Caractérisation de structures explorées dans les simulations de dynamique moléculaire.

Sana Bougueroua

NNT: 2017SACLV099

Caractérisation de structures explorées dans les simulations de dynamique moléculaire

Thèse de doctorat de l'Université Paris-Saclay
Préparée à l'Université Versailles Saint-Quentin en Yvelines
et l'Université d'Evry Val d'Essonne

École doctorale n°580 Sciences et Technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Versailles, le 13 Décembre 2017, par

Sana Bougueroua

Composition du Jury :

Jean-Michel Fourneau Professeur, Université de Versailles Saint-Quentin en Yvelines	Président
Anne-Claude Camproux Professeure, Université Paris Diderot	Rapporteure
Olivier Poch DR CNRS, ICube de l'Université de Strasbourg	Rapporteur
Delphine Flatters Maîtresse de Conférences, Université Paris Diderot	Examinatrice
Dominique Barth Professeur, Université de Versailles Saint-Quentin en Yvelines	Directeur de thèse
Marie-Pierre Gaigeot Professeure, Université d'Evry Val d'Essonne	Directrice de thèse
Franck Quessette Maître de Conférences, Université de Versailles Saint-Quentin en Yvelines	Invité

Titre : Algorithmes pour l'analyse et la prédiction des conformations des systèmes moléculaires en phase gazeuse

Mots clés : algorithmique des graphes, algorithme pour la chimo-informatique, analyse de trajectoires de dynamique moléculaire, simulations de dynamique moléculaire, conformation moléculaire et prédiction conformationnelle.

Résumé : L'objectif de cette thèse est d'analyser et prédire les conformations d'un système moléculaire en combinant la théorie des graphes et la chimie computationnelle.
Dans le cadre des simulations de dynamique moléculaire, une molécule peut avoir une ou plusieurs conformations au cours du temps. Dans les trajectoires de simulation de dynamique moléculaire, on peut avoir des trajectoires n'explorant qu'une seule conformation ou des trajectoires explorant plusieurs conformations, donc plusieurs transitions entre conformations sont observées. L'exploration de ces conformations dépend du temps de la simulation et de l'énergie fixée dans le système. Pour avoir une bonne exploration des conformations d'un système moléculaire, il faut générer et analyser plusieurs trajectoires à différentes énergies. Notre objectif est de proposer un algorithme universel qui permet d'analyser la dynamique conformationnelle de ces trajectoires d'une façon rapide et automatique. Les trajectoires fournissent les positions cartésiennes des atomes du système moléculaire à des intervalles de temps réguliers. Chaque intervalle contenant un ensemble de positions est appelé image. L'algorithme utilise des règles de géométrie (distances, angles, etc.) sur les positions pour trouver les liaisons (liaisons covalentes, liaisons hydrogène et interactions électrostatiques), permettant par la suite d'obtenir le graphe mixte qui modélise une conformation. Nous ne considérons un changement conformationnel que s'il y a un changement dans les liaisons calculées à partir des positions données. L'algorithme permet de donner l'ensemble des conformations explorées sur des trajectoires, la durée d'exploration de chaque conformation, ainsi que le graphe de transitions qui contient tous les changements conformationnels observés.
Chaque conformation possède une énergie potentielle. L'ensemble des énergies des différentes énergies donne une surface à 3N dimensions appelée surface d'énergie potentielle. Des chercheurs, en chimie théorique, s'intéressent d'une part à trouver des conformations à plus basse énergie qui sont considérées comme états stables de la molécule. Ces conformations présentent les minima sur la surface d'énergie potentielle. D'autres part, ils s'intéressent à trouver les points de passage entre deux conformations (les états de transition) qui représentent les maxima de la surface d'énergie potentielle. Les méthodes développées pour calculer ces points nécessitent une connaissance de l'énergie potentielle ce qui est coûteux en temps et en calculs. Notre objectif est de proposer une méthode alternative en utilisant des mesures ad hoc basées sur des propriétés des graphes qu'on a utilisées dans le premier algorithme et sans faire appel à la géométrie ni aux calculs moléculaires. Ces mesures permettent de générer des conformations avec un classement énergétique ainsi de définir le coût énergétique de chaque transition permise. Les conformations possibles avec les transitions représentent respectivement les sommets et les arcs de ce qu'on appelle le "graphe des possibles". Les hypothèses utilisées dans le modèle proposé est que seules les liaisons hydrogène peuvent changer entre les conformations et que le nombre de liaisons hydrogène présentes dans le système permet de déterminer son coût énergétique. L'algorithme d'analyse des trajectoires a été testé sur trois types de systèmes moléculaires en phase gazeuse de taille et de complexité croissantes. Bien que la complexité théorique de l'algorithme soit exponentielle (tests d'isomorphisme) les résultats ont montré que l'algorithme est rapide (quelques secondes).
De plus, cet algorithme peut être facilement adapté et appliqué à d'autres systèmes. Pour la prédiction conformationnelle, le modèle proposé a été testé sur des peptides isolés.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Title : Algorithm to analyse and predict conformations for molecular system in gas phase

Keywords: algorithms for chemo-informatics, graph theory, molecular dynamics trajectory analysis, molecular dynamics simulations, molecular conformation, and conformational prediction

Abstract: This PhD is part of transdisciplinary works, combining graph theory and computational chemistry.
In molecular dynamics simulations, a molecular system can adopt different conformations over time. Along a trajectory, one conformation or more can thus be explored. This depends on the simulation time and energy within the system. To get a good exploration of the molecular conformations, one must generate and analyse several trajectories (this can amount to thousands of trajectories). Our objective is to propose an automatic method that provides rapid and efficient analysis of the conformational dynamics explored over these trajectories. The trajectories of interest here are in cartesian coordinates of the atoms that constitute the molecular system, recorded at regular time intervals (time-steps). Each interval containing a set of positions is called a snapshot. At each snapshot, our developed algorithm uses geometric rules (distances, angles, etc.) to compute bonds (covalent bonds, hydrogen bonds and any other kind of intermolecular criterium) formed between atoms in order to get the mixed graph modelling one given conformation. Within our current definitions, a conformational change is characterized by either a change in the hydrogen bonds or in the covalent bonds. One choice or the other depends on the underlying physics and chemistry of interest. The proposed algorithm provides all conformations explored along one or several trajectories, the period of time for the existence of each one of these conformations, and also provides the graph of transitions that shows all conformational changes that have been observed during the trajectories. A user-friendly interface has been developed, that can de distributed freely.
Our proposed algorithm for analysing the trajectories of molecular dynamics simulations has been tested on three kinds of gas phase molecular systems (peptides, ionic clusters).
This model can be easily adapted and applied to any other molecular systems as well as to condensed matter systems, with little effort.
Although the theoretical complexity of the algorithm is exponential (isomorphism tests), results have shown that the algorithm is rapid.
In addition, we have worked on computationally low cost graph methods that can be applied in order to pre-characterize specific conformations/points on a potential energy surface (it describes the energy of a system in terms of positions of the atoms). These points are the minima on the surface, representing the most stable conformations of a molecular system, and the maxima on that surface, representing transition states between two conformers. Our developed methods and algorithms aim at getting these specific points, without the prerequisite knowledge/calculation of the potential energy surface by quantum chemistry methods (or even by classical representations). By avoiding an explicit calculation of the potential energy surface by quantum chemistry methods, one saves computational time and effort. We have proposed an alternative method using ad doc measures based on properties of the graphs (already used in the first part of the PhD), without any knowledge of energy and/or molecular calculations. These measures allow getting the possible conformations with a realistic energy classification, as well as transition states, at very low computational cost. The algorithm has been tested on gas phase peptides.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Remerciement

Je remercie tout d'abord Dieu, le Tout-Puissant, pour ses faveurs et ses grâces, de m'avoir donné le courage et la patience pour accomplir ce travail.

Je voudrais exprimer ma sincère gratitude à mes superviseurs Dominique Barth, Marie-Pierre Gaigeot et Franck Quessettes. Je les remercie de m'avoir suivie et si bien orienté tout au long de ce travail. J'ai énormément appris de leurs remarques pertinentes, compétences et de leurs expériences pour accomplir cette thèse.

Je remercie également Anne-Claude Camproux et Olivier Poch qui m'ont fait l'honneur d'être rapporteurs de cette thèse. Je les remercie pour leurs disponibilités et leurs compétences pédagogiques et scientifiques.

J'exprime tous mes remerciements à Delphine Flatters et Jean-Michel Fourneau d'avoir accepté d'être parmi les membres de jury pour l'évaluation de cette thèse. Je les remercie vivement pour leurs temps et tout l'intérêt qui ont apporté à mon travail.

Un grand MERCI aux membres du laboratoire DAVID de l'université de Versailles et à ceux du LAMBE de l'université d'Evry pour l'accueil et les conditions du travail, qui m'ont été offertes, pour leur aide et leurs conseils.

Je remercie, particulièrement, Sandrine Vial et Riccardo pour leurs conseils et leurs suggestions.

Enfin, je voudrais exprimer un profond sentiment de gratitude à mes parents, qui ont toujours étaient là pour moi comme un pilier, à qui je dois ma vie, pour leur amour constant, encouragement, soutien moral et bénédictions. Je remercie tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce travail.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

1

Introduction Générale

Les propriétés physiques et chimiques d'une molécule donnée dépendent principalement de sa structure tridimensionnelle. Ces propriétés ne sont pas seulement déterminées par les atomes qu'elle contient, mais aussi par l'ensemble de liaisons formées entre ces atomes, ce qui donne une forme tridimensionnelle. Un simple exemple, le diamant et le graphite (voir figure 1) sont tous les deux composés uniquement de carbone, et pourtant leurs propriétés physiques et chimiques sont totalement différentes. Cette différence est due à la disposition de leurs atomes de carbone [1].

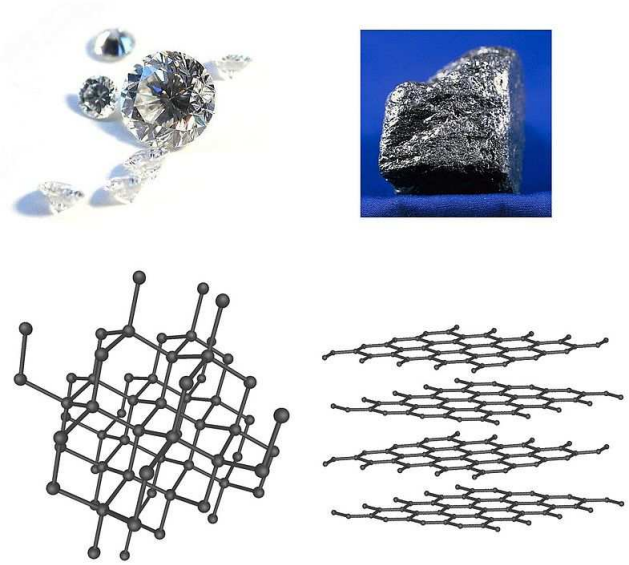

FIGURE 1 – Structure d'un diamant et d'un graphite [1].

Cette forme tridimensionnelle est appelée **conformation**. Une même molécule peut avoir plusieurs conformations. Ces différentes conformations peuvent typiquement résulter d'apparition ou de disparition de liaisons hydrogène ou covalentes, d'interactions électrostatiques, de transferts de proton, de rotation, etc. En effet, tous les atomes de la molécule bougent au cours du temps. Ce mouvement est soit une simple vibration, soit un changement de liaisons/interactions chimiques et dans ce cas on parle d'un changement conformationnel (la molécule change de conformation). La stabilité de la molécule dépend de la stabilité de la conformation qu'elle adopte. Chaque conformation possède une énergie appelée **énergie**

potentielle. C'est une énergie qui reflète les liaisons covalentes, les liaisons hydrogène et interactions formées dans cette conformation. Les conformations les plus stables sont les conformations de plus basse énergie. La stabilité de la molécule joue un grand rôle dans ses fonctions "biologiques". D'où l'importance de connaitre ses conformations, de trouver les plus stables ainsi que de comprendre les changements entre elles (les transitions entre ces conformations).

En chimie théorique et computationnelle, pour trouver les conformations d'une molécule plusieurs simulations de dynamique moléculaires (avec méthodes ab initio ou champs de forces classiques) sont réalisées [2]. Le principe est généralement de faire évoluer les positions des atomes de la molécule au cours du temps. Ces simulations génèrent des **trajectoires** contenant des coordonnées cartésiennes des atomes de la molécule à des intervalles de temps réguliers. Chaque intervalle contenant un ensemble de positions est appelé **image**.

Une fois les trajectoires obtenues, l'étape suivante consiste à les analyser pour identifier les conformations explorées et déterminer la durée d'exploration de chacune de ces conformations. Pour les analyser, les groupes de recherche comme le groupe de théorie et modélisation du LAMBE avec qui cette thèse a été co-réalisée [2] utilisent des outils de visualisation qui permettent d'avoir des représentations tridimensionnelles statiques à chaque image de la trajectoire. Ensuite, ils analysent "**à l'oeil**" les conformations explorées car les logiciels de visualisation ne donnent aucune information sur la dynamique conformationnelle. En d'autres termes, l'utilisation de tels logiciels ne permet pas de décider si la conformation d'une image donnée était la même à l'image précédente. L'analyse à l'oeil est largement faisable pour des molécules petites et structurellement peu complexes (des peptides de 10 à 20 atomes). Pour des peptides [3] plus complexes, cette méthode n'est plus si facilement applicable, et ne permet pas une détermination fine et précise des structures explorées au cours du temps.

D'autre groupes préfèrent utiliser leurs propres codes pour analyser ces trajectoires. Le grand inconvénient de ces codes est qu'ils ne sont applicables qu'aux systèmes pour lesquels ils ont été développés (codes peu transposables). Au sein d'un même groupe et pour un même objectif qui est l'identification des conformations sur les trajectoires, on trouve plusieurs codes complètement différents.

De plus, pour avoir une bonne exploration des conformations d'un système moléculaire, il faut générer et analyser un nombre important de trajectoires à différentes énergies et avec des durées suffisamment longues. Car selon l'énergie (température) fixée et la durée de simulation, les trajectoires peuvent n'explorer qu'une

2. Equipe du laboratoire d'Analyse et Modélisation pour la Biologie et l'Environnement (LAMBE), de l'université d'Evry Val d'Essonne, lien :http://www.lambe.univ-evry.fr/spip.php?article84

3. Dans le cadre de la thèse, nous nous intéressons aux systèmes en phase gazeuse, à savoir des peptides (appelés également polymères) ou des clusters d'eau.

seule conformation. Dans ces cas, avoir une méthode alternative devient nécessaire. Une méthode qui permet d'analyser d'un façon automatique, rapide et pertinente la dynamique conformationnelle sur les trajectoires de simulation de dynamique moléculaire et qui peut être appliquée à n'importe quel système moléculaire.

L'énergie potentielle des conformations peut être représentée par une surface à plusieurs dimensions, appelée **surface d'énergie potentielle**. L'identification des conformations sur les trajectoires de simulation de dynamique moléculaire revient à explorer les bassins conformationnels de cette surface comme sont indiquées en cercles violets sur la figure 2. Le changement de conformations est accompagné par le changement d'énergie potentielle qui évolue et passe par des points caractéristiques qui vont décider des propriétés chimiques du système moléculaire [2]. Ces points sont les minima (voir les points verts sur la figure 2) qui représentent les conformations les plus stables et les maxima ou états de transition (voir les points oranges sur la figure 2) qui représentent les coûts énergétiques de passage d'une conformation à une autre.

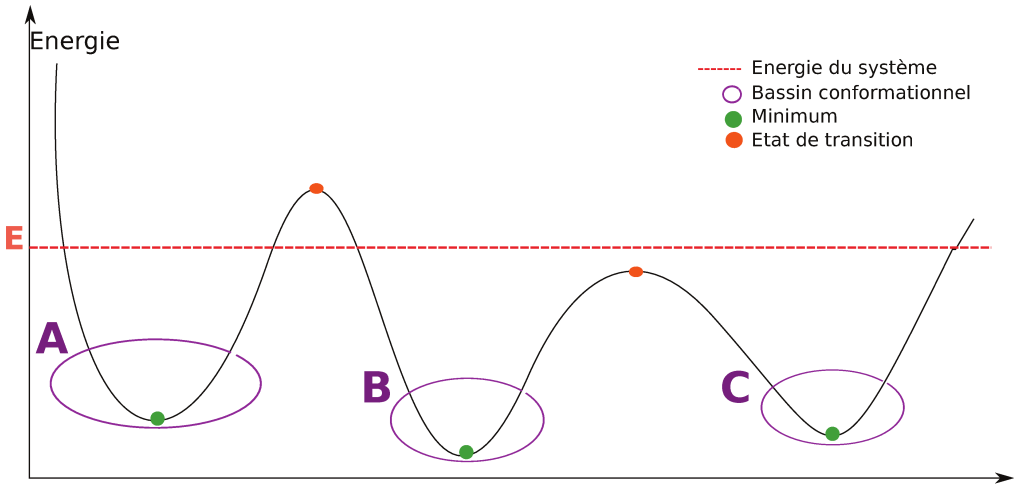

FIGURE 2 – Représentation schématique de la surface d'énergie potentielle.

Des méthodes et des algorithmes ont été développés pour trouver ces points. Pour les minima sur la surface d'énergie potentielle, on parle de méthodes de minimisation. En général, le principe des méthodes de minimisation comme le gradient conjugué ou le Newton-Rhapson [3] est d'appliquer des techniques basées sur le calcul des dérivées de l'énergie potentielle sur un ou plusieurs points de départ issu de simulation (ces points sont également choisi par d'autres méthodes [4]). Pour arriver à un minimum qui est généralement le plus proche de ces points. Ces méthodes nécessitent beaucoup de calculs et de temps, et surtout nécessitent une connaissance de l'énergie potentielle. De plus, les points de départ sont fixés à travers des algorithmes d'optimisation qui ne sont pas toujours efficaces et un mauvais choix des points de départ peut entrainer des faux minima. Par ailleurs, les états de

transitions sont plus compliqués à calculer en comparaison avec les minima sur la surface d'énergie potentielle [5]. De même, les méthodes et les algorithmes développés ne sont facilement applicables et à moindre coûts computationnels qu'aux systèmes de petites tailles (une dizaine d'atomes typiquement) et ils ne garantissent pas de bons résultats tout le temps.

Dans le présent document, nous proposons deux algorithmes principaux. Le premier algorithme a pour objectif d'analyser la dynamique conformationnelle sur les trajectoires de simulation de dynamique moléculaire. Cet algorithme utilise des règles de géométrie (distances, angles, etc.) sur les positions pour trouver les liaisons (liaisons covalentes, liaisons hydrogène et interactions électrostatiques) formées entre les atomes, permettant par la suite d'obtenir le graphe mixte qui modélise une conformation. En se basant sur des concepts de la théorie des graphes, à savoir l'isomorphisme du graphe, l'algorithme fournit la liste des conformations explorées dans la trajectoire ainsi que la durée d'apparition de chacune et le graphe des transitions qui contient tous les changements observés au cours du temps. Le modèle utilisé permet à l'algorithme d'être adapté et appliqué facilement à n'importe quel système moléculaire. De plus, les conformations identifiées par cet algorithme peuvent être considérées comme des points pertinents à minimiser. Pour s'affranchir de la surface d'énergie potentielle, nous proposons un deuxième algorithme qui permet de prédire toutes les conformations possibles qu'une molécule peut adopter ainsi que les transitions possibles entre elle. L'algorithme repose sur des mesures *ah doc* appliquées sur les graphes mixtes des conformations. Il s'agit de construire un graphe appelé **graphe des possibles**. C'est un graphe orienté pondéré, où les sommets représentent les conformations possibles et les arcs les transitions possibles entre elles. Nous définissons deux pondérations sur ce graphe : une pondération sur les sommets (conformations) qui permet de donner une classification à ces conformations et donc d'en pouvoir choisir les plus stables. La deuxième pondération est sur les arcs qui permet de donner le coût énergétique pour passer d'une conformation à une autre, en d'autres termes, cela permet de donner une approximation des états de transition. Ce graphe permet également de fournir des chemins de coût minimum d'une conformation à une autre. Ces chemins peuvent être similaires ou différents des chemins observés dans les trajectoires. Il faut noter que les graphes de transitions observés dans les trajectoires de simulation sont des sous-graphe du graphe des possibles. Un objectif du graphe des possibles est de pouvoir exhiber des différences entre les chemins réels (résultats d'analyse des trajectoires avec l'algorithme) et les chemins théoriques (graphe des possibles).

Dans le chapitre 1, nous introduisons le contexte dans lequel cette thèse a été réalisée, à savoir les systèmes moléculaires auxquels nous nous sommes intéressés, leurs caractéristiques et les changements de conformations dans ces systèmes.

Nous présentons dans le même chapitre, les méthodes existantes, qui sont les plus réputées et employées par nos collaborateurs de théorie et modélisation du LAMBE avec qui cette thèse a été co-réalisée [4], pour l'identification des conformations d'un système moléculaire et la recherche des minima et des états de transition sur la surface d'énergie potentielle. Cela en mettant en évidence les caractéristiques de chacune et les limitations qui sont adressées dans cette thèse. Ensuite, dans le chapitre 2 nous présentons d'une façon formelle les modèles proposés pour représenter les systèmes moléculaires, leurs propriétés et l'ensemble des règles utilisées dans la conception des algorithmes et des méthodes proposés.

Dans les chapitres 3 et 4, nous nous focalisons sur les algorithmes et les méthodes développées dans le cadre de cette thèse. Dans le chapitre 3, nous présentons l'algorithme qui permet d'analyser la dynamique conformationnelle d'un système moléculaire sur une ou plusieurs trajectoires. Nous expliquons les étapes d'identification des conformations dans une trajectoire et nous décrivons comment identifier un changement conformationnel en utilisant des concepts de la théorie des graphes. La prédiction conformationnelle est ensuite introduite dans le chapitre 4, où nous présentons le modèle et les différentes règles utilisées pour le graphe des possibles. Dans ce chapitre, nous décrivons le processus de pondération du graphe des possibles en utilisant des mesures ah doc et comment utiliser cette pondération pour classifier les conformations possibles et trouver des chemins de coût minimum entre ces conformations.

Ensuite, dans le chapitre 5, nous illustrons l'application de nos algorithmes à des systèmes moléculaires en phase gazeuse. Le premier algorithme destiné à l'analyse des trajectoires de simulation de dynamique moléculaire a été testé sur trois types de systèmes : trajectoires de peptides isolés de taille et de complexité croissantes, trajectoires de dissociation de peptides induite par collision et des trajectoires de clusters. L'objectif de prendre des systèmes différents est de montrer le fonctionnement et la flexibilité de l'algorithme ainsi qu'expliquer ses différents résultats. Pour la prédiction conformationnelle, le modèle proposé repose sur la dynamique des liaisons hydrogène. A cet effet, les systèmes moléculaires testés sont les peptides isolés. A la fin de ce chapitre, nous présentons une comparaison pour un système moléculaire entre des chemins observés dans des trajectoires de simulation et les chemins théoriques trouvés dans le graphe des possibles de ce système. Enfin, nous terminons ce travail par une conclusion et quelques perspectives.

4. Une équipe du laboratoire d'Analyse et Modélisation pour la Biologie et l'Environnement (LAMBE), de l'université d'Evry Val d'Essonne, lien :http://www.lambe.univ-evry.fr/spip.php?article84

Chapitre 1

Contexte, problématiques et littérature

Les propriétés physiques et chimiques d'une molécule donnée dépendent principalement de sa structure tridimensionnelle. En effet, ces propriétés ne sont pas seulement déterminées par les atomes qu'elle contient, mais aussi par la disposition spatiale de ces atomes (ce sont des atomes liés entre eux par un ensemble de liaisons covalentes ou hydrogène) qui donne une forme tridimensionnelle définie qu'on appelle **conformation**. Cette conformation n'est pas unique, elle peut changer au cours du temps. Ces changements sont dus aux apparitions/disparitions de ces liaisons. Chaque conformation possède une certaine énergie qui influence la stabilité de la molécule. Plus cette énergie est basse, plus la molécule est stable. La recherche des conformations de plus basses énergies ainsi que l'évolution de la molécule (les chemins entre ces conformations) au cours du temps est d'une grande importance en chimie.

En chimie théorique et computationnelle, trouver les conformations est effectué en général à travers des méthodes de simulation. Le principe est par exemple de faire évoluer les positions des atomes de la molécule au cours du temps. Ces simulations génèrent un ensemble de trajectoires contenant des positions d'atomes à différents instants. Le défi est d'être capable d'identifier d'une façon rapide et efficace, à partir de ces trajectoires, les bassins conformationnels explorés durant les simulations. A une énergie (température) donnée, tous les atomes de la molécule bougent, ce mouvement est soit une simple vibration, c'est-à-dire on reste dans le même bassin conformationnel (même conformation), ou bien il entraine un changement de liaisons et dans ce cas on dit qu'on a un changement de bassin (changement de conformation). Dans le cadre de cette thèse, nous nous intéressons aux changements dans les liaisons chimiques (covalentes et hydrogène). Les bassins conformationnels sont représentés par une surface à plusieurs dimensions, appelée **surface d'énergie potentielle** (voir figure 1.1).

