérifiée. Dans le cas

de la propriété d'ancrage, cela se traduit par une vérification en $O(n)$, où n représente le nombre de sommets d'un graphe.

La même problématique se pose pour la propriété de voisinage (Propriété 3, page 44), qui concerne la similarité des voisinages dans les graphes de cycles, pouvant varier en fonction des cycles polymorphes considérés. Bien que les arêtes de contraintes déjà établies pour l'incompatibilité des voisinages restent valables, à chaque nouvelle fusion, il est nécessaire de vérifier la propriété pour établir les couples qui ne la vérifient plus. Dans ce cas, la complexité est de l'ordre de $O(N^2)$, où N désigne le nombre de graphes différents.

La représentation de ces propriétés implique donc un coût non négligeable dans la construction d'une nouvelle solution améliorante. Pour obtenir un modèle efficace, nous conservons le modèle décrit précédemment, avec les arêtes classiques représentant la coexistence et l'interaction des polycycles, ainsi que les arêtes de contraintes traduisant uniquement la coexistence. Cette approche nous permet de représenter efficacement les partitions en polycycles et de générer des solutions améliorantes lorsque cela est possible.

Une implémentation de cette approche consiste à utiliser une étiquette associée aux arêtes comme une porte logique. Ainsi, une arête étiquettée 1 est une arête classique qui traduit une interaction entre deux polycycles, et une arête étiquettée 0 désigne une arête de contraintes.

Remarque 15. *Dans l'Algorithme 13, la transmission des arêtes d'un polygraphe vers le suivant ne prend pas en compte le type des arêtes. Il faut donc ajouter cette information lors de l'ajout de l'arête dans le nouveau polygraphe. Cela revient à faire les modifications suivantes :*

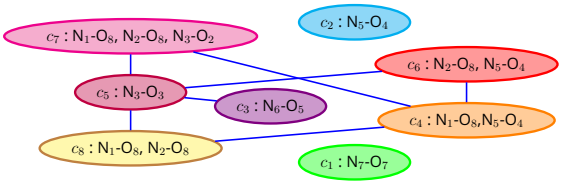
- *la ligne 7 décrit l'ajout de l'arête $[p', r]$, ainsi cette instruction devient «Ajouter $[p', r]$ de type $T([p, r])$ à GP^* si $[p, r] \in I$, sinon $[p', r]$ est de type $T([q, r])$ »*

Reconnaissance d'une solution améliorante. À partir du modèle que nous avons défini, pour que deux polycycles soient compatibles et puissent être fusionnés en un seul polycycle, ils ne doivent pas partager d'arêtes dans le polygraphe. De plus, l'ensemble résultant de leur union doit vérifier les propriétés de voisinage et d'ancrage.

Reprenons l'exemple de la trajectoire hypothétique avec la Figure 4.8 qui illustre la reconnaissance des solutions améliorantes depuis le polygraphe de la Figure 4.7. La Figure 4.12a représente son graphe complémentaire, où chaque arête $[p, q]$ indique l'absence de cette arête dans le polygraphe original. Chaque arête de ce graphe complémentaire représente un couple de polycycles vérifiant la propriété de non-coexistence (Propriété 2, page 44). Le Tableau 4.12b énumère, quant-à-lui, tous les couples vérifiant la propriété de non-coexistence et identifie s'ils sont ou non polymorphes. En conclusion, à partir du polygraphe en singletons, nous avons accès à deux solutions améliorantes.

Pour ce faire nous proposons maintenant des implémentations possibles pour l'identification de couples polymorphes et donc l'existence de solutions améliorantes.

Vérifier si l'union de deux polycycles respecte la propriété d'ancrage (Propriété 4, page 46) ne nécessite pas une procédure très compliquée. Nous considérons une méthode qui fait l'intersection entre l'ensemble des sommets présents dans tous les cycles d'un polycycle p et les sommets qui sont l'extremité d'au moins une arête de l'ensemble d'arêtes H. L'appel Intersection_Sommets(p, H) fait référence à cette procédure. Ainsi, l'union de deux



(a) Graphe complémentaire

p	q	polymorphes ?	propriété non vérifiée
c_3	c_5	Non	Prop. 4
c_4	c_6	Oui	Aucune
c_4	c_7	Non	Prop. 3
c_4	c_8	Non	Prop. 3
c_5	c_6	Non	Prop. 4
c_5	c_7	Oui	Aucune
c_5	c_8	Non	Prop. 4

(b) Compatibilité des couples correspondant à une arête du graphe complémentaire.

Figure 4.8 – Exemple de détection de solutions améliorantes à partir du polygraphe de la Figure 4.7.

polycycles p et q respecte la propriété 4 si et seulement si

$$\text{Intersection\_Sommets}(p, H) \cap \text{Intersection\_Sommets}(q, H) \neq \emptyset.$$

L'Algorithme 14 présente maintenant une fonction qui vérifie si l'union de deux polycycles satisfait la propriété de voisinage (Propriété 3, page 44). Pour ce faire, l'algorithme prend en entrée deux polycycles, une partition en polycycles $\mathcal{P}$ et un ensemble de graphes de cycles $\mathcal{GC}$.

Algorithme 14 Vérifie si l'union de deux polycycles respecte la Propriété 3

Entrée Une partition en polycycles $\mathcal{P}$, deux polycycles p et q tels que p et q sont deux parties distinctes de $\mathcal{P}$ et un ensemble de graphes de cycles $\mathcal{GC}$
Sortie Vrai si la Propriété 3 est vérifiée pour $p \cup q$, et Faux sinon.

```
1:  Fonction Prop_Voisins(P, p, q, GC) : booléen
2:      attendu ←un tableau de taille |P| initialisé à Null.
3:      Pour tout i allant de 1 à |GC| faire
4:          Soit, gc_i = (C_i, E_{C_i}) tel que gc_i ∈ GC.
5:          Si (p ∪ q) ∩ C_i ≠ ∅ alors
6:              Sans perte de généralité, p_i désigne l'identité de (p ∪ q) appartenant
    à C_i.
7:              Pour tout c ∈ E_{C_i} \ p_i faire
8:                  Si [p_i, c] ∈ E_{C_i} alors
9:                      valeur ← 1
10:                 Sinon
11:                     valeur ← 0
12:                 P(c) désigne la partie de P contenant c.
13:                 Si attendu[P(c)] = Null alors
14:                     attendu[P(c)] ← valeur
15:                 Sinon Si attendu[P(c)] ≠ valeur alors
16:                     Renvoyer Faux
17:     Renvoyer Vrai
```

À présent pour déterminer l'existence d'une solution améliorante dans un polygraphe donné, il est nécessaire de vérifier s'il existe un couple de polycycles qui ne partagent pas d'arêtes et dont l'union vérifie les Propriétés 3 et 4.

Ainsi, l'Algorithme 15 présente la fonction Couples_ Polymorphes qui énumère tous les couples de polycycles dont la fusion forme un polycycle. Chacun de ces couples constitue donc une solution améliorante. Cette méthode sera utilisée dans les deux approches que nous proposerons pour identifier les couples compatibles disponibles.

Algorithme 15 L'ensemble des couples de polycycles compatibles

Entrée Un polygraphe $GP = (P, I)$ où P est son ensemble de sommets et I est son ensemble d'arêtes, un ensemble de graphes de cycles $\mathcal{GC}$ et un ensemble d'arêtes H.

Sortie L'ensemble des couples de polycycles (p, q) tels que $p \cup q$ est un ensemble de cycles polymorphes.

1: **Fonction** Couples_polymorphes(GP, $\mathcal{GC}$, H) : ensemble de couples
2: $\mathcal{A} \leftarrow \emptyset$
3: **Pour** i allant de 1 à $|P| - 1$ **faire**
4: **Pour** j allant de $i + 1$ à $|P|$ **faire**
5: $p, q \leftarrow P[i], P[j]$
6: **Si** $[p, q] \notin I$ **alors**
7: **Si** Intersection_Sommets$(p, H) \cap$ Intersection_Sommets$(q, H) \neq \emptyset$ **alors**
8: **Si** Prop_Voisins$(P, p, q, \mathcal{GC})$ **alors** ▷ Algorithme 14
9: Ajouter (p, q) à $\mathcal{A}$.
10: **Renvoyer** $\mathcal{A}$

Solution de départ Dans les deux prochaines sections, nous présentons deux méthodes de parcours des solutions améliorantes pour calculer le polygraphe d'une trajectoire. Ces deux méthodes ont le même point de départ : la partition en singletons. L'Algorithme 16 détaille la construction du polygraphe correspondant à cette partition étant donné un ensemble de graphes de cycles $\mathcal{GC}$ et un ensemble d'arêtes H. Pour ce faire, la procédure nécessite de déterminer le type d'arête existant entre deux sommets.

Algorithme 16 Construction du polygraphe de la partition en singletons

Entrée Un ensemble de graphes de cycles $\mathcal{GC}$

Fonction Polygraphe_Singletons ($\mathcal{GC}$) : un polygraphe
    $\mathcal{C}^* \leftarrow \emptyset$
    Soit $\mathcal{C}^*$ désigne l'ensemble de sommets, et $\mathcal{I}$ désigne l'ensemble d'arêtes du polygraphe. On a $\mathcal{C}^*, \mathcal{I} \leftarrow \emptyset, \emptyset$
    Pour tout i allant de 1 à $|\mathcal{GC}|$ **faire**
        Soit, $gc_i = (C_i, E_{C_i})$ tel que $gc_i \in \mathcal{GC}$.
        Pour tout $p \in C_i$ **faire**
            Si $p \notin \mathcal{C}^*$ **alors**
                Ajouter p à $\mathcal{C}^*$
            Pour tout $q \in (C_i \setminus p) \cap \mathcal{C}^*$ **faire**
                Si $[p, q] \in E_{C_i}$ **alors**
                    ajouter l'arête $[p, q]$ étiquettée 1 à $\mathcal{I}$
                Sinon
                    ajouter l'arête $[p, q]$ étiquettée 0 à $\mathcal{I}$
    Renvoyer $GP = (\mathcal{C}^*, \mathcal{I})$

Remarque 16. *Dans l'Algorithme 16, les arêtes classiques sont étiquettées* 1 *et les arêtes de contraintes sont étiquettées* 0*. Ainsi, le polygraphe utilisé pour l'analyse d'une trajectoire de dynamique moléculaire correspond au sous-graphe ne contenant que des arêtes de type* 1*.*

4.2.2 . Méthode exacte pour résoudre min-PCP

Cette section détaille la procédure pour résoudre de manière optimale le problème min-PCP étant donnée une trajectoire pour laquelle un ensemble de graphes de cycles $\mathcal{GC}$ a été déterminé, ainsi qu'un ensemble d'arêtes H.

À chaque étape, la méthode calcule, à partir d'un polygraphe donné, toutes les solutions améliorantes accessibles. Ainsi, la fonction Couples_Polymorphes renvoie la liste des couples valides et crée pour chacun une solution améliorante, puis l'ajoute à une liste de polygraphes améliorés. Cette étape énumérative est décrite dans l'Algorithme 17.

Algorithme 17 Méthode de construction de toutes les solutions améliorantes à partir d'un polygraphe

Entrée Un polygraphe GP, un ensemble de graphes de cycles $\mathcal{GC}$, un ensemble d'arêtes H.

1: **Fonction** all_amelioration(GP, $\mathcal{GC}$, H)
2: $\mathcal{A} \leftarrow$ Couples_Polymorphes$(GP, \mathcal{GC}, H)$ ▷ Algorithme 15
3: $\mathcal{GP}s \leftarrow \emptyset$
4: **Pour tout** $(p, q) \in \mathcal{A}$ **faire**
5: $gp \leftarrow$ Ameliore_Polygraphe(p, q, GP)
▷ Algorithme 13 modifié comme indiqué dans la Remarque 15
6: Ajouter gp à $\mathcal{GP}s$
7: **Renvoyer** $\mathcal{GP}s$

À partir de cette étape énumérative, nous construisons une méthode exacte qui explore toutes les partitions possibles en partant de la partition en singletons pour trouver une ou plusieurs solutions optimales au problème min-PCP. Il s'agit donc d'une méthode "brute force" qui évalue toutes les possibilités pour s'assurer de trouver la meilleure. Toutes les solutions améliorantes accessibles depuis un graphe sont donc construites et ajoutées à l'ensemble des graphes à traiter, pour repéter l'opération et ce jusqu'à ce qu'il n'y ait plus de solution améliorante. La méthode est décrite par la fonction min-PCP_exact de l'Algorithme 18. Dans cet algorithme, nous utilisons deux listes :

- $new\_\mathcal{GP}$ désigne l'ensemble des solutions à traiter. À chaque étape la première solution de la liste est extraite pour être traitée, et lorsque de nouvelles solutions sont construites celles-ci sont ajoutées à la fin de la liste. Nous traitons donc les solutions dans l'ordre d'une file d'attente.
- $solution\_\mathcal{GP}$ désigne l'ensemble des solutions optimales trouvées. Cette liste est mise à jour au fur et à mesure de l'exécution en ajoutant les solutions qui n'ont pas de solution améliorante et en supprimant des solutions lorsque de meilleures solutions sont trouvées.

Remarque 17. *La condition de la ligne 17 vérifie avant d'ajouter un élément à l'ensemble des polygraphes de l'étape que celui-ci n'a pas déjà été construit et ajouté par une autre solution améliorante.*

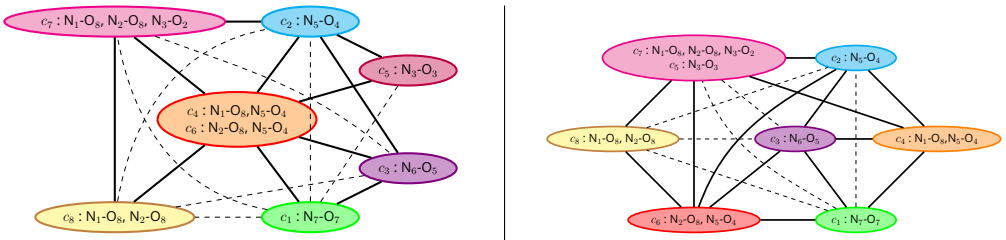
Il s'agit là d'un exemple simple, mais les Figures 4.9 et 4.10 illustrent toutes les solutions explorées à partir du polygraphe de la Figure 4.7. Ce polygraphe constitue le point de départ de la méthode et correspond ici à une partition de cardinalité 8. Dans la Figure 4.9,

Algorithme 18 Méthode exacte pour la résolution de min-PCP

Entrée Un ensemble de graphes de cycles $\mathcal{GC}$, un ensemble d'arêtes H.
Sortie Un ensemble de polygraphes représentant chacun une solution optimale

1: **Fonction** min-PCP_exact($\mathcal{GC}$, H) : un ensemble de polygraphes
2: $solution\_\mathcal{GP}, new\_\mathcal{GP} \leftarrow \emptyset, \emptyset$
3: $GP \leftarrow$ Polygraph_Singletons($\mathcal{GC}$) ▷ Algorithme 16
4: Ajouter GP à $new\_\mathcal{GP}$
5: $min \leftarrow |V_{GP}|$ où V_{GP} désigne l'ensemble de sommets du graphe GP.
6: **Tant que** $new\_\mathcal{GP} \neq \emptyset$ **faire**
7: $GP \leftarrow$ le premier élément de $new\_\mathcal{GP}$
8: supprime GP dans $new\_\mathcal{GP}$
9: $\mathcal{GP}s \leftarrow$ all_amelioration(GP, GC, H) ▷ Algorithme 17
10: **Si** $\mathcal{GP}s \neq \emptyset$ et $|V_{GP}| \leq min$ **alors**
11: **Si** $|V_{GP}| < min$ **alors**
12: $solution\_\mathcal{GP} \leftarrow \emptyset$
13: $min \leftarrow |V_{GP}|$
14: Ajouter GP à $solution\_\mathcal{GP}$
15: **Sinon**
16: **Pour tout** $gp \in \mathcal{GP}s$ **faire**
17: **Si** $gp \notin new\_\mathcal{GP}$ **alors**
18: Ajouter gp à la fin de la liste $new\_\mathcal{GP}$
19: **Renvoyer** $solution\_\mathcal{GP}$

nous observons deux solutions améliorantes. Ces solutions avaient déjà été présentées dans la Figure 4.8 de la section précédente mais n'avaient pas été illustrées sous forme de polygraphe. À partir de ces deux solutions, une solution unique de cardinalité 6 est trouvée. En effet, à partir du polygraphe de la Figure 4.9a, le seul couple polymorphe est (c_5, c_7), et à partir du polygraphe de la Figure 4.9b, le seul couple polymorphe est (c_4, c_6). Cette solution est donc créée deux fois mais ne sera ajoutée à l'ensemble des solutions à traiter qu'une seule fois car il s'agit d'un doublon. Lorsque cette solution est traitée, aucune solution améliorante n'est accessible donc nous avons une solution optimale.



(a) fusion de c_4 et c_6 (b) fusion de c_5 et c_7

Figure 4.9 – Solutions améliorantes correspondant à des partitions de cardinalité 7, obtenues à partir du polygraphe de la Figure 4.7

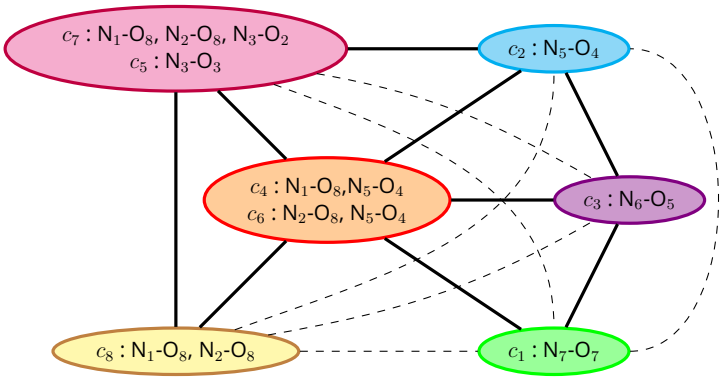


Figure 4.10 – Unique solution améliorante obtenue à partir des polygraphes de la Figure 4.9 et correspondant à une partition de cardinalité 6.

4.2.3 . Méthode heuristique pour tendre vers une solution à min-PCP

Cette section détaille la procédure pour résoudre grâce à une heuristique gloutonne le problème min-PCP étant donné une trajectoire pour laquelle un ensemble de graphes de cycles $\mathcal{GC}$ a été déterminé, ainsi qu'un ensemble d'arêtes H.

Cette procédure, tout comme la méthode énumérative, part de la partition en singletons et construit des solutions améliorantes pour aboutir à une solution. La différence réside dans le fait que la méthode exacte créée toutes les solutions améliorantes accessibles depuis une solution courante, tandis que dans cette méthode heuristique, une seule solution améliorante est créée à partir de la solution courante. Nous devons, parmis toutes les solutions améliorantes accessibles, en choisir une. Pour cela, nous avons défini une fonction de score qui prend en compte le nombre de cycles des polycycles ainsi que le nombre de sommets impliqués dans des liaisons spécifiques partagées par les polycycles.

Pour un polycycle p, on note $V(\bigcap_{c\in p} c)$ l'ensemble des sommets impliqués dans tous les cycles de p. Ainsi, étant donnés deux polycycles p et q, nous avons :

$$score(p,q) = |V(\bigcap_{c\in p} c) \cap V(\bigcap_{d\in q} d) \cap V(H)| \times (|p| + |q|)$$

Remarque 18. *Cette fonction de score a été élaborée et testée au fil des expérimentations, et bien qu'elle se soit avérée efficace, elle reste un élément à faire évoluer dans la perspective d'une étude approfondie.*

L'Algorithme 19 décrit la procédure pour extraire le couple de polycycles ayant le score maximum à partir d'un ensemble de couples de polycycles donné.

Remarque 19. *Étant donnés deux polycycles p et q tels que $p \cup q$ est également un polycycle, nous avons $score(p,q) \geq 2$. En effet, si p et q sont polymorphes alors $V(\bigcap_{c\in p} c) \cap V(\bigcap_{d\in q} d) \cap V(H) \neq \emptyset$ pour que la propriété d'ancrage soit vraie. Les polycycles étant au moins composés d'un cycle chacun, nous avons ce score minimum.*

À chaque étape, une seule meilleure solution est donc construite. Si aucune solution améliorante n'est trouvée à partir du polygraphe courant, alors celui-ci est renvoyé. Cette méthode est décrite par la fonction min-PCP_greedy de l'Algorithme 20.

Nous reprenons l'exemple de la la partition en singletons de la Figure 4.7. Deux couples polymorphes ont été identifiés dans la Figure 4.8, ils ont les scores suivants :

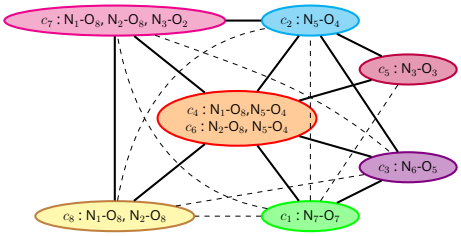
Algorithme 19 Un couple de polycycles de score maximum

Entrée un ensemble de couples de polycycles $\mathcal{A}$
Sortie un couple de polycycle.

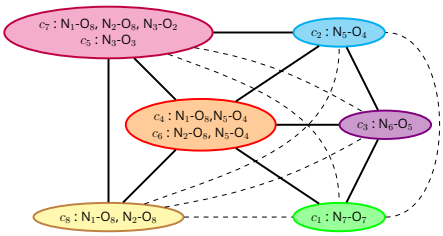
1: **Fonction** Meilleur_couple($\mathcal{A}$) : deux polycycles
   $best\_p, best\_q, best\_s \leftarrow Null, Null, 0$
2: **Pour tout** $(p, q) \in \mathcal{A}$ **faire**
3: $sommets\_p \leftarrow$ Intersection_Sommets(p, H)
4: $sommets\_q \leftarrow$ Intersection_Sommets(q, H)
5: $score \leftarrow |sommets\_p \cap sommets\_q| \times (|p| + |q|)$
6: **Si** $score > best\_s$ **alors**
7: $best\_p, best\_q, best\_s \leftarrow p, q, score$
8: **Renvoyer** $best\_p, best\_q$

- $score(c_4, c_6) = 6$ avec $V(c_4) \cap V(H) = \{N_1, N_5, O_4, O_8\}$ et $V(c_6) \cap V(H) = \{N_2, N_5, O_4, O_8\}$.
- $score(c_5, c_7) = 2$ avec $V(c_5) \cap V(H) = \{N_3, O_3\}$ et $V(c_7) \cap V(H) = \{N_1, N_2, N_3, O_2, O_8\}$.

Le couple (c_4, c_6) est donc sélectionné et le polygraphe de la Figure 4.11a est construit. La Figure 4.12 illustre les solutions améliorantes de ce polygraphe. Nous observons que le seul couple polymorphe est (c_5, c_7), ces polycycles sont fusionnés comme illustré dans la Figure 4.11b. Ce dernier polygraphe sera la réponse de l'algorithme car il n'a aucune solution améliorante. Dans cet exemple, le résultat de la méthode heuristique et de la méthode exacte est le même, mais ce n'est pas toujours le cas.

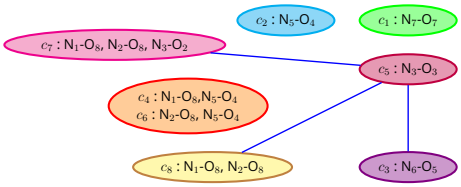


(a) fusion de c_4 et c_6 de score 6.



(b) fusion de c_5 et c_7 de score 2.

Figure 4.11 – Illustration des polygraphes parcourus par la fonction min-PCP_greedy à partir du polygraphe de la Figure 4.7.



(a) Graphe complémentaire.

p	q	polymorphes ?	propriété non vérifiée
c_3	c_5	Non	Prop. 4
c_5	c_7	Oui	Aucune
c_5	c_8	Non	Prop. 4

(b) Compatibilité des couples correspondant à une arête du graphe complémentaire.

Figure 4.12 – Illustration des solutions améliorantes à partir du polygraphe de la Figure 4.11a.

Algorithme 20 Heuristique pour la résolution de min-PCP

Entrée Un ensemble de graphes de cycles $\mathcal{GC}$, un ensemble d'arêtes H.
Sortie Un ensemble de polygraphes représentant chacun une solution optimale

1: **Fonction** min-PCP_greedy($\mathcal{GC}$, H) : un polygraphe
2: $GP \leftarrow$ Polygraph_Singletons($\mathcal{GC}$) ▷ Algorithme 16
$\mathcal{A} \leftarrow$ Couples_Polymorphes($GP, \mathcal{GC}, H$) ▷ Algorithme 15
3: **Tant que** $\mathcal{A} \neq \emptyset$ **faire**
4: $p, q \leftarrow$ Meilleur_couple($\mathcal{A}$) ▷ Algorithme 19
5: $GP \leftarrow$ Ameliore_Polygraphe(p, q, GP)
▷ Algorithme 13 modifié comme indiqué dans la Remarque 15
6: $\mathcal{A} \leftarrow$ Couples_Polymorphes($GP, \mathcal{GC}, H$) ▷ Algorithme 15
7: **Renvoyer** GP

4.3 . Synthèse sur le calcul du polygraphe

Le Chapitre 2 avait introduit le problème fondamental de la construction du polygraphe d'une trajectoire. Ce chapitre était quant-à lui focalisé sur l'étude approfondie de ce problème. Les résultats de ce chapitre, font partie d'une publication soumise à la conférence ESA 2024 [1].

Étant donnés un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ et un ensemble d'arêtes H, nous considérons l'ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$, où pour tout i avec $1 \leq i \leq N$, $B_i \in \mathcal{MCB}(G_i)$. Nous obtenons alors l'ensemble de graphes de cycles $\mathcal{GC} = \{GC_1, GC_2, \ldots, GC_N\}$ où $B_i^* \subseteq B_i$ tel que B_i^* contient tous les cycles de B_i qui utilisent au moins une arête de H. C'est à partir de ces informations que nous recherchons une partition en polycycles $\mathcal{P} =< \mathcal{G}, \mathcal{GC}, H >$ de cardinalité minimale pour construire le polygraphe.

Nous avons présenté le problème Partition de Cycles Polymorphes (PCP) ainsi que le problème de minimisation qui l'accompagne, min-Partition de Cycles Polymorphes (min-PCP). Nous avons établi que même pour des instances similaires à celles d'une trajectoire de dynamique moléculaire, le problème PCP est NP-complet. De plus, nous avons démontré l'inapproximabilité de min-PCP dans le cas général.

La Section 4.2.2 présente une méthode exacte de calcul d'une partition en polycycles $\mathcal{P}$. Nous faisons référence à cette méthode comme la méthode "exacte" du calcul du polygraphe. L'approche proposée commence par une partition en singletons, qui constitue une partition en polycycles. Ensuite, elle parcourt toutes les partitions en polycycles ayant exactement une partie de moins. En répétant cette recherche, nous obtenons un ensemble de partitions toutes de cardinalité m telle qu'il n'existe pas de partition en polycycles $\mathcal{P}$ de cardinalité inférieure à m.

La Section 4.2.3 propose quant à elle la méthode "heuristique". Celle-ci commence également par une partition en singletons mais, elle sélectionne à chaque itération une seule partition en polycycles ayant exactement une partie de moins. Ainsi, bien que la méthode retourne une partition en polycycles, elle ne garantit pas l'optimalité de la solution.

Ces deux méthodes seront étudiées dans le Chapitre 5. Étant donné la diversité des trajectoires moléculaires en termes de dimension du système moléculaire et du nombre

de conformations observées, nous analyserons le coût supplémentaire de la méthode exacte par rapport à la méthode heuristique. Pour cela, nous évaluerons la distance entre la solution obtenue par la méthode exacte et celle proposée par la méthode heuristique pour des trajectoires dans lesquelles le temps d'exécution de la méthode exacte reste raisonnable.

5 - Comparaison des méthodes de calcul du polygraphe étant donnée une trajectoire de dynamique moléculaire

Ce chapitre présente notre évaluation des différentes méthodes décrites dans cette thèse pour le calcul du polygraphe.

Rappelons que nous considérons en entrée un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ et un ensemble d'arêtes H. À partir de ces données, nous calculons un polygraphe qui a pour objectif d'être représentatif de la topologie des éléments de $\mathcal{G}$.

Pour ce faire, nous procédons en deux étapes consécutives comme nous l'avons décrit dans la Chapitre 2. La première étape consiste à définir un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$ où $\forall i$ avec $1 \leq i \leq N, B_i \in \mathcal{MCB}(G_i)$. À partir de ces bases, un ensemble de graphes de cycles $\mathcal{GC} = \{GC_1, GC_2, \ldots, GC_N\}$ est défini, avec $\forall i$ tel que $1 \leq i \leq N, B_i^*$ est l'ensemble de sommets de GC_i qui constitue l'ensemble des cycles de B_i utilisant une arête de H. Ainsi, $B_i^* \subseteq B_i$. La seconde étape consiste alors à calculer une partition en polycycles $\mathcal{P} = \langle \mathcal{G}, \mathcal{GC}, H \rangle$ qui minimise son nombre de parties. Le polygraphe final est obtenu en considérant cette partition comme l'ensemble de ses sommets. Ces deux étapes sont consécutives et interdépendantes. En effet, la sélection des bases de cycles lors de la première étape influence directement la partition obtenue à la seconde.

Les méthodes seront évaluées en termes de qualité et de rapidité d'exécution afin d'établir une méthodologie optimale pour le calcul du polygraphe étant donnée une trajectoire de dynamique moléculaire. Les méthodes sont implementées en python et executées sur un MacBook Pro (2021) muni d'une puce Apple M1 Pro et d'une mémoire unifiée de 16 Go.

Rappelons que notre objectif est de minimiser le nombre de sommets de ce polygraphe.

Pour procéder à l'évaluation de ces méthodes, nous avons deux types de données sur lesquelles nous les avons exécutées :

1. Des trajectoires simulées, issues de modèles physiques complexes, représentent des molécules connues. Leur calcul est coûteux, nous disposons donc seulement d'un petit nombre de ces trajectoires. Elles sont utilisées pour comparer les méthodes de partitionnement en polycycles. En effet, les modèles physiques derrière ces trajectoires assurent la cohérence de l'ensemble des graphes de la trajectoire pour cette étape.

2. Les trajectoires générées, quant à elles, sont des suites de graphes créées pour ressembler à des trajectoires simulées. Elles sont obtenues à faible coût et nous permettent de tester nos méthodes sur des modèles de trajectoires variées. Cependant, ces trajectoires ne représentent pas de véritables phénomènes physiques. Elles sont utilisées pour l'évaluation des méthodes de sélection des cycles. Grâce à elles, nous disposons d'un jeu de données suffisamment conséquent pour observer la tendance générale de ces méthodes pour le calcul du polygraphe.

Le chapitre est organisé comme suit. Tout d'abord, dans la Section 5.1, nous analysons les méthodes de calcul du partionnement en polycycles présentées dans le Chapitre 4. À partir d'un jeu de données préalablement décrit et étant donné un ensemble de bases de cycles, nous commençons par présenter nos résultats sur la méthode exacte de calcul d'une partition en polycycles de cardinalité minimale. Ensuite, nous comparons les résultats obtenus par la méthode heuristique sur les mêmes jeu de données. Nous analysons alors la distance entre la solution obtenue par la méthode heuristique et celle obtenue par la méthode exacte, pour établir l'efficacité de la méthode heuristique pour le calcul du polygraphe.

Ensuite, dans la Section 5.2, nous évaluons les bases obtenues par les méthodes de sélection des bases de cycles minimum présentées dans le Chapitre 3, en fonction du polygraphe finalement obtenu par l'utilisation de la méthode heuristique. Nous examinons les différences sur le polygraphe final en fonction de la méthode de sélection des cycles utilisée. Pour cette évaluation, qui sera plus axée sur des aspects numériques, nous utilisons des suites de graphes générées aléatoirement pour ressembler à des trajectoires simulées. Nous présentons donc, avant notre évaluation, la méthode de construction des trajectoires générées que nous avons utilisée.

Enfin, la Section 5.3 décrit la procédure que nous avons arrêtée pour le calcul du polygraphe, étant donné une trajectoire de dynamique moléculaire.

5.1 . Calcul de la partition en polycycles

Dans cette section, nous évaluons les résultats obtenus par les différentes méthodes de calcul du polygraphe. Pour ce faire, nous étudions les résultats obtenus sur différentes trajectoires de dynamiques moléculaires issues de simulations.

Les trajectoires simulées, bien que plus restreintes que celles générées par nos propres moyens, reposent sur de nombreux calculs physiques précis. Cela permet d'établir de manière plus fiable la cohérence des résultats obtenus par la méthode heuristique en les comparant à ceux obtenus par la méthode exacte. L'utilisation de trajectoires simulées, nous a donc paru, plus appropriée pour valider la robustesse et la précision de la méthode heuristique.

En procédant ainsi, nous visons à illustrer que la méthode heuristique offre une solution de qualité presque comparable à celle de la méthode exacte, tout en étant bien sûr plus efficace en termes de temps de calcul.

Dans la suite, à la Section 5.1.1, nous décrivons les différentes trajectoires simulées sur lesquelles nous avons exécuté les méthodes. Puis, dans la Section 5.1.2, nous étudions les résultats obtenus sur ces différentes trajectoires avec les deux méthodes. Nous concluons cette section, en observant la distance entre la solution de la méthode heuristique et la solution de la méthode exacte, ainsi que le temps d'exécution de chacune des méthodes.

5.1.1 . Données de test : trajectoires simulées

Dans cette section, nous présentons les trajectoires simulées utilisées pour évaluer nos méthodes de calcul du polygraphe.

Ces trajectoires sont le fruit de simulations de dynamique moléculaire, prenant en compte de nombreux paramètres physiques. Plus de détails sur leur obtention seront présentés dans le Chapitre 6, où nous explorerons l'interprétation du polygraphe dans le

contexte moléculaire.

Les trajectoires simulées fournissent une représentation physique réaliste, ce qui les rend idéales pour évaluer la méthode de calcul du polygraphe. En effet, leur utilisation nous permet de comparer la méthode heuristique à la méthode exacte dans un contexte plus fidèle à la réalité moléculaire.

Dans cette section, nous nous concentrons donc sur les caractéristiques générales de ces trajectoires, en les distinguant en fonction de leurs caractéristiques.

Rappelons (voir Chapitre 2, page 33) que nous considérons, chaque trajectoire comme un ensemble de graphes $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, où pour tout i tel que $1 \leq i \leq N$, nous avons $G_i = (V, E \cup H_i)$. L'ensemble des arêtes variables H est tel que $H = \bigcup_{i=1}^{N} H_i$. L'ensemble de cycles de la trajectoire $\mathcal{C}$ est obtenu par l'union des cycles d'intérêts de chacun des graphes de $\mathcal{G}$. Ainsi, nous avons $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$ où pour tout $1 \leq i \leq N$, $B_i^* \subseteq B_i \in \mathcal{MCB}(G_i)$ désigne l'ensemble de cycles d'intérets de G_i, c'est-à-dire l'ensemble des cycles de B_i qui utilisent une arête de H.

Pour définir cet ensemble de cycles de la trajectoire, nous devons être en possession d'un ensemble de bases de cycles minimum $\mathcal{B} = \{B_1, B_2, \ldots B_N\}$ où pour tout $1 \leq i \leq N$, $B_i \in \mathcal{MCB}(G_i)$. Dans le contexte de la comparaison des méthodes de calcul du partionnement en polycycles, nous considérons un ensemble $\mathcal{B}$ obtenu par l'algorithme de Horton (Algorithme 1, page 54). Cette méthode est polynomiale et constitue donc notre première approche quant-à la sélection des bases de cycles.

Remarque 20. *Nous étudions l'apport des autres méthodes de calculs de bases de cycles minimum dans la Section 5.2. Après avoir choisi la méthode de calcul du partitionnement dans la Section 5.1.2.*

Pour compléter cette évaluation, il faudrait appliquer toutes les combinaisons de méthodes pour la sélection des bases de cycles et pour le calcul du partionnement. Néanmoins, comme nous le verrons dans la Section 5.1.2, le coût de la méthode exacte rend son utilisation très limitée.

La Table 5.1 présente les trajectoires simulées que nous utilisons dans ce chapitre. Ce tableau présente les caractéristiques suivantes : le nombre de graphes N, le nombre de sommets $|V|$, le nombre d'arêtes fixes $|E|$, le nombre d'arêtes variables $|H|$ et, le nombre de cycles d'intérêts de la trajectoire $|\mathcal{C}|$.

molécule	N	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert H\rvert$	$\lvert\mathcal{C}\rvert$
Z-Ala$_6$−COOH	53	41	41	10	12
Chondroitin disulfate	64	35	36	13	14
Gramicidine−COOH	430	136	143	18	51

Table 5.1 – Caractéristiques des trajectoires simulées utilisées pour la comparaison des méthodes de partionnement en polycycles.

5.1.2 . Évaluation des méthodes de calculs de la partition en polycycles

Cette section présente l'étude comparée des deux méthodes décrites dans le Chapitre 4. Pour ce faire, la Section 5.1.2 présente les résultats obtenus par l'implémentation de la méthode exacte. Ensuite, la Section 5.1.2 présente les résultats obtenus par la méthode heuristique et les compare à ceux de la méthode exacte.

Pour rappel, la méthode exacte est présentée dans la Section 4.2.2 (page 105), tandis que la méthode heuristique est décrite dans la Section 4.2.3 (page 107). Les deux méthodes partent d'une partition en singletons, c'est-à-dire une partition dans laquelle chaque partie contient exactement un cycle. À chaque itération, nous cherchons une ou plusieurs partitions en cycles polymorphes ayant exactement une partie de moins. Les algorithmes s'arrêtent lorsque la partition courante ne peut plus être améliorée. Ainsi, la méthode exacte considère à chaque itération toutes les solutions correspondant à une cardinalité donnée, garantissant ainsi une solution optimale. Tandis que l'heuristique choisit à chaque itération une seule solution de cardinalité donnée pour poursuivre son exécution, offrant une solution moins coûteuse en temps de calcul, mais sans garantie d'optimalité globale.

Résultats de la méthode exacte

Nous observons les résultats pour les trois trajectoires présentées dans la Section 5.1.1 à la Table 5.1. Dans la Section suivante, les cardinalités des solutions obtenues seront comparées, ainsi que le temps de calcul nécessaire pour les obtenir. Le temps observé ici est uniquement celui de l'algorithme de partitionnement.

Dans la suite de cette section, nous présentons les solutions optimales obtenues pour les trajectoires testées. Ces résultats sont présentés sous la forme de tableaux. Un tableau décrit alors l'ensemble des solutions optimales obtenues étant donnée une trajectoire et, représente alors l'appartenance des cycles de la trajectoire aux différents polycycles de la solution. Ainsi, chaque colonne correspond à une solution optimale trouvée par la méthode exacte, et chaque ligne correspond à un cycle de la trajectoire. Dans les solutions où un cycle est un singleton, cela se représente par le symbole "$-$" dans la case correspondante. En revanche, si un cycle appartient à un polycycle ayant plusieurs identité alors c'est le numéro du polycycles en question qui est indiqué.

Résultats pour le peptide Z-Ala$_6-$COOH : Nous observons les résultats d'une trajectoire issue d'une simulation de dynamique moléculaire d'un peptide d'environ 80 atomes, nommé Z-Ala$_6-$COOH. Les caractéristiques de cette trajectoire sont décrites dans la Table 5.1. Notons que les atomes d'hydrogène ne sont pas pris en compte dans ces graphes, ce qui explique le passage de 80 atomes à seulement 41 sommets.

La méthode exacte a identifié 6 solutions optimales de cardinalité 9 pour les 12 cycles de cette trajectoire. La Table 5.2 présente en détails la répartition des cycles de la trajectoire en polycycles. Tandis que la Figure 5.1 illustre la solution optimale décrite par la première colonne de la Table 5.2. Le temps de calculs sur cette trajectoire a été d'environ 6 secondes.

cycles de la trajectoire	S. 1	S. 2	S. 3	S. 4	S. 5	S. 6
$cycle_1$: N_4-O_4, N_5-O_5	1	1	1	1	1	1
$cycle_2$: N_4-O_4	–	–	–	–	–	–
$cycle_3$: N_7-O_7	–	–	–	–	–	–
$cycle_4$: N_5-O_4	1	1	1	4	4	–
$cycle_5$: N_2-O_2	–	–	–	–	1	–
$cycle_6$: N_3-O_3	–	–	–	–	–	–
$cycle_7$: N_6-O_5	7	7	7	7	7	7
$cycle_8$: N_4-O_4, N_5-O_5, N_6-O_5	7	8	8	4	4	7
$cycle_9$: N_2-O_4	–	8	–	1	–	–
$cycle_{10}$: N_6-O_6	–	–	–	–	–	–
$cycle_{11}$: N_6-O_6, N_7-O_5, N_7-O_7	7	7	7	7	7	7
$cycle_{12}$: N_5-O_5	–	–	8	–	1	1

Table 5.2 – Répartition des cycles de la trajectoire pour les six solutions optimales trouvées. Les polycycles contenant plusieurs cycles sont indiqués par un numéro d'identification, tandis que ceux correspondant à un singleton sont identifiés par le symbole "–".

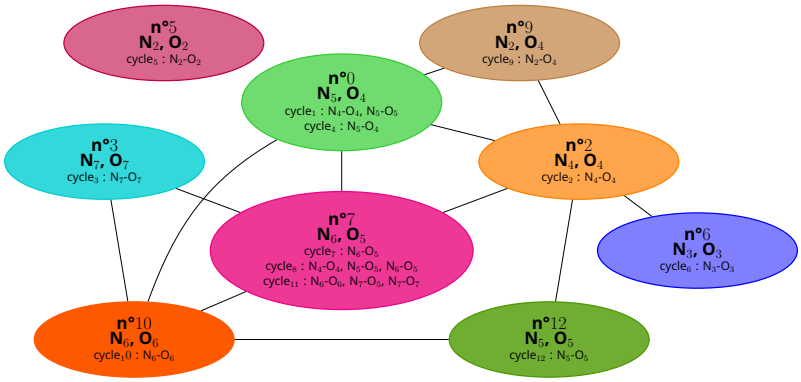


Figure 5.1 – Polygraphe obtenu de la partition en polycycles correspondant à la solution 1 de la Table 5.2.

Résultats pour le peptide Chondroitin disulfate : Les caractéristiques de cette trajectoire sont décrites dans la seconde ligne de la Table 5.1. Pour ce deuxième exemple, nous avons sélectionné une trajectoire montrant de nombreux regroupements. En effet, les 14 cycles de la trajectoires sont partionnés en seulement 6 polycycles. La Table 5.3 détaille la répartition des cycles dans les deux solutions optimales obtenues. Ces solutions sont très proches et se distinguent seulement par le regroupement du $cycle_{15}$ avec le $cycle_3$ dans la solution 1, ou avec le $cycle_{13}$ dans la solution 2.

La méthode exacte sur cette trajectoire a nécessité un temps de calcul d'environ 117 secondes.

Résultats pour le peptide Gramicidine–COOH : Nous considérons pour finir une molécule plus volumineuse. Les caractéristiques de cette trajectoire sont décrites à la troisième ligne de la Table 5.1. En plus de représenter une molécule plus volumineuse en

cycles de la trajectoire	S. 1	S. 2
$cycle_1$: O_5-O_{18}	1	1
$cycle_2$: N_1-O_1	2	2
$cycle_3$: O_2-O_{11}	3	–
$cycle_5$: O_7-O_{15}	5	5
$cycle_6$: N_1-O_1, O_9-O_{17}	6	6
$cycle_7$: O_9-O_{17}	6	6
$cycle_8$: O_7-O_{16}	5	5
$cycle_9$: O_6-O_7	5	5
$cycle_{10}$: O_4-O_{18}	1	1
$cycle_{11}$: O_7-O_9	6	6
$cycle_{12}$: O_7-O_{14}	5	5
$cycle_{13}$: O_2-O_{18}	–	13
$cycle_{14}$: N_1-O_3	2	2
$cycle_{15}$: O_2-O_{12}	3	13

Table 5.3 – Répartition des cycles de la trajectoire parmis les deux solutions optimales trouvées pour la trajectoire de la Chondroitin disulfate. Les polycycles contenant plusieurs cycles sont indiqués par un numéro d'identification, tandis que ceux correspondant à un singleton sont identifiés par le symbole "$-$".

nombre de sommets et d'arêtes, cette trajectoire contient de nombreux graphes différents.

La méthode exacte a identifié 8 solutions optimales de cardinalité 42 pour les 51 cycles identifiés pour cette trajectoire. Le temps de calculs est bien plus conséquent que pour l'exemple précédent du fait du nombre de cycles à considérer. Il a finalement été d'environ 4601 secondes (soit un peu plus d'une heure et quinze minutes).

La Table 5.4 présente en détails la répartition des cycles de la trajectoire en polycycles. Cependant, la visualisation de la répartition des 51 cycles de la trajectoire n'est pas évidente. Nous nous sommes donc limité dans la Table 5.4 aux cycles appartenant à un polycycle contenant au moins deux cycles dans une des solutions obtenues. Ainsi, tous les cycles qui n'apparaissent pas dans le tableau sont des singletons dans toutes les solutions optimales.

On peut noter que de nombreux polycycles sont identiques dans toutes ces solutions optimales, comme par exemple les polycycles 5, 6, 7 et 10.

Nous observons également que certains cycles peuvent appartenir à différent polycycles. Prenons l'exemples des polycycles 4 et 11 dans la Table 5.4. Le polycycle 4 contient le $cycle_4$ dans toutes les solutions, tandis que le polycycle 11 contient toujours le $cycle_{11}$ et le $cycle_{13}$. En revanche, le $cycle_{14}$ appartient au polycycle 4 dans la moitié des solutions et au polcycle 11 dans l'autre moitié. Nous pouvons alors supposer que le $cycle_{14}$ est très flexible, si bien qu'il est polymorphes avec des polycycles différents. Cependant, ces éléments flexibles augmentent le nombre de combinaisons correspondant à un polycycle valide, ce qui se répercute directement sur le nombre de solutions a calculer à chaque étape et donc sur le temps de calcul nécessaire à la méthode.

cycles de la trajectoire	S. 1	S. 2	S. 3	S. 4	S. 5	S. 6	S. 7	S. 8
$cycle_1$: N_8-O_{15} , N_{15}-O_{10}	1	1	1	1	1	1	–	–
$cycle_2$: N_6-O_{13}, N_8-O_{15}	2	2	2	2	2	2	2	2
$cycle_3$: N_8-O_{15}, N_{10}-O_{17}, N_{14}-O_{16}	1	1	1	1	3	3	3	3
$cycle_4$: N_{10}-O_{17}, N_{16}-O_{14}	4	–	4	–	4	–	4	–
$cycle_5$: N_{10}-O_{17}, N_{13}-O_8, N_{16}-O_{14}	5	5	5	5	5	5	5	5
$cycle_6$: N_7-O_2	6	6	6	6	6	6	6	6
$cycle_7$: N_1-O_3, N_7-O_2, N_9-O_4	7	7	7	7	7	7	7	7
$cycle_8$: N_8-O_{15}, N_{10}-O_{17}	1	1	8	8	1	1	3	3
$cycle_9$: N_7-O_2, N_9-O_4	7	7	7	7	7	7	7	7
$cycle_{10}$: N_{15}-O_{10}	10	10	10	10	10	10	10	10
$cycle_{11}$: N_{10}-O_{17}, N_{14}-O_{16}	11	11	11	11	11	11	11	11
$cycle_{12}$: N_8-O_{15}, N_{10}-O_{17}, N_{15}-O_{17}	–	–	8	8	3	3	3	3
$cycle_{13}$: N_{10}-O_{17}, N_{15}-O_{17}	11	11	11	11	11	11	11	11
$cycle_{14}$: N_{10}-O_{17}	4	11	4	11	4	11	4	11
$cycle_{15}$: N_6-O_{13}, N_8-O_{15}, N_{12}-O_{14}	2	2	2	2	2	2	2	2
$cycle_{16}$: N_{12}-O_{14}, N_{15}-O_{10}	10	10	10	10	10	10	10	10
$cycle_{17}$: N_1-O_3, N_7-O_2	6	6	6	6	6	6	6	6
$cycle_{18}$: N_{13}-O_8	5	5	5	5	5	5	5	5

Table 5.4 – Répartition des cycles de la trajectoire pour les huit solutions optimales trouvées pour la trajectoire de la Gramicidine−COOH. Seuls les cycles appartenant à des regroupements apparaissent dans ce tableau. Les polycycles contenant plusieurs cycles sont indiqués par un numéro d'identification, tandis que ceux correspondant à un singleton sont identifiés par le symbole "−".

Limites de l'application de la méthode exacte. Le nombre de solutions considérées dépend de deux facteurs : le nombre de cycles à partitionner et les contraintes entre ces cycles. En effet, si de nombreuses contraintes existent entre les cycles, alors peu de polycycles différents sont accessibles, ce qui réduit le nombre de solutions possibles. Cette observation explique pourquoi, dans les dernières itérations, le nombre de solutions décroît, en raison des contraintes qui accompagnent les polycycles déjà établis. Dans les premières itérations, le nombre de combinaisons est tel que le nombre de solutions croît très rapidement. Puis, au fur et à mesure que les polycycles se forment, les contraintes se multiplient, si bien que dans les dernières itérations, nous retrouvons assez peu de solutions.

Ainsi, dans le cas de l'exemple de la Gramicidine−COOH de nombreuses contraintes existent entre les cycles si bien que 33 cycles sur les 51 de la trajectoire constituent des singletons dans toutes les solutions optimales. Ainsi, même si le nombre de cycles est grand la méthode exacte est applicable grâce à ces contraintes qui limitent le nombre de solutions à construire.

Concernant l'évolution du nombre de solutions construites au cours de la méthode exacte, la Figure 5.2 illustre le nombre de solutions considérées à chaque itération et le nombre de partitions en polycycles qui sont identifiées à partir de ces solutions dans le cas de la trajectoire de la Chondroitin disulfate. Dans cet exemple, nous observons le partionnement d'un ensemble de seulement 14 cycles et, déjà, à l'itération 5, nous considérons

1382 solutions obtenues à partir des 270 partitons de l'itération 4. Nous observons alors 233 partitions en polycycles issues de ces 1382 solutions.

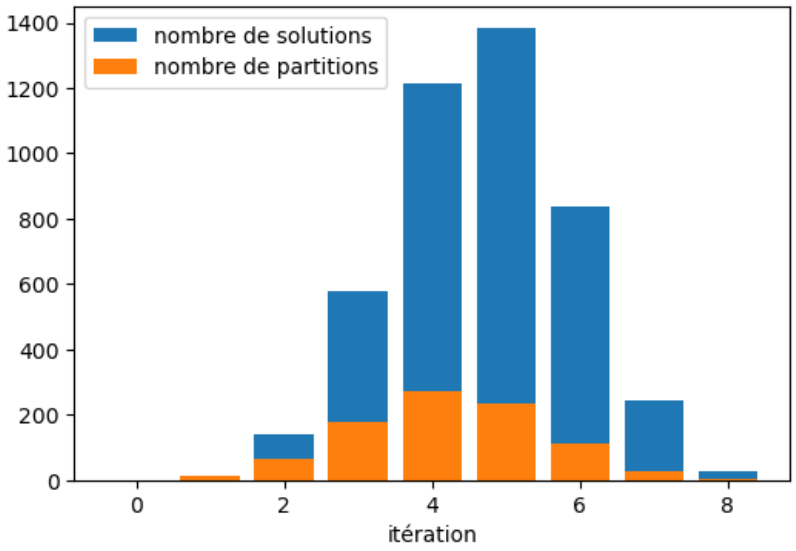


Figure 5.2 – Évolution du nombre de solutions et de partitions en polycycles considérées au cours de l'exécution de la méthode exacte pour la trajectoire de la Chondroitin disulfate.

La méthode exacte considère toutes les combinaisons et a, par conséquent, un temps d'exécution particulièrement long. Elle nécessite également la conservation de nombreuses solutions intermédiaires au cours de son exécution.

Ainsi, l'impact sur le temps de calcul croît en fonction des caractéristiques d'une trajectoire, lesquelles induisent des ensembles de cycles conséquents rendant la méthode inutilisable dans la plupart des cas. C'est pour cela que les exemples que nous avons présentés sont relativement simple, afin d'être en mesure d'y appliquer la méthode exacte.

Comparaison avec la méthode heuristique

Nous observons maintenant les résultats obtenus pour les trois trajectoires de la Table 5.1 avec la méthode heuristique. Sur ces exemples, la solution de la méthode heuristique est l'une des solutions optimales trouvées par la méthode exacte. Ainsi, elle correspond à la solution colorée dans les Tables 5.2, 5.3 et 5.4. La Table 5.5 décrit les résultats des cardinalité des solutions obtenues et du temps de calculs nécessaire pour les deux méthodes. Ainsi, avec la méthodes heuristique, nous avons obtenu pour ces trois exemples, une solution optimale en une fraction du temps nécessaire pour exécuter la méthode exacte.

molécule	$\lvert\mathcal{C}\rvert$	**méthode exacte**		**méthode heuristique**	
		Cardinalité	Temps (sec)	Cardinalité	Temps (sec)
Z-Ala$_6$−COOH	12	9	6	9	$\sim 0,001$
Chondroitin disulfate	14	6	117	6	$\sim 0,002$
Gramicidine−COOH	51	42	4601	42	$< 0,1$

Table 5.5 – Synthèse des résultats obtenus pour les trajectoires simulées avec la méthode heuristique.

Pour conclure, malgré les résultats théoriques de complexité que nous avons établis au Chapitre 4 (page 89), il semble que le problème de partitions en polycycles se résoud plutôt bien pour les trajectoires réelles simulées et que même une heuristique gloutonne se montre efficace.

Ces tests ont été réalisés sur de petites trajectoires et ne démontrent pas pleinement l'efficacité de l'heuristique, mais ils restent encourageants.

Dans la suite, nous adopterons donc cette méthode heuristique pour le calcul du partitionnement.

5.2 . Calcul des bases de cycles minimum

Dans cette section, nous comparons les résultats obtenus par les différentes méthodes de calcul de bases de cycles minimum. Pour cette étude, nous avons généré des suites de graphes analogues à des trajectoires afin de disposer d'un jeu de données plus varié.

Les trajectoires simulées sont peu nombreuses et souvent relativement courtes en raison de leur coût de calcul élevé. Pour obtenir un jeu de données plus vaste, nous avons généré des trajectoires artificielles. Ces dernières ne reposent pas sur des éléments physiques, mais uniquement sur des graphes similaires à ceux observés dans des trajectoires simulées.

Ainsi, dans la Section 5.2.1, nous décrivons la méthode de génération des suites de graphes sur lesquelles nous avons exécuté les différentes méthodes. Puis, dans la Section 5.2.2, nous comparons les résultats obtenus sur ces ensembles de trajectoires générées pour chacune des méthodes de sélection de bases de cycles minimum. Pour évaluer nos approches, nous calculons une partition en polycycles en utilisant la méthode heuristique basée sur les bases de cycles minimum obtenues. Nous comparons ensuite la distance entre les cardinalités des partitions issues des différents ensembles de bases de cycles minimales.