L'identification des conformations sur les trajectoires de simulation de dynamique moléculaire est appelée **exploration d'espace des conformations sur la surface d'énergie potentielle**. Même une simulation extrêmement longue en temps ne garantit pas un changement de bassins conformationnels. Tout dépend de l'énergie du système et des barrières énergétiques à franchir comme illustrée sur la figure 1.1 (la droite tracée en rouge horizontalement) D'où l'importance d'avoir un outil théorique robuste pour identifier les bassins explorés dans les trajectoires

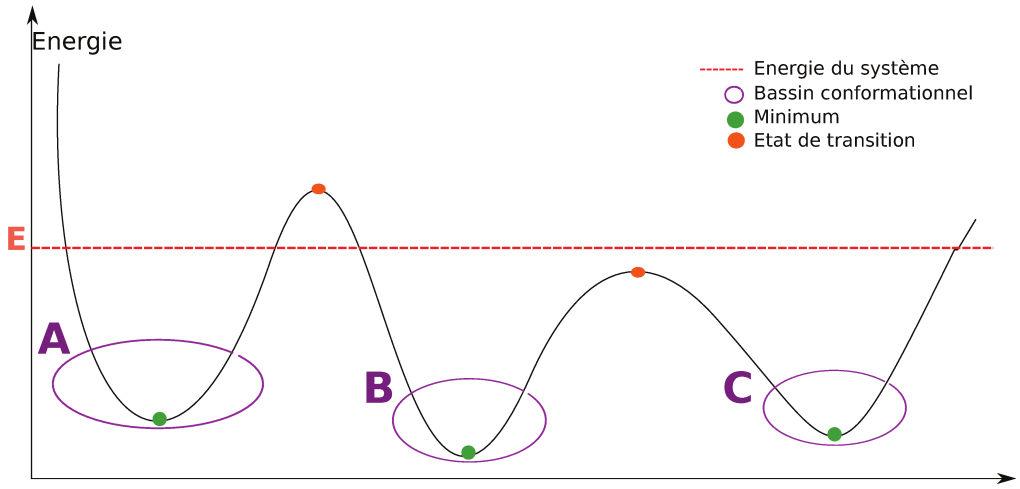

FIGURE 1.1 – Représentation schématique de la surface d'énergie potentielle.

de simulation. De nos jours, certains groupes de recherche utilisent des logiciels de visualisation qui fournissent une représentation statique tridimensionnelle du système moléculaire à chaque instant, ensuite à l'oeil, les chercheurs analysent les changements conformationnels possibles. D'autres groupes analysent les trajectoires avec leurs propres codes, qui sont généralement restreints aux systèmes étudiés (codes peu transposables). Notre but est de proposer un algorithme universel qui peut être appliqué à n'importe quel système moléculaire (cf. chapitre 3).

En plus de l'exploration de la surface d'énergie potentielle, il y a des groupes en chimie théorique et computationnelle qui s'intéressent à des points particuliers sur cette surface. Il s'agit des minima (voir les points verts sur la figure 1.1) qui représentent les conformations les plus stables et des maxima ou états de transition (voir les points oranges sur la figure 1.1) sur la surface d'énergie potentielle, qui représentent les points de passage d'un bassin conformationnel à un autre. En effet, d'une part, la conformation la plus stable dans un bassin conformationnel est celle de plus basse énergie. D'autres part, pour aller d'un bassin conformationnel à un autre il faut une énergie supplémentaire, le point maximum représente l'état de transition. Différentes méthodes ont été développées pour trouver ces points. Les méthodes de recherche des minima sont appelées **méthodes de minimisation** car elles consistent à minimiser l'énergie dans un bassin conformationnel. L'inconvénient principal de ces méthodes est qu'elles nécessitent le calcul de l'énergie potentielle et de ses dérivées, ce qui est coûteux en temps et en calculs. De même pour le calcul des états de transition. Bien que plusieurs méthodes et algorithmes ont été développés dans la littérature, le calcul de ces points reste compliqué voire impossible pour certains systèmes moléculaires, à savoir des systèmes de grande taille où des systèmes à grande flexibilité des squelettes. Le second objectif de cette thèse est de s'affranchir de la connaissance explicite de la géométrie du système moléculaire

et de son énergie pour trouver ces points (cf. chapitre 4).

Le présent chapitre a pour objectif d'introduire, en premier lieu, les différentes notions physico-chimiques qui peuvent faciliter la lecture de ce manuscrit (cf. section 1.1). Ensuite, nous présentons dans la section 1.2 quelques méthodes de simulation et d'exploration de la surface d'énergie potentielle mises en oeuvre dans la communauté de chimie théorique et computationnelle et employées par nos partenaires en théorie et modélisation du laboratoire LAMBE avec qui cette thèse a été co-réalisée. La section 1.3 présente des méthodes de minimisation et de recherche des états de transition. L'objectif de ce chapitre n'est pas de montrer toutes les méthodes qui existent dans la littérature, ni de montrer tous les détails des méthodes, mais plutôt de présenter les méthodes les plus utilisées et de mettre en évidence les caractéristiques de chacune et en voir l'intérêt (et les limites) dans le cadre des objectifs/enjeux de cette thèse.

1.1 Les systèmes moléculaires

La présente thèse est réalisée en collaboration avec une équipe de chimie théorique et computationnelle. A cet effet, les objets manipulés sont les molécules.

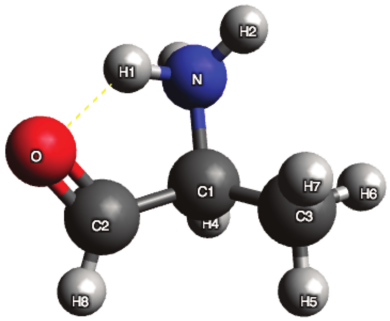

FIGURE 1.2 – Exemple d'une molécule, ici un acide aminé. Les atomes d'oxygène sont représentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique.

Définition 1 (Molécule)**.** *Une molécule est la structure de base d'un composant chimique. Il s'agit d'un ensemble d'atomes liés entre eux. La figure 1.2 montre un exemple de molécule, ici un acide aminé qui contient 4 types d'atomes : l'hydrogène (H, représenté en gris clair sur la figure), le carbone (C, représenté en gris foncé), l'oxygène (O, représenté en rouge) et l'azote (N, représenté en bleu).*

Les atomes d'une molécule sont positionnés dans l'espace et selon leurs emplacements des liaisons chimiques se créent entre eux. Il existe plusieurs types de liaisons. Dans le cadre de cette thèse, nous nous intéressons aux liaisons suivantes :

Définition 2 (Liaison covalente)**.** *C'est la liaison la plus forte qui peut être créée entre les atomes. C'est une liaison chimique dans laquelle deux atomes partagent leurs électrons.*

Remarque 1. *A chaque type d'atome (carbone, hydrogène, etc.) est associée une distance, définie comme la demi-distance d'une liaison covalente entre cet atome et un autre de même type. Cette distance est appelée* ***rayon de covalence*** *et sera utilisée dans le calcul des liaisons covalentes [6].*

De plus, chaque atome est caractérisé par un nombre maximal de liaisons covalentes qui peut former avec d'autre atomes. Ce nombre dépend du nombre d'électrons (les électrons contenus dans la couche de valence) que cet atome possède. Ce nombre varie entre 0 *et* 4 [1].

Définition 3 (Liaison hydrogène)**.** *C'est une liaison faible en comparaison des liaisons covalentes. C'est un type de liaison chimique où un atome d'hydrogène lié par covalence à un atome électronégatif comme l'oxygène ou l'azote (qu'on appelle atome* donneur*) forme une liaison intermoléculaire avec un autre atome électronégatif (qu'on appelle atome* accepteur*).*

Définition 4 (Interactions intermoléculaires électrostatiques)**.** *C'est une interaction forte qui fait intervenir l'attraction ou la répulsion entre atomes chargés négativement (type cation* [2]*) ou atomes chargés positivement (type anion* [3]*).*

En plus des propriétés chimiques en termes de liaisons, la molécule possède une énergie potentielle qui reflète les liaisons covalentes formées et les liaisons hydrogène et interactions présentes. Cette énergie est formellement définie par la résolution de l'équation de Schrödinger (indépendante du temps, ici seulement considérée) : $H\Psi = E\Psi$ où E représente l'énergie potentielle, H l'Hamiltonien électronique du système et Ψ une fonction d'onde électronique du système [2]. Le calcul et l'utilisation de cette énergie sont détaillés dans les sections suivantes.

Définition 5 (Conformation)**.** *Une conformation d'une molécule est la manière dont ses atomes sont positionnés dans l'espace et les liaisons (covalentes, hydrogène et électrostatiques) créés entre eux.*

Dans ce qui suit nous utilisons le terme **système moléculaire** au lieu de molécule. Il s'agit d'un ensemble d'atomes qui peuvent former une ou plusieurs molécules. Toutes les définitions données ci-dessus pour une molécule restent valables pour un système moléculaire.

1. Voir sur le tableau périodique https://fr.wikipedia.org/wiki/Tableau_périodique_des_éléments
2. Liste des cations disponible sur : https://fr.wikipedia.org/wiki/Cation
3. Liste des anions disponible sur : https://fr.wikipedia.org/wiki/Anion

1.1.1 Systèmes moléculaires étudiés

Dans le cadre de cette thèse, nous nous intéressons en particulier aux peptides en phase gazeuse.

Définition 6 (peptides)**.** *Un peptide est un enchainement d'acides aminés reliés entre eux par des liaisons peptidiques (une liaison covalente). Les acides aminés sont appelés également "building blocks". Un peptide peut aussi être appelé* ***polymère*** *d'acides aminés.*

La figure 1.3 montre la formule générale d'un acide aminé. Un acide aminé comporte un groupe carboxyle COOH, un groupe amine, par exemple une amine primaire NH_2 et d'un résidu noté R. Chaque acide aminé a un résidu R différent qui lui confère une propriété physicochimique particulière et ce qui fait la différence entre les acides aminés. Il existe 22 acides aminés naturels[4]. Grâce à l'enchainement de ces acides aminés plus ou moins chargés, les polymères des peptides ont la propriété de se replier et d'adopter différentes conformations qui conditionnent en partie leurs fonctions biologiques.

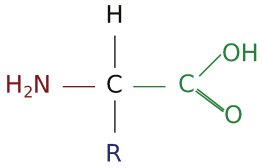

FIGURE 1.3 – Structure d'un acide aminé (building block). Ici représenté dans sa forme neutre (avec terminaisons NH2 et COOH).

Les peptides peuvent également être nommés selon le nombre d'acides aminés portés dans la chaine, à savoir dipeptide pour ceux comportant deux acides aminés, tripeptide pour trois acides aminés et ainsi de suite. Quand le nombre d'acides aminés dépasse 10, le terme polypeptide est généralement utilisé. Les protéines sont des assemblages de polypeptides. La figure 1.4 présente un exemple d'un dipeptide.

De plus, chaque peptide est caractérisé par une **terminaison-N** qui peut être NH_3^+ ou NH_2 et une **terminaison-C** qui peut être COOH ou COO^-. Ces terminaisons varient selon le système et son environnement, c'est-à-dire, s'il est présent en phase gazeuse ou aqueuse (même si des subtilités supplémentaires existent). Dans le dipeptide de la figure 1.4 la terminaison-N est NH_3^+ et la terminaison-C est COOH.

Pour les développements et les tests réalisés au cours de cette thèse, nous avons choisi deux types de systèmes : i) les systèmes moléculaires contenant un peptide isolé avec l'objectif d'analyser la dynamique des liaisons hydrogène présentes dans

4. Liste des acides aminés est disponible sur le lien https://fr.wikipedia.org/wiki/Acide_aminé

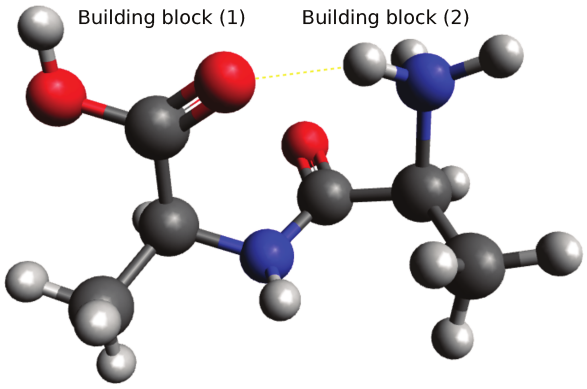

FIGURE 1.4 – Exemple d'un dipeptide. Les atomes d'oxygène sont représentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique.

le système et ii) des systèmes moléculaires contenant un peptide avec un gaz inerte où l'objectif est d'examiner les fragments engendrés par la collision entre ce gaz et le peptide, et de suivre leur évolution au cours du temps.

En plus des peptides, nous nous intéressons à des clusters, en particulier les clusters d'eau en interaction avec un cation.

Définition 7 (Cluster d'eau)**.** *Un cluster d'eau est un ensemble de molécules d'eau (H_2O) liées par des liaisons hydrogène ou des interactions intermoléculaires électrostatiques dans le cas de la présence d'un ion (cation ou anion).*

La figure 1.5 montre un exemple d'un cluster $Li^+(H_2O)_4$ composé de quatre molécules d'eau H_2O et d'un atome de lithium Li^+ au centre. Dans cet exemple, on voit que le cation Li^+ est entouré directement par trois molécules d'eau (ce qui est appelé sphère de solvatation) et qu'une quatrième molécule d'eau est liée au reste du cluster par des liaisons hydrogène (représentées en pointillés).

Dans ce genre de systèmes, la stabilisation de la structure (conformation) résulte principalement de l'interaction entre l'ion du milieu et les molécules autour de lui (molécules d'eau par exemple).

1.2 Problématique (I) : analyse de la dynamique conformationnelle

Les propriétés physiques et chimiques d'un système moléculaire donné dépendent principalement de la conformation qu'il adopte. Cette dernière n'est pas unique, elle peut changer au cours du temps. Pour identifier ces conformations et analyser leur dynamique au cours du temps, des simulations basées sur des expériences informatiques sont utilisées. De nos jours, de nombreuses techniques sont

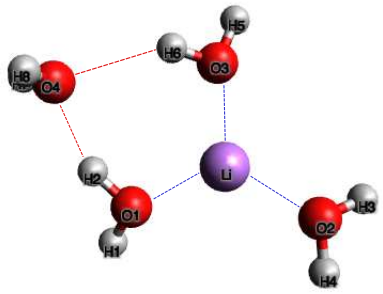

FIGURE 1.5 – Exemple d'un cluster d'eau/lithium. Les atomes d'oxygène sont représentés en rouge, les atomes d'hydrogène en gris clair et le lithium en violet. L'étiquette attribuée à chaque sommet est son type chimique. Les interactions intermoléculaires électrostatiques entre le lithium Li^+ et les molécules d'eau sont représentées en pointillés bleus et liaisons hydrogène en pointillés rouges.

disponibles, parmi lesquelles les méthodes déterministes comme la dynamique moléculaire [7, 8, 9] et les méthodes stochastiques comme le Monte Carlo [10, 2]. Ces différentes approches théoriques peuvent donner une vue globale des conformations que peut adopter un système moléculaire, c'est ce que l'on appelle l'espace des conformations. Dans ce qui suit, nous ne discutons que la dynamique moléculaire. Nous présentons en premier lieu, les principes de la dynamique moléculaire et les trajectoires issues de ses simulations. Ensuite, nous présentons les méthodes et les logiciels utilisés pour identifier les conformations explorées dans les simulations et les difficultés rencontrées par ces méthodes.

1.2.1 La dynamique moléculaire

La dynamique moléculaire comme son nom l'indique consiste à simuler l'évolution temporelle (dynamique) d'un système moléculaire [7, 8, 9]. En effet, selon les positions des atomes, les liaisons et les interactions formées peuvent changer et donc la conformation du système change. Le principe de la dynamique moléculaire consiste à générer des trajectoires d'un ensemble fini d'atomes, constituant un système moléculaire, en intégrant de façon numérique des équations classiques des mouvements. Il s'agit de calculer les positions, dans l'espace, et les vitesses des atomes au cours du temps à une énergie donnée en résolvant numériquement un ensemble d'équations.

Dans un système moléculaire à N atomes, une trajectoire est générée à travers l'équation de Newton du mouvement, qui pour chaque atome i, s'écrit :

$$\vec{F}_i = \sum_{j \neq i}^{N} (\vec{F}_{ext})_{j \to i} = m_i \vec{a}_i = m_i \frac{d^2 \vec{r}_i}{dt^2} = \frac{d\vec{p}_i}{dt}$$

Où :

- $\vec{F_i}$: la somme des forces exercées par les atomes du système moléculaire, sur l'atome i
- m_i : la masse de l'atome i
- $\vec{a}_i$: l'accélération de l'atome i
- $\vec{r}_i$: la position de l'atome i
- $\vec{p}_i$: la quantité de mouvement de l'atome i, $\vec{p}_i = m_i\vec{v}_i$
- $\vec{a}_i = \frac{d^2\vec{r}_i}{dt^2} = \frac{d\vec{v}_i}{dt}$
- $\vec{v}_i = \frac{\vec{p}_i}{m_i}$: vitesse de l'atome i

La force exercée par l'atome j sur l'atome i est égale à l'opposé du gradient du potentiel d'interaction V_{ij} entre les deux atomes i et j :

$$(\vec{F}_{ext})_{j\to i} = -\vec{\nabla}_i V_{ij} = \begin{vmatrix} \frac{-\partial V_{ij}}{\partial x_i} \\ \frac{-\partial V_{ij}}{\partial y_i} \\ \frac{-\partial V_{ij}}{\partial z_i} \end{vmatrix}, \text{ en coordonnées cartésiennes.}$$

La force totale s'exerçant sur l'atome i est :

$$\vec{F_i} = \sum_{j\neq i}^{N} (\vec{F}_{ext})_{j\to i} = \begin{vmatrix} \sum_{j\neq i}^{N} \frac{-\partial V_{ij}}{\partial x_i} \\ \sum_{j\neq i}^{N} \frac{-\partial V_{ij}}{\partial y_i} \\ \sum_{j\neq i}^{N} \frac{-\partial V_{ij}}{\partial z_i} \end{vmatrix} = \begin{vmatrix} \frac{-\partial V}{\partial x_i} \\ \frac{-\partial V}{\partial y_i} \\ \frac{-\partial V}{\partial z_i} \end{vmatrix}, \text{ par convention d'écriture.}$$

Où :

- V est le potentiel d'interaction entre l'atome i et l'ensemble des atomes du système. Il s'agit de l'énergie potentielle du système que nous notons E et qui représente la clé des calculs moléculaires.

En mécanique quantique (la dynamique ab initio par exemple), l'énergie potentielle E est parfaitement définie et résulte de la résolution de l'équation de Schrödinger indépendante du temps :

$$H\Psi = E\Psi$$

L'étude complète d'un système moléculaire nécessite en principe la résolution de l'équation de Schrödinger dépendante du temps, pour un ensemble important d'électrons et de noyaux. Cette équation est dans la grande majorité des cas trop compliquée. Dans la pratique et dans une approximation appelée **Born-Oppenheimer** [2], on résout l'équation de Schrödinger indépendante du temps qui consiste à résoudre l'équation de Schrödinger à positions fixées des atomes :

Où :

- H : le Hamiltonien électronique du système composé de N atomes, $H = -\sum_{i}^{e^-} \frac{\Delta_i}{2} + \sum_{i}^{e^-}\sum_{j}^{e^-} \frac{1}{2}\frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{i}^{e^-}\sum_{k}^{noyaux} \frac{Z_k}{|\vec{r}_i - \vec{R}_k|}$
- Ψ : la fonction d'onde électronique du système
- E : l'énergie potentielle du système (notée précédemment V). Ces deux notations sont indifféremment utilisées par la communauté des théoriciens.

La dynamique ab initio donne des résultats très précis, cependant, elle est coûteuse en termes de temps et de calculs.

Une alternative à la dynamique ab initio est la dynamique classique. Il s'agit de méthodes moins précises mais moins coûteuses computationnellement.

Une expression analytique approximée de l'énergie potentielle est alors employée et regroupe les interactions intra- et inter-moléculaires du système moléculaire. Ces méthodes sont appelées également "champs de forces" [2]. L'énergie potentielle entre les atomes du système moléculaire est calculée en faisant la somme d'une série de fonctions de potentiel.

Malheureusement dans la dynamique classique, l'expression de l'énergie potentielle E n'est pas unique et dépend du champ de forces utilisé. Parmi les expressions les plus utilisées, l'expression suivante :

$$E = \sum_{liaisons} \frac{k_i}{2}\left(l_i - l_{i0}\right)^2 + \sum_{angles} \frac{k_i}{2}\left(\theta_i - \theta_{i0}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left(1 + \cos\left(n\omega - \gamma\right)\right) +$$

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{k_i}{2}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right)$$

Dans cette expression, le premier terme modélise l'interaction intra-moléculaire entre les paires d'atomes qui sont liés par une liaison covalente où k_i et l_{i0} représentent respectivement la constante de force de la liaison et la longueur de la liaison à l'équilibre. Le deuxième terme modélise les interactions intra-moléculaires en termes d'angles, entre chaque triplet d'atomes (a, b, c) telle que a et c sont liés à b. k_i

et θ_{i0} représentent également la constante de force angulaire et l'angle de valence à l'équilibre. Le troisième terme présente les rotations autour des liaisons, où $\frac{V_n}{2}$, n, ω, γ représentent la barrière de torsion, sa périodicité et sa phase respectivement. Enfin le quatrième terme présente les interactions inter-moléculaires entre les atomes qui n'ont pas des liaisons covalentes en commun (les interactions van der Waals et électrostatiques) [2]. La figure 1.6 montre une représentation schématique des quatre termes présents dans l'expression du champ de forces ci-dessus.

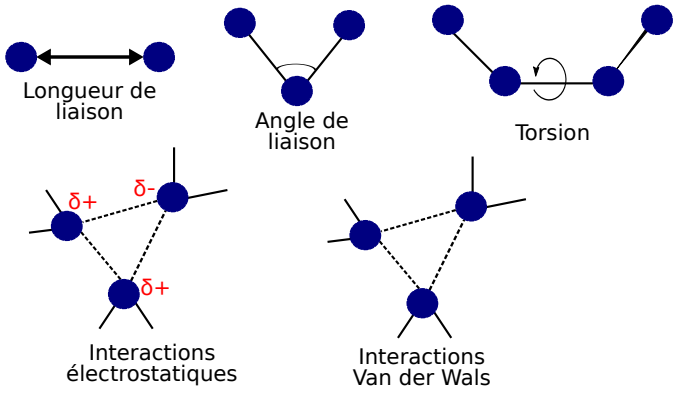

FIGURE 1.6 – Représentation schématique des termes présents dans l'expression du champs de forces. Les interactions intra-moléculaires sont les 3 en haut de la figure et les interactions inter-moléculaires sont les 2 en bas de la figure [2].

Les trajectoires analysées par nos algorithmes sont générées en utilisant la première méthode, c'est-à-dire la dynamique ab initio.

Une fois l'énergie potentielle E calculée, cela permet d'avoir $\vec{\nabla}_i E$ et donc les forces $\vec{F}_i$ qui s'exercent sur chaque atome i. A ce niveau, il est possible de résoudre les équations du mouvement en fonction du temps et trouver les positions et vitesses des atomes.

Pour atteindre cet objectif, des algorithmes numériques de résolution sont employés. L'algorithme le plus employé est celui de Verlet [5] [11] basé sur les développements de Taylor.

A l'instant t, chaque atome i est caractérisé par sa position $\vec{r}_i(t)$, sa vitesse $\vec{v}_i(t)$ et $\vec{F}_i(t)$ la force totale qui s'exerce sur cet atome. A l'instant suivant, $t+\delta t$, l'algorithme de Verlet produit les positions $\vec{r}_i(t+\delta t)$ et les vitesses $\vec{v}_i(t+\delta t)$ comme suit :

$$\begin{cases} \vec{r}_i(t+\delta t) = \vec{r}_i(t) + \vec{v}_i(t)\delta t + \dfrac{\vec{F}_i(t)}{2m_i}\delta t^2 \\ \vec{v}_i(t+\delta t) = \vec{v}_i(t) + \dfrac{\vec{F}_i(t) + \vec{F}_i(t+\delta t)}{2m_i}\delta t \end{cases}$$

Nous remarquons que la nouvelle position $\vec{r}_i(t+\delta t)$ dépend de la position à l'instant précédent $\vec{r}_i(t)$, de la vitesse à l'instant précédent $\vec{v}_i(t)$ et de la force à l'ins-

5. Voir sur : http://www.cinam.univ-mrs.fr/klein/teach/mip/numeriq/node44.html

tant précédent $\vec{F_i}(t)$. La nouvelle vitesse $\vec{v_i}(t+\delta t)$ dépend de la vitesse à l'instant précédent $\vec{v_i}(t)$ mais aussi de la force au nouvel instant $\vec{F_i}(t+\delta t)$. D'où la particularité de la dynamique moléculaire, contrairement aux méthodes stochastiques, la génération de la conformation à l'instant $t+\delta t$ dépend de la conformation générée à l'instant précédent t.

La génération d'une trajectoire peut être résumée comme suit :

1. A l'instant $t=0$, initialiser $\vec{r_i}(0), \vec{v_i}(0)$ pour chaque atome.
2. Résoudre numériquement $H\Psi = E\Psi$ ce qui donne E et Ψ (le Hamiltonien H est connu).
3. Obtenir les forces $\vec{F_i}$ exercées sur chaque atome i en utilisant l'énergie E, juste obtenue en 2.
4. Appliquer l'algorithme de Verlet pour calculer les nouvelles positions $\vec{r_i}(t+\delta t)$ et les nouvelles vitesses $\vec{v_i}(t+\delta t)$ de chaque atome.
5. Continuer la trajectoire au temps $t = t+\delta t$ et revenir à l'étape (2)
6. Si $t = \Delta t$, la durée de simulation voulue est atteinte, on stoppe la simulation.

En plus de l'énergie potentielle, le système moléculaire est caractérisé par deux autres énergies :

- L'énergie cinétique : $E_{Cin} = \sum_{i=1}^{n} \frac{1}{2} m_i v_i^2$, où $\frac{1}{2} m_i v_i^2$ est l'énergie cinétique de chaque atome du système moléculaire et $v_i^2 = ||\vec{v_i}||^2$
- L'énergie totale E_T : cette énergie est la somme de l'énergie potentielle E_{Pot} et de l'énergie cinétique E_{Cin}.

$$E_T = E_{Pot} + E_{Cin}$$

En termes de coûts computationnels, la résolution de l'équation du mouvement de Newton n'est pas coûteuse contrairement à la résolution de l'équation de Schrödinger qui est complexe. La génération d'une trajectoire de dynamique moléculaire ab initio prend en général de quelques jours (trajectoires des molécules isolées) à quelques semaines (trajectoires d'interfaces solide-liquide). Ces trajectoires sont employées en routine dans le groupe de théorie et modélisation du LAMBE avec qui cette thèse a été co-réalisée.