5.2.1 . Données de test : des trajectoires générées

Les trajectoires simulées sont limitées car elles sont coûteuses à calculer. De plus, pour que les résultats soient facilement traitables, nous considérons le plus souvent des molécules de dimension réduite. Nous proposons donc, dans cette section, une méthode de génération permettant de créer des suites de graphes similaires à des trajectoires de dynamique moléculaire.

Notre objectif est de produire des graphes qui, bien que générés sans fondement physique, présentent des caractéristiques similaires à celles observées dans des simulations de dynamique moléculaire. Cette approche permet d'élargir notre jeu de données et de tester les différentes méthodes de calcul des bases de cycles minimales dans des conditions variées.

Nous construisons une suite de graphes de manière aléatoire à partir d'un arbre couvrant en fonction de plusieurs paramètres, qui sont les suivants :

- Nombre de graphes dans la suite, M
- Nombre de sommets du backbone, v
- Nombre d'arêtes fixes supplémentaires du backbone, e
- Nombre d'arêtes variables différentes de la trajectoire, h
- Nombre d'arêtes variables minimum d'un graphe de la trajectoire, min_h
- Nombre d'arêtes variables maximum d'un graphe de la trajectoire, max_h

Nous commençons par construire le backbone commun à tous les graphes de la trajectoire. Pour cela, nous créons un graphe G_I (graphe initial) 4-régulier à v sommets, c'est-à-dire un graphe dans lequel chaque sommet a un degré de 4. Pour construire ce graphe initial, nous utilisons la fonction `random_regular_graph` de la bibliothèque Networkx [25], basée sur la méthode de Steger et Wormald (1999) [47]. Nous déterminons un arbre couvrant de ce graphe initial, auquel nous ajoutons e arêtes de G_I choisies aléatoirement. Le backbone, G_B, ainsi défini est un graphe à v sommets et $v + e - 1$ arêtes. Nous allons maintenant associer aléatoirement un type chimique à chacun de ses sommets. Nous considérons l'ensemble des arêtes de $G_I \setminus G_B$. Parmi cet ensemble, nous tirons aléatoirement h arêtes pour former l'ensemble des liaisons variables possibles, noté H. Pour chaque arête de H, nous définissons les types chimiques des sommets extrémités. Les extrémités de ces arêtes sont référencées comme des azotes ou des oxygènes. Si possible, les deux sommets ne sont pas du même type, mais cela n'est pas obligatoire. Une fois les types des sommets extrémités des h liaisons variables potentiels fixés, les sommets non parcourus, appartenant donc uniquement à des liaisons fixes, sont référencés comme des carbones.

Remarque 21. *Cette étape peut sembler étrange car ces types chimiques ne sont pas basés sur des principes physiques. Cependant, elle est nécessaire pour garantir que l'exécution du programme se déroule sans difficulté. L'implémentation a été conçue initialement pour des trajectoires où les sommets des graphes représentent des types chimiques spécifiques, avec les sommets des liaisons variables étant spécifiquement des atomes d'azote ou d'oxygène.*

À partir du backbone G_B, nous générons M graphes en ajoutant aléatoirement entre min_h et max_h arêtes de H à chaque graphe.

Nous commençons par générer un premier graphe en choisissant le nombre d'arêtes variables nb_h aléatoirement entre min_h et max_h. Ensuite, nb_h arêtes sont choisies parmi

l'ensemble H et ajoutées au backbone pour définir G_1.

Pour définir tous les graphes G_i où $2 \leq i \leq M$, nous procédons un peu différemment pour se rapprocher des propriétés des suites de graphes issues de trajectoires. Rappelons que dans les trajectoires, les graphes consécutifs ne diffèrent généralement que d'une arête. Soit G_i avec $2 \leq i < M$, le dernier graphe de la suite de graphes constituant la trajectoire, nous construisons le graphe G_{i+1}. Nous commençons par définir aléatoirement si le graphe G_{i+1} contient une arête variable en plus ou en moins par rapport à G_i. Le nombre d'arêtes variables étant borné, si G_i contient déjà max_h arêtes variables, alors G_{i+1} en contiendra nécessairement une de moins, et réciproquement si G_i ne contient que min_h arêtes variables, alors G_{i+1} en contiendra nécessairement une de plus. Notons H_i les arêtes variables de G_i, nous avons deux possibilités :

- Si une arête de G_i doit être supprimée, alors une arête $a \in H_i$ est choisie aléatoirement. Le graphe G_{i+1} est obtenu en ajoutant l'ensemble d'arêtes $H_i \setminus \{a\}$ au backbone.
- Si une arête doit être ajoutée à G_i, alors une arête $a \in H \setminus H_i$ est choisie aléatoirement. Le graphe G_{i+1} est obtenu en ajoutant l'ensemble d'arêtes $H_i \cup \{a\}$ au backbone.

Nous observons alors le graphe G_{i+1} obtenu. Si celui-ci est égal à un des graphes déjà calculés, nous retrouvons l'indice de ce graphe pour l'ajouter à la suite. S'il s'agit d'un nouveau graphe, celui-ci est identifié comme tel en lui associant le plus petit indice disponible.

Cette étape est répétée jusqu'à obtenir une suite de M graphes.

La génération aléatoire et les paramètres définis garantissent une diversité de structures, tout en maintenant des propriétés proches des propriétés chimiques réelles, telles que le degré des sommets et la proximité des graphes consécutifs. Cette méthode permet de générer des suites de graphes qui ressemblent à des trajectoires sans contrainte sur la dimension des graphes considérés ou sur la longueur de la suite de graphes.

Dans la Section 5.2.2, nous évaluons nos méthodes de calculs de bases de cycles minimum d'une trajectoire pour la construction du polygraphe. Ainsi, nous considérons trois jeux de trajectoires présentés dans la Table 5.6. Parmis les jeux de données que nous avons générées, nous en avons choisi trois pour illustrer nos résultats dans des situations diverses. Ainsi, nous conduisons une première étude approfondie sur le jeu n°1. C'est un jeu qui contient un grand nombre de trajectoires. Nous étudions ensuite les jeux n°2 et n°3, plus restreints, pour vérifier que nous obtenons les mêmes résultats. En effet, le jeu n°2 représente des graphes deux fois plus grands que ceux du jeu n°1. Ainsi, nous nous sommes restreint à une centaine de trajectoires par limite calculatoire. Ensuite, le jeu n°3 propose, quant-à-lui, des graphes un peu plus petits que le jeu n°1 mais avec plus de liaisons variables simultanées possibles. Nous avons alors 500 trajectoires afin d'assurer la diversité des données. Néanmoins, comme nous le verrons dans la suite, les résultats sont très similaires et ce, peu importe le jeu de données considéré. Il ne nous a donc pas semblé nécessaire d'ajouter plus de trajectoires pour appuyer nos résultats.

5.2.2 . Comparaison des méthodes de sélection des bases de cycles minimum

L'heuristique a montré des résultats satisfaisants dans la Section 5.2 en utilisant l'algorithme classique de Horton pour la sélection des bases de cycles minimum. Nous évaluons ici trois de nos propositions présentées dans le Chapitre 3 :

Jeu	Nombre de trajectoires	M	v	e	h	min_h	max_h
n°1	1000	500	25	3	15	0	5
n°2	100	500	50	3	15	0	10
n°3	500	500	20	3	10	0	10

Table 5.6 – Paramètres utilisés pour générer les trajectoires testées pour la comparaison des méthodes de sélection des bases de cycles minimum.

- Horton modifié (Algorithme 3, page 56) : Cette méthode est très proche de la méthode classique de Horton, mais elle intègre un critère supplémentaire dans l'ordonnancement des cycles. Ainsi, tout comme la méthode classique, elle calcule des bases de cycles minimum très similaires en raison du caractère déterministe de la procédure.
- MCBI (Algorithme 9, page 76) : Cette méthode s'inspire de la résolution locale du problème MCBI pour les graphes consécutifs d'une trajectoire. Ainsi, les bases de cycles minimum des graphes consécutifs maximisent leur interaction. Nous ne considérons pas ici la version globale, car elle est inadaptée aux trajectoires où l'intersection stricte des ensembles de bases de cycles minimum est souvent presque vide.
- Voisinage (Algorithme 12, page 84) : Cette méthode commence par un ensemble de bases de cycles minimum calculé préalablement. Ensuite, elle utilise une heuristique de voisinage pour rechercher un meilleur ensemble de bases de cycles minimum. L'objectif est, ici, de minimiser le nombre de cycles dans l'union des bases.

La cardinalité du polygraphe obtenu par l'utilisation des bases issues de ces trois procédures sera en premier lieu comparé à celle du graphe obtenu à partir des bases obtenues avec la méthode classique de Horton. Nous conclurons ainsi sur l'efficacité de nos propositions pour le calcul du polygraphe.

Notation 5. *Dans cette section, nous faisons référence à la méthode utilisant l'Algorithme de Horton par le sigle "HD" tandis que la méthode basée sur l'algorithme de Horton modifie est désignée par "HDM".*

Les méthodes que nous étudions diffèrent considérablement les unes des autres. En effet, les méthodes de HDM et de MCBI sont directes, car elles calculent un ensemble de bases de cycles minimum en se basant sur une trajectoire donnée. En revanche, la méthode de Voisinage est une méthode d'optimisation qui explore plusieurs ensembles de bases de cycles minimum. Ainsi, nous commencerons notre étude par l'analyse des méthodes directes, avant d'examiner l'apport de la méthode d'optimisation.

Obtention de nos critères d'évaluation Notre critère d'évaluation principal pour ces méthodes de calcul d'un ensemble de bases de cycles minimum est la cardinalité de la partition en polycycles obtenue à partir de cet ensemble. Ainsi, étant donné un jeu de données et une méthode de sélection des bases, nous calculons les partitions en polycycles pour toutes les trajectoires et recupérons ainsi la cardinalité de la partition finale obtenue par la méthode heuristique. Nous construisons ainsi la distribution de la cardinalité des partitions en polycycles pour le jeu considéré en entrée.

Nous considérons plusieurs critères secondaires. Le premier, commun à l'évaluation des méthodes directes et des méthodes de voisinages, est le temps d'exécution total de

la procédure.

Pour les méthodes directes, nous considérons également :

- la cardinalité moyenne des polycycles dans la partition finale
- la cardinalité maximum des polycycles dans la partition finale
- le nombre d'arêtes dans le polygraphe

Ces critères nous permettent d'observer si une des méthodes favorise la diversité au sein des polycycles. Le nombre d'arêtes est, quant-à-lui, un indicateur du nombre de polycycles qui interagissent dans le polygraphe.

Pour les méthodes de voisinages, nous considérons également le nombre d'itérations parmi nos critères. En effet, celui-ci permet d'estimer le nombre de solutions à parcourir avant que l'algorithme termine.

Tous ces critères sont calculés en même temps que la cardinalité de la partition finale et sont représentés sous la forme de distribution, en fonction du jeu de données considéré et de la méthode de calcul des bases de cycles minimum sélectionnée.

Évaluation des méthodes directes

Dans cette section, nous étudions les résultats des méthodes directes. Nous comparons donc les résultats des méthodes HD, HDM, et MCBI sur le jeu de données n°1 décrit à la Section 5.2.1. Nous validons ensuite nos conclusions en observant les résultats des deux autres jeux de données.

La Figure 5.3 et la Figure 5.4 présentent les résultats pour la méthode de Horton, que nous considérons ici comme notre référence, sur le jeu n°1.

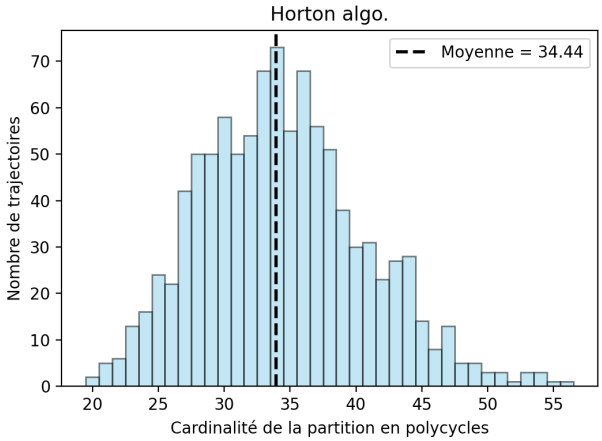
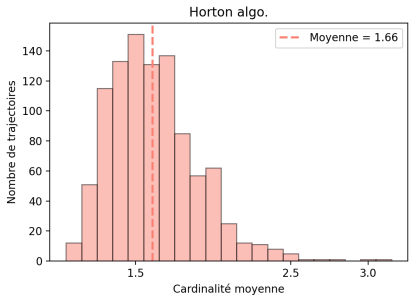
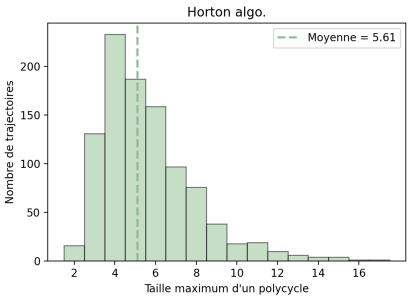


Figure 5.3 – Distribution de la cardinalité des partitions minimales obtenues par l'heuristique de partitionnement appliquée à partir de bases de cycles minimum calculées par la méthode HD. La distribution a été établie sur l'ensemble des 1000 trajectoires qui composent le jeu n°1.

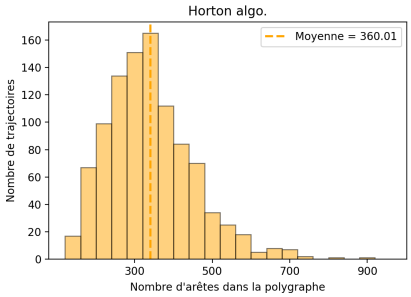
Nous remarquons dans la Figure 5.4d que peu de trajectoires ont un temps d'exécution proche de la moyenne de 5 secondes. En effet, nous observons qu'environ 45% des trajectoires ont un temps d'exécution inférieur à $4,5$ secondes, et environ 40% des trajectoires ont quant-à-elles, un temps d'exécution d'au moins $5,5$ secondes. Ce creux n'a pas de raison particulière d'exister, il s'agit donc d'une particularité du jeu de données n°1.



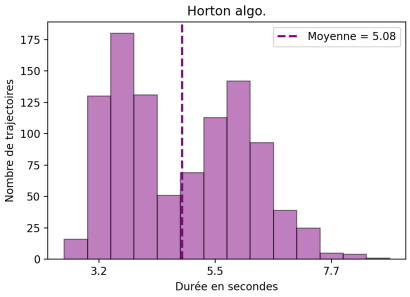
(a) Distribution de la cardinalité moyenne des polycycles d'une partition.



(b) Distribution de la cardinalité maximum des polycycles d'une partition.



(c) Distribution du nombre d'arêtes dans le polygraphe issu de la partition en polycycles.



(d) Distribution du temps d'éxecution de la méthode.

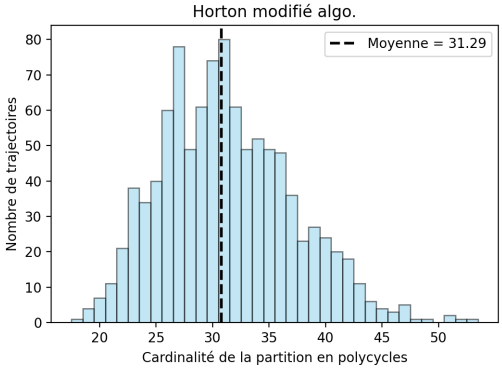
Figure 5.4 – Autre éléments d'évaluation pour la méthode HD. Résultats basés sur les 1000 trajectoires du jeu n°1.

Résultats de l'utilisation de l'Algorithme de Horton modifié La Figure 5.5 présente les résultats obtenus et les comparent aux résultats de la méthode HD. À première vue, la méthode HDM permet une amélioration nette de la cardinalité de la partition en polycycles.

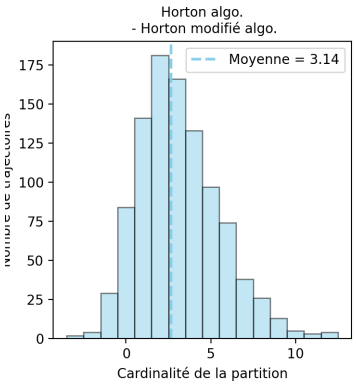
La Figure 5.5a illustre la distribution de la cardinalité des partitions issues des bases calculées par la méthode de Horton modifié. Nous observons que la plage de cette distribution est $[18; 52]$, tandis que pour la méthode HD cette plage est $[20; 56]$. Cela se remarque aussi par la moyenne des cardinalités obtenues. En effet, la moyenne est ici de $31, 29$ contre une moyenne de $34, 44$ indiquée dans la Figure 5.3.

Pour aller plus loin, la Figure 5.5b présente la différence des cardinalités obtenues en comparant les résultats trajectoire par trajectoire. Pour chaque trajectoire, nous avons fait la différence entre la cardinalité obtenue par les bases de HD et celle obtenue par la méthode HDM. Dans la plupart des cas, nous observons que cette différence est de signe positif, ce qui indique que la cardinalité obtenue avec la méthode HDM est meilleure que celle obtenue avec les bases de HD. En moyenne, nous observons une amélioration de la cardinalité de $3, 14$. Cela confirme que la méthode HDM tend à produire des bases de cycles minimum plus optimales, réduisant la cardinalité des partitions en polycycles par rapport à la méthode classique HD.

La Figure 5.6 présente d'autres éléments pour l'évaluation de la méthode HDM. Nous observons que le temps d'exécution de cette méthode est sensiblement le même que celui de la méthode HD, avec une moyenne autour de 5 secondes dans les deux cas, et

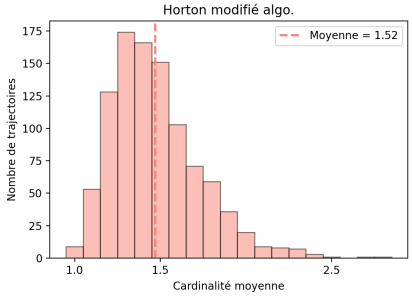


(a) Distribution de la cardinalité des partitions en polycycles.

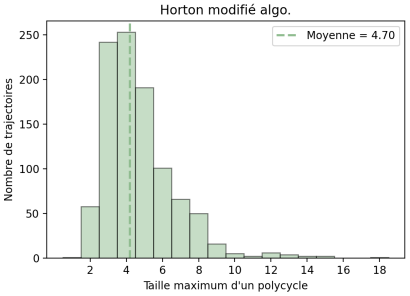


(b) Différence : HD - HDM

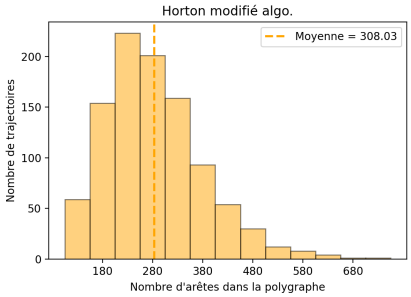
Figure 5.5 – Résultats de la distribution de la cardinalité des partitions minimales obtenues par l'heuristique de partionnement avec des bases de cycles minimum calculées par la méthode HDM. La distribution a été établie sur l'ensemble des 1000 trajectoires qui composent le jeu n°1.



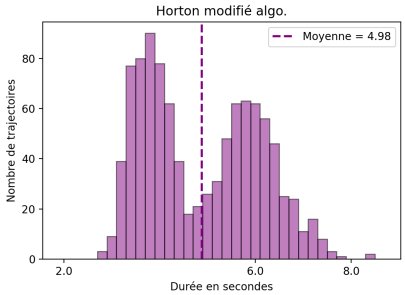
(a) Distribution de la cardinalité moyenne des polycycles d'une partition.



(b) Distribution de la cardinalité maximum des polycycles d'une partition.



(c) Distribution du nombre d'arêtes dans le polygraphe issu de la partition en polycycles.



(d) Distribution du temps d'éxecution de la méthode.

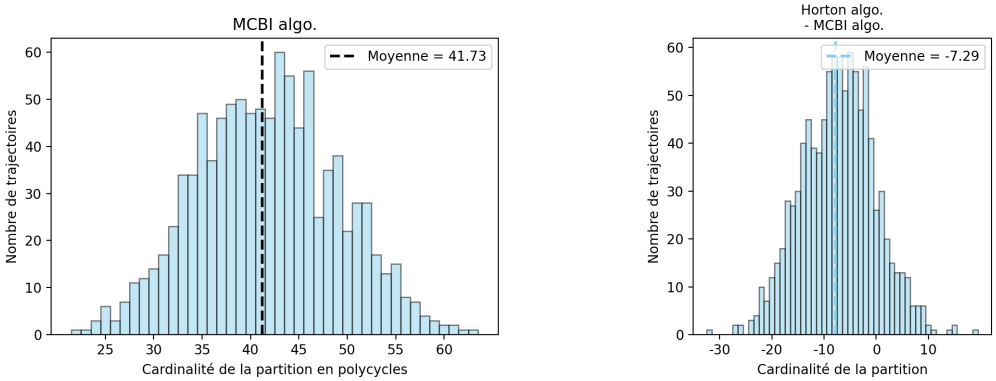
Figure 5.6 – Autre éléments d'évaluation pour la méthode HDM. Résultats basés sur les 1000 trajectoires du jeu n°1.

toujours ce creu propre au jeu de données n°1. De même, en ce qui concerne la cardinalité moyenne au sein des polycycles. Nous observons une différence pour le nombre d'arêtes dans le polygraphe final. Néanmoins, les polygraphes obtenus à partir de HDM contiennent moins de sommets et ont de ce fait moins d'arêtes. Enfin, nous observons

que la cardinalité maximum d'un polycycle est en moyenne plus faible $4,70$ ici, que pour la méthode HD où il est en moyenne de $5,41$. Cela suggère qu'il y a moins d'éléments à regrouper dans les bases de cycles minimum issues de HDM, peut-être parce qu'elles se ressemblent beaucoup plus entre elles.

Ces observations tendent à montrer que la méthode HDM diminue la cardinalité des partitions en polycycles. Cela semble cohérent avec l'hypothèse selon laquelle les bases de cycles minimum produites par HDM sont potentiellement encore plus ressemblantes que celles produites par HD.

Résultats de l'utilisation de MCBI La Figure 5.7 présente les résultats pour cette méthode et les compare à ceux de la méthode de Horton classique. Ces résultats illustrent que la méthode MCBI est le plus souvent inefficace pour le calcul du polygraphe.



(a) Distribution de la cardinalité des partitions en polycycles.

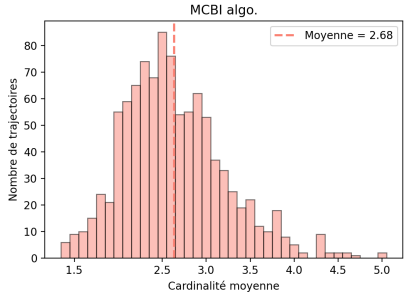
(b) Différence : HD - MCBI

Figure 5.7 – Résultats de la distribution de la cardinalité des partitions minimales obtenues par l'heuristique de partionnement avec des bases de cycles minimum calculées par la méthode MCBI. La distribution a été établie sur l'ensemble des 1000 trajectoires qui composent le jeu n°1.

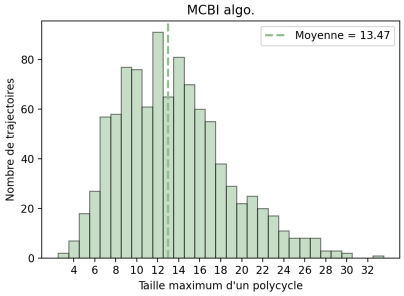
La Figure 5.7a représente la distribution de la cardinalité des partitions issues des bases calculées par la méthode MCBI. Nous observons que la plage de cette distribution est $[22; 63]$, et qu'elle est donc bien supérieure à celle obtenue avec la méthode classique HD. Cela se remarque également par la moyenne des cardinalités obtenues, qui est ici de $41,73$ ce qui est beaucoup plus que la moyenne observée pour la méthode HD. Cela suggère que la méthode MCBI n'est pas favorable à l'obtention d'une partition de faible cardinalité par la méthode heuristique.

La Figure 5.7b présente la différence des cardinalités obtenues en comparant les résultats trajectoire par trajectoire. Le résultat est ici un peu plus positif. En effet, nous faisons la différence entre les bases obtenues par HD et celles obtenues par MCBI, ainsi si cette valeur est négative alors le résultat par MCBI est moins bon que celui de HD. Sur cette figure, dans une grande majorité des cas, les bases obtenues par MCBI induisent des partitions en polycycles de bien plus grande cardinalité. Néanmoins, dans un peu plus de 10% des cas, nous observons qu'une meilleure cardinalité a été obtenue avec la méthode MCBI. La maximisation locale de l'intersection semble donc être généralement

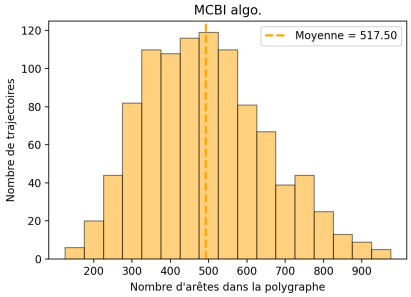
inefficace pour le polygraphe, sauf dans des situations particulières. Une étude approfondie serait nécessaire pour identifier ces trajectoires où une amélioration est observée, afin de comprendre les raisons qui rendent MCBI efficace dans ces exemples et inefficace dans la plupart des trajectoires observées ici.



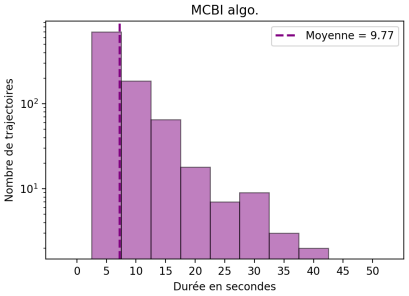
(a) Distribution de la cardinalité moyenne des polycycles d'une partition.



(b) Distribution de la cardinalité maximum des polycycles d'une partition.



(c) Distribution du nombre d'arêtes dans le polygraphe issu de la partition en polycycles.



(d) Distribution du temps d'éxecution de la méthode.

Figure 5.8 – Autres éléments d'évaluation pour la méthode MCBI. Résultats basés sur les 1000 trajectoires du jeu n°1.

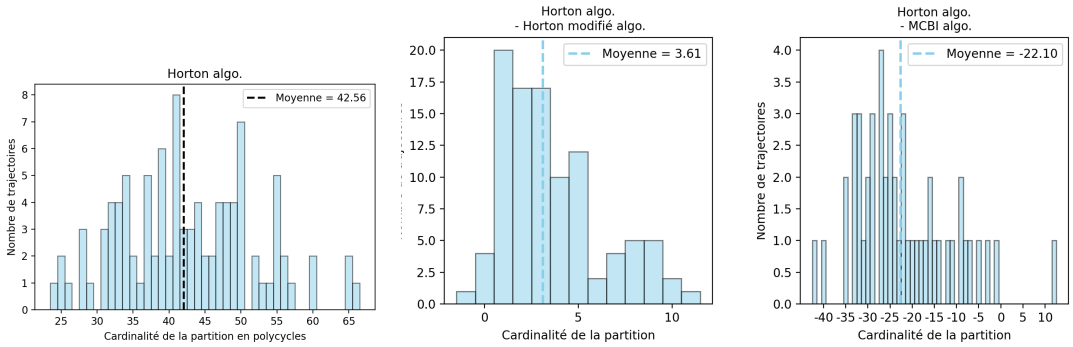
Pour mieux comprendre les particularités des partitions obtenues par des bases issues de MCBI, la Figure 5.8 présente d'autres éléments pour l'évaluation de la méthode.

Nous observons que le temps d'exécution de la méthode MCBI est en moyenne d'environ 10 secondes, ce qui est presque le double de la méthode HD. Cependant, la méthode peut être très longue, comme le montre la Figure 5.8d où plusieurs trajectoires ont nécessité près de 40 secondes d'exécution, ce qui correspond à une augmentation de près de 800% du temps d'exécution. De plus, notons que le creux observé dans les Figures 5.4d et 5.6d n'est plus présent. Celui-ci était donc probablement lié à la méthode déterministe de Horton, que nous réuitilisons dans l'Algorithme de Horton modifié.

Cette variabilité du temps d'exécution, combinée aux résultats moins performants en termes de cardinalité de partition en polycycles, suggère que MCBI est moins efficace et surtout moins stable que les méthodes HD et HDM.

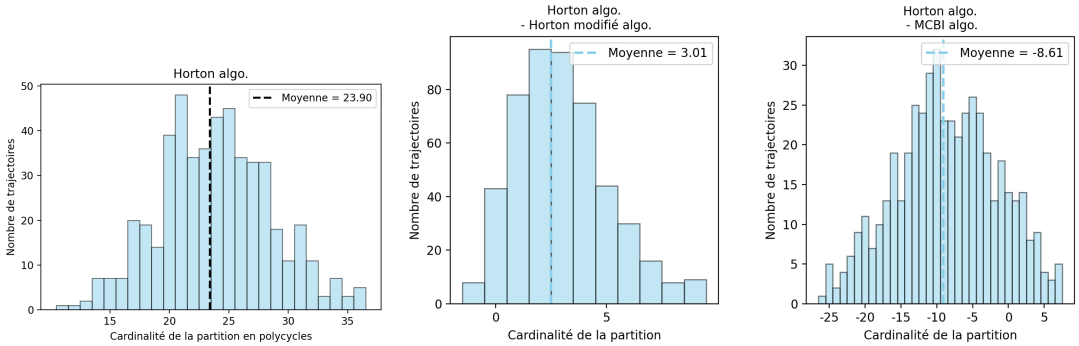
En revanche, les cardinalités des polycycles de la partition finale obtenues avec la méthode MCBI restent étonnantes. Comme, nous l'évoquions, des expérimentations supplémentaires seraient nécessaires pour comprendre quelles trajectoires sont favorables à l'utilisation de cette méthode.

Résultats sur les autres jeux de données Les Figures 5.9 et 5.10 présentent la comparaisons des distributions du nombre de polycycles finaux obtenus en fonction de la méthode de calcul des bases de cycles utilisée. Nous y observons des résultats similaires à ceux du jeu n°1. En effet, la méthode HDM permet dans la plupart des cas d'améliorer le résultat en comparaison de celui obtenu par les bases issues de HD. De même, la méthode MCBI est très instable, elle permet d'accéder à une meilleure solution dans un petit nombre de cas, et reste défavorable pour la majeure partie des trajectoires que nous avons testées.



(a) Distribution de réference avec HD (b) HD - HDM (c) HD - MCBI

Figure 5.9 – Comparaisons du nombre de polycycles obtenu en fonction de l'algorithme de calcul des bases de cycles minimum sur le jeu de données n°2 (100 trajectoires).



(a) Distribution de réference avec HD (b) HD - HDM (c) HD - MCBI

Figure 5.10 – Comparaisons du nombre de polycycles obtenu en fonction de l'algorithme de calcul des bases de cycles minimum sur le jeu de données n°3 (500 trajectoires).

Conclusion sur les méthodes directes Nous avons observé et comparé les résultats des méthodes HDM et MCBI vis-à-vis de la méthode classique HD.

La méthode HDM a montré des résultats très favorables. En effet, cette méthode nécessite un temps d'exécution similaire à celui de la méthode classique tout en améliorant généralement la cardinalité de la partition en polycycles obtenue par l'heuristique. Cette amélioration est significative pour la plupart des trajectoires analysées, soulignant l'efficacité de la méthode HDM dans notre contexte.

En revanche, la méthode MCBI s'est révélée très décevante. Elle est généralement inadaptée pour obtenir une partition en polycycles de cardinalité inférieure. Nous avons donc décidé de ne pas poursuivre cette piste.

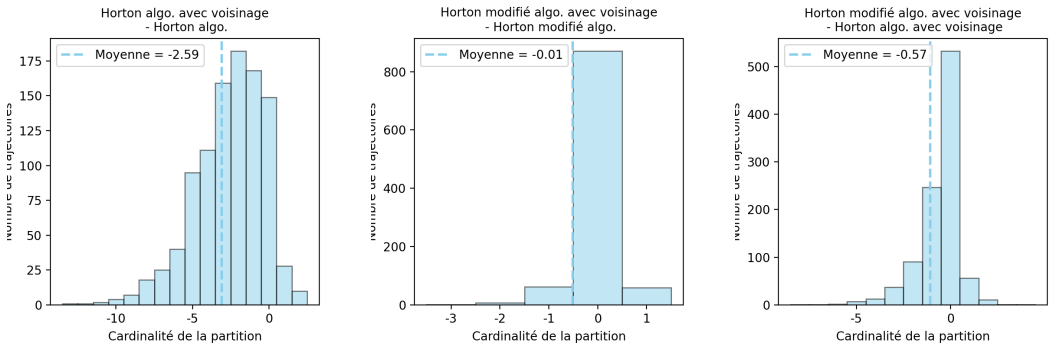
Apport de la méthode de voisinage

Nous présentons maintenant nos résultats concernant l'utilisation de la méthode de voisinage pour améliorer la qualité des bases de cycles minimum, permettant ainsi d'obtenir une partition en polycycles de cardinalité inférieure grâce à la méthode heuristique.

Dans cette section, nous utilisons le même jeu de données que précédemment, mais cette fois-ci, nous appliquons une méthode d'exploration du voisinage à partir des ensembles de bases de cycles minimum obtenus par les méthodes HD et HDM.

Nous commençons par observer les résultats de l'application de l'Algorithme 12, qui décrit une méthode de voisinage pour minimiser la cardinalité de l'union des cycles. Ensuite, nous étudierons si une exploration aléatoire des solutions permet d'accéder à de meilleures solutions que celles obtenues par l'Algorithme 12.

Résultats obtenu de l'application de l'Algorithme 12. Nous examinons, à présent, la solution obtenue par la méthode de voisinage à partir d'un ensemble de bases de cycles minimum. Nous analysons également l'impact du choix des bases de départ sur la solution trouvée par l'algorithme. Pour ce faire, nous observons et comparons les résultats de l'application de la méthode HD seule, puis suivie de l'Algorithme 12, ainsi que de l'application de la méthode HDM seule et suivie de l'Algorithme 12. Nous fixons un nombre maximum d'itérations à $K = 100$.



(a) Voisinage avec HD - HD

(b) Voisinage avec HDM - HDM

(c) Voisinage avec HDM - Voisinage avec HD

Figure 5.11 – Comparaisons de la distance avec la solution renvoyée par l'Algorithme 12 en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 1000 trajectoires du jeu n°1.

La Figure 5.11 présente une comparaison entre les solutions obtenues par l'Algorithme 12 et celles obtenues par la méthode directe seule. Les Figures 5.11a et 5.11b illustrent la différence entre la solution obtenue par l'Algorithme 12 et celle obtenue sans. Ainsi, une différence négative nous indique que la méthode de voisinage est meilleure que la méthode directe seule.

La Figure 5.11a illustre le cas de la méthode HD, nous observons que dans la grande majorité des cas (environ 80%), une meilleure solution est obtenue grâce à la méthode de voisinage. La moyenne de cette différence est de $-2,59$ et dans environ 15% des cas, une solution de même cardinalité à été trouvée par la méthode de voisinage.

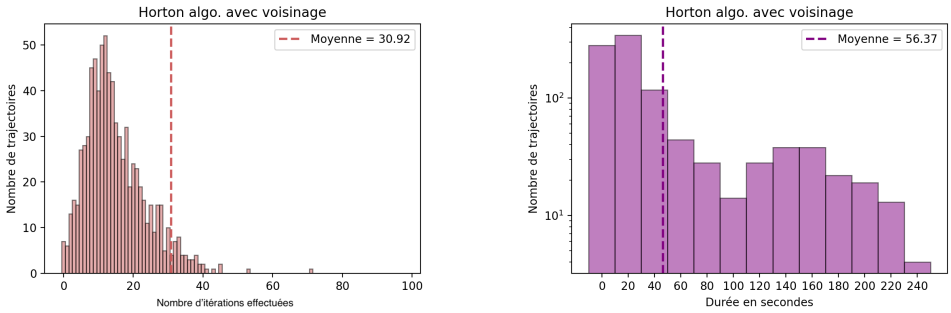
Le cas de la méthode HDM est légèrement différent comme le montre la Figure 5.11b. Nous observons que dans plus de 85% des cas, l'utilisation de la méthode de voisinage n'a pas modifié la cardinalité de la solution optimale. Pour les cas restants, dans la moitié d'entre eux, la solution est améliorée par l'Algorithme 12, et pour l'autre, la solution est déteriorée par l'utilisation de cet algorithme. Notons quand même que les modifications de cardinalité observées sont très faibles, seulement de 1 pour la plupart, qu'elles soient positives ou négatives.

Enfin, la Figure 5.11c met en lumière l'impact de l'ensemble de bases de cycles minimum de départ. Nous observons que dans la très grande majorité des cas, commencer l'algorithme de voisinage par les bases de HDM conduit à des solutions meilleures ou équivalentes aux solutions trouvées en partant des bases de HD. Toutefois, il est à noter que dans quelques cas ($< 8\%$), initier l'Algorithme 12 avec les bases de HD a permis de trouver une partition en polycycles de cardinalité inférieure de 1 ou 2.

La Figure 5.12 présente des critères d'évaluations du coût de l'exécution de l'algorithme de voisinage en utilisant HD pour définir les bases de départ, et de même, dans la Figure 5.13 en utilisant HDM.

Ces figures montrent que le nombre d'itérations effectuées en partant des bases de cycles de HD est bien plus important qu'en partant des bases de cycles de HDM. Cela s'observe également par un temps d'exécution généralement plus long.

Notons que le creux dans le temps d'exécution qui était observé pour les méthodes directes seules est toujours présent ici.

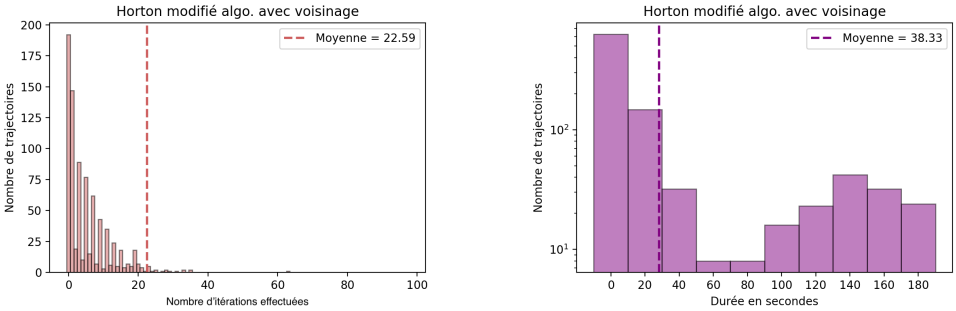


(a) Distribution du nombre total d'itérations effectuées par la méthode de Voisinage.

(b) Distribution du temps d'éxecution de la méthode.

Figure 5.12 – Évaluation du coût d'exécution de l'Algorithme 12 en partant de HD pour définir les premières bases. Résultats établis sur les 1000 trajectoires du jeu n°1.

Pour conclure, l'Algorithme 12 permet généralement d'obtenir des bases de cycles minimum, ce qui réduit la cardinalité de la partition en polycycles issue de l'heuristique de partitionnement. De plus, partir des bases de cycles minimum calculées par HDM permet d'accéder dans la plupart des cas à une meilleure solution et cela plus rapidement, qu'en partant des bases de cycles calculées par HD.



(a) Distribution du nombre total d'itérations effectuées par la méthode de Voisinage.

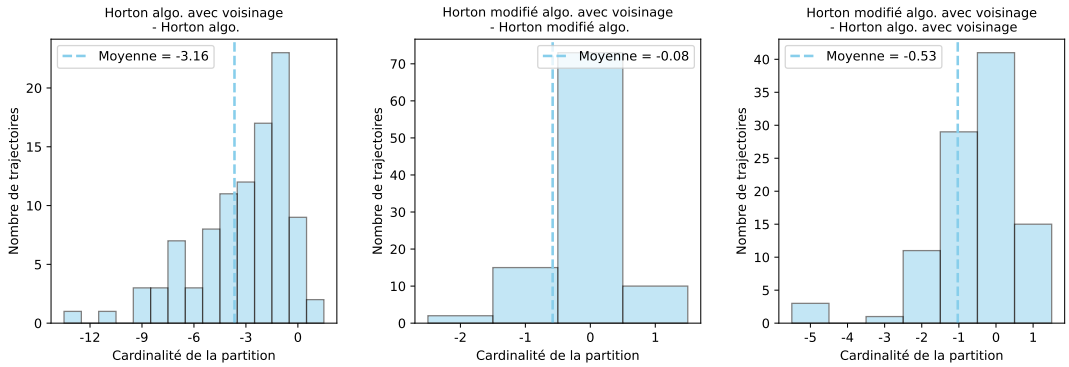
(b) Distribution du temps d'éxecution de la méthode.

Figure 5.13 – Évaluation du coût d'exécution de l'Algorithme 12 en partant de HDM pour définir les premières bases. Résultats établis sur les 1000 trajectoires du jeu n°1.

Cependant, il est important de noter que la partition en polycycles obtenue à partir des bases de HDM correspond déjà à une bonne solution. En effet, l'Algorithme 12 ne permet qu'une amélioration légère, observée dans seulement un petit nombre de cas.

Pour aller plus loin, les Figures 5.14 et 5.15 illustrent les résultats de l'Algorithme 12 appliqué sur les jeux n°2 et n°3.

Nous observons des résultats similaires à ceux du jeu n°1. Ainsi, comme le montre la Figure 5.14a et la Figure 5.15a, en partant des bases de cycles de HD la plupart des tests trouvent une solution meilleure ou équivalent. Cela est également vérifié dans les Figures 5.14b et 5.15b. Néanmoins, dans le cas de HDM l'amélioration est bien moins significative. Pour autant dans les Figures 5.14c et 5.15c montrent que les solutions obtenues en partant de HDM sont dans la majeure partie des cas de cardinalité plus faible ou égale.

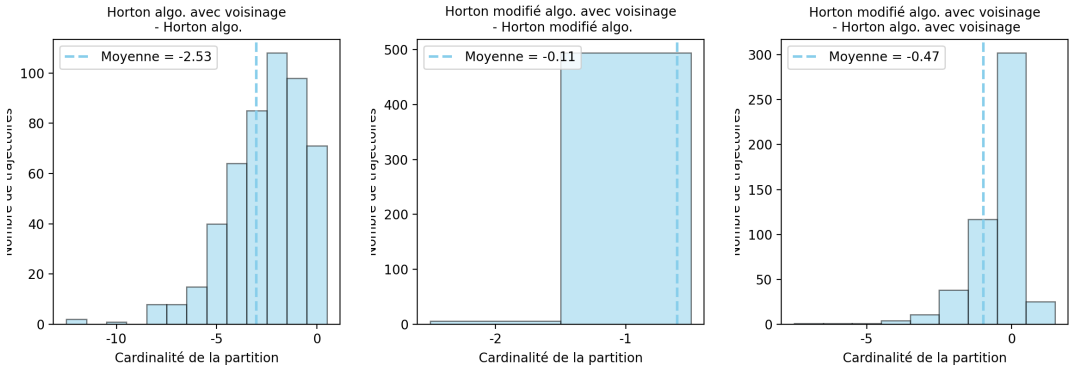


(a) Voisinage avec HD - HD

(b) Voisinage avec HDM - HDM

(c) Voisinage avec HDM - Voisinage avec HD

Figure 5.14 – Comparaisons de la distance avec la solution renvoyée par l'Algorithme 12 en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 100 trajectoires du jeu n°2.



(a) Voisinage avec HD - HD

(b) Voisinage avec HDM - HDM

(c) Voisinage avec HDM - Voisinage avec HD

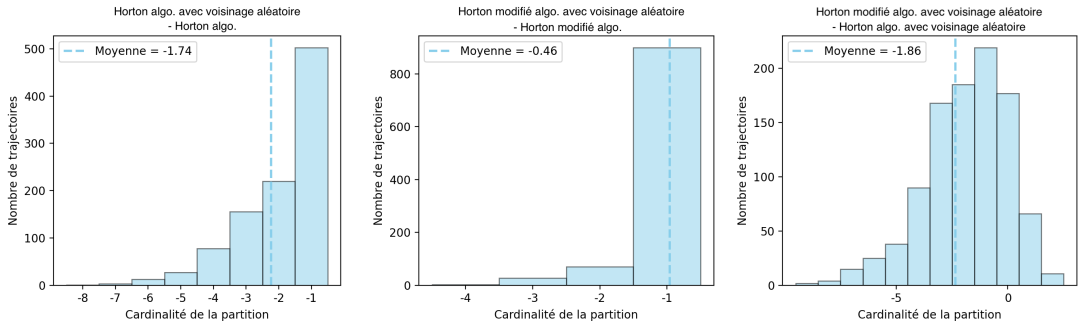
Figure 5.15 – Comparaisons de la distance avec la solution renvoyée par l'Algorithme 12 en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 500 trajectoires du jeu n°3.

Évaluation des meilleures solutions obtenues par une exploration aléatoire

Nous examinons maintenant si de meilleures solutions peuvent être atteintes en utilisant une exploration aléatoire du voisinage.

Pour ce faire, nous modifions légèrement l'Algorithme 12 (Chapitre 3, page 84) pour remplacer la méthode de sélection du couple induisant une nouvelle solution par une sélection aléatoire. Nous restreignons tout de même cette sélection aux couples de cycles pouvant induire une solution différente de la solution courante. Ensuite, nous appliquons l'heuristique de partitionnement pour évaluer la cardinalité de la partition en polycycles obtenue. Nous conservons et renvoyons la meilleure solution rencontrée au cours de l'exécution. Cette exploration prend fin après avoir effectué $K = 100$ itérations ou lorsqu'aucune nouvelle solution n'est accessible depuis la solution courante.

La Figure 5.16 présente les résultats de la comparaison entre la meilleure solution obtenue par l'exploration aléatoire et la solution obtenue par les méthodes directes seules : HD et HDM.



(a) Voisinage aléatoire avec HD - HD

(b) Voisinage aléatoire avec HDM - HDM

(c) Voisinage aléa. avec HDM - Voisinage aléa. avec HD

Figure 5.16 – Comparaison de la distance avec la meilleure solution parcourue aléatoirement en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 1000 trajectoires du jeu n°1.

La Figure 5.16a et la Figure 5.16b montrent que l'exploration aléatoire permet toujours d'améliorer la solution obtenue par la méthode directe. Les résultats de la Figure 5.16b montrent, quant-à eux, que dans la plupart des cas les solutions obtenues en sélectionnant au départ les bases de cycles calculées par HDM sont de cardinalité plus faible.

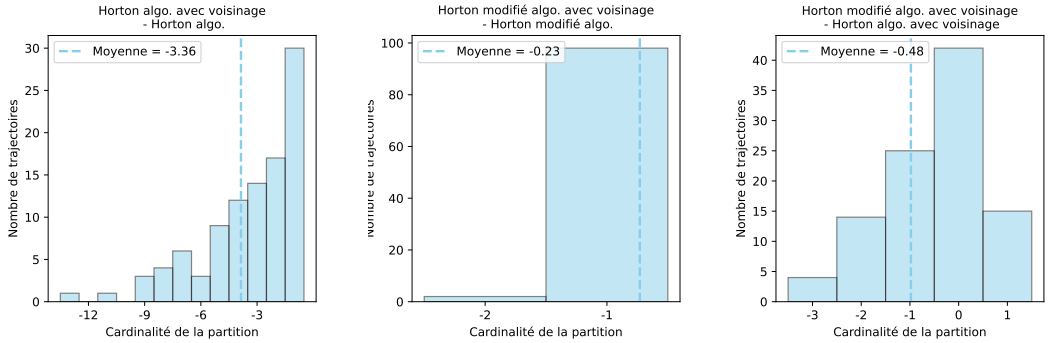
	Algorithme 12	Voisinage aléatoire
HD avec Vois. - HD	$-2,59$	$-1,74$
HDM avec Vois. - HDM	$-0,01$	$-0,46$
HDM avec Vois. - HD avec Vois.	$-0,57$	$-1,86$

Table 5.7 – Synthèse de la moyenne de la différence entre les cardinalités de la meilleure solution parcourue, en fonction de la méthode de sélection des bases pour le jeu n°1.

La Table 5.7 regroupe les résultats des Figures 5.11 et 5.16. Ainsi, nous observons qu'en partant des bases de Horton modifié, l'exploration aléatoire permet d'accéder à des solutions de meilleure cardinalité que celles obtenues par l'Algorithme 12. Notons, tout de même, que les améliorations observées sont très légères. La méthode de voisinage est peut-être encore améliorable mais elle semble déjà très efficace.

Nous observons maintenant les résultats de l'exploration aléatoire pour les deux autres jeux de données. La Figure 5.17 et la Figure 5.18 illustrent des résultats similaires à ceux du jeu de données n°1 de la Figure 5.16. Les Tables 5.8 et 5.9 regroupent les résultats obtenus pour la méthode de voisinage de l'Algorithme 12 et ceux de l'exploration aléatoire.

Nous observons que l'exploration aléatoire permet d'accéder à des solutions un peu meilleures que celles renvoyées par l'Algorithme 12 en partant des bases de Horton modifié. Nous observons dans le jeu n°3, tout comme c'était le cas dans le jeu n°1, que l'exploration aléatoire depuis les bases de Horton conduit en moyenne a des solutions de moins bonne cardinalité.



(a) Voisinage aléatoire avec HD - HD

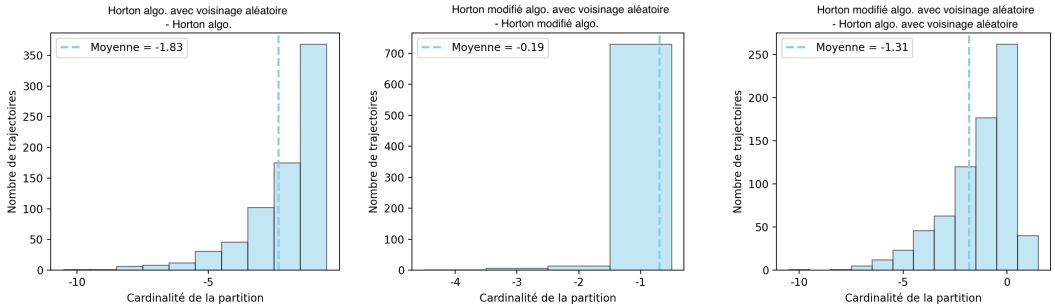
(b) Voisinage aléatoire avec HDM - HDM

(c) Voisinage aléa. avec HDM - Voisinage aléa. avec HD

Figure 5.17 – Comparaison de la distance de la meilleure solution parcourue en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 100 trajectoires du jeu n°2.

	Algorithme 12	Voisinage aléatoire
HD avec Vois. - HD	$-3,16$	$-3,36$
HDM avec Vois. - HDM	$-0,08$	$-0,23$
HDM avec Vois. - HD avec Vois.	$-0,53$	$-0,48$

Table 5.8 – Synthèse de la moyenne de la différence entre les cardinalités de la meilleure solution parcourue, en fonction de la méthode de sélection des bases pour le jeu n°2.



(a) Voisinage aléatoire avec HD - HD

(b) Voisinage aléatoire avec HDM - HDM

(c) Voisinage aléa. avec HDM - Voisinage aléa. avec HD

Figure 5.18 – Comparaison de la distance de la meilleure solution parcourue en fonction de l'ensemble des bases de cycles minimum de départ. Résultats établis sur les 500 trajectoires du jeu n°3.

	Algorithme 12	Voisinage aléatoire
HD avec Vois. - HD	$-2,53$	$-1,83$
HDM avec Vois. - HDM	$-0,11$	$-0,19$
HDM avec Vois. - HD avec Vois.	$-0,47$	$-1,31$

Table 5.9 – Synthèse de la moyenne de la différence entre les cardinalités de la meilleure solution parcourue, en fonction de la méthode de sélection des bases pour le jeu n°3.

Conclusion sur l'apport de la méthode de voisinage L'Algorithme 12 permet une légère amélioration de la cardinalité de la partition finale, par rapport à celle obtenue par la méthode HDM seule. Néanmoins, cette amélioration a un coût significatif en temps d'exécution. De plus, l'exploration aléatoire a montré l'existence de meilleures solutions qui n'ont pas été renvoyées par l'Algorithme 12. Pour autant, la solution renvoyée par la méthode HDM seule est toujours compétitive, car l'amélioration apportée par l'exploration du voisinage reste très légère, en moyenne la cardinalité de la solution est améliorée de moins de $0,5$ sur l'ensemble des trajectoires testées.

Nous n'avons pas réalisé une exploration exhaustive de toutes les solutions possibles. Ainsi, nos conclusions se basent uniquement sur les solutions qui nous étaient accessibles. Une étude approfondie des voisinages et de notre modèle est nécessaire pour tirer des conclusions définitives sur l'apport d'une méthode de voisinage pour améliorer la solution du partitionnement en polycycles. Cependant, le coût supplémentaire, combiné à, *a priori* la faible marge d'amélioration offerte par rapport à la méthode HDM, remet en

question l'apport de cette approche pour le calcul du polygraphe.

5.3 . Procédure de calcul du polygraphe d'une trajectoire

Ce chapitre a comparé les différentes méthodes proposées pour répondre aux étapes de la construction du polygraphe.

Dans la Section 5.1, nous avons établi que la méthode heuristique apportait une bonne solution au problème $min-$PCP, tout en étant très rapide. Cette méthode était comparée à la méthode exacte, particulièrement coûteuse en raison du grand nombre de solutions évaluées. Ainsi, la méthode gloutonne décrite par l'Algorithme 20 (page 109) est la méthode que nous sélectionnons pour cette étape.

Dans la Section 5.2, nous avons observé l'influence du choix des bases de cycles sur le résultat du partitionnement obtenu par l'Algorithme 20. Nous avons établi que la méthode de Horton modifié est efficace pour minimiser la cardinalité de la partition obtenue par l'heuristique. De plus, comme nous l'avons vu avec la méthode de voisinage, cette méthode peut-être améliorée par l'exploration du voisinage, néanmoins cette amélioration n'est pas très significative et nécessite beaucoup de temps de calcul supplémentaire. Ainsi, la méthode de Horton modifié seule telle que décrite par l'Algorithme 3 (page 56) est celle que nous sélectionnons pour cette étape.