Les trajectoires ainsi déterminées sont utilisées pour étudier la structure moléculaire, à savoir la dynamique conformationnelle, c'est-à-dire l'identification des conformations d'un système moléculaire dans une période de temps et avoir une vue sur la conformation moyenne. Dans les systèmes complexes comme les interfaces *solide-liquide*, les trajectoires servent à voir l'orientation des molécules d'eau, les liaisons hydrogène créées entre ces molécules, les liaisons hydrogène entre le

solide et le liquide, etc. Les trajectoires peuvent être utilisées pour d'autres objectifs comme la spectroscopie vibrationelle [12].

1.2.2 Surface d'énergie potentielle

Nous avons vu dans la section précédente que l'énergie potentielle résulte des interactions entre les atomes qui sont une fonction des positions de ces atomes. Cette énergie peut être représentée par une courbe ou surface appelée **surface d'énergie potentielle** ou **hypersurface**.

Pour un système moléculaire de N atomes, cette surface est décrite dans un espace à $3N - 6$ dimensions, en enlevant les 3 dimensions associées à des translations d'ensemble du système (selon l'axe x, ou y ou z) et les 3 dimensions associées à des rotations d'ensemble du système autour des axes x, y et z [2]. Pour une molécule constituée de 3 atomes (la molécule d'eau H_2O par exemple), l'énergie va dépendre de 3 paramètres : les 2 distances inter-atomiques de la liaison covalente $O - H$ et l'angle formé par les 3 atomes. On a donc une fonction du type $E = f(O - H, O - H, \widehat{HOH})$ à trois dimensions.

Quand le nombre d'atomes augmente, il est impossible de visualiser toute la surface d'énergie potentielle. Pour remédier à ce problème, la méthode utilisée consiste à ne visualiser qu'une partie de la surface en utilisant une seule coordonnée, par exemple la distance entre deux atomes ou un angle dièdre, ou une combinaison de paramètres etc. Cette coordonnée est appelée **coordonnée de réaction**. La figure 1.7 présente un exemple de surface d'énergie potentielle pour une molécule d'eau (H_2O) [6], en utilisant la longueur d'une liaison $O - H$ et un angle de liaison $\widehat{HOH}$ pour une représentation que nos yeux et cerveaux puissent interpréter.

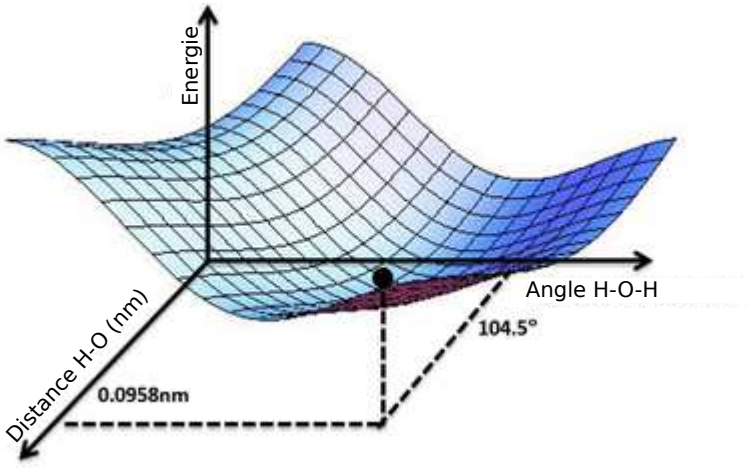

FIGURE 1.7 – Surface d'énergie potentielle d'une molécule d'eau.

6. Source : https://fr.wikipedia.org/wiki/Surface_d%27énergie_potentielle

La figure 1.8 montre un exemple schématique d'une surface d'énergie potentielle regardée en fonction d'une coordonnée de réaction choisie ici arbitrairement.

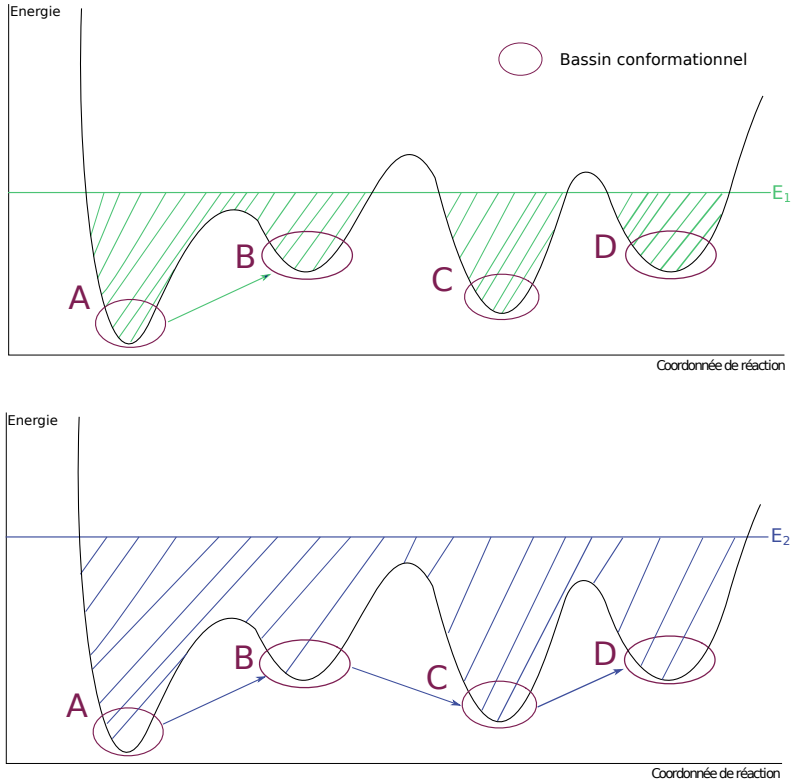

FIGURE 1.8 – Un exemple schématique d'une surface d'énergie potentielle.

Nous remarquons qu'un système peut avoir plusieurs bassins conformationnels. Chaque bassin représente une conformation, la différence entre les structures du même bassin est un changement léger de positions d'atomes (vibration d'atomes). Le but des simulations de dynamique moléculaire est d'avoir une bonne exploration de la surface d'énergie potentielle, c'est-à-dire d'explorer le maximum de bassins différents pour un système donné. Selon l'énergie fixée dans le système (E dans la figure 1.8 avec un une droite tracée horizontalement à cette valeur) et le temps de la simulation nous pouvons avoir des trajectoires explorant un seul bassin ou plusieurs. En effet, à basse énergie (température), on explore potentiellement des bassins séparément les uns des autres et très peu les transitions, comme montré dans la figure 1.8 par le niveau E_1. Les trajectoires vont donc avoir très peu de dynamique conformationnelle. Cependant, à haute énergie (température), on pourra explorer plusieurs bassins dans une même trajectoire. Cela dépendra également du temps de simulation. Considérons par exemple le niveau E_2 de la figure 1.8, si la trajectoire est assez longue, nous pourrons explorer les conformations A, B, C et D. Les énergies fixées dans le système pour passer d'un bassin à un autre sont appelées aussi **barrières d'énergie**.

En général, à une énergie E fixée, on ne peut explorer que les zones hachurées sur la figure 1.8 de la surface d'énergie potentielle.

1.2.3 La dynamique conformationnelle explorée

Une fois les trajectoires obtenues, la deuxième étape consiste à les analyser pour identifier les bassins conformationnels explorés et les durées d'exploration de chacun. La génération de plusieurs positions au cours du temps n'implique pas nécessairement des changements conformationnels. A une énergie (température) donnée, tous les atomes de la molécule bougent, ce mouvement est soit une simple vibration, c'est-à-dire on reste dans le même bassin conformationnel (même conformation), ou bien il entraine un changement de liaisons et dans ce cas on dit qu'on a un changement de bassin (changement de conformation).

Pour analyser les trajectoires et identifier les conformations explorées, il existe deux grandes classes de méthodes utilisées à présent :

- Outils d'analyse statique et de visualisation dynamique.
- Algorithmes d'analyse d'évolution des structures moléculaires.

Outils d'analyse statique et de visualisation dynamique

Un ensemble de logiciels comme **Avogadro** [13], **VMD** [14] et des logiciels en libre accès ont été développés pour l'analyse et la visualisation des structures moléculaires. Ils fournissent des représentations tridimensionnelles statiques à chaque instant de la trajectoire. Ils permettent de donner les liaisons covalentes de la molécule avec les distances et les angles entre les différents atomes. Ils permettent également de visualiser la molécule tout au long de la trajectoire sous forme de succession d'images. Cependant, ils ne donnent aucune information statistique sur la molécule ou sur la dynamique conformationnelle. A titre d'exemple, l'utilisation d'un tel logiciel ne permet pas de décider si la conformation à l'instant t est la même que celle observée dans un instant précédent. L'analyse de la dynamique conformationnelle est donc effectuée manuellement. Quand il s'agit d'un système et d'une trajectoire de petites tailles, l'analyse de la dynamique se fait à l'oeil. Dans d'autres cas, des codes personnalisés, généralement destinés à un seul type de systèmes moléculaires, sont utilisés pour analyser les changements de conformations. L'utilisation de ces méthodes pour analyser une seule trajectoire peut prendre jusqu'à plusieurs jours voire quelques semaines.

Algorithmes d'analyse d'évolution des structures moléculaires

A coté des logiciels, des algorithmes ont été développés pour automatiser l'analyse de la dynamique conformationnelle à partir des trajectoires. Récemment, deux algorithmes ont été proposés : MoleculaRnetworks [15] et ChemNetworks [16]. Ils

sont basés sur Google PageRank [17] dont le principe est de donner une classification des objets selon sa popularité. Le PageRank est utilisé également car il permet de donner une solution unique pour les arrangements polyédriques d'atomes. Les résultats publiés montrent que les deux algorithmes sont efficaces. Cependant, ils sont restreints aux systèmes de clusters (d'après les articles publiés) et ils utilisent une base de données externe pour comparer les résultats trouvés [17].

1.2.4 Limites des méthodes existantes

L'analyse des trajectoires avec les logiciels de visualisation dynamique permet de donner une vision globale sur les conformations en donnant une représentation statique tridimensionnelle du système moléculaire à chaque instant de la trajectoire. L'analyse de la dynamique conformationnelle est assurée, d'une part, par l'oeil, ce qui est faisable que pour les systèmes et les trajectoires de petites tailles. Cette analyse donne des résultats peu précis et parfois erronés. D'autres part, certains groupes analysent cette dynamique avec leurs propres codes qui sont généralement non publiés. Le problème majeur de ces codes est qu'ils sont généralement restreints aux systèmes étudiés (codes peu transposables) et ne sont utilisés que par les personnes qui les ont développés. Par exemple, au sein de l'équipe de théorie et modélisation du LAMBE avec qui cette thèse a été co-réalisée, il y a trois groupes qui étudient trois types de systèmes différents et de complexités différentes. Ces groupes utilisent des codes différents pour un même objectif qui est l'analyse de la dynamique conformationnelle.

De même pour les algorithmes développés, le problème est (selon les articles publiés) qu'ils sont également restreints à des types particuliers de systèmes moléculaires. De plus, certains algorithmes comme le MoleculaRnetworks utilisent une base de données externe pour trouver les résultats finaux.

Notre objectif est donc de s'inspirer des méthodes existantes et d'utiliser les concepts de théorie des graphes pour proposer un algorithme universel pour l'analyse des trajectoires (sans faire appel à des ressources externes). Un algorithme qui peut être appliqué à n'importe quel système moléculaire et qui peut être utilisé par plusieurs groupes de la communauté de chimie théorique et computationnelle. Cet algorithme fera l'objet du chapitre 3.

1.3 Problématique (II) : prédiction conformationnelle

Ce que l'on vient de voir dans la section 1.2 revient à explorer l'espace des conformations, ce qui est appelé également (par abus de langage) l'exploration de la surface d'énergie potentielle. Dans la communauté de chimie théorique, il y a une deuxième sous communauté qui ne s'intéresse qu'à des points particuliers sur

la surface d'énergie potentielle. Ces points sont de deux types : des minima sur la surface d'énergie potentielle (voir les points en vert $m1$, $m2$ et $m3$ sur la figure 1.9) et des états de transition (voir les points en orange $ET1$ et $ET2$ sur la figure 1.9). Ces points sont particuliers car les minima représentent les conformations les plus stables (conformations de plus basses énergies) et les états de transition représentent les passage d'un bassin conformationnel à un autre.

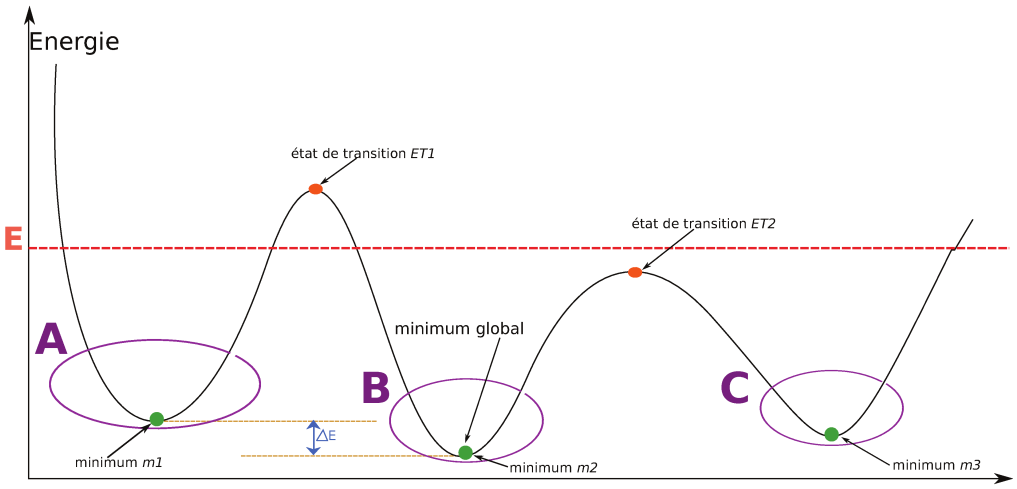

FIGURE 1.9 – Représentation schématique des points caractéristiques sur la surface d'énergie potentielle d'un système moléculaire donné.

Une simulation de dynamique moléculaire contenant plusieurs bassins conformationnels va contenir plusieurs minima ainsi que plusieurs états de transitions, comme schématisé dans la figure 1.9.

Un des challenges de la chimie théorique et computationnelle est de trouver ces points sur une surface de dimension $3N - 6$ en se basant sur des approches mathématiques.

Définition 8 (Minima et états de transition)**.** *Pour décrire les points spécifiques, à savoir les minima et les états de transition sur la surface d'énergie potentielle, la dérivée seconde de la surface par rapport aux coordonnées cartésiennes est généralement utilisée. On définit une structure correspondante à un minium d'énergie n'ayant que des dérivées secondes positives. On définit une structure correspondante à un état de transition n'ayant qu'une dérivée seconde négative. Toutes les autres étant positives.*

Nous présentons dans les sections suivantes les méthodes de recherche des minima et des états de transitions sur la surface d'énergie potentielle généralement employées dans la littérature. Nous rappelons que l'objectif de ce chapitre n'est pas de montrer toutes les méthodes qui existent dans la littérature, ni de montrer tous les détails des méthodes, mais plutôt de présenter les plus réputées et employées

dans la communauté de chimie théorique et computationnelle et par nos partenaires théoriciens de cette thèse et de mettre en évidence les caractéristiques de chacune.

1.3.1 Méthodes de calcul des minima sur la surface d'énergie potentielle

La minimisation est une étape primordiale en modélisation moléculaire. Celle-ci permet de donner les conformations les plus stables du système moléculaire.

Etant donné une surface d'énergie potentielle, le principe de la minimisation est de prendre un point au hasard de cette surface qu'on appelle point de départ (cela correspond à une structure moléculaire générée par la simulation par exemple) puis de déplacer les atomes afin d'obtenir la conformation la plus stable possible (qui possède une énergie plus basse que le point de départ). Comme son nom l'indique, la minimisation va déplacer les atomes afin de diminuer au fur et à mesure l'énergie de la molécule jusqu'à obtenir une énergie la plus basse, on finira alors par se trouver au fond d'un bassin sur la surface d'énergie potentielle. La figure 1.10 schématise le principe de la minimisation. La difficulté est de trouver la bonne direction à emprunter sur la surface d'énergie, afin de produire la conformation ayant l'énergie la plus basse possible. Pour cela dans la minimisation, plusieurs méthodes sont utilisées d'une façon hiérarchique.

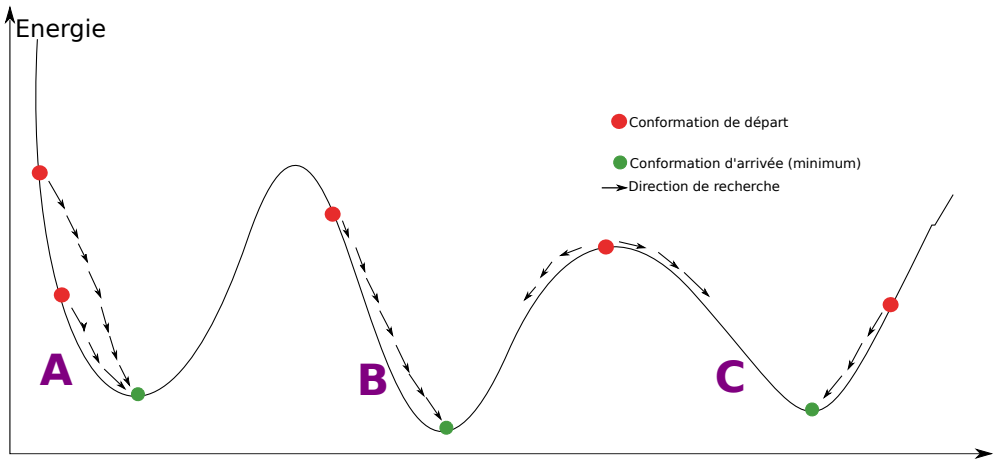

FIGURE 1.10 – Un exemple schématique du principe de la minimisation.

Comme dans les simulations, il existe des méthodes de minimisation déterministes et d'autres stochastiques. Parmi les méthodes stochastiques les plus rencontrées : le recuit simulé et les algorithmes génétiques [18, 19]. Ces méthodes sont dites des méthodes de recherche globale, car elles démarrent de plusieurs points de départ pour arriver ensuite à plusieurs minima. Elles utilisent un composant aléatoire, ce qui rend facile de dire qu'elles sont très inefficaces par rapport aux

méthodes déterministes qui effectuent une minimisation dans un puit local.

Dans le cadre de cette thèse, nous nous intéressons aux méthodes déterministes. Ce sont des méthodes dites de recherche locale car contrairement aux méthodes stochastiques elles démarrent d'un seul point de départ pour trouver un minimum qui est le plus proche du point de départ. Elles regroupent principalement les méthodes de gradient et les méthodes de recherche directe. Les méthodes de gradient comme le steepest-descent, le gradient conjugué et le Newton-Rhapson [3] sont intrinsèquement locales parce qu'elles suivent un chemin continu depuis leur point de départ jusqu'à ce qu'un point le plus bas possible soit atteint [19].

Le steepest-descent et le gradient conjugué reposent principalement sur la dérivée première de l'énergie potentielle seulement. L'avantage de ces méthodes est qu'elles acceptent un point de départ se situant loin du minimum. Cependant, l'inconvénient du steepest-descent est que les directions de recherche ne sont pas optimisées, ce qui rend la convergence vers le minimum plus difficile. Cette méthode est en général utilisée comme première étape de minimisation avant de passer à une autre méthode plus performante. Le gradient conjugué ressemble au steepest-descent, sauf qu'il utilise un autre principe dans les directions de recherche, ce qui les rend plus optimisées. Dans le gradient conjugué, le minimum est trouvé en un nombre d'étapes réduit par rapport au steepest-descent, cependant il y a plus d'étapes que les méthodes à dérivées secondes.

Dans le Newton-Rhapson, en plus de la dérivée première de l'énergie potentielle, la méthode utilise également des informations provenant de la dérivée seconde. La méthode repose sur des directions de recherche optimisées ce qui permet de trouver le minimum en très peu d'étapes. Cependant, les calculs sont plus complexes que le steepest-descent et le gradient conjugué et ils n'acceptent pas des points initiaux qui sont loin du minimum. Une nouvelle méthode a été développée récemment, la L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) [20]. Elle utilise moins de calculs que Newton-Raphson, cependant, elle est moins précise et plus lente que le steepest-descent et le gradient conjugué.

Les méthodes de recherche directe n'utilisent pas les dérivées contrairement aux méthodes de gradient. Elles sont connues d'être des méthodes rapides et précises par rapport aux méthodes stochastiques. Les méthodes de recherche directe peuvent être divisées en trois groupes : les méthodes de recherches par motifs généralisés comme Hooke et Jeeves (pattern search) [21]. La spécificité de ces méthodes est que les directions de recherche ne changent pas avec les itérations de l'algorithme. Le deuxième groupe est celui des méthodes à directions conjuguées (algorithme de Powell et ses variantes [21]). Comme son nom l'indique, la méthode appartenant à ce groupe réalise des minimisations unidimensionnelles suivant des directions conjuguées. Enfin, le dernier groupe est celui des méthodes basées sur la

figure géométrique d'un simplexe comme la méthode de Nelder-Mead [21, 22]. Ce type de méthodes est facile à mettre en oeuvre, cependant, elle est peu performante quand le système est de grande taille.

Pour améliorer la recherche des minima, des méthodes hybrides ont été développées qui associent des méthodes de recherche globale à des méthodes de recherche locale. La façon la plus simple de réaliser cette association est d'effectuer les recherches en série, c'est-à-dire qu'une optimisation globale à coût limitée est d'abord exécutée, ensuite cette solution est raffinée par une recherche locale. Il s'agit des méthodes appelées de recherche multidirectionnelle. Ce sont des méthodes adaptées pour être efficaces en calcul sur machines parallèles et possèdent des propriétés de convergence [23].

1.3.2 Méthodes de calculs des états de transition sur la surface d'énergie potentielle

De même que pour trouver les minima sur la surface d'énergie potentielle, il existe des algorithmes spécifiques pour trouver les états de transition. Généralement, le principe est de choisir intuitivement une géométrie qui pourrait être un état de transition, ensuite appliquer un algorithme comme Berny [24], EF (eigenvector following) [25] ou la méthode dimer [26] pour optimiser l'état de départ et trouver un état de transition final. Ces méthodes sont appelées *single-ended methods* car elles dépendent seulement de la structure de départ. L'autre type de méthodes est appelée *double-ended methods*, en plus de la structure de départ, les algorithmes appartenant à cette classe nécessitent une connaissance de la structure finale. Parmi ces méthodes the linear synchronous transit method [27], the nudged elastic band (NEB) method [28] et the string method [29] ont été développées. De nos jours, il n'existe aucune méthode automatique qui permet de trouver les états de transition sans connaissance de la conformation de départ et/ou la conformation finale. La figure 1.11 schématise le principe général pour trouver les états de transition avec les deux classes.

Récemment, une nouvelle méthode a été développée par Martinez-Nunez [30] pour trouver les états de transition avec une recherche directe sur les trajectoires de simulations de dynamique moléculaire, en utilisant la méthode TSSCDS (Transition State Search using Chemical Dynamics Simulations) [5] itérativement. Le principe est de prendre une trajectoire de simulation de dynamique moléculaire initialisée à un minimum donné. La méthode TSSCDS est appliquée sur cette trajectoire. Des états de transition sont calculés ainsi que des nouveaux minima sont obtenus. Ces minima sont utilisés pour générer d'autres trajectoires. La méthode TSSCDS est appliquée sur ces nouvelles trajectoires, ce qui engendre d'autres états de transition et d'autres minima. Et ainsi de suite, jusqu'à qu'il n'y ait plus de nouveaux états de

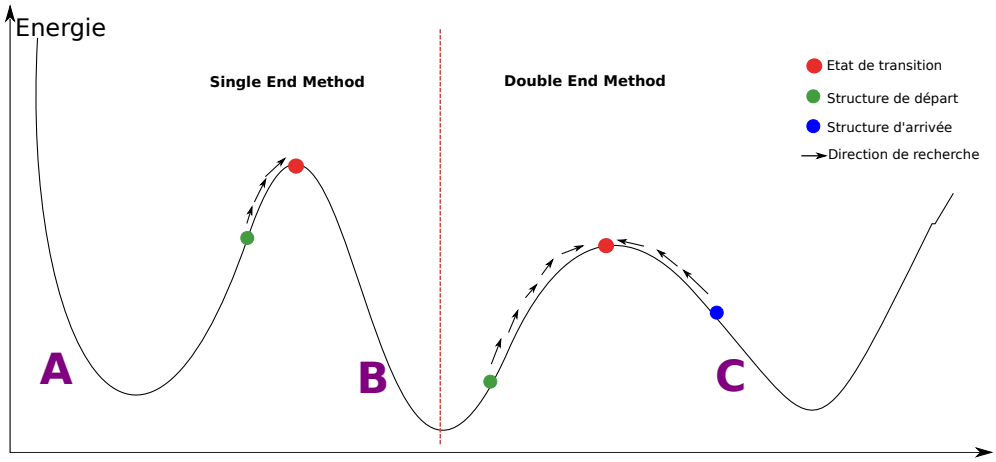

FIGURE 1.11 – Un exemple schématique du principe de calcul des états de transition. La figure à gauche montre le principe des méthodes dites *single-ended methods* et la figure à droite les méthodes dites *double-ended methods*

transition et minima qui apparaissent. Par exemple, dans un système moléculaire contenant un peptide de 8 atomes avec un gaz inerte où l'objectif est d'examiner les fragments engendrés par la collision entre ce gaz et le peptide, cette méthode a donné 66 minima et 276 états de transition [30].