Étant donné une trajectoire $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, rappelons que chaque graphe de $\mathcal{G}$ est décrit comme $G_i = (V, E \cup H_i)$ pour tout $1 \leq i \leq N$, et que H désigne l'ensemble des liaisons variables de $\mathcal{G}$, soit $H = \bigcup_{i=1}^{N} H_i$. Le polygraphe GP de $\mathcal{G}$ est calculé comme suit :

1. **Calcul des bases de cycles minimum** : L'ensemble des bases de cycles minimum $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$ est obtenu en calculant les B_i avec $1 \leq i \leq N$ par l'Algorithme de Horton modifié (Algorithme 3, page 56), avec G_i comme entrée.
2. **Sélection des cycles d'intérêt** : L'ensemble des cycles d'intérêt de la trajectoire $\mathcal{C} = \bigcup_{i=1}^{N} B_i^*$ est défini tel que, avec pour tout $1 \leq i \leq N$, $B_i^* \subseteq B_i$ et B_i^* contient tous les cycles de B_i qui utilisent au moins une liaison de H.
3. **Construction des graphes de cycles minimum** : L'ensemble des graphes de cycles minimum $\mathcal{GC} = \{GC_1, GC_2, \ldots, GC_N\}$ tel que $\forall i$ avec $1 \leq i \leq N$, $GC_i = (B_i^*, E_{B_i^*})$, où pour $c_1, c_2 \in B_i^*$, nous avons $[c_1, c_2] \in E_{B_i^*}$ si et seulement si $c_1 \cap c_2 \neq \emptyset$.
4. **Partition en polycycles** : La partition en polycycles $\mathcal{P} \in < \mathcal{G}, \mathcal{GC}, H >$ est obtenue par l'heuristique de partitionnement (Algorithme 20, page 109).
5. **Construction du polygraphe** : Le polygraphe $GP = (\mathcal{P}, I)$ est défini tel que pour tout $p_1, p_2 \in \mathcal{P}$, $[p_1, p_2] \in I$ si et seulement si il existe $GC \in \mathcal{GC}$ avec $c_1, c_2 \in GC$ tel que $c_1 \in p_1$, $c_2 \in p_2$, et $[c_1, c_2]$ est une arête de GC.

6 - Analyse et comparaison de trajectoires de dynamique moléculaire par le polygraphe

Dans ce chapitre, nous explorons les méthodes d'utilisation du polygraphe pour l'analyse des trajectoires de dynamique moléculaire. Le polygraphe est en soi un outil d'analyse des trajectoires. Par conséquent, sa composition et son évolution à travers les sous-polygraphes des conformations tout au long de la trajectoire fournissent déjà des informations intéressantes. Un aspect que nous étudions tout particulièrement dans la Section 6.2. De plus, lorsque les trajectoires représentent la même molécule, les polycycles et donc les polygraphes représentant ce système sont comparables. Ainsi, nous pouvons tirer des conclusions sur l'impact d'un paramètre sur le système si nous disposons des trajectoires appropriées. Par exemple, en calculant plusieurs trajectoires qui ne diffèrent que par la température de simulation, nous pouvons directement étudier l'influence de la température sur le polygraphe calculé et donc sur la trajectoire. Ce point sera observé en détail dans la Section 6.3, dans laquelle nous détaillons la méthode de comparaison des polygraphes et illustrons nos résultats sur un ensemble de trajectoires d'un même système moléculaire. Avant cela, dans la Section 6.1, nous présentons le processus d'obtention des différentes trajectoires sur lesquelles nous avons conduit nos expérimentations.

Remarque 22. *Une première version de nos résultats a été présentée dans [2]. Cependant, étant donné que la méthode a été approfondie, certains résultats ont évolué.*

6.1 . Données expérimentales

Dans cette section, nous détaillons la méthode avec laquelle les trajectoires que nous utilisons dans ce chapitre ont été obtenues. Comme nous l'avons expliqué dans le Chapitre 1, une trajectoire de dynamique moléculaire est, étant donné une molécule, une simulation du mouvement de ses atomes au cours du temps en fonction d'un environement fixe.

Nous nous concentrons ici sur des trajectoires en phase gazeuse impliquant chacune une seule petite molécule isolée. En phase gazeuse, l'environnement est réduit au minimum, ainsi nous observons l'influence de la température sur la molécule. Pour nos expérimentations, nous avons calculé plusieurs trajectoires de dynamique moléculaire *ab initio* et semi-empirique. Les simulations *ab initio* impliquent une résolution directe des équations de Schrödinger pour modéliser la dynamique électronique, tandis que les simulations semi-empiriques n'en utilisent qu'une approximation.

Nous considérons quatre systèmes moléculaires distincts :

- le peptide Z-Ala$_6-$COOH ($C_{26}H_{39}N_7O_8$), noté Z-Ala$_6$.
- le peptide Gramicidine$-$COOH ($C_{99}H_{140}N_{20}O_{17}$), noté Gramicidine.
- le peptide ECCA* ($C_{16}H_{24}N_5O_{13}S_2$), composé de quatre acides aminés : un acide glutamique (E), deux cystéine (C) et un alanine (A).
- le polysaccharide Chondroitin disulfate ($C_{14}H_{20}NO_{18}S_2$), noté CS2S4S.

Le choix de ces systèmes a été guidé par plusieurs considérations. Gramicidine est un peptide contenant un nombre d'atomes plus élevé que les autres systèmes sélectionnés.

De plus, sa dynamique induit des repliements qui génèrent des cycles voisins les uns des autres. Or, cela n'est pas favorable au polymorphisme tel que nous l'avons défini, car les cycles se découpent en plusieurs cycles présents simultanément. Nous n'avons donc pas une évolution de la forme du cycle pour définir un polycycle. De ce fait, le polygraphe n'est *a priori* pas un outil très adapté pour ce système. Un autre exemple est Z-Ala$_6$, un petit peptide flexible pour lequel nous avons pu calculer diverses trajectoires. Ce système est donc tout indiqué pour la Section 6.3.

Nos collaborateurs du laboratoire LAMBE ont réalisé des simulations de dynamique moléculaire *ab initio* (AIMD) sur Z-Ala$_6$ et Gramicidine, ainsi que des simulations semi-empiriques (SEMD) sur Z-Ala$_6$, CS2S4S et ECCA*.

Les simulations AIMD utilisent une représentation électronique basée sur la théorie de fonctionnelle [1] de densité (DFT). La DFT est une méthode de calcul quantique permettant l'étude de la structure électronique. Ces simulations ont été réalisées avec des pas de temps de 0.4 femtosecondes pour Gramicidine et 0.5 femtosecondes pour Z-Ala$_6$.

Les simulations de dynamique semi-empirique ont été effectuées aux niveaux de théorie AM1 [13] et PM6 [48]. Ces méthodes semi-empiriques sont basées sur des paramètres ajustés empiriquement et sont couramment utilisées pour prédire les propriétés des molécules lorsque les calculs *ab initio* complets sont trop coûteux en termes de ressources de calcul. L'AM1 est particulièrement adaptée aux systèmes de taille moyenne et peut fournir des résultats précis pour les molécules organiques, tandis que la PM6 repose sur une paramétrisation plus étendue, ce qui la rend généralement plus précise que les méthodes plus anciennes comme l'AM1, mais elle est plus coûteuse en termes de temps de calcul.

La Table 6.1 détaille les différentes trajectoires calculées pour chaque système moléculaire.

Dans les prochaines sections, nous détaillons les résultats que nous avons pu obtenir pour les différentes trajectoires de dynamique moléculaire en notre possession. En utilisant le polygraphe calculé par la procédure établie dans le Chapitre 5.

Remarque 23. *Dans le contexte de l'analyse de trajectoires, nous choisissons de considérer le sous-polygraphe composé des polycycles qui apparaissent pendant au moins 1% du temps total de la simulation. Cette approche nous permet d'éviter de prendre en compte des artefacts de la simulation au même titre que des événements marquants des changements significatifs.*

6.2 . Analyse de trajectoires de dynamique moléculaire

Cette section illustre notre méthode d'analyse et de caractérisation par le polygraphe et les polycycles à travers des exemples concrets. Nous introduisons ici des outils complémentaires développés autour du polygraphe afin d'enrichir notre analyse.

6.2.1 . Trajectoire semi-empirique du peptide ECCA*

1. Une fonctionnelle est une fonction qui prend des fonctions comme arguments et renvoie des scalaires.

molécule	simulation	température (K)
$Z\text{-}Ala_6$	DFT	150
		450
		600
	PM6	150
		450
		600
	AM1	150
		450
		600
Gramicidine	DFT	50
		150
		600
ECCA*	AM1	150
	PM6	50
		150
CS2S4S	AM1	150
		600
	PM6	150
		600

Table 6.1 – Détails des différentes trajectoires calculées.

Cette section présente l'analyse d'une trajectoire semi-empirique du peptide ECCA*. Pour observer les périodes d'apparition des différents polycycles tout au long de la trajectoire, nous définissons un outil appelé le *chronogramme des polycycles*. ECCA* est un peptide constitué de 60 atomes ($C_{16}H_{24}N_5O_{13}S_2$) dont une représentation 3D est affichée sous différents angles dans la Figure 6.1.

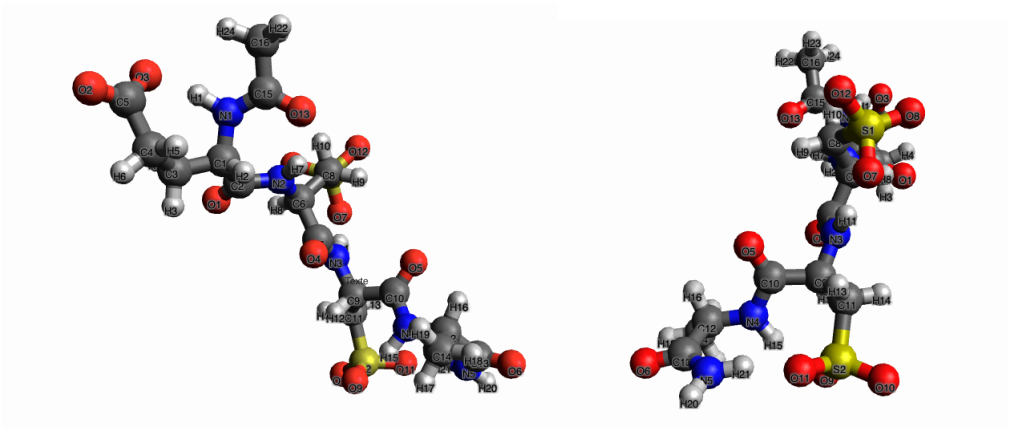


Figure 6.1 – Deux vues d'une représentation 3D de ECCA* ($C_{16}H_{24}N_5O_{13}S_2$). La couleur des atomes définit leur type : gris sombre pour le carbone, bleu pour l'azote, rouge pour l'oxygène, blanc pour l'hydrogène et jaune pour le soufre.

La trajectoire a été générée à une température d'environ 150 K, s'étalant sur une période d'observation d'environ 400 ps. Malgré cette longue durée, nous ne devrions pas

observer de nombreux changements conformationnels en raison de la faible température du système. L'analyse topologique [10, 11] des $1.032.433$ images composant cette trajectoire a révélé l'existence de 14 conformations différentes au cours de la simulation. En outre, la procédure a identifié un total de 5 liaisons hydrogène.

L'analyse des cycles a permis d'identifier un total de 6 cycles d'intérêt pour cette trajectoire. Le polygraphe final, représenté dans la Figure 6.2, a quant à lui réussi à regrouper deux de ces cycles pour définir 5 polycycles distincts. Cette réduction nous laisse penser que malgré la faible variabilité de la trajectoire, le polygraphe reste applicable. De plus, celui-ci est bien représentatif de la structure de la molécule étudiée. En effet, nous observons que plusieurs polycycles du polygraphe final ne sont pas connectés, ce qui est cohérent avec la structure du backbone du peptide. En effet, les polycycles n°0 et n°1 sont positionnés d'un côté du backbone, tandis que les polycycles n°3 et n°4 sont de l'autre côté. Enfin, le polycycle n°2 est central mais ne partage aucune arête avec les autres polycycles.

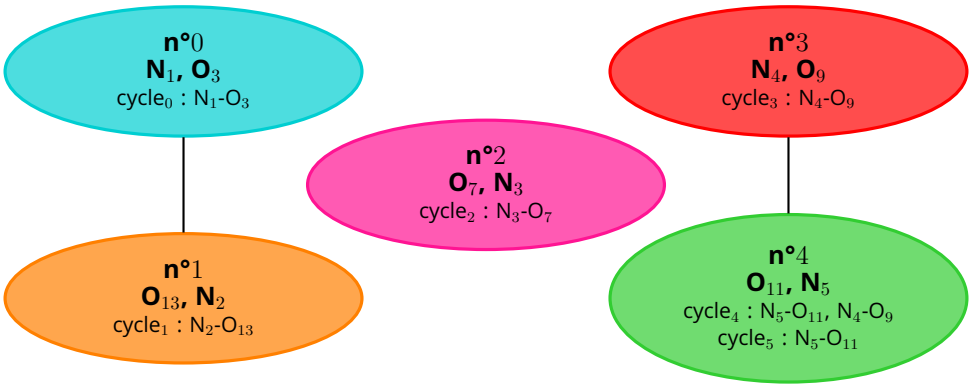


Figure 6.2 – Le polygraphe final obtenu pour la trajectoire PM6 de ECCA* à 150 K. Chaque sommet est étiqueté par un numéro, la liste des atomes communs à tous les cycles du polycycle et participant à des liaisons hydrogènes, ainsi que la liste des cycles inclus.

Le chronogramme des polycycles La Figure 6.3 représente les périodes d'apparition des cinq polycycles dans un graphique appelé *chronogramme*. Dans ce graphique, l'abscisse correspond au temps et l'ordonnée correspond aux polycycles. Un point apparaît à la coordonnée (x, y) si l'une des instances du polycycle y apparaît au temps x. Pour les trajectoires longues, comme c'est le cas ici, nous avons traité les périodes d'apparition selon la procédure suivante pour plus de fluidité :

- Les périodes d'apparition trop courtes sont supprimées. Ces périodes correspondent à des périodes dont la longueur est inférieure à 1000 snapshots, ce qui représente 400 fs sur la trajectoire de ~ 400 ps.
- Les périodes d'apparition consécutives, séparées par des périodes courtes d'absence, sont fusionnées. Il s'agit de périodes d'absence de moins de 500 snapshots.

Remarque 24. *Les seuils proposés ici ont donné de bons résultats sur les trajectoires que nous avons considérées. Néanmoins, ces valeurs doivent être adaptées en fonction de la longueur de la trajectoire donnée et de la précision des résultats attendus.*

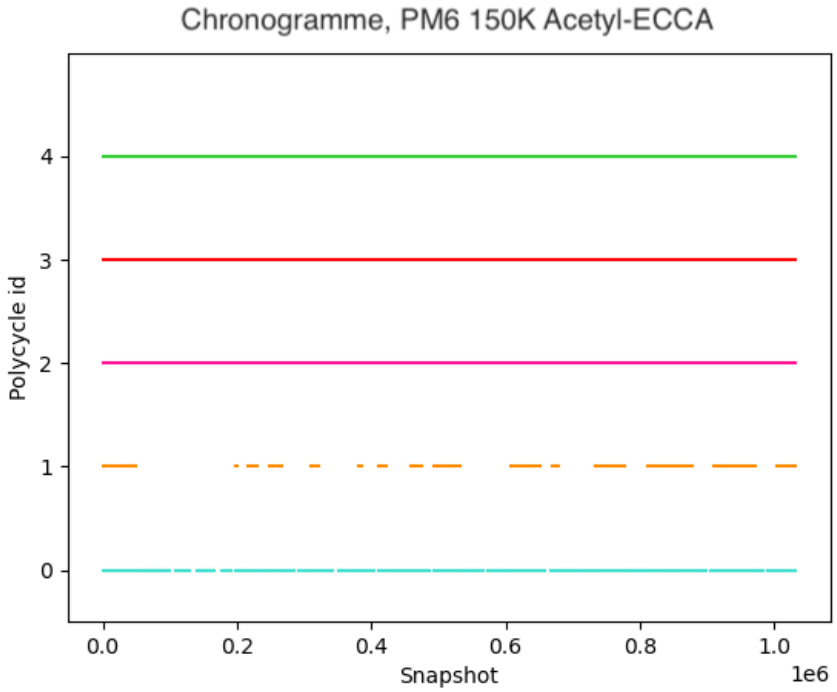


Figure 6.3 – Chronogramme des apparitions des polycycle au cours de la trajectoire PM6 de ECCA* à 150 K. Les noms et les couleurs des polycycles correspondent à ceux utilisés dans le polygraphe de la Figure 6.2.

La Figure 6.3 représente le chronogramme obtenu pour la trajectoire du peptide. Chaque polycycle est étiqueté selon le numéro qu'il a dans le polygraphe illustré dans la Figure 6.2.

À partir du chronogramme, nous observons que la structure de la molécule est très stable tout au long de la simulation. Globalement, malgré les 14 conformations différentes, tous les polycycles sont toujours présents. Le moins stable est le polycycle n°1 qui présente de courtes périodes d'absence. Néanmoins, les autres polycycles sont toujours présents et nous n'observons pas de transitions vers une période marquée par l'absence de ce polycycle. Nous concluons donc que la méta-structure de la molécule est stable tout au long de la trajectoire et est directement représentée par le polygraphe.

6.2.2 . Trajectoire empirique du peptide Z-Ala$_6$

Cette section présente l'analyse d'une trajectoire empirique du peptide Z-Ala$_6$. Dans cet exemple, la dynamique du système est telle que nous proposons une estimation des bassins parcourus à partir du chronogramme des polycyles. En effet, dans l'exemple précedent nous ne pouvons définir qu'un seul bassin, ici en revanche, nous observons plusieurs bassins différents. Z-Ala$_6$ est un peptide constitué de 80 atomes ($C_{26}H_{39}N_7O_8$) dont une représentation 3D est affichée sous différents angles dans la Figure 6.4.

Nous examinons une trajectoire à environ 450 K, d'une durée relativement courte d'environ 6 picosecondes. L'analyse topologique des 12294 images constituant cette trajectoire a révélé la présence de 93 conformations différentes au cours de la simulation. De plus, seulement 9 liaisons hydrogène ont été identifiées parmi ces 93 conformations. Il est

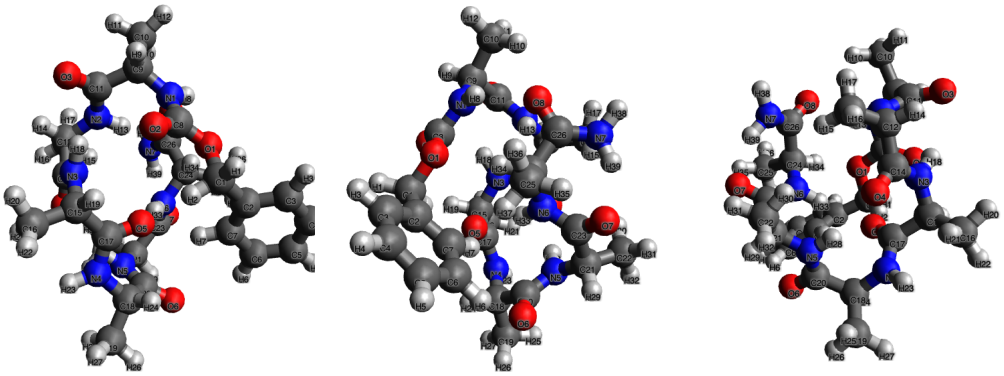


Figure 6.4 – Trois vues d'une représentation 3D de Z-Ala$_6$ ($C_{26}H_{39}N_7O_8$). La couleur des atomes définit leur type : gris sombre pour le carbone, bleu pour l'azote, rouge pour l'oxygène, blanc pour l'hydrogène et jaune pour le soufre.

également important de noter que ces conformations contiennent au plus 6 liaisons hydrogène simultanément, avec une moyenne d'environ $3,3$ liaisons par conformation. Ces premiers résultats indiquent une molécule dont la structure reste relativement complexe tout au long de la trajectoire malgré la température élevée de simulation.

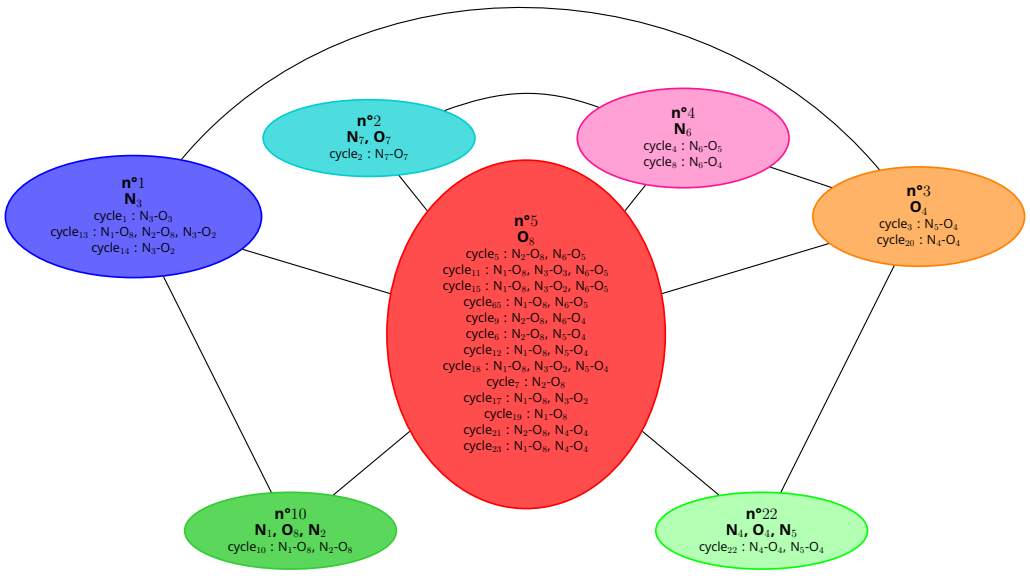


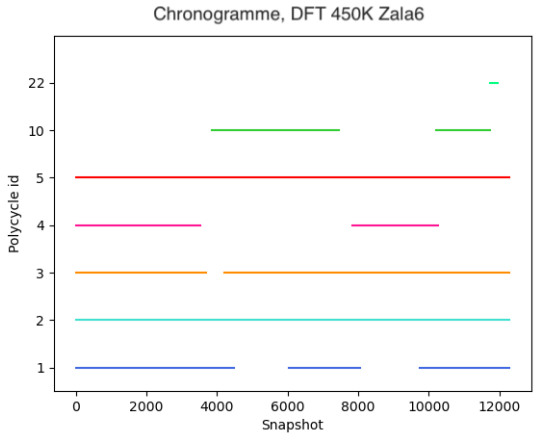
Figure 6.5 – Le polygraphe final obtenu pour la trajectoire DFT de Z-Ala$_6$ à 450 K. Chaque sommet est étiqueté par un numéro, la liste des atomes communs à tous les cycles du polycycle et participant à des liaisons hydrogènes, ainsi que la liste des cycles inclus.

L'analyse des cycles a révélé un total de 23 cycles d'intérêt pour cette trajectoire. Le polygraphe final, illustré dans la Figure 6.5, a réussi à regrouper ces cycles en 7 polycycles distincts. De nombreux regroupements ont été effectués, et nous observons la présence d'un polycycle central dans le polygraphe, qui est adjacent à tous les autres polycycles et qui compte 13 identités. Ce polycycle central semble jouer un rôle crucial dans la structure

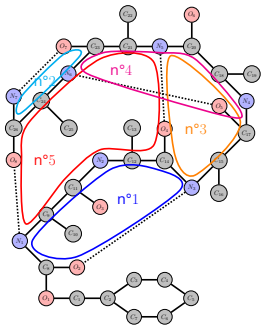
moléculaire. Il est également important de noter que le polygraphe final forme un graphe connexe, ce qui signifie que tous ces polycycles coexistent et partagent des liaisons dans les conformations observées.

La Figure 6.6a présente le chronogramme de la trajectoire. Nous pouvons observer que le polycycle n°5, central dans le polygraphe, est présent tout au long de la trajectoire. De même, le polycycle n°2 et, dans une certaine mesure, le polycycle n°3 (bien qu'il soit absent pendant quelques images) font partie intégrante de la structure de Z-Ala$_6$. La Figure 6.6b met en évidence la coexistence de ces trois polycycles dans le graphe de la molécule.

De plus, nous constatons que le polycycle n°4 ne coexiste jamais avec le polycycle n°10. Cette observation suggère une incompatibilité structurelle entre ces deux polycycles. Chacun est donc associé à une ou plusieurs structures différentes. De même, le polycycle n°22 ne coexiste ni avec le polycycle n°4, ni avec le polycycle n°10. Cette analyse nous permet d'identifier plusieurs structures distinctes dans la trajectoire.



(a) Chronogramme de la trajectoire



(b) Exemple illustrant la coexistence de cinq polycycles dans le graphe moléculaire.

Figure 6.6 – Chronogramme des apparitions des polycycle au cours de la trajectoire DFT de Z-Ala$_6$ à 450 K. Le graphe de droite illustre la coexistence des polycycles n°2, n°3 et n°5 présent tout au long de la trajectoire. Les noms et les couleurs des polycycles correspondent à ceux utilisés dans le polygraphe de la Figure 6.5.

À partir de ces observations, nous proposons de découper le chronogramme en cinq périodes, comme illustré dans la Figure 6.7. Nous considérons que les polycycles d'une période sont ceux qui sont majoritairement présents au cours de celle-ci, même s'ils ne le sont pas en continu. Ainsi, nous avons défini les périodes suivantes :

- Période 1 : ~ 3900 snapshots ; polycycles n°1, n°2, n°3, n°4 et n°5.
- Période 2 : ~ 4000 snapshots ; polycycles n°1, n°2, n°3, n°5 et n°10.
- Période 3 : ~ 2300 snapshots ; polycycles n°2, n°3, n°4 et n°5.
- Période 4 : ~ 1600 snapshots ; polycycles n°1, n°2, n°3, n°5 et n°10.
- Période 5 : ~ 1200 snapshots ; polycycles n°1, n°2, n°3, n°5 et n°22.

Les périodes 2 et 4 sont donc considérées comme identiques. En revanche, la période 1 et la période 3 se distinguent par l'absence du polycycle n°1 dans la seconde. Les différences principales entre les périodes résident dans les interruptions dues aux polycycles n°4,

n°10 et n°22. Chaque période ainsi identifiée pourrait éventuellement correspondre à un bassin conformationnel de la surface d'énergie potentielle. Pour en avoir la certitude, il est nécessaire de revenir à l'énergie potentielle de la molécule sur ces périodes. Il est toutefois possible que certaines de ces périodes appartiennent au même bassin conformationnel, mais ces périodes restent un bon point de départ pour l'étude des différentes structures de la molécule.

Dans la Figure 6.8, chaque sous-polygraphe représente une méta-structure de Z-Ala$_6$ au cours de la période correspondante. Nous utilisons le terme "méta-structure" car chaque période peut contenir potentiellement plusieurs structures qui correspondent au sous-polygraphe proposé ou à un sous-graphe de celui-ci. Ces sous-polygraphes nous donnent un aperçu des contraintes structurelles dominantes pendant chaque période, nous permettant ainsi de visualiser les changements majeurs.

De plus, les conformations associées aux *minima* d'énergie sont souvent caractérisées par un nombre maximal de liaisons hydrogène. Ces liaisons jouent un rôle crucial dans la stabilisation de la structure moléculaire. Par conséquent, bien que la période puisse inclure d'autres structures moins contraignantes que celles représentées par le sous-polygraphe, il est probable que ce dernier reflète la structure associée à un minimum local d'énergie.

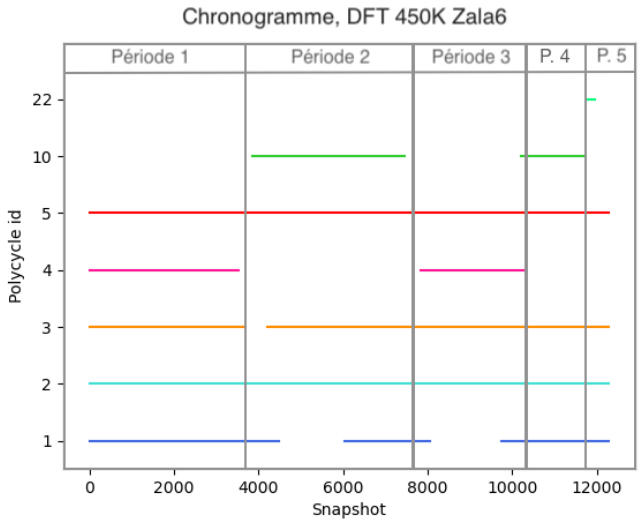
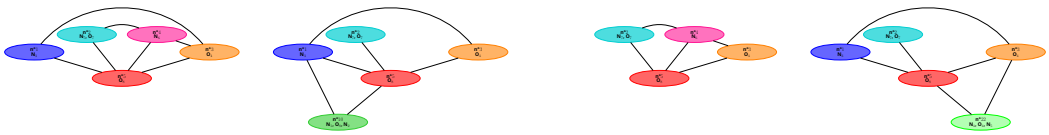


Figure 6.7 – Proposition de découpage de la trajectoire en cinq périodes.



(a) Période 1. (b) Périodes 2 et 4. (c) Période 3. (d) Période 5.

Figure 6.8 – Illustration des sous-polygraphes associés à chacune des périodes identifiées dans la Figure 6.7.

Enfin, notons que les périodes 3 et 4 sont bien plus courtes que leurs prédécesseurs.

Cette observation suggère que l'agitation moléculaire a probablement favorisé un changement conformationnel significatif, conduisant ainsi à la période 5. Cependant, pour étayer cette hypothèse, nous avons besoin de davantage d'informations sur la dynamique de la molécule. Dans la Section 6.3.2, nous comparerons plusieurs trajectoires de Z-Ala$_6$. Ces comparaisons nous permettront d'identifier les éléments structuraux conservés entre les trajectoires, nous permettant ainsi de croiser ces différents résultats pour mieux comprendre la dynamique de la molécule.

Dans le cas présent, cette comparaison nous apprendra que dans une trajectoire de température plus basse (environ 150 K), les polycycles n°4 et n°10 coexistent. Par conséquent, nous pouvons raisonnablement déduire que les périodes 1, 2, 3 et 4 correspondent au même bassin conformationnel. En revanche, le polycycle n°22 est spécifique à cette trajectoire et appartient donc probablement à un bassin différent.

6.2.3 . Trajectoire semi-empirique du polysaccharide Chondroitine disulfate

Cette section présente l'analyse d'une trajectoire semi-empirique PM6 du peptide du polysaccharide CS2S4S. Cette dynamique présente des caractéristiques complexes, avec de nombreux éléments coexistants à différents moments de la trajectoire. Nous examinerons également l'impact de notre seuil de 1% pour la prise en compte des polycycles dans le polygraphe final, et explorerons les conséquences d'un seuil plus restrictif sur cette analyse.

Ce polysaccharide est composé de 55 atomes ($C_{14}H_{20}NO_{18}S_2$), et sa représentation 3D est visualisée sous différents angles dans la Figure 6.9.

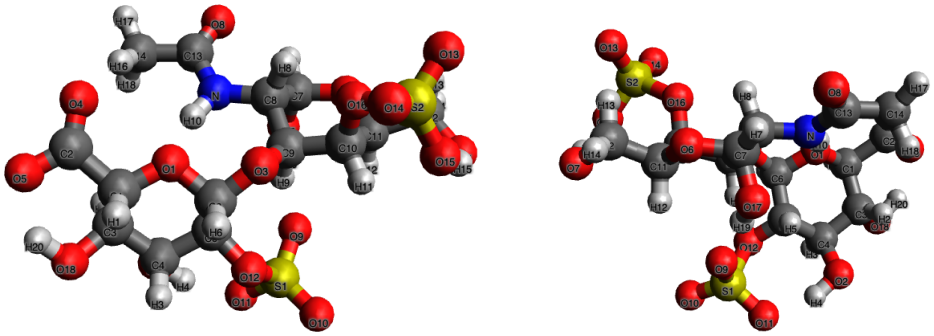


Figure 6.9 – Deux vues d'une représentation 3D de la CS2S4S ($C_{14}\ H_{20}\ N\ O_{18}\ S_2$). La couleur des atomes définit leur type : gris sombre pour le carbone, bleu pour l'azote, rouge pour l'oxygène, blanc pour l'hydrogèneet jaune pour le soufre.

La trajectoire a été générée à une température d'environ 600 K, dans le but de provoquer de nombreux changements conformationnels sur la durée observée d'environ $300,0$ picosecondes. L'analyse topologique des 747135 images composant cette trajectoire a révélé l'existence de 441 conformations différentes au cours de la simulation. En outre, la procédure a identifié un total de 32 liaisons hydrogène. Cette grande variabilité, tant au niveau des liaisons hydrogène que des conformations, suggère que CS2S4S est un peptide très flexible.

L'analyse des cycles a identifié un total de 41 cycles, qui ont ensuite été regroupés en 16 polycycles. Cependant, dans la Figure 6.10, qui représente le polygraphe caractéristique de la trajectoire, seuls 14 polycycles sont présents. Cela s'explique par le fait que

le polygraphe final est limité aux polycycles présents pendant au moins 1% du temps total de la trajectoire. Cette observation montre à quel point certains cycles peuvent être sporadiques.

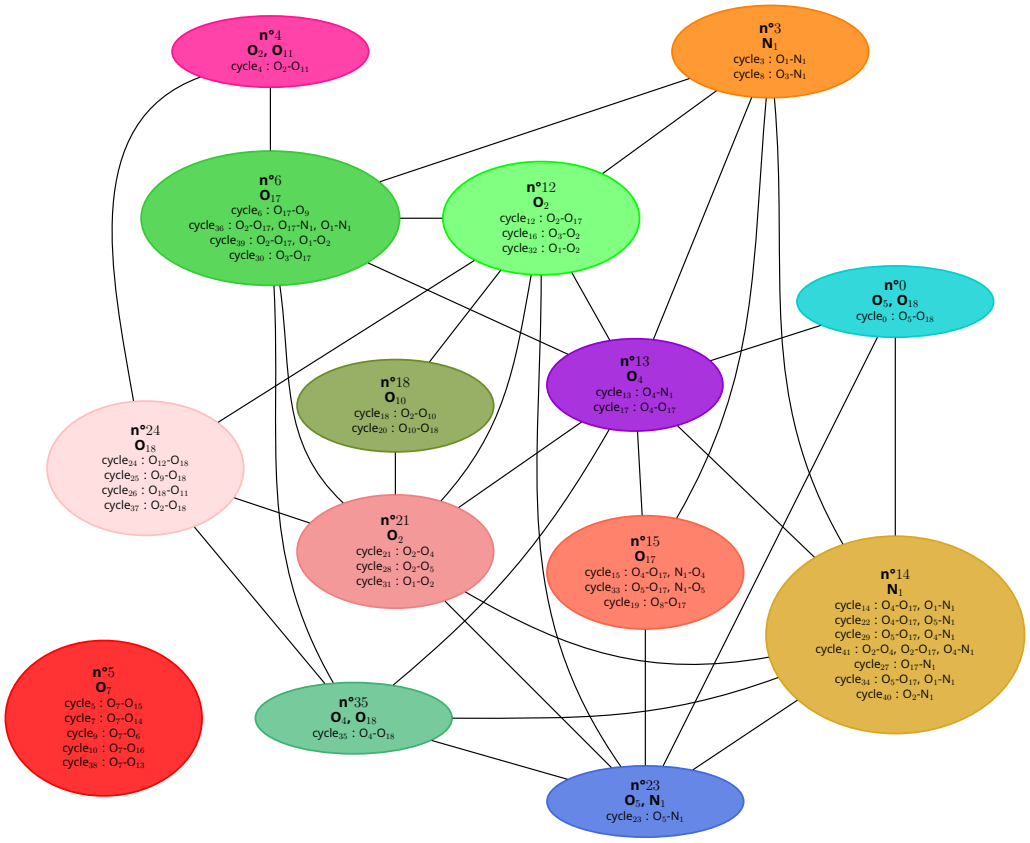


Figure 6.10 – Le polygraphe final obtenu pour la trajectoire PM6 de CS2S4S à 600 K. Chaque sommet est étiqueté par un numéro, la liste des atomes communs à tous les cycles du polycycle et participant à des liaisons hydrogènes, ainsi que la liste des cycles inclus.

Dans ce polygraphe de la trajectoire, nous observons que les polycycles contiennent principalement plusieurs cycles. Il semble donc qu'il y ait peu de cycles suffisamment spécifiques pour ne pas avoir au moins un équivalent au cours de la trajectoire.

Nous notons également que les cycles sont très dynamiques, ce qui rend le chronogramme assez difficile à lire même avec notre crible de 1%.

Pour mieux comprendre l'impact de ce crible sur l'analyse, les Figures 6.11, 6.12, 6.13, 6.14, et 6.15 (page 147 à 150) présentent le polygraphe et le chronogramme pour des seuils minimum d'apparition compris entre 0% et 10%. Dans ces figures, nous proposons une version synthétique du polygraphe. Dans cette représentation, la liste des cycles composant chaque polycycle n'est pas incluse. Chaque sommet est désigné par "Poly$X$", où X représente son numéro, la liste des atomes impliqués dans les liaisons hydrogène partagées par tous les cycles de l'ensemble, et le nombre de cycles de l'ensemble. De plus, la taille de chaque sommet est proportionnelle au nombre d'identités qu'il contient.

Dans la Figure 6.11, nous observons les résultats lorsque aucun crible n'est appliqué, donc tous les polycycles sont inclus dans l'analyse. Le chronogramme associé présente quant à lui de nombreuses lignes avec des apparitions ponctuelles, et aucun découpage ne semble ressortir à première vue de ces événements.

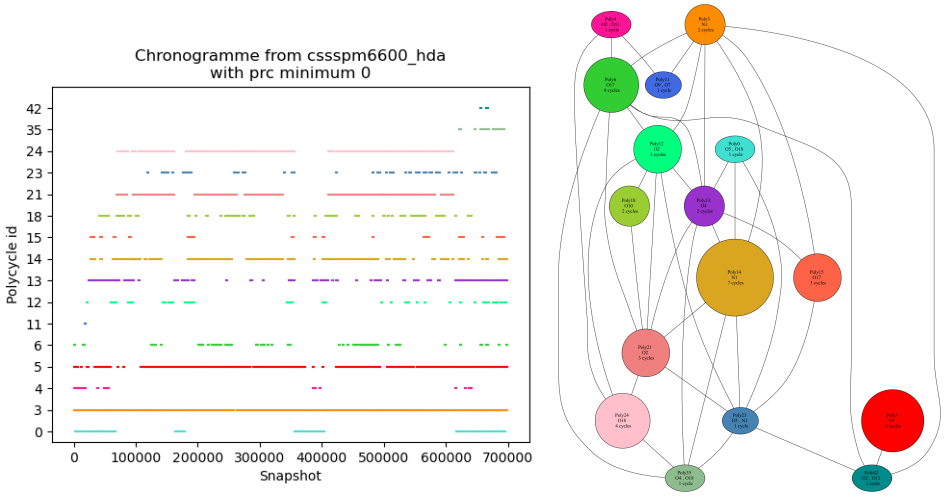


Figure 6.11 – Polygraphe et chronogramme dont le seuil d'apparition des polycycles est au moins de 0%.

Dans la Figure 6.12, le crible de 1% est appliqué, ce qui entraîne la suppression de deux polycycles par rapport à ceux calculés initialement. Les polycycles supprimés sont les numéros 11 et 42. Le chronogramme de la Figure 6.12 illustre clairement l'absence de ces polycycles dans la dynamique globale. Cette observation valide l'utilité de ce crible pour simplifier l'accès à un chronogramme plus pertinent.

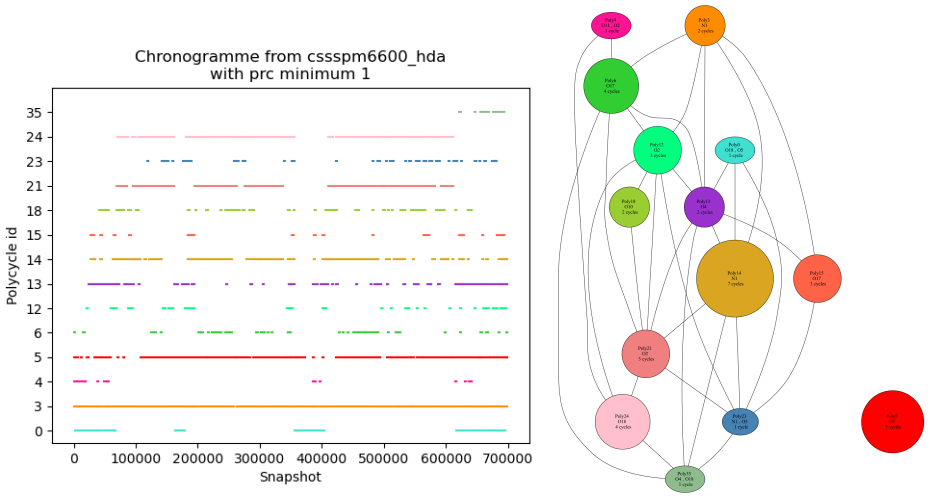


Figure 6.12 – Polygraphe et chronogramme dont le seuil d'apparition des polycycles est au moins de 1%.

Cependant, malgré l'application de ce crible, le chronogramme présenté dans la Figure 6.12 reste assez complexe à lire. Nous pouvons néanmoins noter que certains polycycles, tels que les numéros 0, 3, 5, 21 et 24, présentent des périodes d'apparition bien définies. Ces périodes semblent indiquer une incompatibilité entre le polycycle 0 et les polycycles 21 et 24, suggérant l'existence de différentes structures associées à ces polycycles. Cependant, le polycycle 0 est présent à de nombreuses périodes, ce qui indique des oscillations entre ces différentes structures.

De plus, les polycycles numéros 3 et 5 sont présents tout au long de la trajectoire ce qui nous suggère qu'ils jouent un rôle central. Cette observation associée aux oscillations observées notament sur le polycycle 0 indique que la dynamique correspond soit à un seul bassin conformationnel, soit à plusieurs bassins très proches énergétiquement, ce qui expliquerait des transitions fréquentes entre eux.

Dans les prochaines figures, en augmentant le crible, nous allons sélectionner moins de polycycles. Cela devrait éclaircir le paysage du chronogramme et nous permettre d'observer de façon plus évidente les différences entre les structures explorées.

Ainsi, dans la Figure 6.13, nous présentons les résultats avec un crible de 2%. Ce crible supprime un polycycle supplémentaire, réduisant ainsi le polygraphe à 13 sommets.

Ensuite, dans la Figure 6.14, nous observons l'impact d'un crible de 5%, ce qui réduit le polygraphe à 11 sommets. Bien que le chronogramme soit allégé, certaines apparitions ponctuelles de polycycles demeurent difficiles à associer aux éventuelles structures déterminées par les polycycles 0 et 21.

Enfin, la Figure 6.15 présente un crible de 10%, réduisant le polygraphe à seulement 8 polycycles. Ce polygraphe n'est donc constitué qu'avec la moitié des polycycles initiale-

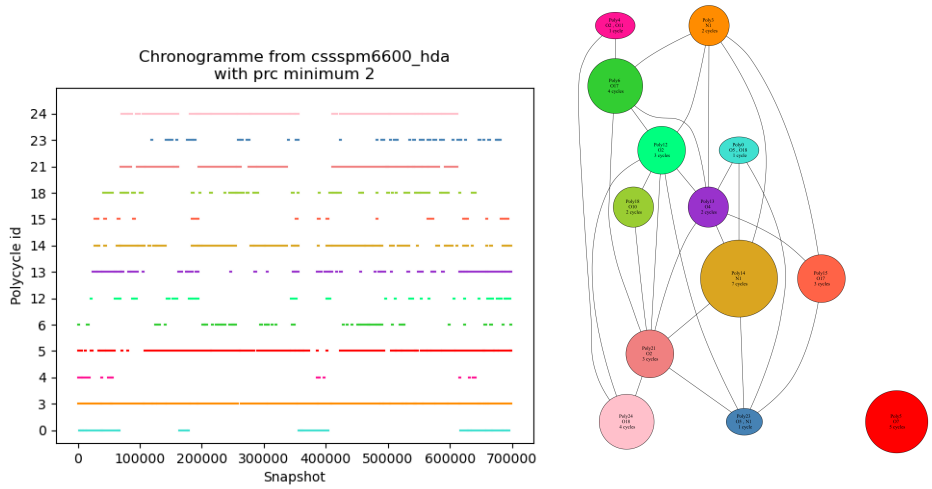


Figure 6.13 – Polygraphe et chronogramme dont le seuil d'apparition des polycycles est au moins de 2%.

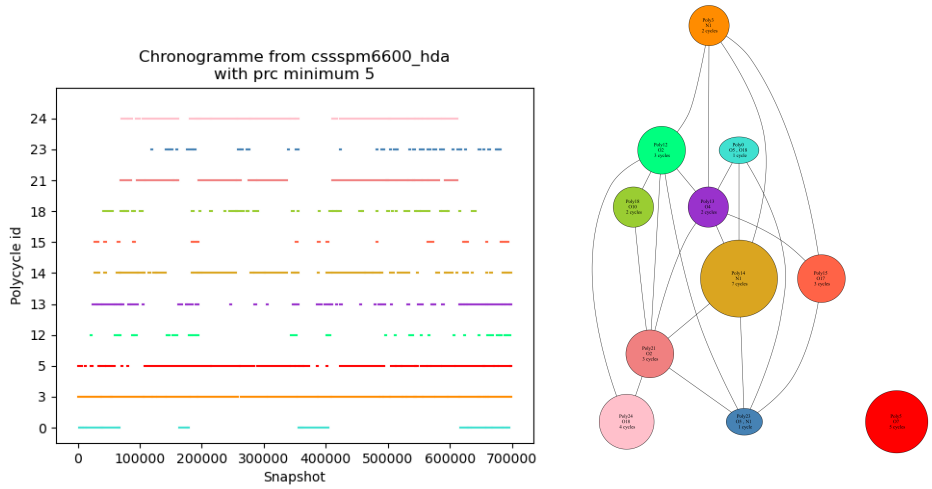


Figure 6.14 – Polygraphe et chronogramme dont le seuil d'apparition des polycycles est au moins de 5%.

ment calculés, ce qui souligne la volatilité de ces cycles et explique les difficultés initiales à déterminer des structures claires. Dans ce dernier chronogramme, nous observons l'incompatibilité *a priori* de seulement deux structures, celle utilisant le polycyle 0 et celle utilisant les polycycles 21 et 24. De plus, nous observons clairement la présence relativement stable des polycycles numéros 3, 5, 13 et 14 tout au long de la trajectoire. Ils forment donc un ensemble de polycycles au coeur de toutes les structures explorées par la dynamique.

Pour conclure, nous avons observé que la dynamique de CS2S4S est bien moins variée structurellement que nous l'attendions. Les cycles bougent et évoluent beaucoup, mais la molécule semble être restée dans le même bassin conformationnel, ou tout du moins dans la même famille de bassins. La structure de la molécule est globalement stable tout au long de la trajectoire comme le traduit la persistence de plusieurs polycycles même avec le seuil d'apparition d'au moins 10%.

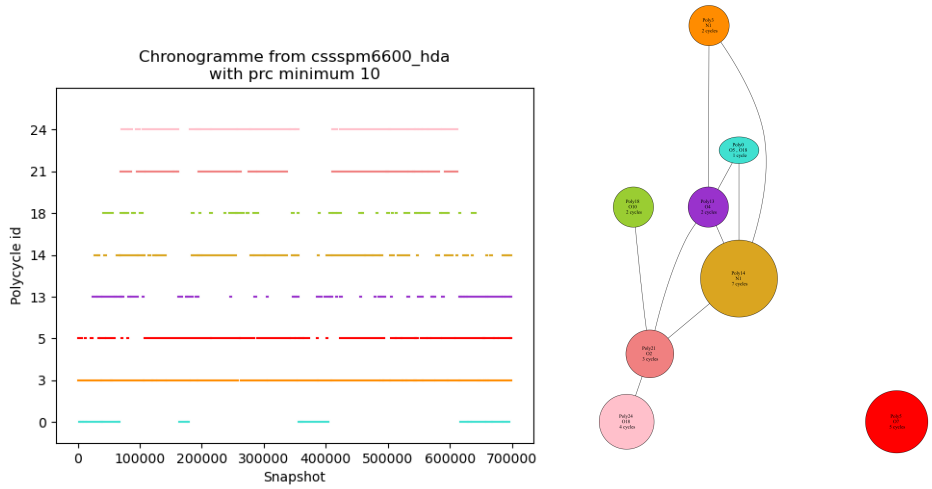


Figure 6.15 – Polygraphe et chronogramme dont le seuil d'apparition des polycycles est au moins de 10%.

6.3 . Comparaison de trajectoires de dynamique moléculaire

Cette section présente les résultats d'analyses impliquant la comparaison de plusieurs trajectoires de dynamique moléculaire. L'analyse d'une trajectoire seule ne nous fournit qu'une vue de la dynamique moléculaire dans un environement donné. Or, la surface d'énergie potentielle est très vaste, nous étudions donc souvent plusieurs trajectoires d'une même molécule obtenues avec différentes variables. Il est donc nécessaire de pouvoir comparer ces trajectoires pour identifier l'impact sur la structure moléculaire des variables utilisées et donc sur l'exploration de la surface d'énergie potentielle.

Pour cela, la Section 6.3.1 introduit une méthode afin d'établir la correspondance entre des polycycles provenant de différentes trajectoires représentant le même système moléculaire dans différentes conditions de simulation. Notons que cette comparaison est orientée car une des trajectoires est considérée comme référence. Une fois cette correspondance établie, les polygraphes caractéristiques des trajectoires deviennent comparables, tout comme les chronogrammes qui leur sont associés. Ainsi, nous pouvons étudier l'impact d'un paramètre de simulation sur la dynamique d'une molécule.

La Section 6.3.2 et la Section 6.3.3 présentent des exemples d'analyse.

6.3.1 . Correspondance entre polygraphes

Cette section explique comment établir la correspondance entre les polycycles de deux polygraphes issus de trajectoires différentes.

Tout d'abord, précisons que cette méthode n'est valide que si les polygraphes ont été obtenus pour le même système moléculaire. En effet, nous allons nous appuyer sur la composition des polycycles pour établir s'ils sont liés, et cela n'est possible que si les atomes sont les mêmes.

Nous établissons la correspondance entre les polycycles d'un des polygraphes et ceux calculés dans l'autre polygraphe. Ainsi, il s'agit d'une correspondance orientée où l'un des deux polygraphes est considéré comme le polygraphe de référence et l'autre est considéré comme le polygraphe "comparé". Pour chacun des polycycles du polygraphe comparé, nous recherchons à savoir s'il possède un correspondant dans le polygraphe de

référence et, le cas échéant, à identifier lequel ou éventuellement lesquels.

Considérons deux polygraphes $GP_1 = (\mathcal{P}_1, I_1)$ et $GP_2 = (\mathcal{P}_2, I_2)$ tels que GP_1 est le polygraphe de référence. Un polycycle $p_2 \in \mathcal{P}_2$ correspond à un polycycle $p_1 \in \mathcal{P}_1$ si et seulement si, l'une des propositions suivantes est vérifiée :

1. Une identité de p_1 apparaît dans p_2, soit $p_1 \cap p_2 \neq \emptyset$. Rappelons qu'une identité désigne un cycle appartenant à un polycycle.
2. Il existe un sommet communs aux cycles de p_1 également communs aux cycles de p_2, et tous les voisins de p_2 dans le polygraphe comparé ont des correspondants qui sont les voisins de p_1 dans le polygraphe de référence. Ce second point représente une similarité des voisinages de p_1 et p_2 qui s'appuie sur la même idée que les propriétés du polymorphisme (Chapitre 2). Ainsi, nous avons $(\bigcap_{c_1 \in p_1} V(c_1)) \cap (\bigcap_{c_2 \in p_2} V(c_2)) \neq \emptyset$ et $\forall p'_2 \in \Gamma_{\mathcal{P}_2}(p_2), \exists p'_1 \in \Gamma_{\mathcal{P}_1}(p_1)$ tel que $p'_1 \cap p'_2 \neq \emptyset$.

Remarque 25. *Pour plus de souplesse, il est envisageable de remplacer la correspondance des voisinages par une mesure de similarité des voisinages. Ainsi, si une similarité minimale est atteinte alors la correspondance est établie.*

Nous définissons le graphe biparti CORRESP(GP_1, GP_2) dans lequel $\mathcal{P}_1 \cup \mathcal{P}_2$ est l'ensemble des sommets et dans lequel il existe une arête $[p_1, p_2]$ avec $p_1 \in \mathcal{P}_1$ et $p_2 \in \mathcal{P}_2$ si et seulement si p_2 correspond à p_1.

Le poids d'une arête représente la ressemblance entre les polycycles des extremités. Ce poids est compris entre 0 et 1. Ainsi, pour les arêtes issues de la première proposition nous utilisons l'indice de Jaccard, une métrique largement reconnue en statistiques pour mesurer la similitude entre deux ensembles. Cet indice est défini comme le rapport entre la taille de l'intersection et la taille de l'union des ensembles comparés. Pour les arêtes issues de la seconde proposition leur poids est fixé à $0,01$.

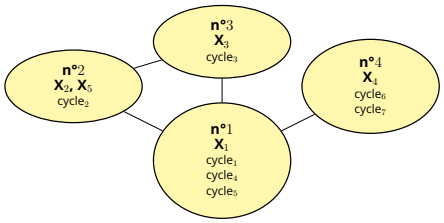
Ainsi, le problème qui nous intéresse est un problème de couverture des arêtes du graphe biparti CORRESP(GP_1, GP_2) tel que pour chaque sommet $p_2 \in \mathcal{P}_2$ ayant un degré supérieur ou égal à 1, exactement une arête adjacente à p_2 est couverte, et qui maximise la somme du poids des arêtes couvertes.

Remarque 26. *Expérimentalement, nous avons observé que plusieurs polycycles de $\mathcal{P}_2$ peuvent correspondre au même polycycle de $\mathcal{P}_1$. Ainsi, un sommet $p_1 \in \mathcal{P}_1$ peut avoir plusieurs arêtes incidentes dans la couverture que nous recherchons.*

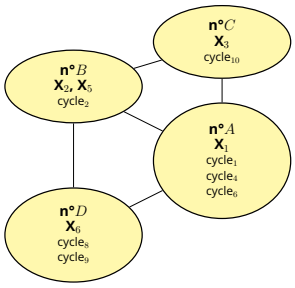
Résoudre ce problème prend un temps polynomial car il suffit de sélectionner pour chaque sommet de $p_2 \in \mathcal{P}_2$, l'arête de poids maximum lui étant adjacente. Néanmoins, cette complexité ne dépend pas que du nombre de sommets et d'arêtes de CORRESP(GP_1, GP_2) car nous utilisons des informations sur les cycles contenus dans les polycycles.

La Figure 6.16 illustre, à l'aide d'un exemple, les correspondances établies entre les polycycles de deux polygraphes. Les polycycles du polygraphe GP_1 sont numérotés de 1 à 4, tandis que ceux du polygraphe GP_2 sont annotés de A à D.