La méthode TSSCDS se divise en deux étapes principales : la première consiste à générer des structures sur la surface d'énergie potentielle en utilisant la dynamique ab initio. Ensuite, l'algorithme BBFS (bond breaking/forming search) [5] est appliqué. Le BBFS permet de choisir des structures qui s'approchent à des états de transition. Il se base sur la géométrie des structures obtenues dans la trajectoire de simulation. L'algorithme calcule les distances entre les atomes et définit les liaisons qui se cassent et se forment. Pour décider qu'il y a eu un changement conformationnel, BBFS utilise des distances normalisées entre les paires d'atomes. Ensuite pour trouver les états de transition finaux, l'algorithme EF (Eigenvector Following) est utilisé [5].

Les états de transitions sont plus compliqués à calculer en comparaison avec les minima sur la surface d'énergie potentielle. Les méthodes et les algorithmes développés ne sont facilement applicables et à moindre coûts computationnels qu'aux systèmes de petites tailles (une dizaine d'atomes typiquement) et ils ne garantissent pas de bons résultats tout le temps.

1.3.3 Limites des méthodes existantes

Le développement des méthodes et des algorithmes pour trouver les points caractéristiques sur une surface d'énergie potentielle d'un système moléculaire ne cesse de progresser.

Le point commun de ces méthodes est qu'elles nécessitent toutes le calcul de

l'énergie potentielle, ce qui est coûteux en temps et en calculs comme nous l'avons vu dans les sections 1.2.1 et 1.2.2.

De plus, une difficulté des méthodes de minimisation ou de calcul des états de transition réside dans le choix des points de départ. Dans la minimisation, les points initiaux des méthodes de recherche locale sont fixés généralement à travers les méthodes de regroupement (clustering ou méthodes aléatoires [4]). Les méthodes de clustering sont des méthodes qui permettent de donner des points de départ pertinents, cependant elles sont peu performantes pour les fonctions ayant de nombreux minima. D'un autre coté, l'inconvénient des méthodes aléatoires est que les points pris sont susceptibles de converger plusieurs fois vers les mêmes minima. De même dans le calcul des états de transition, les points de départ sont fixés à travers des algorithmes d'optimisation qui ne sont pas toujours efficaces. Dans le cas d'un mauvais choix, de grosses erreurs peuvent être obtenues concernant les structures trouvées.

Un objectif de cette thèse est donc, d'une part, d'analyser les bassins explorés par les simulations de dynamique moléculaire et de choisir les points pertinents à minimiser, et d'autres part de prédire les parties non observées de la surface d'énergie potentielle en utilisant des mesures ah doc et des concepts de la théorie des graphes. L'idée de la deuxième partie de cette thèse est de s'affranchir complètement de la connaissance de l'énergie potentielle pour trouver toutes les conformations possibles d'un système moléculaire et les transitions possibles entre elles.

1.4 Conclusion

La connaissance de la structure tridimensionnelle ou la conformation d'un système moléculaire est primordiale pour déduire ses propriétés physiques et chimiques. Une conformation est déterminée non seulement par les atomes qu'elle contient, mais aussi par l'ensemble des liaisons formées entre ces atomes suite à leurs positionnements dans l'espace. Cette conformation n'est pas unique, elle peut changer au cours du temps. Un défi de la chimie théorique et computationnelle est d'identifier les conformations que peut adopter un système moléculaire au cours du temps. Un autre défi est de connaitre les bassins les plus stables en énergie.

Pour atteindre cet objectif, des simulations de dynamique moléculaire (ou éventuellement avec des méthodes aléatoires) sont effectuées. Selon l'énergie fixée dans le système moléculaire et le temps de simulation, nous pouvons explorer un ou plusieurs bassins conformationnels (conformations). Pour avoir une bonne exploration, il faut plusieurs trajectoires à des températures différentes et avec des durées longues. En termes de temps de calcul, la génération d'une trajectoire de dynamique moléculaire ab initio prend en général, de quelques jours (trajectoires des molécules isolées) à quelques semaines (trajectoires d'interfaces solide-liquide). Il

faut être capable d'identifier d'une façon rapide et efficace les bassins conformationnels explorés, la durée passée dans chaque bassin ainsi que d'identifier les changements de bassins sur les trajectoires obtenues. Les méthodes existantes pour faire cette analyse varient entre des logiciels de visualisation statique, qui nécessitent des constructions manuelles pour l'analyse de la dynamique conformationnelle, ou des algorithmes et des codes peu transposables. Ces algorithmes sont généralement restreints à un système moléculaire particulier. D'où l'intérêt et la nécessité d'un algorithme général pour analyser la dynamique conformationnelle sur les trajectoires de simulation de dynamique conformationnelle et applicable à n'importe quel système moléculaire. L'algorithme proposé et développé dans cette thèse est présenté dans le chapitre 3.

Nous avons vu également dans le présent chapitre qu'il y a une deuxième communauté qui ne cherche pas à explorer l'espace des conformations, mais cherche à trouver des points particuliers sur la surface d'énergie potentielle, qui sont les minima (les conformations les plus stables) et les états de transition (les points de passage d'un bassin conformationnel à un autre).

La difficulté des méthodes développées pour trouver ces point réside, d'un côté, dans la résolution des différentes équations qui sont longues en temps de calculs et pas toujours efficaces, et d'un autre côté, ces méthodes nécessitent une connaissance préalable de l'énergie potentielle (coûteuse en temps de calculs). Dans le cadre de cette thèse, nous proposons une méthode alternative (cf. chapitre 4) pour prédire les conformations possibles d'un système moléculaire : une méthode qui permet de s'affranchir de la connaissance de l'énergie potentielle et qui se base sur des concepts de la théorie des graphes et des mesures *ah doc*.

Avant d'entamer les différentes méthodes et algorithmes proposés dans le cadre de cette thèse, nous avons jugé nécessaire de présenter au préalable les différents modèles et règles utilisés pour la conception de ces derniers (cf. chapitre 2).

Chapitre 2

Modélisation

Dans le chapitre précédent, nous avons présenté ce qu'un système moléculaire d'un point de vue chimique, à savoir sa composition en termes d'atomes, de liaisons et d'interactions. De même les propriétés physiques et chimiques des systèmes moléculaires et leurs évolution au cours du temps. Dans ce chapitre, nous présentons d'une façon formelle les modèles proposés pour représenter ces systèmes moléculaires et leurs propriétés. Nous présentons ainsi un ensemble de règles utilisées dans la conception des solutions proposées dans le cadre de la thèse.

2.1 Caractéristiques d'un atome

Un système moléculaire est composé d'un ensemble d'atomes. Nous notons V_M cet ensemble. Chaque atome est caractérisé par son type chimique. Par exemple, dans une molécule d'eau nous trouvons de deux types chimiques : des atomes de type hydrogène et des atomes de type oxygène. L'ensemble $T = \{\mathsf{H}, \mathsf{C}, \mathsf{O}, \mathsf{N}, \mathsf{F}...\}$ représente les différents types d'atomes. Nous définissons une fonction ϕ qui associe à chaque atome son type chimique :

$$\begin{array}{rccc} \phi & : & V_M & \rightarrow & T \\ & & a & \mapsto & \phi(a) \end{array}$$

Selon le type chimique, un atome peut former une ou plusieurs liaison(s) covalente(s) avec d'autres atomes de la molécule. Cependant, chaque atome a un nombre maximal de liaisons covalentes qu'il peut former (cf. section 1.1). La fonction Ψ permet d'associer à chaque atome le nombre maximal de liaisons covalentes qui peut former :

$$\begin{array}{rccc} \Psi & : & V_M & \rightarrow & [\![0, 4]\!] \\ & & a & \mapsto & \Psi(a) \end{array}$$

Enfin, à chaque atome a est associé une distance appelée rayon de covalence (cf. section 1.1). Selon le type chimique, cette distance change. Nous définissons la fonction $covr$ qui donne pour chaque atome son rayon de covalence :

$$\begin{array}{rccc} covr & : & V_M & \rightarrow & \mathbb{R} \\ & & a & \mapsto & covr(a) \end{array}$$

Par exemple, pour un atome a d'hydrogène nous avons :

- $\phi(a) = \mathsf{H}$,
- $\Psi(a) = 4$,
- $covr(a) = 0.31$

En annexe B.1, nous présentons les caractéristiques des atomes pris en compte dans nos algorithmes.

Dans ce qui suit, nous décrivons l'utilisation de ces caractéristiques pour calculer les ensembles de liaisons créées entre les atomes et définir les conformations des systèmes moléculaires.

2.2 Modélisation d'un système moléculaire

Un système moléculaire est un ensemble d'atomes identiques ou non, unis les uns aux autres par le biais de liaisons et interactions chimiques, à savoir les liaisons covalentes, les liaisons hydrogène et les interactions intermoléculaires électrostatiques (cf. chapitre 1). Le système peut contenir une ou plusieurs molécules. Théoriquement, il peut y avoir un nombre infini de liaisons et interactions entre les atomes. Dans ce cas, nous pouvons modéliser le système moléculaire comme suit :

Définition 9 (Modélisation d'un système moléculaire théorique)**.** *Un système moléculaire théorique est un quadruplet* $Mol = (V_M, Cov, Hydro, Inter)$ *tel que :*

- V_M *: l'ensemble des atomes constituant le système moléculaire.*
- $Cov = \{[a, b], a \in V_M, b \in V_M\}$ *: l'ensemble des liaisons covalentes possibles, d'un point de vue chimique, entre ces atomes. C'est un ensemble de* ***paires*** *d'atomes (il n y pas d'ordre).*
- $Hydro \subseteq \{a \in V_M, \phi(a) \in \{\mathsf{O}, \mathsf{N}, \mathsf{F}\}\} \times \{a \in V_M, \phi(a) \in \{\mathsf{O}, \mathsf{N}, \mathsf{F}\}\}$ *: l'ensemble des liaisons hydrogène possibles, d'un point de vue chimique. C'est un ensemble de* ***couples*** *(paires ordonnées). Dans chaque couple de* $Hydro$*, le premier élément est appelé* ***donneur*** *et le deuxième élément du couple est appelé* ***accepteur*** *(cf. chapitre 1).*
- $Inter \subseteq \{[a, b,], a \in V_M, \phi(a) \in \{\mathsf{Li}, \mathsf{Ar}\}, b \in V_M\}$ *: l'ensemble des interactions intermoléculaires électrostatiques possibles, d'un point de vue chimique. C'est un ensemble de* ***paires*** *d'atomes (il n y pas d'ordre).*

Pour définir un système moléculaire réel, les atomes sont placées dans l'espace (cf. chapitre 1). L'ensemble des atomes avec leurs positions représente une image du système moléculaire.

Définition 10 (Image)**.** *Soit un système moléculaire* $Mol = (V_M, Cov, Hydro, Inter)$*. Une image du système moléculaire* Mol *est définie par le couple* $I = (Mol, P_I)$ *tel que :*

- Mol *: le système moléculaire.*
- $P_I : V_M \to \mathbb{R}^3$ *:* $\forall a \in V_M, P(a)_I = (x_a, y_a, z_a)$ *est la position cartésienne de l'atome* a*.*

Les positions des atomes permettent de définir un ensemble fini de liaisons et d'interactions entre les atomes (cf. section 2.3). Cet ensemble présente une conformation du système moléculaire.

Définition 11 (Conformation)**.** *Soit une image* $I = (Mol, P_I)$ *du système moléculaire* $Mol = (V_M, Cov, Hydro, Inter)$*. Une conformation est un* ***graphe mixte*** $G = (V, E_C, A_H, E_I)$*, tel que :*

- V *: l'ensemble d'atomes du système moléculaire tel que* $\forall t \in T$*,* $|V|_{a,\phi(a)=t}$=$|V_M|_{a,\phi(a)=t}$*. Chaque atome est un sommet de* G*.*
- $E_C \subset Cov$ *: l'ensemble des liaisons covalentes (paires). Chaque liaison covalente est une* ***arête*** *de* G*.*
- $A_H \subset Hydro$ *: l'ensemble des liaisons hydrogène (couples). Chaque liaison hydrogène est un* ***arc*** *de* G*.*
- $E_I \subset Inter$ *: l'ensemble des interactions intermoléculaires électrostatiques (paires). Chaque interaction intermoléculaire électrostatique est une* ***arête*** *de* G*.*

A partir du graphe mixte de la conformation $G = (V, E_C, A_H, E_I)$, nous définissons les sous-graphes suivants :

- $G_{E_C} = (V_C = \{a \in V, \exists b : [a, b] \in E_C\}, E_C)$: le sous-graphe non-orienté de G réduit aux arêtes de E_C et à leurs extrémités.
- $G_{A_H} = (V_H = \{a \in V, \exists b : (a, b) \in A_H\}, A_H)$: le sous-graphe orienté de G réduit aux arcs de A_H et à leurs extrémités.
- $G_{E_I} = (V_I = \{a \in V, \exists b : [a, b] \in E_I\}, E_I)$: le sous-graphe non-orienté de G réduit aux arêtes de E_I et à leurs extrémités.

2.3 Détermination des liaisons et interactions entre atomes

A partir d'une image $I = (Mol = (V_M, Cov, Hydro, Inter), P_I)$, nous utilisons deux opérations principales pour définir les liaisons et les interactions entre les atomes du système moléculaire :

- $\mathsf{dist}(P(a)_I, P(b)_I) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$: la distance euclidienne entre deux atomes a et b de V_M.

- $\mathsf{ang}(P(a)_I, P(b)_I, P(c)_I) = acos(\frac{(x_a-x_b)*(x_c-x_b)+(y_a-y_b)*(y_c-y_b)+(z_a-z_b)*(z_c-z_b)}{\mathsf{dist}(P(a)_I,P(b)_I)\times\mathsf{dist}(P(b)_I,P(c)_I)})$: l'angle entre trois atomes a, b et c de V_M (l'angle $\widehat{abc}$).

Une conformation $G = (V, E_C, A_H, E_I)$ est obtenue à partir d'une image $I = (Mol = (V_M, Cov, Hydro, Inter), P_I)$ en appliquant des règles géométriques sur les atomes selon le type de liaison :

Définition 12 (Liaison covalente)**.** *Soient a et b deux atomes de l'image I. La paire $[a, b]$ appartient à E_C si et seulement si :*

1. $[a, b] \in Cov$,
2. $\mathsf{dist}(P(a)_I, P(b)_I) \leq covr(a) + covr(b)$ [1]

Définition 13 (Liaison hydrogène)**.** *Soient d et a deux atomes de l'image I. Le couple (d, a) appartient à A_H si et seulement si :*

1. $(d, a) \in Hydro$,
2. $\exists h \in V$, $\phi(h) = H$ *et* $[a, h] \in E_C$,
3. $\mathsf{dist}(P(a)_I, P(h)_I) \leq D_H$,
4. $\pi - \alpha_H \leq \mathsf{ang}(P(d)_I, P(h)_I, P(a)_I) \leq \pi - \alpha_H$.

L'atome d est appelé ***donneur*** *de la liaison hydrogène et l'atome a l'****accepteur****. L'ensemble des liaisons hydrogène dépend de l'ensemble des liaisons covalentes fixées.*

Définition 14 (Interaction intermoléculaire électrostatique)**.** *Soient a et b deux atomes de l'image I. La paire $[a, b]$ appartient à E_I si et seulement si :*

1. $[a, b] \in Inter$,
2. $\mathsf{dist}(P(a)_I, P(b)_I) \leq D_I$.

Les paramètres de distance D_H, D_I et les rayons de covalence, ainsi que l'angle α_H, sont des paramètres de la chimie théorique. Les valeurs par défaut utilisées dans nos calculs sont présentées en annexe B.1.

En plus des règles géométriques, il faut respecter quelques propriétés chimiques.

Propriété 1. *Soit une conformation $G = (V, E_C, A_H, E_I)$. Les ensembles* **C**, **D**, **A** *et* **H** *sont définies sur l'ensemble d'atomes V telles que :*

- $a \in V, \mathbf{C}(a) = \{b \in V | [a, b] \in E_C\}$ *: ensemble d'atomes qui forment une covalence avec l'atome a.*
- $a \in V, \mathbf{D}(a) = \{b \in V | (a, b) \in A_H\}$ *: ensemble d'atomes qui forment une liaison hydrogène avec l'atome a, tel que a est donneur.*

1. L'annexe B.1 présente les rayons de covalence utilisés dans nos calculs.

- $a \in V, \mathbf{A}(a) = \{b \in V | (b, a) \in A_H\}$ *: ensemble d'atomes qui forment une liaison hydrogène avec l'atome* a*, tel que* a *est accepteur.*
- $a \in V, \phi(a) = H, \mathbf{H}(a) = \{(b, c) \in A_H | [a, b] \in E_C\}$ *: ensemble des liaisons hydrogène où l'atome* a *est l'hydrogène de la liaisons.*

Pour tout atome de V*, ces ensembles ont les caractéristiques suivantes :*

- $0 \leq |\mathbf{C}(a)| \leq \Psi(a)$ *: un atome peut avoir au plus* $\Psi(a)$ *liaisons covalentes (cf. annexe B.1).*
- $0 \leq |\mathbf{D}(a)| \leq 2$ *: un atome peut être donneur au plus* 2 *fois simultanément.*
- $0 \leq |\mathbf{A}(a)| \leq 2$ *: un atome peut être accepteur au plus* 2 *fois simultanément.*
- $0 \leq |\mathbf{H}(a)| \leq 1$ *: un atome hydrogène peut former au plus une et une seule liaison hydrogène simultanément.*

Pour respecter ces caractéristiques dans le calcul des liaisons et interactions, la distance est utilisée. Nous choisissons toujours les atomes qui forment des liaisons les plus proches en distance géométrique. C'est-à-dire, si un atome a peut faire une seule liaison covalente et que deux atomes b et c satisfont la condition d'une liaison covalente mais que b est plus proche de a en distance, nous choisissons l'atome b au lieu de l'atome c.

Une fois les règles sont appliquées et les propriétés sont respectées, l'objectif principal est que la conformation puisse définir le système moléculaire en termes de liaisons et d'interactions entre les atomes, indépendamment des positions.

2.4 Exemple d'un système moléculaire

Soit un système moléculaire contenant 14 atomes comme suit :

$$V_M = \{\mathsf{C}1, \mathsf{C}2, \mathsf{C}3, \mathsf{H}1, \mathsf{H}2, \mathsf{H}3, \mathsf{H}4, \mathsf{H}5, \mathsf{H}6, \mathsf{H}7, \mathsf{H}8, \mathsf{N}1, \mathsf{O}1, \mathsf{O}2\}$$

Les liaisons possibles d'un point de vu chimique indépendamment des positions sont :

- $Cov = \{[\mathsf{C}1, \mathsf{C}2], [\mathsf{C}1, \mathsf{C}3], [\mathsf{C}1, \mathsf{H}1], ...\}$: il y a $\binom{14}{2} = 91$ liaisons covalentes possibles.
- $Hydro = \{(\mathsf{N}1, \mathsf{O}1), (\mathsf{N}1, \mathsf{O}2), (\mathsf{O}1, \mathsf{N}1), (\mathsf{O}1, \mathsf{O}2), ...\}$ il y a $2 \times \binom{3}{2} = 6$ liaisons hydrogène possibles.
- $Inter = \emptyset$

Nous supposons avoir les positions (x, y, z) suivantes :

N1	8.8860670552	−2.0307639360	−8.2638659864
H1	9.8125715069	−1.3585327814	−8.5866534187
C1	8.5602732025	−1.1599094306	−6.9924582229
H2	9.1038097822	−2.9671468690	−7.9954706924
H3	8.1136112388	−2.0106525150	−8.9305990369
H4	7.6810685889	−0.5752342124	−7.1204499660
C2	9.7767661304	−0.2221891587	−6.9718680887
C3	8.5689766647	−2.1634789273	−5.7409827102
O1	10.4932433066	−0.1960166586	−8.0202701211
H5	8.3554625351	−1.6318672791	−4.7906285871
H6	7.8474348242	−2.9336956441	−5.8906104504
H7	9.5868724410	−2.6937436814	−5.6014284125
O2	9.9679949131	0.4604031106	−5.8053468226
H8	9.5573426099	0.1114460341	−4.9587311231

La figure 2.1 présente le graphe mixte de la conformation obtenue.

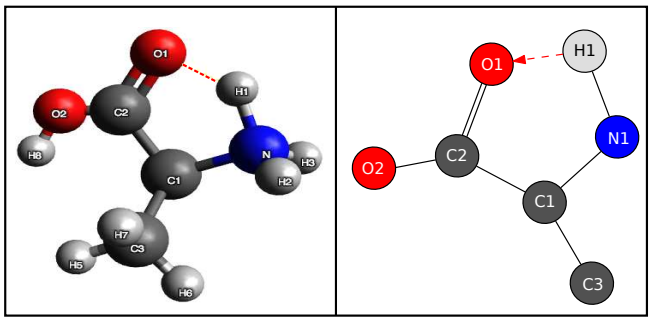

FIGURE 2.1 – Graphe mixte d'une conformation. La figure à gauche présente une représentation 3D d'une conformation d'un système moléculaire et la figure à droite son graphe mixte. Dans le graphe mixte, les liaisons covalentes sont présentées en arêtes noires et les liaisons hydrogène en arcs rouges. Pour rendre le graphe mixte plus lisible, nous présentons uniquement les atomes d'hydrogène qui sont impliqués dans des liaisons hydrogène. La couleur des sommets dépend du type d'atome : les atomes d'oxygène sont présentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair.

La conformation $G = (V, E_C, A_H, E_I)$ obtenue en appliquant les règles des liaisons ci-dessus est :

- $V = \{\mathsf{C1}, \mathsf{C2}, \mathsf{C3}, \mathsf{H1}, \mathsf{H2}, \mathsf{H3}, \mathsf{H4}, \mathsf{H5}, \mathsf{H6}, \mathsf{H7}, \mathsf{H8}, \mathsf{N1}, \mathsf{O1}, \mathsf{O2}\}$
- $E_C = \{[\mathsf{C1}, \mathsf{C2}], [\mathsf{C1}, \mathsf{N1}], [\mathsf{C1}, \mathsf{H4}], [\mathsf{C1}, \mathsf{C3}], [\mathsf{C2}, \mathsf{O1}], [\mathsf{C2}, \mathsf{O2}], [\mathsf{C3}, \mathsf{H5}], [\mathsf{C3}, \mathsf{H6}], [\mathsf{C3}, \mathsf{H7}], [\mathsf{N1}, \mathsf{H1}], [\mathsf{N1}, \mathsf{H2}], [\mathsf{N1}, \mathsf{H3}], [\mathsf{O2}, \mathsf{H8}]\ \}$
- $A_H = \{(\mathsf{N1}, \mathsf{O1})\}$

- $E_I = \emptyset$

2.5 Conclusion

Nous avons présenté dans ce chapitre comment un système moléculaire et ses conformations sont modélisés des graphes. Nous avons également expliqué les règles que nous utilisons pour calculer les liaisons et les interactions formées entre les atomes, à partir des images de trajectoires. Cette modélisation et les différentes règles sont utilisées dans les deux algorithmes proposés au cours de cette thèse, à savoir l'algorithme d'analyse de la dynamique conformationnelle sur les trajectoires de simulation, qui est présenté dans le chapitre 3 et l'algorithme de la prédiction conformationnelle qui fera l'objet du chapitre 4.

Chapitre 3
Analyse conformationnelle des trajectoires

Dans le cadre des simulations de dynamique moléculaire, un nombre important de trajectoires est généré. L'objectif des simulations est d'explorer les conformations qu'un système moléculaire peut adopter et analyser les changements (les transitions) entre elles. En effet, les atomes du système moléculaire bougent au cours du temps et donc changent de position. Cette dynamique des atomes peut entrainer un changement de liaisons (covalentes, hydrogène ou intermoléculaires électrostatiques) ce qui représente un changement conformationnel. Il faut être capable d'identifier ces changements conformationnels dans les trajectoires sachant qu'une trajectoire représente l'évolution du système moléculaire au cours du temps sous forme de changements de coordonnées cartésiennes des atomes qui le composent.

Nous avons vu dans le chapitre précédent (cf. section 1.2), qu'il existe des logiciels d'analyse statique qui permettent de visualiser la structure tridimensionnelle du système moléculaire à chaque instant de la trajectoire mais ne donnent aucune information supplémentaire sur la dynamique conformationnelle. Les conformations sont obtenues "à l'oeil" s'il s'agit des systèmes moléculaires de petite taille et peu complexes ou à travers des bouts de codes personnalisés pour le système étudié.

L'objectif de notre travail est de proposer un algorithme qui permet d'analyser les trajectoires, d'identifier d'une façon efficace et rapide les conformations explorées et les transitions entre elles et qui peut être applicable à n'importe quel système moléculaire.

Dans les sections 3.2 et 3.3, nous expliquons les étapes d'identification des conformations dans une trajectoire et nous décrivons comment identifier un changement conformationnel en utilisant des techniques de théorie des graphes. Ensuite, la section 3.4 présente les étapes de l'algorithme et la section 3.5 présente sa complexité. Dans la section 3.6, nous montrons la possibilité d'analyser les mouvements rotationnels dans les systèmes moléculaires en utilisant des notions de théorie des graphes. Nous verrons également, dans cette section, que l'algorithme peut être utilisé pour analyser simultanément plusieurs trajectoires. Enfin, pour évaluer les performances de l'algorithme, plusieurs tests ont été effectués sur 3 types de systèmes, à savoir les peptides isolés, les dissociations du peptide induites par collision (CID) et les clusters (cf. section 1.1.1). Les résultats sont présentés dans la section 3.7.

3.1 Trajectoire d'un système moléculaire

Etant donné un système moléculaire $Mol = (V_M, Cov, Hydro, Inter)$, les atomes sont positionnés dans l'espace ce qui constitue une image de ce système (cf. section 2.2). Dans le cadre des simulations de dynamique moléculaire (cf. chapitre 1), une énergie (température) est fixée. Cette énergie (température) permet de changer les positions des atomes au cours du temps. Pour observer et analyser ces changements, des images des positions d'atomes sont sauvegardées à des intervalles de temps réguliers, ce qui constitue une trajectoire :

Définition 15 (Trajectoire)**.** *Une trajectoire $\mathcal{I}$ est une séquence d'images $I_1, I_2, ..., I_S$ du système moléculaire, prises à des intervalles de temps réguliers.*

Hypothèse 1. *Sur une trajectoire $\mathcal{I}$, l'ensemble d'atomes est le même sur toute la trajectoire. Il s'agit de l'ensemble V_M d'atomes du système moléculaire Mol que cette trajectoire représente.*

Selon l'énergie du système et le temps de simulation, les positions d'atomes changent et donc, dans une trajectoire, les liaisons et les interactions (cf. section 2.3) calculées sur la séquence d'images peuvent changer d'une image à une autre. Par conséquent, une trajectoire peut contenir une ou plusieurs conformations. L'identification d'une conformation revient à déterminer l'ensemble des liaisons covalentes, des liaisons hydrogène et des interactions intermoléculaires électrostatiques selon les règles décrites dans le chapitre précédent (cf. section 2.3).

3.2 Identification des conformations sur une trajectoire

Dans une trajectoire $\mathcal{I}$ de S images, les règles proposées dans le chapitre précédent permettent de construire pour chaque image $I_i = (Mol, P_{I_i})$ sa conformation $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$. Dans la construction de la conformation G_i, le calcul de l'ensemble des sommets V_i est linéaire. Cependant, le calcul des liaisons et d'interactions (E_{C_i}, A_{H_i} et E_{I_i}) est quadratique en nombre d'atomes dans le système moléculaire.

L'approche basique consiste à appliquer les règles géométriques (cf. section 2.3) sur tous les couples d'atomes du système et à chaque image de la trajectoire. Cependant, les trajectoires analysées telles qu'elles sont générées par les simulations de dynamique moléculaire (cf. chapitre 1) ont la propriété suivante :

Hypothèse 2. *D'une image à la suivante, les atomes bougent généralement peu.*

En utilisant cette hypothèse, il n'est pas nécessaire de recalculer, à chaque image, tous les ensembles de liaisons. En effet, nous utilisons des règles basées sur cette

hypothèse pour réduire le nombre de calculs effectués pour identifier les conformations dans une trajectoire.

3.2.1 L'ensemble référent de conformations

Dans une trajectoire $\mathcal{I}$ de S images, à chaque image I_i est associé son graphe de conformation G_i ($1 \leq i \leq S$). Soit $\mathcal{G}$ l'ensemble de ces conformations. Parmi ces conformations certaines sont isomorphes.

Deux graphes sont isomorphes G_i et G_j si et seulement s'il existe un appariement (une bijection) entre les sommets de G_i et G_j qui préserve les relations d'adjacence présentes dans G_i et G_j. Une définition formelle est donnée dans la section 3.3.1.

En d'autres termes, nous pouvons avoir un sous-ensemble de conformations $G_{R_1}, G_{R_2}, ..., G_{R_S}$ ($1 \leq R_S \leq S$), où chaque conformation G_i ($1 \leq i \leq S$) à une occurence (à un isomorphe près) dans ce sous-ensemble. Cet ensemble est appelé **ensemble référent** $\mathcal{G}_R$.

Définition 16 (Ensemble référent)**.** *Soient $\mathcal{G}$ l'ensemble des conformations associées aux images d'une trajectoire $\mathcal{I}$. L'ensemble $\mathcal{G}_R \subseteq \mathcal{G}$ est ensemble référent de $\mathcal{G}$ si :*

$$\forall G \in \mathcal{G}, \exists G_r \in \mathcal{G}_R; G \text{ et } G_R \text{ sont isomorphes}.$$

L'objectif est de trouver un sous ensemble d'image $I_{R_1}, I_{R_2}, ..., I_{R_S}$ ($1 \leq R_S \leq S$) tel que $\mathcal{G}_R$ obtenu, de ces images, est un ensemble référent. Ces images sont appelées **images de référence**. Il faut également minimiser l'ensemble d'images de référence.

Remarque 2. *Une trajectoire $\mathcal{I}$ de S images peut contenir une ou plusieurs images de référence. Soit la séquence d'images de référence I_{R_1}, I_{R_2}, ..., I_{R_S} ($1 \leq R_S \leq S$) identifiée sur la trajectoire $\mathcal{I}$ et soit la fonction δ qui associe à chaque R_j son numéro i dans l'ensemble des images de la trajectoire :*

$$\begin{array}{rccc} \delta & : & [\![1, S]\!] & \rightarrow & [\![1, S]\!] \\ & & R_j & \mapsto & \delta(R_j) = i \end{array}$$

Le choix des images de référence parmi toutes les images de la trajectoire se base sur l'hypothèse 2 liée aux déplacements des atomes. Le changement d'image de référence dépend du changement des **orbites** de ces atomes que nous définissons dans la section suivante.

3.2.2 Calculs des liaisons hydrogène en utilisant des orbites

Etant donnée une image $I_i = (Mol, P_{I_i})$, un calcul élémentaire de l'ensemble de liaisons hydrogène A_{H_i} de la conformation $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ consiste à parcourir la liste des atomes d'hydrogène, et pour chaque atome d'hydrogène vérifier

s'il existe un atome qui peut être accepteur (cf. section 2.3) en parcourant **tous** les atomes de V_i.

Cependant, les atomes qui peuvent potentiellement former une liaison hydrogène avec un atome d'hydrogène donné sont les atomes proches de lui en termes de distance géométrique. A cet effet, dans notre approche, seuls les atomes appartenant au voisinage de cet atome d'hydrogène sont pris en compte, au lieu de tester tous les atomes de la molécule. L'ensemble de ces atomes est appelé **orbite** de cet atome.

Définition 17. *Soit une image $I = (Mol, P_I)$. L'orbite $\mathcal{O}_h$ d'un atome d'hydrogène $h \in V_M$ ($\phi(h) =$ H) est définie comme suit :*

$$\mathcal{O}_h = \{a \in V_M, \phi(a) \in \{O, N, F\}, \mathsf{dist}(P(a)_I, P(h)_I) \leq \alpha_{\mathcal{O}_H} D_H\}$$

Soient deux image I_i et I_j ($1 \leq j \leq i$ S) d'une trajectoire $\mathcal{I}$ de S images dont l'ensemble d'atomes est V. Nous choisissons deux atomes a et b de V tels que :

$$\forall c \in V : \mathsf{dist}(P(c)_{I_i}, P(c)_{I_j}) \leq \mathsf{dist}(P(a)_{I_i}, P(a)_{I_j})$$

$$\forall c \in V, c \neq a : \mathsf{dist}(P(c)_{I_i}, P(c)_{I_j}) \leq \mathsf{dist}(P(b)_{I_i}, P(b)_{I_j})$$

Dans ce qui suit, nous notons $amax_{i,j}$ la distance $\mathsf{dist}(P(a)_{I_i}, P(a)_{I_j})$ et $bmax_{i,j}$ la distance $\mathsf{dist}(P(b)_{I_i}, P(b)_{I_j})$

Ces distances présentent les deux atomes qui se sont déplacées le plus à l'image I_i relativement à l'image I_j. En se basant sur ces distances nous avons le lemme suivant :

Lemme 1. *Soit I_R une image de référence d'une trajectoire $\mathcal{I}$ de S images ($1 \leq R \leq S$). Si $I_{\delta(R)+1}$, $I_{\delta(R)+2}$,...,I_{k-1} ne sont pas des images de référence, et si $amax_{\delta(R),k} + bmax_{\delta(R),k} < D_R$ ($D_R = (\alpha_{\mathcal{O}_H} - 1) \times D_H$) alors les orbites de I_R sont valables pour I_k.*

Preuve 1. ***Cas (1) :*** *Montrons que pour tout atome a et pour tout atome b qui n'est pas dans l'orbite de a à l'image I_R, les atomes a et b ne forment pas de liaison à l'image I_k*

1. *Le diametre de l'orbite a est $\alpha_{\mathcal{O}_H} D_H$ et donc à l'image I_R,* $\mathsf{dist}(P(a)_{I_R}, P(b)_{I_R}) > \alpha_{\mathcal{O}_H} D_H$
2. *A l'image I_k les atomes a et b se sont rapprochés au plus de $D_R = (\alpha_{\mathcal{O}_H} - 1) \times D_H$*

Donc à l'image I_k :

$$\begin{aligned} \mathsf{dist}(P(a)_{I_k}, P(b)_{I_k}) &> \alpha_{\mathcal{O}_H} D_H - (\alpha_{\mathcal{O}_H} - 1) \times D_H \\ &> D_H \end{aligned} \tag{3.1}$$

Donc a et b ne peuvent pas former de liaison.

***Cas (2) :** Montrons que pour tout atome a et pour tout atome b qui est dans l'orbite de a à l'image I_R et qui forme une liaison avec a, l'atome b reste dans l'orbite de a à l'image I_k*

1. *A l'image I_R nous avons $\mathsf{dist}(P(a)_{I_R}, P(b)_{I_R}) < D_H$*
2. *A l'image I_k les atomes a et b se sont éloignés au plus de $D_R = (\alpha_{\mathcal{O}_H} - 1) \times D_H$*

Donc à l'image I_k :

$$\begin{aligned} \mathsf{dist}(P(a)_{I_k}, P(b)_{I_k}) &< D_H + (\alpha_{\mathcal{O}_H} - 1) \times D_H \\ &< \alpha_{\mathcal{O}_H} D_H \end{aligned} \tag{3.2}$$

Donc b reste dans l'orbite de a à l'image I_k.

La distance D_H est la distance utilisée pour le calcul des liaisons hydrogène (cf. section 2.3). Le coefficient $\alpha_{\mathcal{O}_H}$ [31] est utilisé pour définir la distance maximale pour accepter un atome dans l'orbite. Le choix de ce coefficient est détaillé dans ce qui suit.

Remarque 3. *Le lemme 1 donne une condition suffisante pour recalculer les orbites. En effet, pour une image de référence I_R et une image I_k, si $amax_{\delta(R),k} + bmax_{\delta(R),k} < D_R$, on est certain qu'il ne sert à rien de recalculer les orbites. Parcontre, si $amax_{\delta(R),k} + bmax_{\delta(R),k} \geq D_R$, alors on recalculera les orbites et on changera donc l'image de référence. L'interêt est que la condition utlisée pour le calcul des orbites et d'images de référence soit très facilement calculable.*

La figure 3.1 montre un exemple de changement d'images de référence avec un changement d'orbites.

Afin de fixer une valeur au coefficient $\alpha_{\mathcal{O}_H}$, nous avons choisi des trajectoires de systèmes moléculaires telles que les changements conformationnels sont dus aux changements dans les liaisons hydrogène seulement (les changements dans les autres types de liaisons sont ignorés). Pour ces trajectoires, nous avons analysé l'évolution du nombre d'images de référence et la taille moyenne des orbites en variant le coefficient entre 0 et 10. Nous avons également comparer le temps

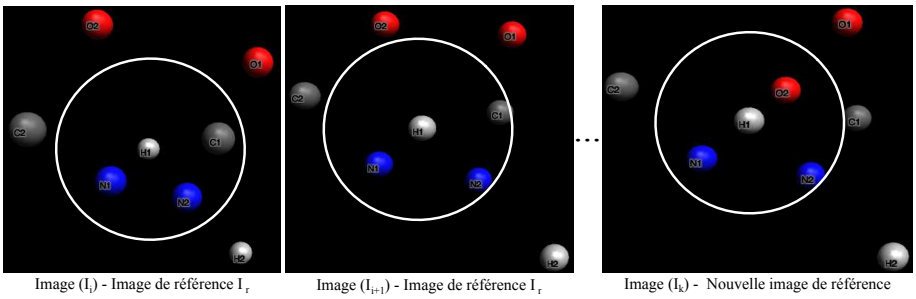

FIGURE 3.1 – Orbite d'un atome d'hydrogène. La figure présente la dynamique d'une orbite de l'atome d'hydrogène H1 suite au déplacement des atomes. L'orbite est présentée en cercle blanc.

d'exécution du programme en prenant les mêmes valeurs du coefficient $\alpha_{\mathcal{O}_H}$. Les résultats obtenus sont détaillés dans la section 3.7.1.

Il n'existe pas de valeur optimale pour $\alpha_{\mathcal{O}_H}$, car cela dépend intervalles des trajectoires (cf. chapitre 1) choisies pour les tests. Dans le choix d'une valeur pertinente, nous visons à minimiser le nombre d'images de référence et la taille d'orbites tout en ayant non seulement un bon temps d'exécution mais surtout garantir que toutes les conformations explorées dans la trajectoire apparaissent au moins une fois.

Le calcul du déplacement des atomes relativement à l'image de référence courante se fait en même temps que la lecture de l'image courante. A cet effet, la condition utilisée pour les orbites n'est pas coûteuse.

Nous utilisons le centre de masse du système moléculaire[1] pour éviter de confondre le déplacement des atomes avec la translation du système moléculaire, Nous calculons le centre de masse du système moléculaire à chaque image de la trajectoire. Ce centre de masse devient le nouveau repère des atomes (nouvel axe). Cette définition n'a pas d'impact sur la complexité de l'algorithme car les calculs se font en même temps que la lecture des coordonnées cartésiennes des atomes de l'image courante.

L'utilisation des orbites permet de réduire le nombre de comparaisons. Dans une trajectoire d'un heptapeptide composé de 74 atomes, par exemple, au lieu de parcourir les 74 atomes, nous parcourons au maximum 7 atomes seulement pour chaque atome d'hydrogène ($< 10\%$ de la taille de la molécule) et sur une trajectoire de 26601 images de ce peptide on calcule 4 fois seulement les orbites (cf. section 3.7.2).

Dans les systèmes moléculaires, nous pouvons trouver une dynamique de liaisons hydrogène mais aussi une dynamique des autres liaisons, à savoir les liaisons

1. Les coordonnées du centre de masse d'une molécule sont définies comme suit : $x_m = \sum x_i/n; y_m = \sum y_i/n; z_m = \sum z_i/n$; où n est le nombre d'atomes de la molécule.

covalentes et les interactions intermoléculaires électrostatiques. En se basant sur le même principe utilisé pour les orbites des liaisons hydrogène, nous définissons les orbites d'atomes pour les liaisons covalentes et les interactions intermoléculaires électrostatiques avec des distances de seuils différentes.

3.2.3 Calculs des liaisons covalentes et des interactions intermoléculaires électrostatiques en utilisant des orbites

Les liaisons covalentes sont des liaisons plus fortes que les liaisons hydrogène, mais dans certains systèmes moléculaires ces liaisons peuvent apparaitre et disparaitre au cours de la trajectoire. Dans ce cas, nous devons calculer à chaque fois l'ensemble des liaisons covalentes pour assurer d'avoir le bon ensemble. Le calcul classique est qu'à chaque image nous parcourons toutes les paires d'atomes et nous vérifions s'ils peuvent former une liaison covalente (cf. section 2.3). De la même façon que les liaisons hydrogène, nous définissons des orbites pour les atomes du système, autour desquels nous calculons les liaisons covalentes.

Définition 18. *Soit une image $I = (Mol, P_I)$. L'orbite $\mathcal{O}_c$ d'un atome $c \in V_M$ est définie comme suit :*

$$\mathcal{O}_c = \{b \in V_M, \mathsf{dist}(P(b)_I, P(c)_I) \leq \alpha_{\mathcal{O}_C} D_C\}$$

Enfin, un autre type de liaisons pris en compte par notre approche sont les interactions intermoléculaires électrostatiques.Ces interactions existent dans certains systèmes moléculaires comme les clusters (cf. section 1.1.1) et concernent un ensemble particulier d'atomes (lithium, argon, etc.). Nous notons V_E cet ensemble.

Définition 19. *Soit une image $I = (Mol, P_I)$. L'orbite $\mathcal{O}_a$ d'un atome $a \in V_E$, est définie comme suit :*

$$\mathcal{O}_a = \{b \in V_M, \mathsf{dist}(P(b)_I, P(a)_I) \leq \alpha_{\mathcal{O}_I} D_I\}$$

Les distances D_C et D_I représentent respectivement la distance moyenne d'une liaison covalente et la distance moyenne d'une interaction intermoléculaire électrostatique (cf. section 2.3). Les coefficients $\alpha_{\mathcal{O}_C}$ et $\alpha_{\mathcal{O}_I}$ sont utilisés pour définir la distance maximale pour accepter un atome dans l'orbite. La valeur de ces coefficients a été fixée de la même manière que le coefficient utilisé dans le calcul des liaisons hydrogène.

Le changement d'orbites entraine le changement d'images de référence. La distance de seuil D_R utilisée pour changer l'orbite et donc l'image de référence pour les liaisons covalentes est $(\alpha_{\mathcal{O}_C} - 1) \times D_C$ et pour les interactions intermoléculaires électrostatiques $(\alpha_{\mathcal{O}_I} - 1) \times D_I$. Les valeurs des distances utilisées pour le calcul des

orbites sont données en annexe B.1. Nous utilisons la même preuve présentée dans la section précédente pour démontrer que les orbites des liaisons covalentes et des interactions intermoléculaires électrostatiques donnent des ensembles référents.

3.2.4 Calcul hiérarchique

Dans le chapitre 1, nous avons vu que les liaisons covalentes qui sont formées entre les atomes sont les liaisons les plus fortes. Selon ces liaisons, des liaisons hydrogène et des interactions intermoléculaires électrostatiques sont formées, ce qui donne une conformation à ce système. Nous utilisons les notations $\mathcal{O}_H$, $\mathcal{O}_C$ et $\mathcal{O}_I$ pour désigner les orbites pour les liaisons hydrogène, les orbites pour les liaisons covalentes et les orbites pour les interactions intermoléculaires électrostatiques respectivement.

Nous pouvons définir une hiérarchie dans le calcul des orbites et d'images de références comme suit :

- **Cas (1)** : le changement d'orbites des liaisons covalentes $\mathcal{O}_C$ entraine un changement de tous les autres orbites ($\mathcal{O}_H$ et $\mathcal{O}_I$) et donc les images de référence correspondantes.
- **Cas (2)** : le changement d'orbites des liaisons hydrogène $\mathcal{O}_H$ n'entraine aucun autre changement dans les autres orbites ($\mathcal{O}_H$ et $\mathcal{O}_I$) et seule l'image de référence pour les orbites des liaisons hydrogène est modifiée.
- **Cas (3)** : le changement d'orbites des interactions intermoléculaires électrostatiques $\mathcal{O}_I$ ressemble au cas (2), seule l'image de référence pour les orbites de ces interactions est modifiée..

Les interactions intermoléculaires électrostatiques dépendent d'un ensemble particulier d'atomes. A cet effet, ces interactions ne sont pas présentes dans tous les systèmes moléculaires contrairement aux liaisons covalentes et liaisons hydrogène. Dans ce genre de systèmes, le cas (3) est ignoré et aucun calcul des interactions intermoléculaires électrostatiques n'est effectué, ce qui réduit le nombre de comparaisons au cours de la trajectoire. De plus, dans certaines trajectoires il n y a pas de dynamique de liaisons covalentes ; elles sont fixes au cours de la trajectoire. Dans ces systèmes, nous calculons les orbites des atomes autour des liaisons covalentes ainsi que les liaisons covalentes une seule fois dans la première image de la trajectoire ensuite nous vérifions que le cas (2) (les orbites autour des atomes d'hydrogène). Ces systèmes sont détaillés dans le chapitre 5.

3.3 Analyse de la dynamique conformationnelle

L'objectif de notre approche n'est pas simplement d'identifier les conformations les unes indépendamment des autres, mais aussi d'étudier et analyser les transitions

entre elles. A cet effet, il faut définir la similarité des conformations et la nature des transitions entre ces conformations.

3.3.1 Isomorphisme entre deux conformations quelconques

Afin d'identifier une transition entre deux conformations, il faut d'abord définir ce qu'est la **similarité des conformations**. Pour cela, nous considérons l'isomorphisme [32, 33, 34, 35] entre les graphes défini ainsi :

Définition 20. *Sur une trajectoire $\mathcal{I}$, deux graphes de conformations $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ et $G_j = (V_j, E_{C_j}, A_{H_j}, E_{I_j})$ sont **isomorphes** si et seulement s'il existe une bijection $\theta_{i,j} : V_i \to V_j$ telle que :*

1. $\forall a \in V_i$, *on a* $\phi(a) = \phi(\theta(a))$
2. $[a, b] \in E_{C_i} \Leftrightarrow [\theta_{i,j}(a), \theta_{i,j}(b)] \in E_{C_j}$
3. $(a, b) \in A_{H_i} \Leftrightarrow (\theta_{i,j}(a), \theta_{i,j}(b)) \in A_{H_j}$
4. $[a, b] \in E_{I_i} \Leftrightarrow [\theta_{i,j}(a), \theta_{i,j}(b)] \in E_{I_j}$

En d'autres mots, deux graphes sont isomorphes si et seulement s'il existe un appariement (une bijection) entre les sommets de G_i et G_j qui préserve les relations d'adjacence présentes dans G_i et G_j. La fonction $\theta_{i,j}$ est appelée fonction d'isomorphisme.

Plusieurs approches algorithmiques existent pour résoudre le problème d'isomorphisme de graphes [36, 32, 37, 38, 39, 40]. Ce problème consiste, étant donné deux graphes, à décider s'ils sont isomorphes ou non. Dans notre cas, nous avons choisi d'utiliser l'approche de **Mckay** [33] (les algorithmes **nauty** et **Trace**) qui consiste à calculer des identifiants canoniques (signatures) pour les deux graphes en question. Ces deux graphes sont isomorphes si et seulement s'ils ont la même signature.

L'algorithme de **Mckay** repose sur trois idées principales :

- La similarité du nombre de sommets par degrés,
- La génération d'un arbre de recherche pour les permutations possibles.
- L'utilisation de l'automorphisme du graphe pour appliquer des élagages sur l'arbre de recherche.

Etant donnée une matrice d'adjacence d'un graphe, on cherche la permutation par ligne et par colonne de la matrice telle que le nombre binaire obtenu par la lecture de la matrice ligne par ligne soit **minimum**. L'approche consiste à trouver cette permutation par une méthode de séparation et évaluation (branch and bound) optimisée. En effet, l'efficacité de l'approche réside dans la génération de l'arbre de recherche. Des élagages de l'arbre sont appliqués selon un ensemble de règles

définies sur les noeuds de l'arbre. Ces règles permettent non seulement de générer qu'une partie de l'arbre mais assurent le choix unique de la signature [33, 34, 35].

La figure 3.2 présente un exemple de deux graphes G_1 et G_2 qui sont isomorphes, bien qu'avant le processus de canonisation les graphes ne sont pas identiques (par exemple sommets 10 et 2 sont liés dans le graphe G_2 alors qu'ils ne le sont pas dans le graphe G_1). Après ce processus, nous obtenons deux graphes identiques comme le montre les graphes à gauche de la figure.

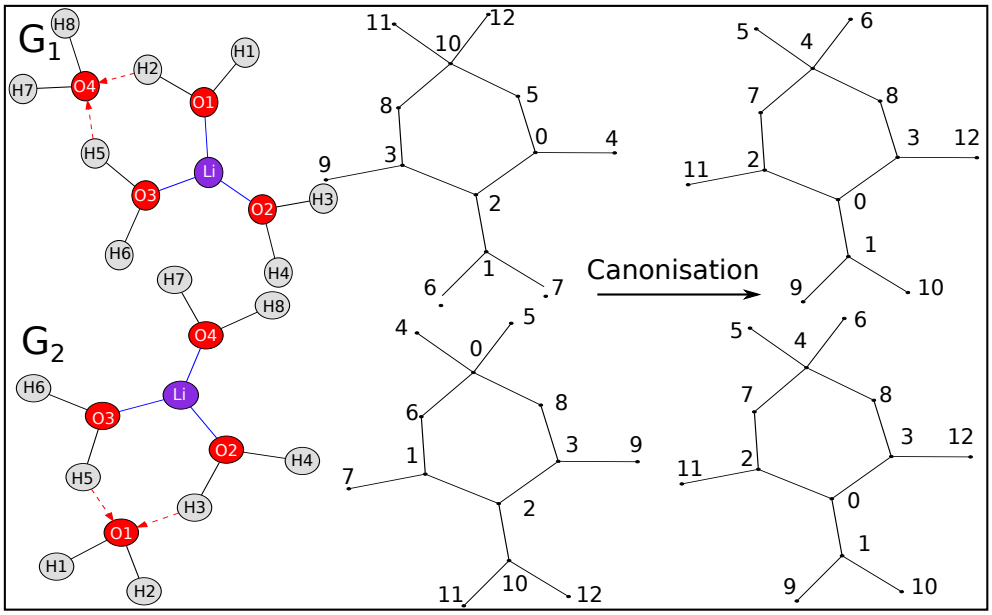

FIGURE 3.2 – Exemple de deux conformations isomorphes. La partie gauche de la figure présente les graphes mixtes des deux conformations, au milieu les graphes avec numération des sommets selon l'ordre d'apparition dans la trajectoire d'entrée et la partie droite présente les deux graphes correspondants après canonisation.

Les signatures obtenues pour les graphes de la figure 3.2 sont :

$$G_1 \ : \ 2|\ 1\ 0\ 3\ 10|\ 11\ 12\ 5\ 8\ 6\ 7\ 4\ 9$$

$$G_2 \ : \ 2|\ 10\ 1\ 3\ 0|\ 4\ 5\ 6\ 8\ 11\ 12\ 7\ 9$$

Pour distinguer les sommets, nous pouvons également définir un **partitionnement initial** (coloration) des sommets du graphe selon un critère donné. Au cours de la trajectoire, deux atomes de même nature peuvent jouer le même rôle sans avoir le même identifiant (numéro par ordre d'apparition). A cet effet, nous effectuons un partitionnement basé sur le type chimique de l'atome : deux atomes ont la même couleur s'ils appartiennent au même type chimique. Par exemple, dans les graphes précédents nous obtenons trois partitions initiales : la première partition $\{2\}$ contient l'atome de lithium, la deuxième $\{0, 1, 3, 10\}$ contient les atomes de d'oxygène et la dernière partition contient le reste des sommets qui sont les atomes

d'hydrogène $\{4, 5, 6, 7, 8, 9, 11, 12\}$. L'ordre des sommets obtenu dans les signatures respecte le partitionnement initial [33, 35].