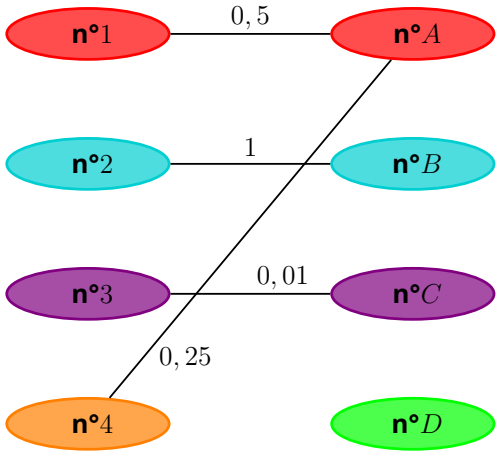
- Le polycycle A a deux arêtes qui correspondent à la première proposition. Le poids de ces arêtes est donné par l'indice de Jaccard. Ainsi, le poids de l'arête $[1, A]$ est égal à $\frac{|\{cycle_1, cycle_4\}|}{|\{cycle_1, cycle_4, cycle_5, cycle_6\}|} = \frac{2}{4}$, et le poids de l'arête $[4, A]$ est égal $\frac{1}{4}$. C'est donc l'arête $[1, A]$ qui appartient à la couverture maximale. Ainsi, le polycycle A correspond au polycycle 1.



(a) Polygraphe de réference GP_1



(b) Polygraphe comparé GP_2



(c) Graphe biparti CORRESP(GP_1, GP_2)

Figure 6.16 – Exemple de la correspondance entre les polycycles de deux polygraphes. Les polycycles dans les polygraphes GP_1 et GP_2 sont décrits par un identifiant, la liste des atomes partagés par tous les cycles du polycycles, et la liste des cycles qui composent le polycyle. Les cycles sont numérotés de telle sorte qu'ils portent le même numéro si et seulement si ils sont identiques.

- Le polycycle B a une seule arête adjacente de poids 1. En effet, les polycycles 2 et B sont parfaitement identiques.
- Le polycycle C a une seule arête adjacente de poids $0,01$ car elle est issue du voisinage similaire avec le polycycle 3 et non d'une correspondance directe entre ces polycycles. Cette correspondance indirecte est établie car le polycycle C est voisin des polycycles A et B, et qu'il partage des atomes avec le polycycle 3 comme l'atome labellisé X_3.
- Le polycycle D n'a aucune arête adjacente. Ce polycycle est voisin des polycycles A et B mais ne partage aucun atome avec le polycyle 3. Le polycycle D n'a donc aucun polycycle correspondant dans le polygraphe de référence.

Dans la suite, nous utilisons un code couleur pour illustrer les correspondances établies entre les polycycles provenant de différents polygraphes. Pour ce faire, chaque polycycle du polygraphe de référence est associé à une couleur spécifique. Ensuite, chaque polycycle du polygraphe comparé est coloré en fonction du polycycle auquel il correspond dans le polygraphe de référence. Si un polycycle du polygraphe comparé n'a pas de correspondant, il est coloré avec une couleur qui n'est pas utilisée dans le polygraphe de référence.

6.3.2 . Analyse des trajectoires de Z-Ala$_6$

Dans cette section, nous étudions plusieurs trajectoires du peptide Z-Ala$_6$. Nous disposons de cinq trajectoires, dont trois sont basées sur la méthode empirique DFT mais

à différentes températures (150K, 450K et 600K), et deux sont semi-empiriques basées sur les méthodes PM6 et AM1 à 450K. Toutes ces trajectoires partent de la même structure. Ainsi, dans la Section 6.3.2, nous analysons l'impact de la température sur les bassins explorés en examinant les trois trajectoires empiriques DFT. Tandis que dans la Section 6.3.2, nous examinons les variations dues à la méthode de simulation choisie en analysant les trois trajectoires à 450K. Pour rappel, Z-Ala$_6$ est un peptide constitué de 80 atomes ($C_{26}H_{39}N_7O_8$) dont nous avons déjà examiné en détail la trajectoire DFT à 450K dans la Section 6.2.2.

Pour établir ces correspondances, nous considérons comme polygraphe de référence celui de la trajectoire obtenue par DFT, ou lorsque la même méthode de simulation est utilisée, celle obtenue avec la température la plus basse.

Polygraphe de référence		**Polygraphe comparé**	
méthode	température	méthode	température
DFT	150K	DFT	450K
DFT	450K	DFT	600K
DFT	450K	PM6	450K
DFT	450K	AM1	450K

Table 6.2 – Liste des comparaisons de polygraphes effectuées pour Z-Ala$_6-$COOH

Afin de facilement comparer toutes ces trajectoires, nous utilisons les mêmes couleurs pour chaque polygraphe. En d'autres termes, le code couleur du polygraphe à 450K n'est pas réinitialisé pour la comparaison avec la trajectoire à 600K. De cette manière, nous pouvons observer grâce au code couleur que des polycycles présents à 150K sont toujours présents à 600K.

Plusieurs trajectoires de Z-Ala$_6$ à différentes températures

Cette section présente les résultats de l'analyse de trois trajectoires empiriques de Z-Ala$_6$ à des températures différentes. L'objectif de cette étude est de comprendre quels éléments sont conservés entre les différentes trajectoires et quelles sont les implications de ces observations sur la structure de la molécule.

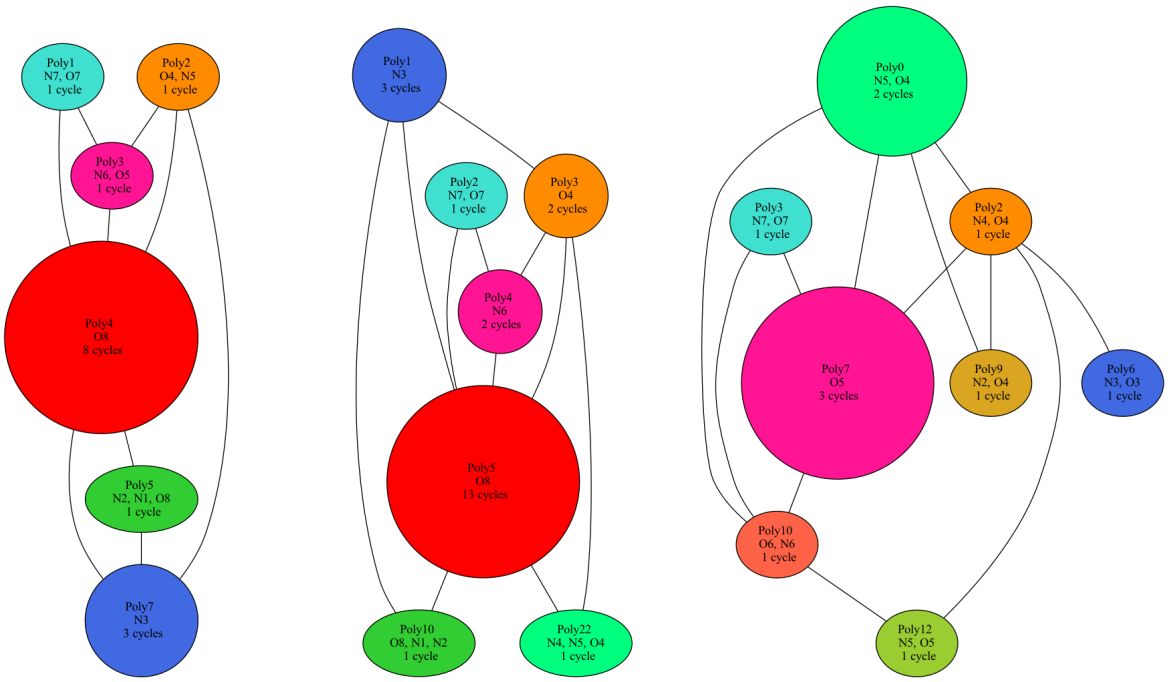
La Table 6.3 propose une synthèse des résultats de l'analyse par le polygraphe pour les trois trajectoires. Nous observons que la trajectoire à 450K est celle qui présente le plus de variabilité en termes de conformations explorées. Ce résultat est assez surprenant. En effet, étant donné que la température est un facteur favorable à l'agitation moléculaire, on pourrait s'attendre à ce que la trajectoire à 600K soit *a priori* la plus variée. La comparaison des polygraphes permettra de mieux comprendre les différences et les similitudes entre ces trajectoires.

La Figure 6.17 présente les polygraphes obtenus par l'analyse de ces trois trajectoires, et l'application de notre méthode de correspondance des polycycles.

Nous observons que le polygraphe de la trajectoire à 150K est un sous-graphe du polygraphe à 450K. Cela signifie que toutes les structures observées à 150K sont également observées à 450K. De plus, le polygraphe à 450K contient un polycycle absent de la trajectoire à 150K, indiquant une exploration plus étendue dans cette trajectoire. Dans ces deux trajectoires, à 150K et 450K, un polycycle central avec de nombreuses identités

température	liaisons hydrogène	conformations	cycles	polycycles
150K	7	36	15	6
450K	9	93	23	7
600K	10	53	12	9

Table 6.3 – Synthèse des résultats de l'analyse par le polygraphe des trois trajectoires empirique de Z-Ala$_6$. Le tableau présente pour chacune, le nombre de liaisons hydrogène, le nombre de conformations différentes, le nombre de cycles d'intérêts et enfin le nombre de polycycles finaux.



(a) à 150K (b) à 450K (c) à 600K

Figure 6.17 – Polygraphes des trois trajectoires DFT de Z-Ala$_6$ à des températures différentes. La coloration des polycycles indique si les polycycles se correspondent malgré les différences de températures.

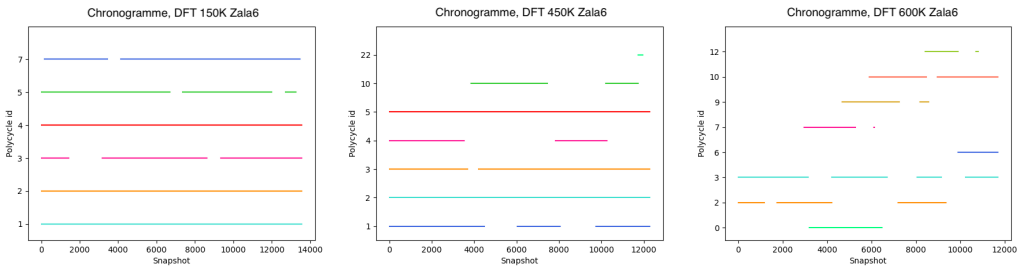
est présent. Ce polycycle constitue un repliement majeur de la molécule dans toutes les structures explorées par les deux trajectoires. Cependant, ce polycycle est absent de la trajectoire à 600K, traduisant ainsi un changement profond dans les conformations observées.

Le polycycle vert, correspondant au numéro 5 dans la trajectoire à 150K et au numéro 10 dans la trajectoire à 450K, est également absent du polygraphe à 600K. Ce polycycle représente un petit cycle caractéristique des structures à l'énergie la plus basse parmi celles que nous avons étudiées. Nous supposons alors que les conformations explorées par la trajectoire à 600K sont plus éloignées des minima d'énergie.

De plus, trois polycycles sont présents dans le polygraphe de la trajectoire à 600K et n'ont aucun correspondant dans le polygraphe de la trajectoire à 450K. Ainsi, il semble que le passage de la température de 450K à 600K marque un changement dans l'exploration de la surface d'énergie potentielle.

Ce changement est observable dans les chronogrammes de la Figure 6.18 qui présente les chronogrammes associés aux polygraphes de la Figure 6.17. Nous observons que tous les polycycles de la trajectoire à 150K sont présents dans la trajectoire à 450K. Comme nous l'avions mentionnée, dans la Section 6.2.2, pour conclure sur les différentes structures explorées par la trajectoire à 450K. Ainsi, les polycycles n°4 et n°10 coexistent dans le chronogramme de la trajectoire à 150K, alors que ce n'est pas le cas dans le chronogramme de la trajectoire à 450K. Cela nous laisse supposer que la structure moléculaire est moins stable à 450K qu'elle ne l'était à 150K. Nous nous attendons alors au même type de liens entre les chronogrammes à 450K et 600K, or ce n'est pas le cas. Les polycycles n°9, n°10 et n°12 sont spécifiques à la trajectoire à 600K, et correspondent dans le chronogramme aux trois lignes les plus hautes. Ces polycycles apparaissant dans la seconde moitié de la trajectoire, nous supposons qu'ils marquent l'exploration de nouveaux bassins sur la surface d'énergie potentielle.

De plus, les polycycles présents dans la première moitié de la trajectoire à 600K sont également présents à 450K. Ils correspondent donc *a priori* aux mêmes bassins conformationnels. Néanmoins, l'agitation moléculaire est telle que la plupart des conformations observées ne contiennent que deux ou trois cycles, ce qui ne permet pas réellement d'étudier les structures moléculaires. Ainsi, la température de 600K permet d'explorer la surface d'énergie, mais pour observer la topologie des structures moléculaires, il est préférable d'étudier des trajectoires obtenus à plus basse température avec des conformations de départ différentes. Le chronogramme à 600K permet alors d'identifier les périodes de changement, pour trouver des conformations de départ correspondant à des bassins conformationnels différents.



(a) à 150K (b) à 450K (c) à 600K

Figure 6.18 – Chronogrammes des trois trajectoires DFT de Z-Ala$_6$ à des températures différentes. La coloration des polycycles indique si les polycycles se correspondent malgré les différences de températures.

Pour conclure, la trajectoire à 150K montre une même structure moléculaire pendant toute sa durée. Cette structure est retrouvée dans la majeure partie de la trajectoire à 450K, mais de manière moins stable comme nous pouvons le voir sur le chronogramme. La fin de la trajectoire à 450K est caractérisée par l'apparition d'un nouveau polycycle, marquant ainsi la définition d'une nouvelle structure. Enfin, toutes ces structures sont explorées dans la première moitié de la trajectoire à 600K, mais de manière très instable, ce qui se traduit par des conformations qui ne représentent que partiellement ces structures. La seconde moitié de la trajectoire présente plusieurs nouveaux polycycles, indiquant ainsi que de nouveaux bassins ont pu être atteints par cette trajectoire.

Plusieurs trajectoires de Z-Ala$_6$ à une même température

Cette section présente les résultats de l'analyse de trois trajectoires à une même température de Z-Ala$_6$, obtenues avec des méthodes différentes. L'objectif de cette étude est d'observer l'impact de la méthode de simulation sur l'estimation de la surface d'énergie potentielle.

La Figure 6.4 propose une synthèse des résultats de l'analyse par le polygraphe pour les trois trajectoires. Les trajectoires semi-empiriques sont moins coûteuses que la trajectoire empirique, ce qui permet de les étendre sur une période beaucoup plus longue. En effet, tandis que la trajectoire DFT comprend seulement environ 12300 images, les trajectoires semi-empiriques en comprennent plus de 900000. Cette différence rend difficile la comparaison des résultats, car il est évident que l'espace exploré par les trajectoires semi-empiriques est bien plus vaste. Ainsi, nous considérons également une version tronquée des trajectoires semi-empiriques afin d'observer, les trois trajectoires sur une période d'exploration similaire.

méthode	longueur	liaisons hydrogène	conformations	cycles	polycycles
DFT	12294	9	93	23	7
PM6	1000000	23	690	89	20
PM6 tronquée	13044	13	174	39	12
AM1	914512	42	769	145	26
AM1 tronquée	13009	13	160	47	13

Table 6.4 – Synthèse des résultats de l'analyse par le polygraphe des trois trajectoires à 450K de Z-Ala$_6$. Les résultats pour les trajectoires semi-empiriques sont présentés à la fois en considérant la trajectoire dans son ensemble, et en considérant une version tronquée des premières images. Le tableau présente pour chacune, le nombre d'images analysées, le nombre de liaisons hydrogène, le nombre de conformations différentes, le nombre de cycles d'intérêts et enfin le nombre de polycycles finaux.

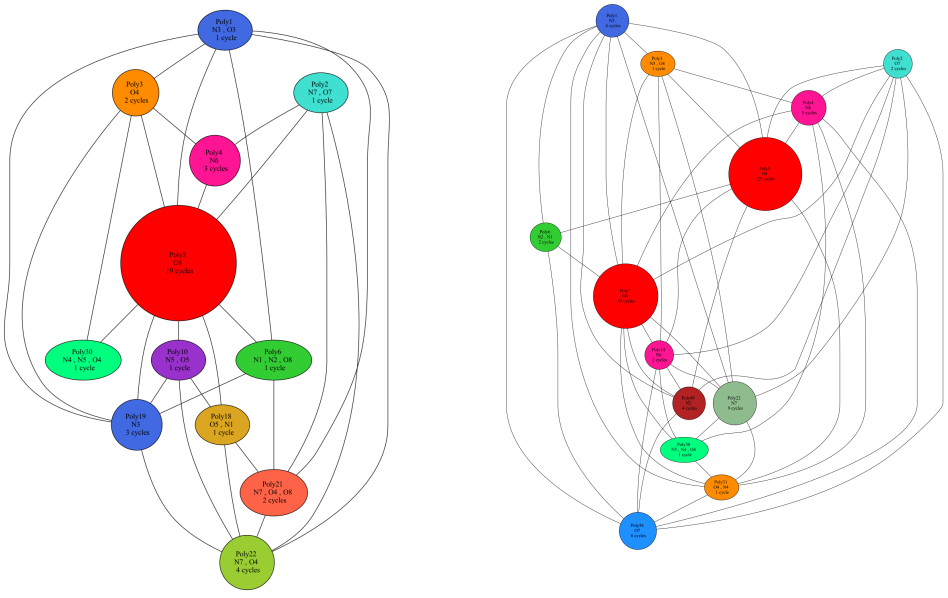
Les trajectoires tronquées correspondent à une petite fraction des trajectoires semi-empiriques, environ $1,3\%$ pour la méthode PM6 et environ $1,4\%$ pour la méthode AM1. Malgré cela, les polycycles obtenus sur ces courtes périodes correspondent à au moins la moitié des polycycles calculés pour les trajectoires complètes. Cependant, ce rapport n'est pas vérifié pour les autres paramètres. Nous l'observons, tout particulièrement, avec la trajectoire AM1 pour laquelle sa version tronquée contient seulement 20% des conformations, 32% des cycles et 50% des polycycles. Pour la trajectoire PM6, sa version tronquée présente déjà plus de la moitié des liaisons hydrogène observées pour cette méthode, mais elle ne compte que pour environ un quart des conformations et moins de la moitié des cycles d'intérêt. Notre hypothèse à la suite de ces premiers résultats est que cette petite portion de trajectoire permet de définir les briques structurelles majeures de la molécule, tandis que le reste de l'exploration permettra d'ajouter des éléments structuraux particuliers et de continuer à approfondir et varier les identités possibles pour les briques structurelles centrales. Cela suggère que sur une courte période, la simulation capture les motifs structuraux essentiels de la molécule.

En ce qui concerne la distance par rapport aux résultats empiriques, lorsque nous comparons les résultats obtenus sur des durées du même ordre, la différence ne semble

pas si grande. Il y a 7 polycycles pour la DFT, contre 12 pour la PM6 tronquée et 13 pour la trajectoire AM1 tronquée. De même, le nombre de liaisons hydrogène augmente légèrement mais reste du même ordre que le nombre de polycycles finaux. La plus grande variabilité se trouve au niveau des conformations et des cycles d'intérêts, qui sont bien plus nombreux que pour la DFT, mais sont peut-être également plus sporadiques.

La Figure 6.19 représente les polygraphes obtenus pour la trajectoire PM6 complète et pour sa version tronquée. Ces polygraphes ont été comparés avec le polygraphe de la trajectoire DFT à 450K, illustré dans la Figure 6.17b de la section précédente. Nous observons facilement que le polygraphe de la DFT est un sous-graphe du polygraphe de la PM6 tronquée. Nous notons plusieurs nouveaux polycycles qui peuvent être dus à de nouvelles zones explorées ou être le fruit des approximations de la méthode.

Concernant la trajectoire complète, le polygraphe de la DFT est également un sous-graphe du polygraphe obtenu. Néanmoins, ce polygraphe est plus complexe que celui de la version tronquée. En effet, le polycycle central de la DFT est toujours présent mais sous la forme de deux polycycles distincts. Pour trancher entre un phénomène dû aux approximations de la méthode et l'exploration de nouveaux bassins, il sera nécessaire d'appliquer des méthodes de recherche de minima d'énergie à partir de plusieurs conformations. Notons également que beaucoup de polycycles ne sont pas représentés sur le polygraphe car ils apparaissent moins de 1% du temps total de la trajectoire la trajectoire (20 polycycles calculés et seulement 13 représentés).



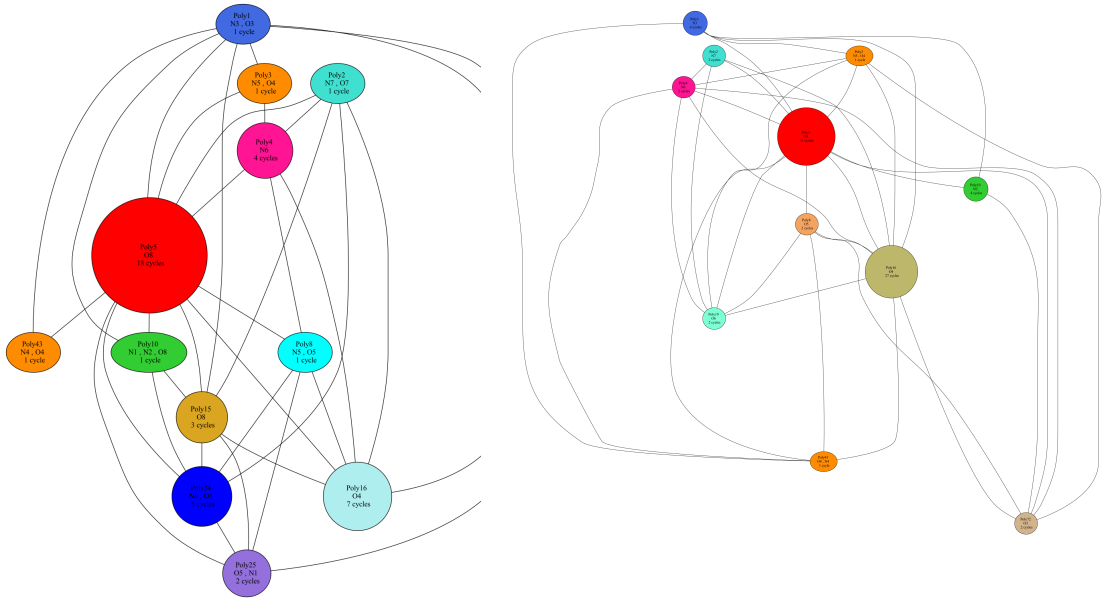
(a) PM6 tronquée (b) PM6

Figure 6.19 – Polygraphes obtenus pour la trajectoire PM6 à 450K du peptide Z-Ala$_6$−COOH complète et pour sa version tronquée.

La Figure 6.20 représente les polygraphes obtenus pour la trajectoire AM1 complète et pour sa version tronquée. Ces polygraphes ont été comparés avec le polygraphe de la trajectoire DFT à 450K. Nous observons que cette fois, même pour la version tronquée de la trajectoire, le polygraphe ne contient pas le polygraphe de DFT parmi ses sous-graphes.

En effet, le polycycle n°22 du polygraphe DFT n'est pas présent dans cette trajectoire. Il semble donc que des éléments structuraux n'aient pas été retrouvés. Notons également la présence de cinq nouveaux polycycles, contre quatre dans la trajectoire PM6 tronquée.

Nous observons que dans la trajectoire complète, parmi les 26 polcycles calculés, seulement 11 apparaissent au moins 1% du temps de la trajectoire et sont donc représentés dans le polygraphe. Cela suggère que ces polycycles sont des structures très sporadiques ou transitoires dans la dynamique de la molécule, et que leur impact est négligeable sur la vision globale des structures de la molécule. Nous avons constaté une tendance similaire dans la trajectoire PM6, bien que de manière moins prononcée, car plus de la moitié des polycycles étaient tout de même représentés dans le polygraphe. Cette observation est cohérente avec le fait que la méthode PM6 est plus récente et plus avancée que la méthode AM1, qui est plus ancienne et considérée comme moins précise.

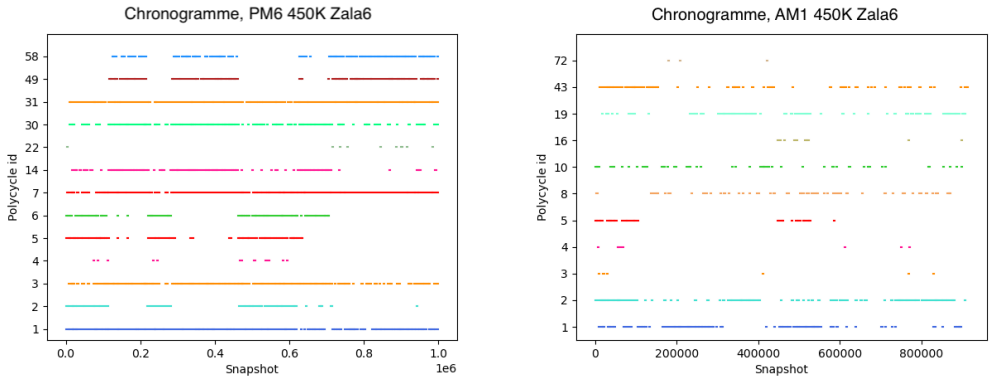


(a) AM1 tronquée (b) AM1

Figure 6.20 – Polygraphes obtenus pour la trajectoire AM1 à 450K du peptide Z-Ala$_6$−COOH complète et pour sa version tronquée.

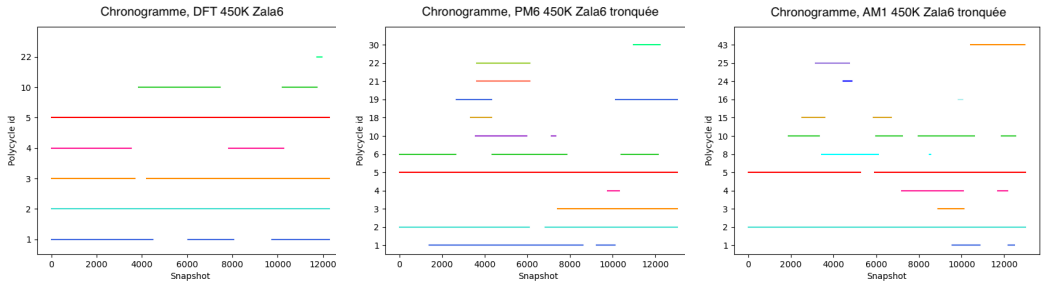
La Figure 6.21 illustre les chronogrammes obtenus pour les deux trajectoires semi-empiriques complètes. Malgré le seuil de 1%, les polycycles de la trajectoire AM1 présentent des périodes d'apparition très courtes de manière consécutive. Cela suggère de nombreux changements conformationnels qui peuvent ne pas être très représentatifs en raison de leur instabilité. En revanche, sur le chronogramme de la trajectoire PM6, nous observons plusieurs polycycles avec des périodes d'apparition très longues. Par exemple, les polycycles n°1, n°3, n°7, n°30 et n°31 sont présents plus ou moins tout au long de la trajectoire, indiquant une structure sous-jacente très conservée par la molécule. Nous remarquons également l'absence de coexistence de certains groupes de polycycles, tels que les polycycles n°49 et n°58, qui n'apparaissent pas en même temps que les polycycles n°5 et n°6, ce qui suggère l'exploration de plusieurs structures différentes.