L'isomorphisme obtenu pour la figure 3.2 par l'algorithme de McKay est le suivant :

$$(0,10)\ (1,0)\ (2,2)\ (3,3)\ (4,11)\ (5,12)\ (6,5)\ (7,4)\ (8,8)\ (9,9)\ (10,1)\ (11,6)\ (12,7)$$

Dans les couples des sommets ci-dessus, le premier élément présente un sommet du graphe G_1 et le deuxième élément présente le sommet correspondant dans le graphe G_2.

Enfin, la différence entre les algorithmes nauty et Trace est que nauty fait un parcours en profondeur de l'arborescence (avec l'intuition d'arriver plus vite à la signature), quant à Traces il fait une exploration en largeur (ce qui augmente l'élagage). Ces algorithmes sont capables de résoudre le problème d'isomorphisme pour la très grande majorité des graphes, et n'ont un temps d'exécution exponentiel que pour très peu de graphes.

3.3.2 Transition entre deux conformations consécutives

Dans une séquence de conformations $G_1, G_2, ..., G_k$ consécutives, ce qui change est l'ensemble de liaisons calculées d'une conformation à une autre. Une **transition entre deux conformations** consécutives $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ et $G_{i+1} = (V_{i+1}, E_{C_{i+1}}, A_{H_{i+1}}, E_{I_{i+1}})$ est un changement (apparition ou disparition) dans l'ensemble des liaisons suite au changement géométrique des atomes :

- **Changement de liaisons covalentes** :
  - **Apparition** d'une nouvelle liaison covalente $[a, b]$ si $[a, b] \in E_{C_{i+1}}$ et $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{C_i}$. Ce changement est noté *C-A*.
  - **Disparition** d'une liaison covalente déja existante $[a, b]$ si $[a, b] \in E_{C_i}$ et $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{C_{i+1}}$. Ce changement est noté *C-D*.
- **Changement de liaisons hydrogène** :
  - **Apparition** d'une nouvelle liaison hydrogène (a, b) si $(a, b) \in A_{H_{i+1}}$ et $(\theta_{i,i+1}(a), \theta_{i,i+1}(b)) \notin A_{H_i}$. Ce changement est noté *H-A*.
  - **Disparition** d'une liaison hydrogène existante (a, b) si $(a, b) \in A_{H_i}$ et $(\theta_{i,i+1}(a), \theta_{i,i+1}(b)) \notin A_{H_{i+1}}$. Ce changement est noté *H-D*.
  - **Transfert de proton** dans la liaison hydrogène (a, b) si $(a, b) \in A_{H_i}$ et $(\theta_{i,i+1}(b), \theta_{i,i+1}(a)) \in A_{H_{i+1}}$. C'est-à-dire l'atome b et l'atome a échangent leur rôle : l'atome a devient le donneur et l'atome b devient l'accepteur (changement d'orientation de l'arc). Ce changement est noté *H-T*
- **Changement d'interactions intermoléculaires électrostatiques** :

- **Apparition** d'une nouvelle interaction intermoléculaire électrostatique $[a, b]$ si $[a, b] \in E_{I_{i+1}}$ et $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{I_i}$. Ce changement est noté *I-A*.
- **Disparition** d'une interaction intermoléculaire électrostatique existante $[a, b]$ si $[a, b] \in E_{I_i}$ et $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{I_{i+1}}$. Ce changement est noté *I-D*.

3.3.3 Graphe de transitions

A partir des conformations explorées dans une trajectoire et des transitions identifiées entre elles, nous construisons le graphe des transitions qui décrit le changement conformationnel observé au cours de la trajectoire d'une façon abstraite.

Définition 21. *Soit $\mathcal{I}$ une trajectoire de S images. Le graphe de transitions $\mathcal{G}_\mathcal{I} = (\mathcal{G}, \mathcal{A}, \rho, \tau)$ de la trajectoire $\mathcal{I}$ est un graphe orienté pondéré tel que :*

- *$\mathcal{G} = \{G_1, G_2, ..., G_C\}$: l'ensemble des sommets de $\mathcal{G}_\mathcal{I}$. Chaque sommet représente une conformation explorée dans la trajectoire $\mathcal{I}$.*
- *$\mathcal{A}$: l'ensemble des arcs entre les sommets de $\mathcal{G}_\mathcal{I}$. Chaque arc représente une transition observée au cours de la trajectoire $\mathcal{I}$.*
- *ρ et τ sont les pondérations sur les arcs de $\mathcal{G}_\mathcal{I}$ telles que :*
  - *$\rho : \mathcal{A} \rightarrow$ {C-A, C-D, H-A, H-D, H-T, I-A, I-D }, $\rho(G_i, G_j)$ est la nature de la transition (G_i,G_j). Par exemple $\rho(G_i, G_j)$ =H-A s'il y a une apparition d'une liaison hydrogène.*
  - *$\tau : \mathcal{A} \rightarrow$ N, $\tau(G_i, G_j)$ est la fréquence de la transition de G_i à G_j. Par exemple $\tau(G_i, G_j) = 3$ si au cours de la trajectoire la conformation G_i est passée à la conformation G_j trois fois.*

Le graphe des transitions permet de donner une vision globale de la dynamique de la molécule (*une vision conformationnelle moyenne*) indépendamment de l'échelle temporelle. Il permet également d'avoir des statistiques sur la dynamique conformationnelle du système analysé.

3.4 Algorithme pour l'analyse conformationnelle des trajectoires

Dans cette section, nous présentons l'algorithme qui permet d'énumérer et d'identifier à partir d'une séquence d'images I_1,I_2,...,I_S d'une trajectoire $\mathcal{I}$ l'ensemble des conformations non isomorphes $\mathcal{G} = \{G_1, G_2, ..., G_C\}$ explorées telles que $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$, et fournit le graphe des transitions associé $\mathcal{G}_\mathcal{I} = (\mathcal{G}, \mathcal{A}, \rho, \tau)$.

A chaque conformation G_i est associé l'ensemble des images où elle apparait $\mathcal{I}_{G_i} = \{I^1_{G_i}, I^2_{G_i}, ..., I^k_{G_i}; 1 \leq k \leq S\}$.

L'algorithme comporte deux étapes principales :

1. **Initialisation** : pour la première image I_1 de la trajectoire :
   (a) Calculer les orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$,
   (b) Construire G_1 à partir de I_1 en calculant les ensembles E_{C_1}, A_{H_1} et E_{I_1},
   (c) Définir I_1 comme étant la première image de référence pour les orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$,
   (d) Initialiser $\mathcal{I}_{G_1}$ l'ensemble des images, où G_1 apparait, à I_1,
   (e) Initialiser l'ensemble des conformations $\mathcal{G}$ à G_1,
   (f) Initialiser l'ensemble des transitions $\mathcal{A}$ à l'ensemble vide.
2. **Analyse de la dynamique conformationnelle** : pour chaque image I_i de la trajectoire :
   (a) Si le déplacement des atomes (max_1+max_2) est supérieur à la distance de seuil $(\alpha_{\mathcal{O}_C} - 1) \times D_C$ alors recalculer les orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$ et l'image I_i devient la nouvelle image de référence pour les orbites.
   (b) Sinon, si max_1+$max_2 > (\alpha_{\mathcal{O}_H}-1)\times D_H$ ou max_1+$max_2 > (\alpha_{\mathcal{O}_I}-1)\times D_I$:
   - Si $max_1 + max_2 > (\alpha_{\mathcal{O}_H} - 1) \times D_H$ alors recalculer les orbites $\mathcal{O}_H$ et changer l'image de référence de ces orbites par l'image courante I_i,
   - Si $max_1 + max_2 > (\alpha_{\mathcal{O}_I} - 1) \times D_I$ alors recalculer les orbites $\mathcal{O}_I$ et changer l'image de référence de ces orbites par l'image courante I_i.

   (c) Calculer les ensembles de liaisons et d'interactions E_{C_i}, A_{H_i} et E_{I_i} autour des orbites calculées (construire G_i).
   (d) Tester si G_i est isomorphe à G_{i-1} :
   - Si G_i est égale à G_{i-1}, en comparant les matrices d'adjacence en même temps que le calcul des E_{C_i}, A_{H_i} et E_{I_i}, la conformation G_{i-1} est maintenue à l'image I_i.
   - Si les matrices d'adjacence ne sont pas identiques, le test d'isomorphisme (l'approche de McKay) est appliqué (décrit dans l'étape ci-après (e)).

   (e) Tester l'isomorphisme avec les conformations déjà explorées : tester s'il existe une conformation G_j dans $\mathcal{G}$ isomorphe à G_i :
   - Si $\exists G_j \in \mathcal{G}$ telle que G_j isomorphe à G_i, alors l'image I_i présente une nouvelle apparition d'une conformation déjà existante G_j, et donc ajouter I_i à I_{G_j}.
   - Si $\nexists G_j \in \mathcal{G}$ telle que G_j isomorphe à G_i, alors l'image I_i présente une nouvelle conformation explorée, et donc ajouter G_i à $\mathcal{G}$ et initialiser $\mathcal{I}_{G_i}$ à I_i.

(f) Ajouter (G_{i-1}, G_i) à $\mathcal{A}$,

A la fin de l'analyse de la trajectoire, l'ensemble des conformations non isomorphes explorées est obtenu avec leurs séquences d'images où elle apparaissent. L'algorithme fournit aussi le graphe des transitions entre ces différentes conformations. Les algorithmes sont présentés dans l'annexe C.

3.5 Complexité théorique de l'algorithme

Pour une trajectoire de S images, où chaque image est composée de N atomes ($N << S$), la complexité théorique dans le pire cas est exponentielle en nombre de conformations. En effet, le pire cas est d'avoir à chaque image une conformation différente.

Cette complexité théorique est calculée comme suit :

- **Construction des conformations** :
  - Lire les images de la trajectoire revient à parcourir séquentiellement le fichier d'entrée de cette trajectoire. Cette lecture a une complexité de $O(S \times N)$.
  - Les calculs des orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$ ont une complexité $O(N^2)$, $O(n_H \times N)$ et $O(n_i \times N)$ respectivement, où n_H et n_I sont le nombre d'atomes d'hydrogène et le nombre d'atomes concernés par les interactions intermoléculaires électrostatiques respectivement. Les calculs d'orbites ne se font que sur les images de référence. Soient R_C le nombre d'images de référence pour les liaisons covalentes, R_H le nombre d'images de référence pour les liaisons hydrogène et R_I le nombre d'images de référence pour les interactions intermoléculaires électrostatiques. La complexité de cette étape est donc $O(R_C \times (N^2 + n_H \times N + n_i \times N) + R_H \times n_H \times N + R_I \times n_i \times N)$.
  - Calculer les ensembles des liaisons et interactions autour des orbites a une complexité de $O(S \times (N \times |\mathcal{O}_C| + n_H \times |\mathcal{O}_H| + n_I \times |\mathcal{O}_I|)$ où n_H et n_I sont le nombre d'atomes d'hydrogène et le nombre d'atomes concernés par les interactions intermoléculaires électrostatiques respectivement. Nous verrons dans la section 3.7 qu'en pratique les tailles des orbites $|\mathcal{O}_C|$, $|\mathcal{O}_H|$ et $|\mathcal{O}_I|$ sont négligeables devant N.
- **Isomorphisme entre les conformations** : il n'a pas été prouvé si le problème d'isomorphisme est dans P [41, 42]. L'un des algorithmes les plus performants actuellement est de complexité quasi-polynomial ($exp((logn)^{O(1)})$) [43]. Dans notre approche, on applique deux niveaux de comparaisons entre conformations :

- *Egalité entre G_i et G_{i-1}* : dans ce test, nous n'appliquons pas un algorithme dédié à l'isomorphisme de graphes mais nous faisons une comparaison des matrices d'adjacence entre la conformation courante G_i et la conformation précédente G_{i-1}. Ce test est immédiat car la comparaison est effectuée en même temps que le calcul des ensembles de liaisons E_{C_i}, A_{H_i} et E_{I_i}. Ce test est important, car selon l'hypothèse 1, les atomes sont conservés sur la trajectoire.

- *Isomorphisme entre G_i et les conformations de l'ensemble $\mathcal{G}$* : si le premier test est négatif (les matrices d'adjacence ne sont pas identiques), un test d'isomorphisme est appliqué entre cette conformation et les conformations précédemment identifiées. Car on peut avoir des atomes similaires (des atomes de même type chimique qui jouent le même rôle, mais qui n'ont pas les mêmes identifiants sur les deux conformations). Dans ce test on applique l'algorithme de Mckay [33, 34]. Nous calculons la signature de la conformation G_i avec une complexité théorique exponentielle et nous la comparons aux signatures déjà obtenues. Le pire cas est d'avoir une conformation à chaque image, ce qui donne une complexité exponentielle à notre algorithme.

Nous rappelons que l'utilisation des orbites et des images de référence réduit le nombre de calculs mais ne réduit pas la complexité de l'algorithme dans le pire cas. Nous verrons dans la section 3.7, qu'en pratique, cette complexité n'est jamais atteinte car le deuxième test d'isomorphisme est rarement appliqué.

3.6 Extensions de l'algorithme

L'algorithme permet d'identifier l'ensemble des conformations à partir d'une séquence d'images d'une trajectoire. Des traitements supplémentaires ont été ajoutés à l'algorithme pour différents objectifs que nous allons présenter ci-après.

3.6.1 Identification des conformations stables

Lors de l'analyse d'une trajectoire, certaines conformations n'apparaissent que sur des durées très courtes. Ces conformations sont considérées comme états transitoires et sont peu importants pour les chimistes. A cet effet, nous faisons un tri des conformations obtenues entre conformations stables (conformations qui durent suffisamment sur la trajectoire) et états transitoires.

Définition 22. *Soient une trajectoire $\mathcal{I}$ de S images et G une conformation explorée dans cette trajectoire. Une durée de la conformation G est le nombre d'images de l'image où elle apparait jusqu'à l'image où une autre conformation apparait.*

Définition 23. *Soient une trajectoire $\mathcal{I}$ de S images et G une conformation explorée dans cette trajectoire. La conformation G est un **état transitoire** si et seulement si toutes les durées d'apparition de G sont inférieures à une durée de seuil T_r (voir annexe B.1 pour la valeur de T_r)*

La figure 3.3 montre un exemple d'une trajectoire de 10200 images (*i.e.* $4ps$). Sur cette trajectoire, l'algorithme a identifié 6 conformations dont 4 sont des conformations stables (G_1, G_2, G_3 et G_4) et 2 sont des états transitoires (G_5 et G_6).

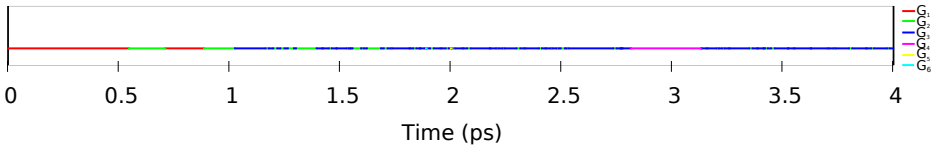

FIGURE 3.3 – Courbe d'évolution des conformations au cours du temps. Les deux états transitoires apparaissent une seule fois et sont présentés en jaune et en bleu clair.

3.6.2 Etudes des mouvements rotationnels dans les systèmes moléculaires

Au-delà de la dynamique des liaisons, il peut y avoir des rotations autour de certaines **liaisons covalentes**. C'est un mouvement circulaire autour d'un axe défini par une liaison covalente.

Soit $\mathcal{G}$ l'ensemble des graphes de conformations. Étant donné $G \in \mathcal{G}$, le graphe $G_R = (V_C, E_C^{IN})$ est le sous-graphe de G réduit aux arêtes (liaisons covalentes) de E_C tel que V_C est l'ensemble de sommets incidents à au mois une arête de E_C^{IN} et :

$$E_C^{IN} = \{[a,b] \in E_C : \phi(a) \neq \mathsf{H}, \phi(b) \neq \mathsf{H}, deg(a) > 1 \, et \, deg(b) > 1\}$$

Autrement dit, nous nous intéressons qu'aux liaisons covalentes internes, en négligeant les liaisons formées avec des atomes d'hydrogène.

Nous partitionnons $\mathcal{G}$ en classes d'équivalence : $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_k$ telles que deux graphes G_1 et G_2 de $\mathcal{G}$ sont dans la même classe d'équivalence $\mathcal{G}_i$ si et seulement si ont le même graphe réduit G_{R_i}.

Pour identifier les axes de rotation dans les graphes de conformations, nous nous intéressons aux **isthmes** dans le graphe réduit de chaque classe d'équivalence.

Définition 24. *Soit un graphe non orienté $G = (V, E)$ où V est l'ensemble de sommets et E l'ensemble d'arêtes. Une arête $[a, b] \in E$ est un **isthme** si sa suppression augmente le nombre de composantes connexes de G [44].*

Pour toute classe d'équivalence $\mathcal{G}_i$, pour tout isthme $[a, b]$ du graphe réduit G_{R_i} :

- $[a, b]$ est **un axe de rotation simple** si et seulement si $\forall G_j \in \mathcal{G}_i$, $[a, b]$ est un isthme.
- $[a, b]$ est **un axe de rotation conformationnel** si et seulement si $[a, b]$ n'est pas un axe de rotation simple et $\exists G_j \in \mathcal{G}_i$ où $[a, b]$ est un isthme.

La figure 3.4 montre un exemple d'axe de rotation conformationnel. L'isthme $[C_1, N_2]$ reste un isthme dans le graphe à droite de la figure 3.4 mais il ne l'est pas dans le graphe à gauche de la figure et donc $[C_1, N_2]$ est considéré comme un axe de rotation conformationnel.

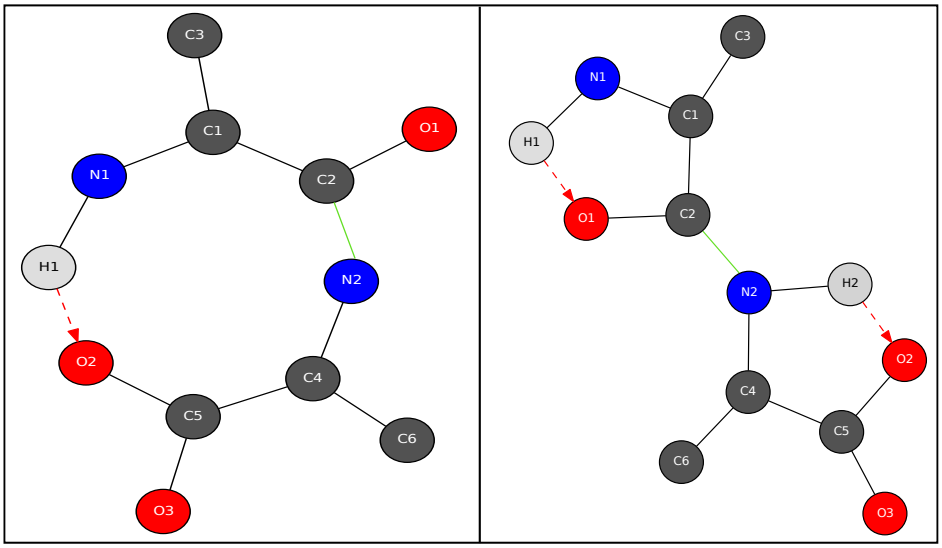

FIGURE 3.4 – Exemple d'axe de rotation. L'axe de rotation conformationnel est présenté en arête verte.

L'identification des rotations dans un système moléculaire constitue une information supplémentaire sur sa dynamique conformationnelle qui est intéressante, car elle peut expliquer les changements des liaisons qui ont eu lieu au cours de la trajectoire.

D'un point de vu algorithmique, nous utilisons un algorithme de parcours en largeur du graphe, pour identifier les axes de rotations. L'algorithme est donné en annexe C.

3.6.3 Analyse de plusieurs trajectoires simultannément

Selon le temps d'exploration et l'énergie fixée dans le système, une trajectoire peut contenir une ou plusieurs conformations. Pour avoir une bonne exploration de la surface d'énergie potentielle, c'est-à-dire explorer le maximum de conformations et de transitions possibles, il faut analyser un nombre important de trajectoires qui peut atteindre des milliers de trajectoires pour un même système moléculaire, car plusieurs trajectoires peuvent avoir les mêmes ensembles de conformations explo-

rées. Pour éviter d'analyser séparément les trajectoires puis comparer les conformations explorées ce qui est en pratique coûteux en temps, nous avons ajouté à notre algorithme la possibilité d'analyser plusieurs trajectoires simultanément.

Soient n trajectoires, notées $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n$. A chaque trajectoire $\mathcal{I}_i$ $(1 \leq i \leq n)$ est associé son graphe de transitions $\mathcal{G}_{\mathcal{I}_i} = (\mathcal{G}_i, \mathcal{A}_i)$ où $\mathcal{G}_i$ l'ensemble des conformations explorées dance cette trajectoire. Le graphe $\mathcal{G}_{\mathcal{I}} = (\mathcal{G}, \mathcal{A}) = \bigcup_{i \in [\![1,n]\!]} \mathcal{G}_{\mathcal{I}_i}$ est le graphe de transitions issu de l'analyse de toutes les trajectoires. Pour obtenir ce graphe, l'algorithme analyse les trajectoires $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n$ en utilisant l'algorithme décrit dans la section 3.4. A chaque analyse d'une trajectoire $\mathcal{I}_i$, l'ensemble des conformations utilisé dans le test d'isomorphisme (l'isomorphisme entre la conformation courante et les conformations déjà explorées) est $\mathcal{G}$, tel que $\mathcal{G} = \bigcup_{j \in [\![1,i-1]\!]} \mathcal{G}_j$. Autrement dit, l'isomorphisme se fait avec toutes les conformations explorées dans toutes les trajectoire analysées précédemment.

L'analyse simultanée de toutes les trajectoires permet non seulement d'analyser plus rapidement un nombre élevé de trajectoires et d'énumérer les conformations explorées, mais aussi pour chaque trajectoire de fournir la dynamique conformationnelle du système moléculaire (les graphes de transitions).

Prenons l'exemple des trajectoires de dissociation du peptide induite par collision (cf. section 1.1.1). Dans ce système moléculaire, nous avons initialement une conformation à un peptide en un seul fragment (une seule composante connexe dans le graphe mixte) et un gaz inerte (par exemple un atome d'argon). Le système est soumis à plusieurs valeurs de température (soit m valeurs de température). Pour chaque valeur de température i, une trajectoire $\mathcal{I}_i$ est générée contenant S images (S étant fixé au départ). L'objectif de l'algorithme est d'analyser ces trajectoires et d'observer les différents fragments obtenus suite à la collision entre le gaz et le peptide. L'ensemble des conformations $\mathcal{G}$ contient donc les ensembles de fragments obtenus sur toutes les trajectoires. L'algorithme classe les trajectoires selon les graphes des transitions $\mathcal{G}_{\mathcal{I}_i} = (\mathcal{G}_i, \mathcal{A}_i)$ associés à chaque trajectoire $\mathcal{I}_i$ $(1 \leq i \leq n)$. Cela permet de voir la fréquence de chaque chemin emprunté par la conformation initiale et de déterminer l'effet de la température sur la dynamique conformationnelle du système. Le graphe des transitions $\mathcal{G}_{\mathcal{I}} = (\mathcal{G}, \mathcal{A})$ permet de voir la corrélation entre les différentes trajectoires (l'intersection des chemins empruntés).

3.7 Evaluation et performance de l'algorithme

L'algorithme avec ses extensions a été implémenté en langage C. Nous rappelons que les trajectoires analysées concernent principalement les systèmes en phase gazeuse (cf. chapitre 1) et plus précisément les systèmes suivants :

- Système avec peptide isolé : il s'agit d'une seule molécule rigide dans le sys-

tème. En général, le changement conformationnel pour ce genre de systèmes repose sur la dynamique des liaisons hydrogène et il ne contient pas d'interactions intermoléculaires électrostatiques.

- Système de dissociation du peptide induite par collision : dans ce système, en général, au début de la trajectoire la molécule est en un seul fragment, ensuite elle passe à une molécule avec plusieurs fragments suite à une collision entre un gaz inerte comme une atome d'argon et la molécule. Dans ce système il y a une dynamique des liaisons hydrogène et également une dynamique des liaisons covalentes et les interactions intermoléculaires électrostatiques ne sont pas considérées.
- Cluster : c'est un système contenant un ion avec une ou plusieurs molécules (exemple des molécules d'eau). Dans ce système, on s'intéresse principalement aux interactions créées entre cet ion et le reste des molécules (interactions intermoléculaires électrostatiques). Il peut y avoir une dynamique des autres types de liaisons.

Pour évaluer les performances de l'algorithme proposé, nous l'avons testé sur plusieurs trajectoires de ces systèmes. L'objectif de ces tests est, d'une part, de fixer les coefficients d'optimisation utilisés pour les calculs d'orbites comme le montre la section 3.7.1. D'autre part, les tests permettent d'évaluer l'algorithme en termes de temps d'exécution. Nous verrons dans la section 3.7.2 que les résultats obtenus en termes de nombre de conformations explorées et de temps d'exécution confirment que la complexité théorique n'est jamais atteinte.

Toutes les trajectoires analysées ainsi que les résultats présentés dans cette section sont disponibles sur l'interface web que nous avons développée et sont accessibles sur le lien `http://hydrochronographe.prism.uvsq.fr`[2]. Ainsi que, les valeurs des paramètres utilisés dans les tests sont données en annexe B.1.