La Figure 6.22 présente les chronogrammes obtenus pour les trois trajectoires de Z-Ala$_6$ à 450K, comprenant environ 13000 images, ce qui correspond à l'intégralité de la trajectoire DFT, ainsi qu'aux versions tronquées des trajectoires semi-empiriques.



(a) PM6 (b) AM1

Figure 6.21 – Chronogrammes des trajectoires semi-empiriques à 450K de Z-Ala$_6$ complètes. La coloration des polycycles indique si les polycycles correspondent malgré les différences de températures.



(a) DFT (b) PM6 tronquée (c) AM1 tronquée

Figure 6.22 – Chronogrammes des trois trajectoires à 450K de Z-Ala$_6$. Pour les trajectoires semi-empiriques, la version tronquée est présentée pour une meilleure comparabilité. La coloration des polycycles indique si les polycycles correspondent malgré les différences de températures.

Concernant le chronogramme de la trajectoire PM6 tronquée, une structure sous-jacente similaire à celle de la trajectoire DFT est observée, bien que les apparitions et les périodes ne soient pas aussi nettes que dans la trajectoire empirique. Il est également à noter que plusieurs nouveaux polycycles coexistent avec les structures connues de la trajectoire DFT.

Concernant le chronogramme de la trajectoire AM1 tronquée, nous retrouvons très partiellement la structure sous-jacente de la trajectoire empirique. De plus, les apparitions, même sur cette courte période, sont assez sporadiques, montrant à la fois peu de continuité et peu de distinctions nettes entre les éventuelles structures explorées.

Pour conclure, les observations que nous avons faites sur ces trajectoires sont cohérentes avec ce que nous savons des méthodes de simulation. En effet, les trajectoires semi-empiriques sont moins précises que les trajectoires empiriques. De plus, nous avons observé plusieurs exemples illustrant que la méthode PM6 est plus fiable que la méthode AM1. Il sera nécessaire de faire appel à des méthodes de recherche des minima d'énergie pour approfondir la comparaison de ces trajectoires.

6.3.3 . Analyse des trajectoires de Gramicidine

Cette section présente les résultats de l'analyse de deux trajectoires empiriques de Gramicidine à des températures de 150K et 600K.

Gramicidine est un peptide constitué de 136 atomes ($C_{99}H_{140}N_{20}O_{17}$), dont une représentation 3D est présentée sous différents angles dans la Figure 6.23. L'objectif de cette étude est d'observer les résultats obtenus par la méthode du polygraphe sur un peptide plus volumineux que ceux analysés précédemment dans ce chapitre. De plus, la forme de la molécule entraîne la formation de cycles côte à côte, ce qui est très peu favorable au modèle du polycycle.

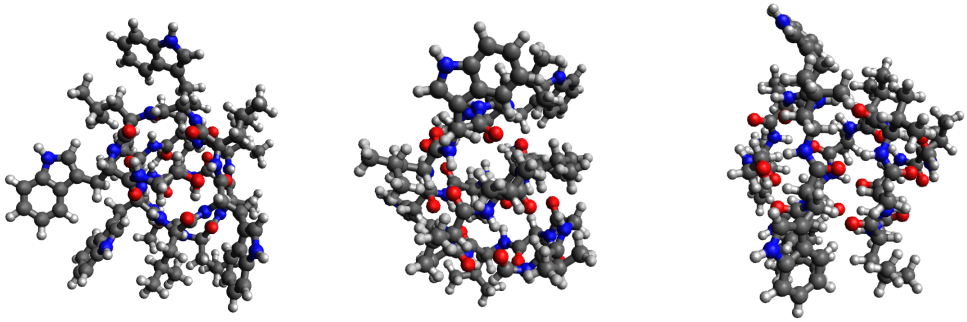


Figure 6.23 – Trois vues d'une représentation 3D de Gramicidine ($C_{99}H_{140}N_{20}O_{17}$). La couleur des atomes définie leur type : gris sombre pour le carbone, bleu pour l'azote, rouge pour l'oxygène, blanc pour l'hydrogène et jaune pour le soufre.

La Figure 6.24 présente un résumé des résultats de l'analyse par le polygraphe pour les deux trajectoires. Ces premiers résultats indiquent que très peu de regroupements ont été effectués, que ce soit à 150K ou à 600K. En effet, dans les deux cas, le polygraphe est constitué d'une quarantaine de polycycles, pour moins de cinquante cycles d'intérêt.

température	liaisons hydrogène	conformations	cycles	polycycles
150K	18	430	45	41
600K	18	561	48	42

Figure 6.24 – Resumé des résultats de l'analyse par le polygraphe des deux trajectoires de Gramicidine. Le tableau présente pour chacune, le nombre de liaisons hydrogène, le nombre de conformations différentes, le nombre de cycles d'intérêts et enfin le nombre de polycycles finaux.

La Figure 6.25 représente le polygraphe de la trajectoire à 150K, tandis que la Figure 6.26 représente celui de la trajectoire à 600K. En raison de leur taille, ces polygraphes sont assez difficiles à interpréter. Cependant, nous pouvons tout de même observer une densité significative, chaque polycycle étant connecté à plusieurs autres. Cette densité est particulièrement marquée sur le polygraphe à 600K, qui montre une augmentation des arêtes malgré l'ajout d'un seul polycycle supplémentaire.

De plus, parmi les 41 polycycles de la trajectoire à 150K, seuls deux ne sont pas retrouvés parmi les 42 polycycles de la trajectoire à 600K. Cela suggère que malgré l'augmentation de la température, la structure moléculaire globale de la trajectoire à 600K reste très proche de celle à 150K.

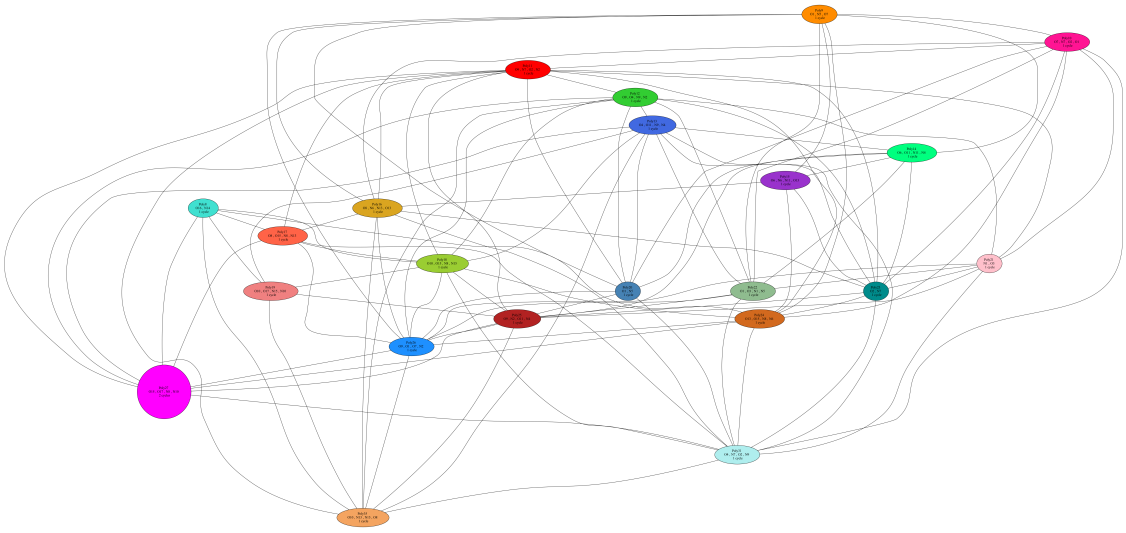


Figure 6.25 – Polygraphe obtenus pour la trajectoire DFT à 150K de la Gramicidine−COOH.

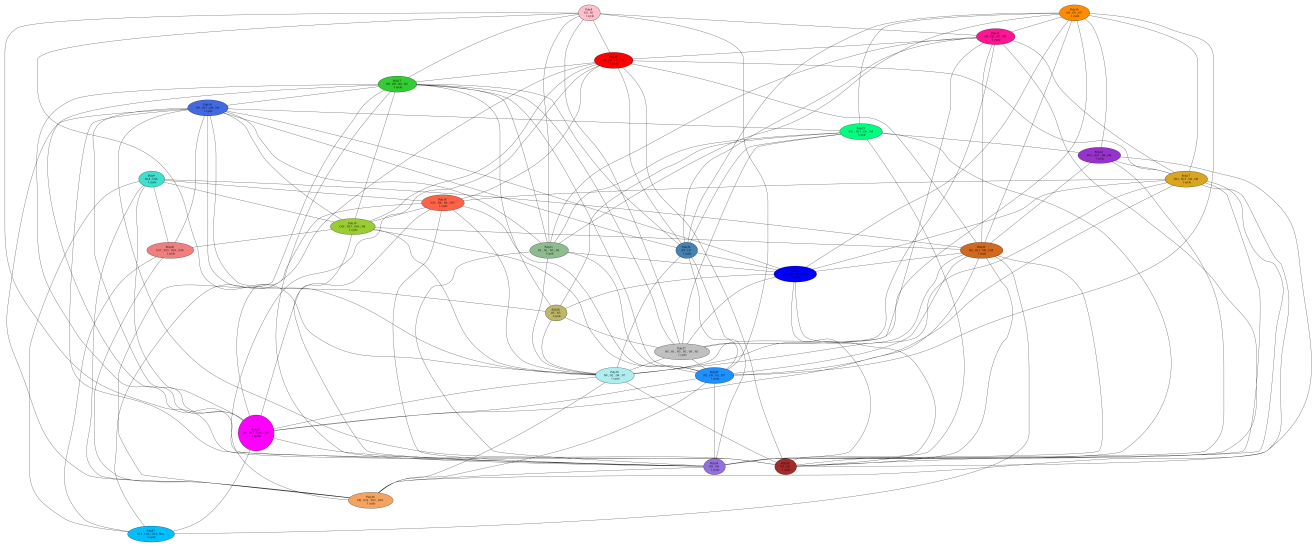
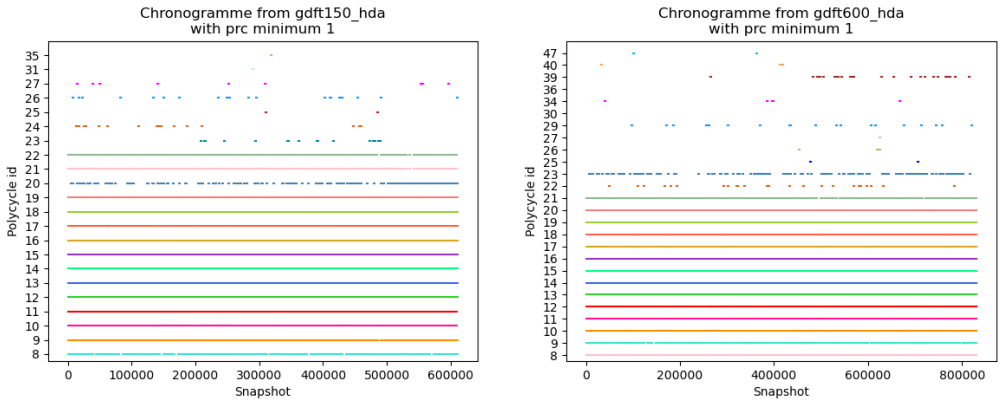


Figure 6.26 – Polygraphe obtenus pour la trajectoire DFT à 600K de la Gramicidine−COOH.

La Figure 6.27 présente les chronogrammes des deux trajectoires du peptide Gramicidine. Comme nous l'avons déjà observé, de nombreux polycycles de la trajectoire à 150K sont conservés dans la trajectoire à 600K. Ces chronogrammes confirment cette observation en montrant que la majeure partie de la structure est conservée entre les deux trajectoires.

Malgré la dimension du peptide Gramicidine et son inadéquation structurelle vis à vis de la méthode du polygraphe, nous parvenons néanmoins à identifier une structure sous-jacente présente tout au long de la trajectoire. Cependant, en ce qui concerne les polycycles variables, les informations disponibles sont limitées, et nous ne parvenons pas



(a) 150K (b) 600K

Figure 6.27 – Chronogrammes obtenus pour les trajectoires de la Gramicidine.

à observer des périodes distinctes marquant des structures moléculaires différentes.

6.4 . Synthèse sur l'analyse et la comparaison de trajectoires

Dans ce chapitre, nous avons observé les possibilités d'analyse et de comparaison offertes par le polygraphe.

Ainsi, nous avons plusieurs trajectoires de dynamique moléculaire obtenues avec différents paramètres et différentes méthodes de simulation pour lesquelles nous avons calculé le polygraphe.

Dans un premier temps, nous avons analysé des trajectoires de dynamique moléculaire seule. Nous avons alors défini le chronogramme pour représenter l'évolution dans le temps des différents polycycles du polygraphe. Cette analyse porte sur la reconnaissance des différentes structures qui peuvent s'apparenter à des bassins conformationnels.

Certains de ces résultats apparaissent dans la première version de l'analyse que nous avons fait dans le cadre d'une publication [2].

Dans un second temps, nous avons comparer des trajectoires de dynamique moléculaire. Pour ce faire, nous introduisons une méthode d'appariement des polycyles provenant de deux trajectoires d'une même molécule. Cette méthode orientée considère un des deux polygraphes comme reférence. Ainsi, nous avons observé l'impact de la température sur la dynamique moléculaire d'un peptide. Puis, l'impact de la méthode de simulation sur la dynamique moléculaire de ce même peptide.

Il serait intéressant dans une prochaine étude d'observer si les résultats diffèrent lorsque la comparaison est faite en inversant le sens de la comparaison.

Aussi, une perspective d'amélioration porte sur une méthode d'appariement plus générale pour les cas qui nécessitent de comparer un grand nombre de trajectoires. Ainsi, le graphe biparti serait remplacé par un graphe multi-parties dans lequel chaque partie correspond à l'ensemble des polycycles d'un polygraphe. Néanmoins, une étude supplémentaire est nécessaire pour définir le problème de couverture que nous cherchons à résoudre dans ce contexte.

Conclusion

Dans cette thèse, nous proposons une nouvelle méthode pour l'analyse et la comparaison de trajectoires de dynamique moléculaire basée sur l'algorithmique de graphes.

Dans le Chapitre 1, nous avons présenté le modèle des dynamiques moléculaires et les méthodes d'analyses traditionnelles. Nous avons également expliqué les raisons de notre orientation vers l'algorithmique de graphes pour sortir des modèles basés sur les calculs de l'énergie potentielle, ainsi que le choix des cycles comme un élément topologique représentatif de la dynamique.

Nous avons poursuivi dans le Chapitre 2 en introduisant notre méthode dans les grandes lignes. Nous y présentons les définitions nécessaires à la mise en place de notre modèle, ainsi que les premières pistes sur l'utilisation du polygraphe pour l'analyse de trajectoires. Cela donne une vue d'ensemble de la méthode avant d'aborder les problèmes sous-jacents aux différentes étapes de celle-ci.

Ainsi, le Chapitre 3 présente le problème de la sélection des cycles, par le biais notamment du problème d'intersection des bases de cycles minimum (MCBI), tandis que le Chapitre 4 présente le problème du partitionnement des cycles en polycycles (PCP). Nous avons étudié la complexité de ces problèmes et proposé différentes méthodes de résolution que nous avons ensuite évaluées et comparées dans le Chapitre 5.

Ces tests nous ont permis d'arrêter un protocole de construction d'un polygraphe à partir de l'ensemble des graphes d'une trajectoire de dynamique moléculaire. Enfin, le Chapitre 6 a mis en œuvre ce protocole pour établir le polygraphe de plusieurs trajectoires de dynamique moléculaire. Ces trajectoires ont été analysées à l'aide des outils associés au polygraphe afin d'identifier les structures importantes de la dynamique. Parmi les outils associés au polygraphe, nous retrouvons le chronogramme, qui permet d'obtenir une vue d'ensemble de la dynamique des polycycles, et la méthode de correspondance entre les polycycles obtenus pour des trajectoires différentes. Combinés, ces outils nous permettent de comparer ces trajectoires et de reconnaître les structures spécifiques aux paramètres des trajectoires.

La méthode que nous proposons est novatrice et permet une analyse structurelle des trajectoires de dynamiques moléculaires qui dépasse la granularité atomique en se concentrant sur des aspects topologiques plus larges. Cette approche permet de tirer des conclusions rapidement sans recourir à l'énergie potentielle. De plus, elle permet de reconnaître des structures d'intérêts sur lesquelles des analyses plus traditionnelles peuvent ensuite être effectuées.

Le polygraphe permet de modéliser la topologie des polycycles des trajectoires, offrant de vastes perspectives d'utilisation. Un aspect important à aborder à l'avenir est le découpage de la trajectoire en structures d'intérêts. Actuellement, le découpage proposé sur le chronogramme est effectué manuellement mais nous souhaiterions automatiser ce processus. Cependant, cette tâche est difficile car il est nécessaire de reproduire la vision globale du chronogramme que nous utilisons pour établir les structures d'intérêts. Or, le plus souvent, un algorithme considère et traite les données de façon locale.

Comme abordé dans les chapitres spécifiques, plusieurs pistes peuvent être explo-

rées pour approfondir l'analyse théorique des problèmes que nous avons définis. Une étude intéressante consisterait à comprendre précisément pourquoi les bases de cycles minimum obtenues par la méthode de Horton, et plus encore par la méthode de Horton modifié, sont si favorables à la réduction du nombre de polycycles dans la partition issue de leur union.

Toujours sur ce problème de la sélection des bases de cycles, nous pouvons envisager la mise en oeuvre d'une construction parallélisée des bases de cycles. Comme le montre la différence de resultats obtenus entre les méthodes de Horton et de Horton modifié, l'ordre de sélection des cycles dans la base impacte directement la cardinalité de la partition finale. Il semble donc naturel étudier l'ordre optimal dans lequel les sommets sont à considérer pour obtenir un ensemble de bases de cycles optimales pour le partitionnement en polycycles.

De plus, malgré nos hypothèses, nous n'avons pas encore élucidé les raisons exactes de l'insuccès du modèle MCBI. Le modèle s'est montré beaucoup plus défavorable que ce que nous avions anticipé alors qu'il n'est pas si éloigné de l'algorithme de Horton dans sa construction. Pour répondre à cela, une inverstigation plus poussée serait nécessaire afin de mieux comprendre quels sont les cycles défavorables et quand est-ce qu'ils sont choisis.

Bibliographie

[1] Aboulfath, Y., Barth, D., Mautor, T., Watel, D., and Weisser, M.-A. Polymorphic cycle basis in a sequence of graphs to analyze the structural evolution of a molecular dynamic trajectory. *submitted to ESA 2024*.

[2] Aboulfath, Y., Bougueroua, S., Cimas, A., Barth, D., and Gaigeot, M.-P. Time-resolved graphs of polymorphic cycles for h-bonded network identification in flexible biomolecules. *J. Chem. Theor. Comput., DOI : 10.1021/acs.jctc.3c01031* (2024).

[3] Abu-Khzam, F. N., Bonnet, E., and Sikora., F. On the complexity of various parameterizations of common induced subgraph isomorphism. *In J. Kratochv il, M. Miller, and D. Froncek, editors, Combinatorial Algorithms - 25th International Workshop, IWOCA 2014, volume 8986 of LNCS, page 1-12 Springer* (2014).

[4] Balaban, A. T. Topological indices based on topological distances in molecular graphs. 199–206.

[5] Balducci, R., and Pearlman, R. S. Efficient exact solution of the ring perception problem. *Journal of Chemical Information and Computer Sciences 34*, 4 (1994), 822–831.

[6] Bellare, M., Goldreich, O., and Sudan, M. Free bits, pcps, and nonapproximability—towards tight results. *SIAM Journal on Computing 27*, 3 (1998), 804–915.

[7] Berge, C. *Graphs and hypergraphs*. North-Holland Pub. Co., 1973.

[8] Berger, F., Flamm, C., Gleiss, P. M., Leydold, J., and Stadler, P. F. Counterexamples in chemical ring perception. *Journal of chemical information and computer sciences 44*, 2 (2004), 323–331.

[9] Berger, F., Gritzmann, P., and de Vries, S. Computing cyclic invariants for molecular graphs. *Networks 70*, 2 (2017), 116–131.

[10] Bougueroua, S., Quessette, F., Barth, D., and Gaigeot, M.-P. Gateway : Graph theory based software for an automatic analyses of molecular conformers generated over time. *ChemRxiv, DOI : 10.26434/chemrxiv-2022-1d5x8* (2022).

[11] Bougueroua, S., Spezia, R., Pezzotti, S., Vial, S., Quessette, F., Barth, D., and Gaigeot, M.-P. Graph theory for automatic structural recognition in molecular dynamics simulations. *J. Chem. Phys. 149*, 18 (2018), 184102.

[12] de Pina, J. C. *Applications of shortest path methods*. na, 1995.

[13] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. am1 : a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society 107*, 13 (1985), 3902–3909.

[14] Diestel, R. *Graph Theory*, 5th ed. Springer, Heidelberg, 2017.

[15] Diudea, M. V., Gutman, I., and Jantschi, L. *Molecular topology*. Nova Science Publishers Huntington, NY, USA, 2001.

[16] Edmonds, J. Minimum partition of a matroid into independent subsets. *J. Res. Nat. Bur. Standards Sect. B 69* (1965), 67–72.

[17] Edmonds, J. Matroids and the greedy algorithm. *Mathematical Programming 1*, 1 (1971), 127–136.

[18] Edmonds, J. *Submodular Functions, Matroids, and Certain Polyhedra*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 11–26.

[19] Frank, A. A weighted matroid intersection algorithm. *Journal of Algorithms 2*, 4 (1981), 328–336.

[20] Fujita, S. A new algorithm for selection of synthetically important rings. the essential set of essential rings for organic structures. *Journal of Chemical Information and Computer Sciences 28*, 2 (1988), 78–82.

[21] Gasteiger, J., and Jochum, C. An algorithm for the perception of synthetically important rings. *Journal of Chemical Information and Computer Sciences 19*, 1 (1979), 43–48.

[22] Gianfrotta, C. *Modélisation, analyse et classification de motifs structuraux d'ARN à partir de leur contexte, par des méthodes d'algorithmique de graphes*. PhD thesis, Université Paris-Saclay, 2022.

[23] Gleiss, P. M. Short cycles. *Universität Wien* (2001).

[24] Gutman, I., and Estrada, E. Topological indices based on the line graph of the molecular graph. *Journal of Chemical Information and Computer Sciences 36*, 3 (01 1996), 541–543.

[25] Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference* (Pasadena, CA USA, 2008), G. Varoquaux, T. Vaught, and J. Millman, Eds., pp. 11 – 15.

[26] Hinsen, K. The molecular modeling toolkit : A new approach to molecular simulations. *Journal of Computational Chemistry 21*, 2 (2024/02/21 2000), 79–85.

[27] Horton, J. D. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing 16*, 2 (1987), 358–366.

[28] Humphrey, W., Dalke, A., and Schulten, K. Vmd : Visual molecular dynamics. *Journal of Molecular Graphics 14*, 1 (1996), 33–38.

[29] James Speight, P. D. *Lange's Handbook of Chemistry, Sixteenth Edition*, 16th ed. / ed. McGraw-Hill Education, New York, 2005.

[30] Karp, R. Reducibility among combinatorial problems. *Complexity of computer computations* (1972), 85–103.

[31] Karp, R. M. *Reducibility among Combinatorial Problems*. Springer US, 1972, pp. 85–103.

[32] Kavitha, T., Mehlhorn, K., and Michail, D. New approximation algorithms for minimum cycle bases of graphs. *Algorithmica 59*, 4 (2011), 471–488.

[33] Kavitha, T., Mehlhorn, K., Michail, D., and Paluch, K. A faster algorithm for minimum cycle basis of graphs. In *Automata, Languages and Programming* (2004), J. Díaz, J. Karhumäki, A. Lepistö, and D. Sannella, Eds., Springer Berlin Heidelberg, pp. 846–857.

[34] Kavitha, T., Mehlhorn, K., Michail, D., and Paluch, K. E. An $\tilde{o}(m^{2}n)$algorithm for minimum cycle basis of graphs. *Algorithmica 52*, 3 (2008), 333–349.

[35] Kruskal, J. B. On the shortest spanning tree of a graph and the traveling salesman problem. *theAmerican Mathematical Society 7* (1956), 48–50.

[36] Lawler, E. L. Matroid intersection algorithms. *Mathematical Programming 9*, 1 (1975), 31–56.

[37] Leach, A. *Molecular Modelling : Principles and Applications*. Pearson education. Longman, 1996.

[38] Leach, A. *Molecular Modelling : Principles and Applications*. Prentice Hall, 2001.

[39] Lewars, E. *Computational Chemistry : Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Springer Netherlands, 2010.

[40] May, J. W., and Steinbeck, C. Efficient ring perception for the chemistry development kit. *Journal of Cheminformatics 6* (2014), 1–12.

[41] Mehlhorn, K., and Michail, D. Implementing minimum cycle basis algorithms. *ACM J. Exp. Algorithmics 11* (feb 2007), 2.5–es.

[42] Nouleho ilemo, S. *Algorithmique de graphes pour la similarité structurelle de molécules et de réactions*. PhD thesis, 2020. Thèse de doctorat dirigée par Barth, Dominique Informatique université Paris-Saclay 2020.

[43] Nouleho Ilemo, S., Barth, D., David, O., Quessette, F., Weisser, M.-A., and Watel, D. Improving graphs of cycles approach to structural similarity of molecules. *Plos one 14*, 12 (2019), e0226680.

[44] Randic, M., Hansen, P. J., and Jurs, P. C. Search for useful graph theoretical invariants of molecular structure. *Journal of Chemical Information and Computer Sciences 28*, 2 (1988), 60–68.

[45] Rouvray, D. H. Graph theory in chemistry. *R. Inst. Chem., Rev. 4* (1971), 173–195.

[46] Sega, M., Hantal, G., Fábián, B., and Jedlovszky, P. Pytim : A python package for the interfacial analysis of molecular simulations. *Journal of Computational Chemistry 39*, 25 (2024/02/21 2018), 2118–2125.

[47] STEGER, A., and WORMALD, N. C. Generating random regular graphs quickly. *Combinatorics, Probability and Computing 8*, 4 (1999), 377–396.

[48] Stewart, J. J. P. Optimization of parameters for semiempirical methods v : Modification of nddo approximations and application to 70 elements. *Journal of Molecular Modeling 13*, 12 (2007), 1173–1213.

[49] Sysko, M. The subgraph isomorphism problem for outerplanar graphs. *Theoretical Computer Science 17(1)* (1982), 91–97.

[50] VISHVESHWARA, S., BRINDA, K. V., and KANNAN, N. Protein structure : Insights from graph theory. *Journal of Theoretical and Computational Chemistry 01*, 01 (2002), 187–211.

[51] Vismara, P. Union of all the minimum cycle bases of a graph. *the electronic journal of combinatorics* (1997), R9–R9.

[52] Watel, D., Aboulfath, Y., Barth, D., Mautor, T., and Weisser, M.-A. Maximizing minimum cycle bases intersection. *IWOCA 2024, proceedings will be published in the "Lecture Notes in Computer Science"*.

[53] Welsh, D. J. *Matroid theory*. Courier Corporation, 2010.

[54] Whitney, H. On the abstract properties of linear dependence. *American Journal of Mathematics 57*, 3 (1935), 509–533.

[55] Wong, H.-W., Li, X., Swihart, M. T., and Broadbelt, L. J. Encoding of polycyclic si-containing molecules for determining species uniqueness in automated mechanism generation. *Journal of Chemical Information and Computer Sciences 43*, 3 (05 2003), 735–742.