3.7.1 Choix du paramètre d'optimisation

Nous avons, dans la section 3.2, indiqué que pour identifier les conformations explorées dans une trajectoire, nous utilisons la notion d'orbites. Le calcul de liaisons et d'interactions se fait en utilisant les voisinages des atomes. Ces orbites ainsi que les images de référence correspondantes dépendent d'un coefficient qui change selon le type de liaison ou d'interaction calculée. Nous présentons dans cette section comment fixer le coefficient utilisé pour les orbites des liaisons hydrogène qui est $\alpha_{\mathcal{O}_H}$.

Afin de donner une valeur à ce coefficient, nous avons analysé l'évolution du nombre d'images de référence et la taille moyenne des orbites en variant le coefficient entre 0 et 10.

2. Une description de l'interface web est donnée en annexe E.

Nous remarquons de la figure 3.5 que pour les valeurs de $\alpha_{\mathcal{O}_H}$ proches de zéro, nous changeons plus souvent l'image de référence et la taille moyenne des orbites tend vers zéro. Par contre, pour des grandes valeurs de $\alpha_{\mathcal{O}_H}$ le nombre d'images de référence diminue et la taille moyenne des orbites augmente jusqu'à une taille maximale qui n'est que le nombre total d'atomes du système.

La valeur choisie pour $\alpha_{\mathcal{O}_H}$ doit non seulement minimiser le nombre d'images de référence calculée et la tailles des orbites mais surtout doit garantir que toutes les conformations explorées dans la trajectoire soient identifiées par l'algorithme (voir le lemme **??** et sa preuve). Pour cela, nous avons vérifié pour chaque valeur de $\alpha_{\mathcal{O}_H}$ que le nombre de conformations identifiées est égale au nombre de conformations identifiées pour $\alpha_{\mathcal{O}_H} = 0$.

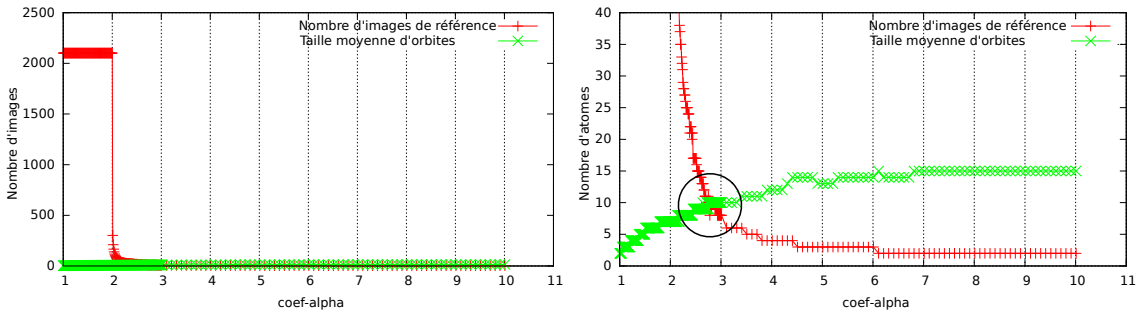

FIGURE 3.5 – Evolution du nombre d'images de référence et la taille moyenne des orbites en fonction du coefficient $\alpha_{\mathcal{O}_H}$. La courbe en rouge présente l'évolution du nombre d'images de référence et la courbe en vert présente l'évolution de la taille moyenne des orbites d'atomes d'hydrogène.

Nous avons également comparer le temps d'exécution du programme sur plusieurs trajectoires en prenant les mêmes valeurs du coefficient $\alpha_{\mathcal{O}_H}$ $[\![0, 10]\!]$
. Les résultats obtenus étaient similaires pour les différentes trajectoires. La figure 3.6 montre un exemple de ces courbes. Nous remarquons que le temps d'exécution est très élevé quand le coefficient s'approche de zéro. Cela est du au calcul fréquent des images de référence et donc le calcul des ensembles de liaisons et des orbites. Puis le temps d'exécution diminue pour re-augmenter quand le coefficient est grand car le nombre de comparaisons effectuées est élevé quand la taille des orbites augmente.

En analysant les différentes courbes, nous avons remarqué qu'une bonne valeur du coefficient $\alpha_{\mathcal{O}_H}$ est comprise entre 2 et 4 expérimentalement.

En utilisant le même principe, nous avons effectué des tests pour trouver les bonnes valeurs de $\alpha_{\mathcal{O}_C}$ et $\alpha_{\mathcal{O}_I}$ pour calculer respectivement les orbites des liaisons covalentes et d'interactions intermoléculaires électrostatiques et leurs images de référence.

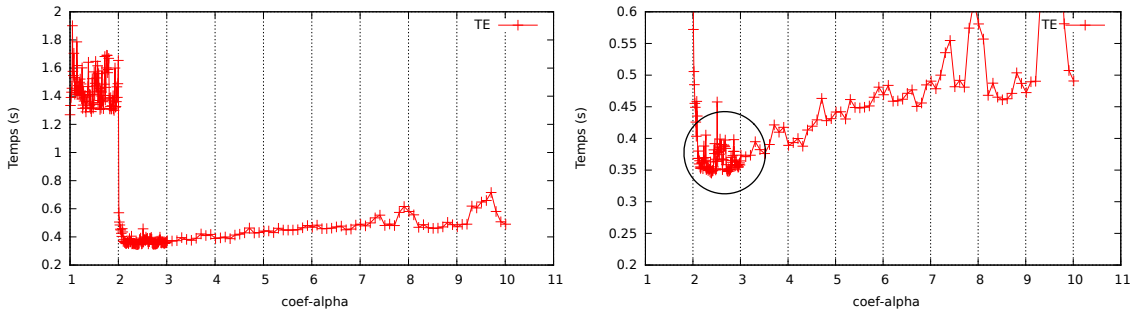

FIGURE 3.6 – Evolution du temps d'exécution en fonction du coefficient $\alpha_{\mathcal{O}_H}$. La figure à droite présente un zoom sur la partie optimale de la courbe de la figure à gauche.

3.7.2 Evaluation du temps de calcul

Nous avons vu dans la section 3.5 que, dans le cas le plus défavorable, la complexité théorique de l'algorithme est exponentielle. Cela est du principalement aux tests d'isomorphisme. Cependant, le tableau 3.1 montre que le nombre de conformations identifiées est négligeable en comparaison avec la taille de la trajectoire, ce qui revient à dire que le test d'isomorphisme entre la conformation courante et les conformations déjà explorées se fait peu de fois et avec peu de conformations.

Dans le tableau 3.1, nous résumons les résultats des tests effectués sur les trajectoires des systèmes décrit ci-dessus. Ces trajectoires diffèrent en termes du nombre d'images, de la taille des image (nombre d'atomes) et du type du système étudié. Durant les tests, nous avons analysé les paramètres suivants :

- Le nombre d'images de référence calculées pour les liaisons hydrogène (R_H).
- La taille maximale des orbites d'atomes d'hydrogène (nombre d'atomes maximum dans une orbite, $N_{\mathcal{O}_H}$).
- Le nombre de conformations stables identifiées ainsi que le nombre d'états transitoires (C_s et C_t).
- Le temps d'exécution (TE), en secondes, sachant que les tests ont été effectués sur un ordinateur qui a les caractéristiques suivantes : MacBook Pro, modèle 2012, processeur 2,9 GHz Intel Core i7 et une RAM de 8Go.

D'une façon générale, les résultats montrent que le nombre d'images de références calculées ainsi que le nombre de conformations identifiées sont négligeables en comparaison avec les tailles des trajectoires (nombre d'images S).

Si nous considérons la ligne du tableau 3.1, indiquée en gras (la dernière ligne de la trajectoire du système $C_6H_{13}N_2O_3$), seules **5** conformations (entre conformations stables et états transitoires) ont été identifiées sur **27000** images. Cela montre que le calcul des ensembles de liaisons et des orbites est effectué très peu de fois. De même, le test d'isomorphisme entre la conformation courante et les conformations

TABLE 3.1 – Résultats de l'analyse de la dynamique conformationnelle avec l'algorithme proposé

Formule empirique de la molécule*	Nombre d'atomes (N)	Nombre d'images (S)	Nombre d'images de référence (R_H)	Taille max. d'orbites ($N_{\mathcal{O}_H}$)	Nombre de conformations stables (C_s)	Nombre d'états transitoires (C_t)	Temps d'exécution (s)	Ref.
Peptides protonés isolés								
$C_6H_{13}N_2O_3$	24	5900 10200 10200 **27000**	17 7 9 **6**	3 3 2 **2**	3 4 3 **3**	3 2 2 **2**	3 0.88 0.76 **1.83**	[45]
$C_9H_{18}N_3O_4$	34	2101 4184 6001 6599 7325	2 3 7 8 6	4 2 2 2 3	4 2 3 1 1	0 9 5 3 3	0.31 0.75 0.83 0.87 0.94	[46]
		7500	8	3	2	2	0.98	
$C_{21}H_{38}N_7O_8$	74	26601 59201	4 4	7 6	9 12	35 68	7.53 15.76	[47]
$C_{26}H_{39}N_7O_8$	80	75000	4	6	2	19	20.87	-
Systèmes de dissociation induite par collision								
$C_4H_{10}N_3O_2Ar$	20	51	26	2	3	0	0.02	[48]
Clusters								
$Li^+(H_2O)_4$	13	50001	2 17 39	3 3 3	1 1 2	0 2 3	2.29 2.51 2.77	[49]

*Certaines trajectoires concernent la même molécule sauf qu'elles sont générées dans différentes conditions thermodynamiques.

déjà identifiées est appliqué au plus pour 5 conformations. Ces résultats affirment qu'en pratique la complexité théorique n'est jamais atteinte.

Dans le cadre d'évaluation des performances, il est aussi important de vérifier la fiabilité de l'algorithme en termes de conformations fournies. Quelques trajectoires ont été déjà analysées et publiées [45, 46, 49, 48]. Dans ce cas, la validation des résultats trouvés par notre algorithme se fait par comparaison avec ces trajectoires. Par exemple, pour les systèmes $C_6H_{13}N_2O_3$ et $C_4H_{10}N_3O_2Ar$ nous avons retrouvé les mêmes conformations publiées [45, 48]. De plus, l'algorithme ne fournit pas seulement les conformations stables mais aussi d'autres informations supplémentaires pertinentes, à savoir les états transitoires, les transitions observées entre ces différentes conformations (stables et états transitoires), l'évolution au fil du temps des différentes liaisons, etc.

Par contre, pour le système $C_{21}H_{38}N_7O_8$ l'algorithme a identifié, en outre des conformations publiées [47], d'autres conformations. Cela a été expliqué par le fait que la publication était basée sur une partie de la molécule et non pas sur toute la molécule, contrairement à notre algorithme, qui fait une analyse de tout le système.

3.8 Conclusion

Une trajectoire de simulation de dynamique moléculaire peut contenir une ou plusieurs conformations du système moléculaire analysé. A travers une modélisation en graphes, nous avons conçu un algorithme qui analyse ces trajectoires afin d'identifier ces conformations ainsi que les transitions entre elles. L'algorithme combine les propriétés chimiques du système et les notions de la théorie de graphes, pour analyser la dynamique conformationnelle sur les trajectoires. Il utilise une méthode d'égalité des matrices d'adjacence et une méthode d'isomorphisme de graphes efficace pour différencier les conformations et générer les transitions entre elles. L'algorithme tient compte de trois types de liaisons : les liaisons covalentes, les liaisons hydrogène et les interactions intermoléculaires électrostatiques. Le modèle proposé permet de s'étendre à d'autres types de liaisons. Cela consiste à ajouter d'autres ensembles, définissant les autres types de liaisons au graphe mixte de la conformation et puis ajouter les tests d'isomorphisme correspondants à l'algorithme.

L'algorithme a été testé sur trois types de systèmes moléculaires en choisissant des trajectoires de taille et de complexité croissantes. Les résultats montrent qu'en pratique l'algorithme est efficace malgré sa complexité théorique exponentielle.

L'analyse des trajectoires revient à explorer la surface d'énergie potentielle et à identifier les changements de bassins conformationnels au cours du temps (cf. section 1.2.2). Comme présenté dans le chapitre précédent, en chimie théorique, il y a une deuxième communauté qui s'intéresse à des points particuliers sur la surface

d'énergie potentielle. Il s'agit des minima sur la surface d'énergie potentielle qui représentent les conformations les plus stables de la molécule et les maxima sur cette surface qui représentent les points de passage d'une conformation à une autre. Les méthodes utilisées pour trouver ces points nécessitent le calcul d'énergie potentielle du système ce qui est coûteux en termes de temps et matériel pour les méthodes précises (dynamique moléculaire). Concernant les méthodes les moins précises les résultats sont parfois erronés ce qui peut donner des faux minima/maxima. , Le chapitre suivant présente une méthode alternative pour trouver ces points en utilisant des méthodes ah doc, tout en s'affranchissant du calcul d'énergie.

Chapitre 4

Prédiction conformationnelle

Dans le chapitre précédent, nous avons présenté un algorithme qui permet d'identifier les conformations sur une ou plusieurs trajectoires de simulation de dynamique moléculaire. En d'autres termes, l'algorithme permet l'exploration d'espace des conformations sur la surface d'énergie potentielle à travers ces trajectoires (cf. chapitre 1).

En chimie théorique et computationnelle, il existe d'autres groupes de recherche qui ne s'intéressent pas à explorer l'espace des conformations mais à trouver des points particuliers sur la surface d'énergie potentielle (cf. chapitre 1). Il s'agit des minima sur cette surface qui représentent les conformations les plus stables de la molécule et les maxima qui représentent les points de passage d'une conformation à une autre (états de transition) (cf. section 1.3). toutefois, pour avoir ces points il faut disposer de plusieurs trajectoires avec plusieurs niveaux d'énergie. Cela nécessiterait des simulations excessivement longues en temps de calculs et une statistique très importante sur le nombre de simulations à réaliser. De plus, même une simulation extrêmement longue en temps ne garantit pas l'exploration de beaucoup de bassins conformationnels (voir les résultats d'analyse des trajectoires dans le chapitre 3). Les méthodes proposées pour trouver ces points nécessitent la connaissance de la géométrie du système moléculaire et le calcul d'énergie pour identifier ces points (minima et états de transition), ce qui est également coûteux en temps et en calculs.

Dans ce chapitre, nous présentons une méthode alternative pour prédire les conformations possibles d'un système moléculaire. L'idée consiste à construire un graphe orienté pondéré appelé **graphe des possibles** à partir d'une conformation initiale d'un système moléculaire. Les conformations (les sommets du graphe) ainsi que les transitions (les arcs du graphe) entre elles, sont construits sans faire appel ni aux calculs moléculaires, ni à la géométrie de la molécule. Les pondérations proposées permettent, d'un coté de donner une classification aux conformations possibles (pondération sur les sommets du graphe des possibles) pour pouvoir ensuite choisir les conformations les plus stables, et d'un autre coté, elles permettent de donner une vison à gros grain sur les coûts énergétiques qu'il faut pour passer d'une conformation à une autre (pondération sur les arcs).

Nous nous intéressons dans ce chapitre aux peptides isolés (cf. section 1.1.1), où le changement conformationnel dépend uniquement de la dynamique des liaisons

hydrogène.

La section 4.1 présente le modèle et les différentes règles utilisées pour le graphe des possibles. Ensuite, nous expliquons, dans la section 4.2, le processus de pondération du graphe des possibles en utilisant des mesures ah doc et dans la section 4.3, la classification des conformations. La section 4.4 se focalise sur l'utilisation du graphe des possibles, à savoir trouver des chemins de coût minimum entre les conformations, et explique le lien avec l'analyse des trajectoires présentée dans le chapitre précédent. Enfin, la section 4.5 présente quelques résultats numériques des tests de la méthode sur des peptides isolés.

4.1 Graphe des possibles

Nous avons vu dans le chapitre précédent que, malgré les tailles élevées des trajectoires (des milliers d'images), peu de conformations sont explorées, avec une moyenne inférieure à 5 conformations par trajectoire. En outre, générer plusieurs trajectoires pour un même système moléculaire ne garantit pas d'explorer des bassins conformationnels différents. Par exemple, sur 4 trajectoires d'un tripeptide ($C_9H_{18}N_3O_4$) de tailles 7500, 4184, 6599 et 7325 images respectivement, seules 10 conformations (conformations stables et états transitoires) ont été identifiées avec l'algorithme d'analyse des trajectoires (cf. section 3.4).

Le graphe des possibles, présenté dans ce chapitre, permet d'avoir une exploration de toutes les conformations théoriquement possibles d'un système moléculaire en se basant sur la dynamique des liaisons hydrogène, sans connaitre la géométrie tridimensionnelle du système moléculaire.

Les sommets de ce graphe présentent les conformations et les arcs présentent les transitions possibles entre elles.

Définition 25 (Graphe des possibles)**.** *Le graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$, est un graphe orienté pondéré tel que :*

- *$\mathcal{G} = \{G_0, G_1, ..., G_C\}$: l'ensemble des sommets de $\mathcal{G}_P$ où chaque sommet est une conformation possible $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$(cf. section 4.1.1).*
- *$\mathcal{A} \subseteq \mathcal{G} \times \mathcal{G}$: l'ensemble des arcs de $\mathcal{G}_P$ où chaque arc est une transition possible (cf. section 4.1.2).*
- *$\Omega : \mathcal{G} \to \mathbb{R}$ une pondération sur l'ensemble des sommets $\mathcal{G}$ (cf. section 4.2.2). Dans ce qui suit, la terminologie **niveau énergétique** est utilisée pour désigner la pondération d'une conformation.*
- *$\omega : \mathcal{A} \to \mathbb{R}$ une pondération sur l'ensemble des arcs $\mathcal{A}$ (cf. section 4.2.3). De même, la terminologie **coût énergétique** est utilisée pour désigner la pondération d'un arc.*

Remarque 4. *Etant donné une trajectoire $\mathcal{I}$ d'un système moléculaire, le graphe de transitions $\mathcal{G}_\mathcal{I}$ issu de cette trajectoire représente un sous-graphe du graphe des possibles $\mathcal{G}_P$. Nous appelons ce sous-graphe un* ***graphe expérimental*** *du système moléculaire.*

4.1.1 Ensemble des conformations possibles

Dans une molécule donnée où les liaisons covalentes sont fixes et les interactions ioniques n'existent pas, la dynamique conformationnelle repose sur la dynamique des liaisons hydrogène. La construction des conformations possibles revient à générer toutes les combinaisons possibles entre les liaisons hydrogène. En théorie, si M est le nombre de liaisons hydrogène qui peuvent se former dans le système, le nombre de conformations possibles est 2^M. Le nombre est non seulement exponentiel mais aussi beaucoup de conformations ne sont pas réalistes. Afin de diminuer ce nombre, nous introduisons des contraintes structurelles sur les graphes des conformations.

Soient un système moléculaire $Mol = (V_M, Cov, Hydro, Inter)$, et $G = (V, E_C, A_H, E_I)$ une conformation de ce système tels que :

- $|Hydro| = M$
- $A_H = \emptyset$
- $E_I = \emptyset$

Nous rappelons que E_C, A_H et E_I sont respectivement les ensembles de liaisons covalentes, de liaisons hydrogène et d'interactions intermoléculaires électrostatiques et le graphe G_{E_C} est le sous-graphe non-orienté de G réduit aux arêtes de E_C et à leurs extrémités.

Avant d'introduire les étapes pour calculer les conformations possibles, nous définissons les dépendances suivantes sur l'ensemble des liaisons hydrogène $Hydro$ en se basant sur les règles décrites en section 2.2 :

- **La non-existence (**$N \subset Hydro \times Hydro$**)** : une liaison hydrogène (d_1, a_1) implique la non-existence d'une autre liaison hydrogène (d_2, a_2) que nous noterons $N((d_1, a_1), (d_2, a_2))$, si une des conditions est vérifiée :
  - Le donneur d_1 (respectivement l'accepteur a_1) et l'accepteur a_2 (respectivement le donneur d_2) sont identiques .
  - Le donneur d_1 et le donneur d_2 sont identiques et il existe un unique atome h tel que $\phi(h) = \mathsf{H}$ et $(d_1, h) \in E_C$.
  - La liaison hydrogène (d_1, a_1) domine la liaison hydrogène (d_2, a_2).
- **Domination (**$D \subset Hydro \times Hydro$**)** : une liaison hydrogène (d_1, a_1) domine une liaison hydrogène (d_2, a_2) que nous noterons $D((d_1, a_1), (d_2, a_2))$, si la plus

courte chaine entre d_2 et a_2 dans G_{E_C} est totalement inclue dans la plus courte chaine entre d_1 et a_1. Nous rappelons qu'une chaine entre deux atomes a et b de V est une suite $(a_1, a_2, ...a_n)$ de sommets de G_{E_C} telle que $a_1 = a$, $a_n = b$ et chaque deux sommets consécutifs a_i et a_{i+1} sont reliés par une arête de G_{E_C}, cette chaine est une plus courte si toutes les chaines entre a et b sont de taille supérieure(cf. annexe A). La plus courte chaine n'est pas nécessairement unique, dans ce cas, nous choisissons une plus courte chaine aléatoirement). Le principe est que l'apparition d'une liaison hydrogène (d_1, a_1) fige tous les axes de rotations (cf. section 3.6.2) appartenant au chemin de d_1 à a_1.

- **L'indépendance ($I \subset Hydro \times Hydro$)** : si les dépendances précédentes ne sont pas vérifiées entre deux liaisons (d_1, a_1) et (d_2, a_2), ces liaisons sont dites donc indépendantes et nous notons $I((d_1, a_1), (d_2, a_2))$.

Etant donné un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$, le calcul de $\mathcal{G}$ est **itératif**, c'est-à-dire que nous calculons les conformations avec 1 liaison hydrogène, ensuite les conformations avec 2 liaisons hydrogène en utilisant les conformations et ainsi de suite jusqu'à que nous pouvons plus ajouter de nouvelles liaisons hydrogène.

Nous notons $\mathcal{G}^k = G_1^k, G_2^k, ..., G_{m_k}^k$ ($0 \leq k \leq M$, $0 \leq m_k \leq \binom{k}{M}$), l'ensemble des conformations possibles ayant k liaisons hydrogène présentes. L'ensemble $\mathcal{G}^0$ contient une seule conformation. Il s'agit de la conformation de départ G.

La construction de $\mathcal{G}^k$ à partir de l'ensemble $\mathcal{G}^{k-1}$ ($1 \leq k \leq M$) revient à ajouter une liaison hydrogène de $Hydro$ pour chaque conformation de $\mathcal{G}^{k-1}$ comme suit :

1. Pour toute conformation $G_i^{k-1} = (V_i, E_{C_i}, A_{H_i}, E_{I_i}) \in \mathcal{G}^{k-1}$
   (a) Pour toute liaison hydrogène $(d, a) \in Hydro$ telle que $(d, a) \notin A_{H_i}$:
      i. Soit G_j^k une conformation obtenue de G_i^{k-1}, telle que $A_{H_j} = A_{H_i} \cup (d, a)$,
      ii. S'il existe une conformation G^k dans $\mathcal{G}^k$ isomorphe à G_j^k alors retourner à l'étape (a).
      iii. Si $\forall (d\prime, a\prime) \in A_{H_i}$: $I((d\prime, a\prime), (d, a))$ et il existe au moins deux axes de rotation conformationnels dans une plus courte chaine entre d et a dans le graphe G_i^{k-1} alors ajouter G_j^k à $\mathcal{G}^k$.
      iv. Retourner à l'étape (a).
2. Si $\mathcal{G}^k = \emptyset$ alors c'est la fin de l'algorithme et $\mathcal{G} = \bigcup_{i \in [\![0, k-1]\!]} \mathcal{G}^i$

Lemme 2. *$\forall G_i^k \in \mathcal{G}^k$ ($0 \leq k \leq M$, $0 \leq i \leq \binom{k}{M}$), pour toute liaison hydrogène (a, b) présente dans G_i^k, le graphe obtenu en supprimant (a, b) est un graphe de $\mathcal{G}^{k-1}$.*

Preuve 2. *Soient $G_i^k \in \mathcal{G}^k$ ($1 \leq k \leq M$). Cette conformation G_i^k vérifie que $\forall (d, a) \in A_{H_i}$, $(d\prime, a\prime) \in A_{H_i}$: $I((d\prime, a\prime), (d, a))$ et pour toute liaison hydrogène (d, a) il*

existe deux axes de rotation conformationels dans une plus courte chaine entre d et a dans le graphe G_i^{k-1} sans (d, a).

On suppose qu'il existe une conformation G_j^{k-1} telle que $|A_{H_j}| = |A_{H_i} \cap A_{H_j}| - 1$ et qu'elle ne peut pas exister. Donc, il existe deux liaisons hydrogène (d, a) et $(d\prime, a\prime)$ de A_{H_j} telle que $I((d\prime, a\prime), (d, a))$ n'est pas vérifiée ou pour une liaison (d, a) il n'existe pas deux axes de rotation conformationnels dans une plus courte chaine entre d et a dans le graphe G_i^{k-2}. Or $A_{H_j} \subset A_{H_i}$, d'où la contradiction.

Pour rendre l'algorithme plus efficace en particulier en termes d'espace mémoire, nous ne sauvegardons que les conformations de l'étape courante et celles de l'étape précédente et nous utilisons une table de hashage pour la recherche des conformations. Pour la recherche des chaines et des axes de rotations, nous utilisons les mêmes algorithmes que ceux de l'analyse conformationnelle des trajectoires (cf. chapitre 3), c'est-à-dire un algorithme du parcours en largeur du graphe et un algorithme de recherche des composantes connexes. L'algorithme général pour construire l'ensemble des conformations possibles du graphe des possibles est donné en annexe C.

4.1.2 Ensemble des transitions possibles

En se basant sur le principe des transitions entre deux conformations consécutives défini dans la section 3.3.2, nous définissons **une transition** entre deux conformations possibles comme suit :

Définition 26. *Soient un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$, $Hydro$ l'ensemble des liaisons hydrogène possibles et deux conformations possibles $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ et $G_j = (V_j, E_{C_j}, A_{H_j}, E_{I_j})$ de $\mathcal{G}$. La transition $(G_i, G_j) \in \mathcal{A}$ si et seulement si une des deux conditions suivants est respectée :*

- $|A_{H_j}| = |A_{H_j} \cap A_{H_i}| + 1$ *, c'est-à-dire une apparition d'une liaison hydrogène.*
- *Ou $|A_{H_i}| = |A_{H_i} \cap A_{H_j}| + 1$, c'est-à-dire une disparition d'une liaison hydrogène.*

4.2 Pondération du graphe des possibles

Un aspect important de la dynamique moléculaire est d'être capable d'identifier les conformations les plus stables. Il s'agit des conformations du plus bas niveau d'énergie. Afin d'atteindre cet objectif deux pondérations sont appliquées sur le graphe des possibles. Ces pondérations concernent les conformations (les sommets du graphe), ce qui permet d'avoir une classification, et les transitions (les arcs du graphe), ce qui permet de choisir des chemins meilleurs que d'autres entre les conformations. Dans ce qui suit, nous utilisons le terme **niveau énergétique** pour désigner la pondération sur un sommet du graphe des possibles et le terme **coût énergétique** pour désigner la pondération sur un arc du graphe des possibles.

Avant d'introduire ces pondérations, nous devons définir la pondération sur les sommets et les arêtes d'une conformation (graphe mixte).

4.2.1 Pondération sur le graphe mixte d'une conformation

Dans une conformation, chaque atome participe à sa stabilité. Un atome impliqué dans une liaison hydrogène n'a pas le même impact qu'un atome libre (un atome qui n'est pas impliqué dans des liaisons hydrogène). Cet impact dépend des contraintes structurelles imposées par les liaisons hydrogène. En effet, les liaisons hydrogène entrainent des repliements en cycles au niveau du graphe mixte. Plus le nombre de liaisons hydrogène est grand, plus le nombre de cycles augmente, moins il y a des axes de rotations (les isthmes) (cf. section 3.6.2). Par conséquent, quand le nombre de cycles devient important, les atomes seront contraints par ces cycles et donc auront moins de liberté de rotation pour former d'autres liaisons hydrogène. Dans notre démarche nous s'intéressons aux plus courts cycles dans un graphe(cf. annexe A).

Remarque 5. *Etant donnée une conformation, un cycle dans cette conformation peut contenir une ou plusieurs liaisons hydrogène. De plus, un atome peut appartenir à un ou plusieurs cycles. Un atome appartenant à un cycle de taille élevée est moins contraint qu'un atome appartenant à un cycle de petite taille.*

La figure 4.1 montre des plus courts cycles engendrés par 3 liaisons hydrogène dans une conformation. Par exemple, le cycle C_6 contient deux liaisons hydrogène.

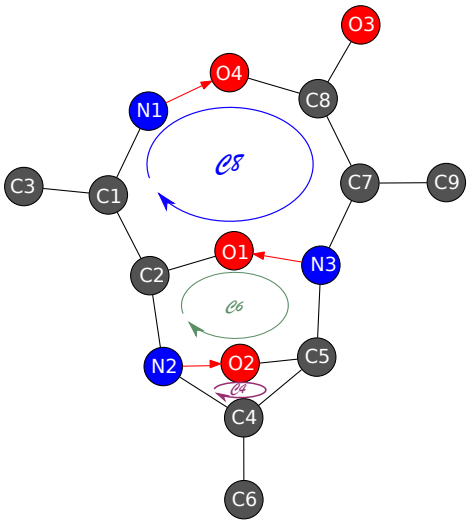

FIGURE 4.1 – Exemple de plus courts cycles. La figure montre trois plus courts cycles engendrés par trois liaisons hydrogène : un cycle de taille 8, un de taille 6 et un autre de taille 4. Les liaisons hydrogène sont présentées en arcs rouges.

Nous appelons l'impact d'un atome d'une conformation "**pénalité de l'atome**".

Cet impact est utilisé dans le calcul du coût énergétique d'une transition (cf. section 4.2.3).

Définition 27 (Pénalité d'un atome)**.** *Soient une conformation* $G = (V, E_C, A_H, E_I)$. $\forall a \in V$*, l'ensemble* $\mathcal{C}_a$=$\{C_1, C_2, ..., C_K\}$ *est l'ensemble des cycles de* G *tel que* $\forall C_i \in \mathcal{C}_a$, $a \in C_i$*. La pénalité de l'atome* a *est définie comme suit :*

$$pen_G(a) = \begin{cases} 0 & \text{si } \mathcal{C}_a = \emptyset \\ \sum\limits_{C_i \in \mathcal{C}_a} \frac{\beta}{|C_i| \times |\mathcal{C}_a|} & \text{sinon.} \end{cases}$$

La constante β *est liée au gain énergétique d'une liaison hydrogène. Cette constante est détaillée dans la section suivante 4.2.2.*

Remarque 6. *La recherche des cycles dans une conformation consiste à chercher à partir de chaque sommet les plus courts cycles, ensuite faire l'union des cycles trouvés. L'ensemble des cycles trouvés à partir d'un sommet est inclu dans l'ensemble des cycles auxquels il appartient.*

Si nous reprenons l'exemple de la figure 4.1, l'atome O2 donne un plus court cycle de taille 4 : $C_1 = (\mathsf{N2}, \mathsf{O2}, \mathsf{C5}, \mathsf{C4}, \mathsf{N2})$ mais $\mathcal{C}_{\mathsf{O2}}$ contient deux cycles. L'atome O2 appartient également à $C_2 = (\mathsf{N2}, \mathsf{C2}, \mathsf{O1}, \mathsf{N3}, \mathsf{C5}, \mathsf{O2}, \mathsf{N2})$ de taille 6, sa pénalité est donc $\frac{5\beta}{24}$. Par contre, l'atome C9 a 0 de pénalité car il n'appartient à aucun cycle.

D'un point de vue algorithmique, nous considérons une base de cycles dans le graphe de conformation $G = (V, E_C, A_H, E_I)$. Pour cela nous utilisons l'algorithme d'Horton [50] :

1. Pour tout a dans V :
2. Pour toute arête $[b, c]$ dans $E_C \cup A_H$:
   (a) Chercher une plus courte chaine entre a et b, soit $P(a, b)$,
   (b) Chercher une plus courte chaine entre a et c, soit $P(a, c)$,
   (c) Si $P(a, b) \cap P(a, c) = a$ donc il y a un cycle $C = P(a, b) \cup P(a, c) \cup (b, c)$.

Nous ne considérons pas les cycles qui sont une composition des cycles plus petits. Pour de tels cycles, nous appliquons l'élimination de Gauss [51]. L'idée est d'utiliser une matrice où les colonnes représentent les sommets de la conformation et les lignes représentent les cycles trouvés. Chaque case représente "1" ou "0" pour indiquer, respectivement, si le sommet appartient au cycle ou non. Les cycles sont ordonnés par ordre croissant de taille. L'élimination se fait à l'aide d'opérations binaires sur les lignes, à savoir le "OU exclusif" entre deux lignes. Si à une étape de l'élimination une ligne devient est à "0", le cycle correspondant est éliminé de l'ensemble des cycles trouvés.

4.2.2 Niveau énergétique d'une conformation possible

L'objectif de la prédiction conformationnelle proposée est de trouver les conformations possibles d'un système moléculaire en s'affranchissant des calculs d'énergie basé sur la géométrie comme indiqué dans le premier chapitre (cf. section 1.3), qui sont chers en temps et en matériel. Dans notre approche, l'estimation des niveaux d'énergie est basée sur des propriétés structurelles sur les graphes mixtes et les liaisons hydrogène de cette conformations. En effet, chaque liaison hydrogène contribue à la stabilité du système moléculaire. Plus le nombre de liaisons hydrogène est grand, plus le système moléculaire est stable. Cependant, quand le nombre de liaisons hydrogène augmente, les interactions entre les liaisons hydrogène augmentent (le partage des atomes) et donc le système moléculaire est soumis à des contraintes structurelles qui peuvent réduire sa stabilité. A cet effet pour définir le niveau énergétique d'une conformation, nous utilisons deux mesures :

- Le gain énergétique engendré par la présence des liaisons hydrogène, il est proportionnel au nombre de liaisons hydrogène.
- La dépense énergétique engendrée par les interactions entre les liaisons hydrogène proportionnel au nombre d'atomes partagé entre ces liaisons.

Définition 28 (Gain énergétique des liaisons hydrogène)**.** *Soit une conformation* $G = (V, E_C, A_H, E_I)$*. Le gain engendré par les liaisons hydrogène est défini comme suit :*

$$Eng_{LH}(G) = \beta \times |A_H|$$

Où β *est une constante liée à la structure de la molécule.*

Dans un peptide, il s'agit du nombre de building blocks (cf. section 1.1.1) qui le composent. Si k est le nombre du building blocks alors $\beta = 10 * k$.

Définition 29 (Dépense énergétique des liaisons hydrogène)**.** *Etant donnée une conformation* $G = (V, E_C, A_H, E_I)$*, nous définissons* $x_{(i)}$ *le nombre de sommets (atomes) appartenant à* i *liaisons hydrogène (*$0 \leq i \leq |A_H|$*). La dépense énergétique des liaisons hydrogène est définie par :*

$$pen_{LH}(G) = \begin{cases} 0 & \textit{si } |A_H| = 0, \\ \frac{\sum_{i=1,j=1}^{|A_H|} x^i_{(j)}}{|A_H|} & \textit{sinon.} \end{cases}$$

Si nous reprenons la figure 4.1, la conformation contient 3 liaisons hydrogène. Il s'agit d'un peptide avec 3 building blocks (cf. section 1.1.1). Le gain et la dépense énergétiques des liaisons hydrogène sont calculés comme suit :

- $Eng_{LH}(G) = (3 \times 10) \times 3 = 90$

- $pen_{LH}(G) = 8^1 + 2^2 + 3^3$

A partir du gain et de la dépense énergétiques des liaisons hydrogène, nous définissons le niveau énergétique d'une conformation (pondération d'une conformation) comme suit :

Définition 30 (Niveau énergétique)**.** *Soit une conformation $G = (V, E_C, A_H, E_I)$. Le niveau énergétique de G est défini par :*

$$\Omega(G) = -Eng_{LH}(G) + pen_{LH}(G)$$

Pour garder une cohérence avec les méthodes existantes (cf. chapitre 1), le gain est présenté en négatif et la dépense en positif. Les conformations les plus stables sont les conformations à plus bas niveau d'énergie.

4.2.3 Coût énergétique d'une transition possible

Nous avons vu dans le chapitre 1, que passer d'une conformation à une autre nécessite de passer une barrière d'énergie (cf. section 1.3), cette barrière change selon le type de transition (cf. section 4.1.2). En effet, le coût d'apparition et de disparition des liaisons hydrogène n'est pas toujours le même. Pour qu'une liaison hydrogène puisse apparaitre dans une conformation, il faut suffisament d'axes de rotation conformationnels (voir les règles de la section 4.1.1). Cette apparition entraine une création d'un nouveau cycle ou plus et les d'atomes présents sur ces nouveaux cycles subiront de nouvelles contraintes. Pour définir le coût énergétique d'une transition, nous utilisons les pénalités des sommets (cf. section 4.2.1). Comme indiqué dans la section 4.2.1, le coût énergétique d'une liaison hydrogène est réparti uniformément sur l'ensemble d'atomes qu'elle contient. La pénalité d'un atome dans une conformation G_i n'est pas forcement le même dans une autre conformation G_j. Cela dépend du nombre de liaisons présentes dans la conformation et les cycles obtenus par ces liaisons. Le coût de la transition représente la différence maximale entre les pénalités de d'arrivée et de départ des atomes.

Définition 31 (Coût énergétique d'une transition)**.** *Soient un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ et deux conformations $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ et $G_j = (V_j, E_{C_j}, A_{H_j}, E_{I_j})$ de ce graphe telles que $(G_i, G_j) \in \mathcal{A}$. Le coût énergétique de la transition (G_i, G_j) est défini comme suit :*

$$\omega((G_i, G_j)) = \begin{cases} 0 & \textit{si } |A_{H_i}| > |A_{H_j}|, \\ \max\limits_{a \in V_i}(pen_{G_i}(a) - pen_{G_j}(a)) & \textit{sinon.} \end{cases}$$

La figure 4.2 montre un exemple de coût énergétique pour passer d'une conformations avec deux liaisons hydrogène (la conformation (A) sur la figure) à une

conformation avec trois liaisons hydrogène (la conformation (B) sur la figure). L'ajout de la liaison hydrogène $(\mathrm{N2}, \mathrm{O2})$ a coûté 83.33.

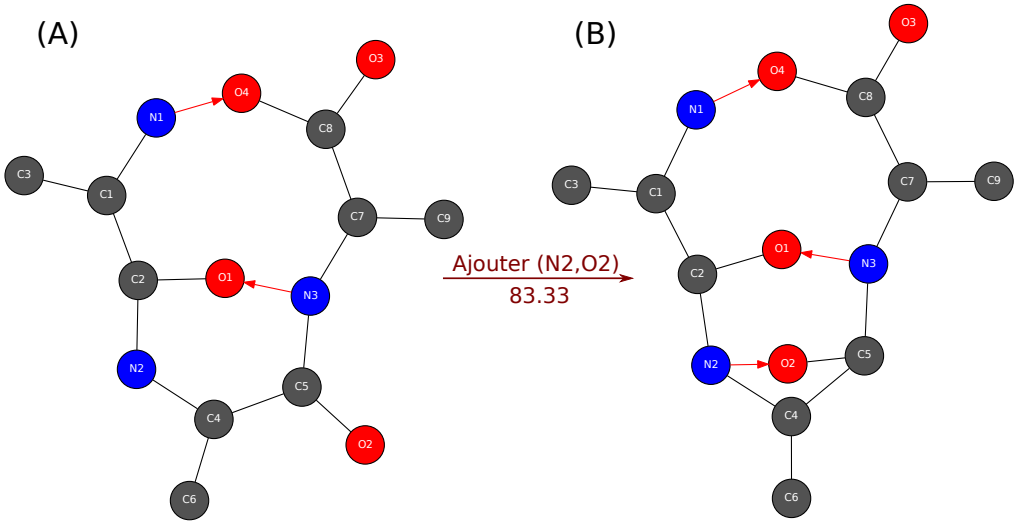

FIGURE 4.2 – Exemple de coût énergétique d'une transition entre deux conformations. L'ajout de la liaison hydrogène $(\mathrm{N2}, \mathrm{O2})$, pour passer de la conformation (A) à la conformation (B), coûte 83.33. Les liaisons hydrogène sont présentées en arcs rouges.

4.3 Classification des conformations possibles

Le calcul des niveaux énergétiques des conformations permet de donner une classification à gros grain de ces conformations.

Le but de la classification est d'obtenir des familles séparées de conformations selon les niveaux d'énergie. Cela permet, d'une part, d'avoir une vue globale sur les conformations les plus stables et d'autre part, permet d'éliminer d'autres conformations non désirées qui sont les conformations avec un grand niveau d'énergie.

Les figures 4.3 et 4.4 montrent un exemple de classification des conformations pour deux peptides à complexité croissante, un tripeptide de 34 atomes et un peptide composé de 80 atomes. La classification est effectuée selon les niveaux d'énergie définis dans la section précédente (cf. section 4.2.2). L'axe des abscisses présente les labels des conformations. Une coloration est attribuée aux conformations selon le nombre de liaisons hydrogène présentes. Deux conformations sont représentées par la même couleur, si elles ont le même nombre de liaisons hydrogène présentes.

Nous remarquons que pour le tripeptide il y a 4 familles de conformations et ces familles sont reparties de la même manière qu'une classification basée sur le nombre de liaisons hydrogène. Autrement dit, la première famille de conformations comporte les conformation avec une seule liaison hydrogène, il s'agit des conformations les moins stables. Ensuite, il y a la deuxième famille et la troisième famille

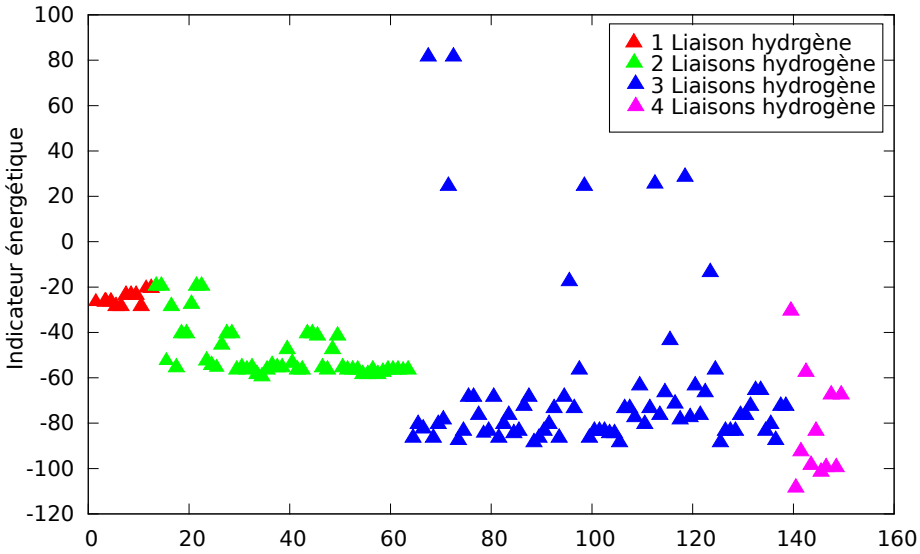

FIGURE 4.3 – Classification des conformations d'un tripeptide. La figure montre qu'en général le nombre de liaisons hydrogène emporte sur la classification des conformations.

de conformations qui comportent les conformations avec deux et trois liaisons hydrogène respectivement. Ce sont des conformations un peu plus stables mais les conformations les plus stables sont les conformations avec quatre liaisons hydrogène et qui constituent la quatrième famille de conformations. Dans cette classification, nous remarquons que certaines conformations avec trois liaisons hydrogène ont un niveau énergétique haut. Cela est dû aux contraintes sur les atomes qui sont élevées.

Cependant, la figure 4.4 montre qu'il y a un grand chevauchement au niveau des conformations pour le pepitde composé de 80 atomes. Il y a des conformations avec six liaisons hydrogène qui sont au même niveau énergétique que des conformations avec deux liaisons hydrogène seulement. Les résultats sont pertinents, car ils confirment l'hypothèse qu'avoir une nombre élevé de liaisons hydrogène présentes n'implique pas nécessairement une conformation plus stable (cf. section 4.2.2).

Dans le chapitre suivant, nous présentons plus en détails les résultats de la classification des conformations selon le modèle proposé pour calculer le niveau d'énergie.

4.4 Recherche des chemins entre deux conformations possibles

Une fois construit, le graphe des possibles peut être vu comme un paysage, où le niveau d'énergie représente l'altitude. Plus le niveau d'énergie est élevé, plus nous découvrons de nouveaux vallées (minima) et de nouveaux pics (maxima). Le défi

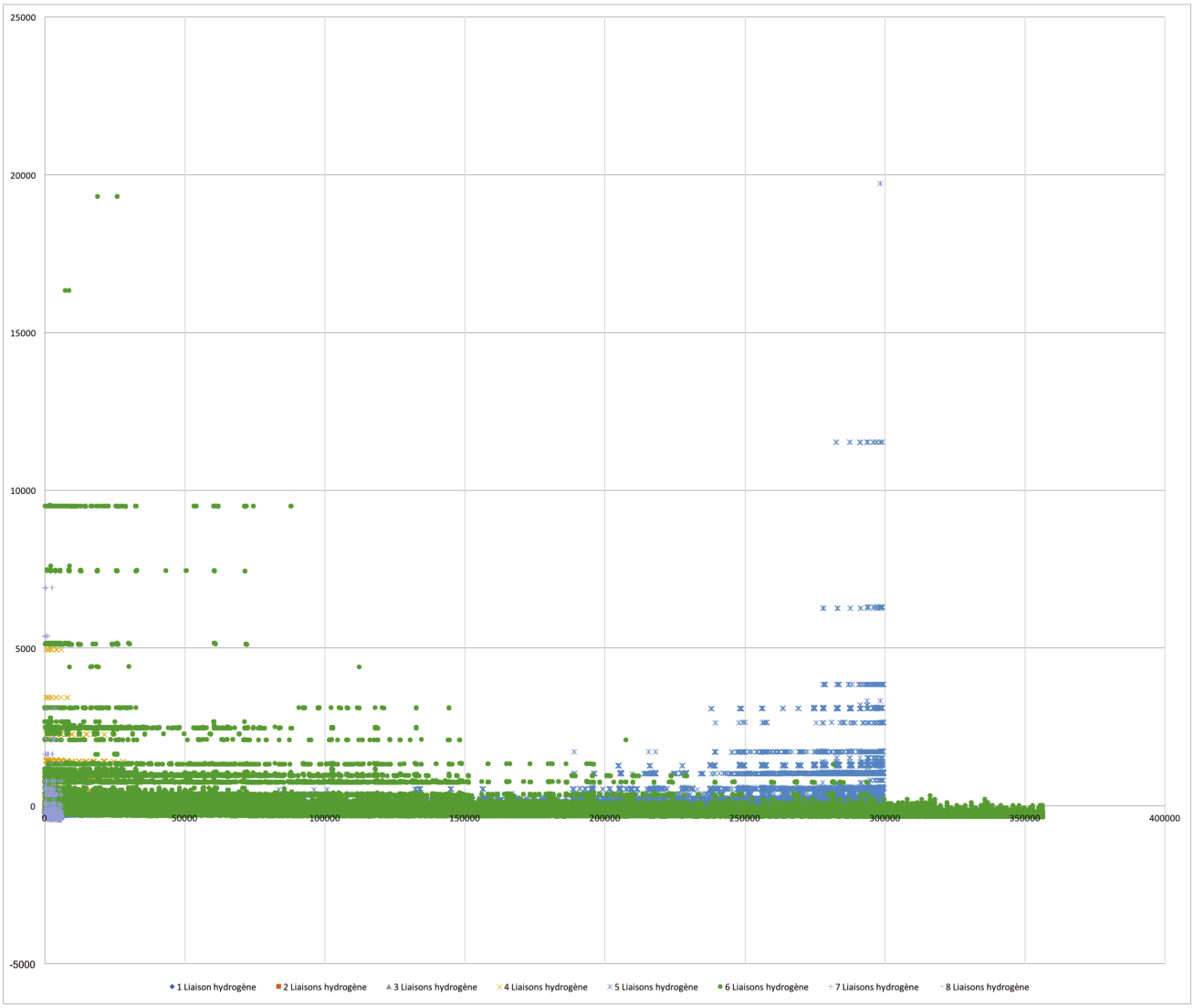

FIGURE 4.4 – Classification des conformations d'un peptide composé de 80 atomes. La figure montre qu'il y a des chevauchements entre les conformations avec un nombre différent de liaisons hydrogène présentes.

est d'être capable de trouver des chemins d'une conformation à une autre avec des altitudes pas trop élevées et en passant par un minimum de vallées.

D'une part, nous avons vu que sur la surface d'énergie potentielle, aller d'une conformation à une autre a un coût énergétique qui diffère selon le type de transition. D'autre part, pour aller d'une conformation G_i à une conformation G_j, nous pouvons passer par différents chemins selon l'énergie (température) fixée dans le système (cf. chapitre 1). Un chemin d'une conformation G_i à une conformation G_j observé dans une trajectoire peut être inaccessible dans une autre.

Une utilisation du graphe des possibles est de trouver les chemins à coût minimum entre les conformations. En effet, il se trouve qu'un chemin observé dans une trajectoire entre deux conformations (cf. chapitre 3) soit différent que le che-

min à coût minimum donné par le graphe des possibles. Un objectif est de pouvoir exhiber ces différences entre les chemins réels (résultats d'analyse des trajectoires avec l'algorithme présenté en chapitre 3) et les chemins théoriques (graphe des possibles).

Définition 32 (Chemin d'une conformation à une autre)**.** *Soit un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$. Un chemin d'une conformation à une autre G_i et G_j de $\mathcal{G}$ est une suite $P(G_i, G_j) = (G_1, G_2, ...G_n)$ de sommets de $\mathcal{G}$ telle que G_1 isomorphe à G_i, G_n isomorphe à G_j et chaque deux sommets consécutifs G_i et G_{i+1} sont reliés par un arc $\mathcal{G}_P$.*

Plusieurs paramètres peuvent être utilisés pour choisir les chemins à coût minimum :

1. Les niveaux énergétiques des conformations appartenant au chemin.
2. Le nombre des liaisons hydrogène des conformations appartenant au chemin.
3. Le coût énergétique maximal des transitions dans le chemin.
4. Le nombre de conformations traversées dans le chemin.

Construire le graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ selon le nombre de liaisons hydrogène revient à construire des sous-graphes hypercube $\mathcal{G}_P^k$ [44], k étant le nombre de liaisons hydrogène présentes ($0 \leq k \leq |Hydro|$). La dimension de l'hypercube obtenu est proportionnel au nombre de liaisons hydrogène des conformations possibles.

Si nous utilisons les niveaux énergétiques des conformations, nous obtenons un graphe des possibles déconnecté avec des niveaux énergétiques bas. Plus le niveau énergétique est haut, plus le graphe des possibles est connecté (le nombre de composantes connexes du graphe diminue).

La figure 4.5 montre un exemple d'évolution des composantes connexes du graphe des possibles d'un peptide composé de 80 atomes selon le niveau énergétique des conformations possibles. Cette figure indique que le nombre de composantes connexes est croissant au début jusqu'à arriver à un maximum de 23576 composantes connexes ensuite ce nombre diminue jusqu'à arriver à une seule composante connexe et reste une seule même si le nombre de conformations augmente.

Dans notre approche, nous utilisons le coût énergétique et le nombre de conformations traversées pour choisir les chemins à coût minimum.

Définition 33 (Coût d'un chemin d'une conformation à une autre)**.** *Soient un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ et $P(G_i, G_j) = (G_1, G_2, ...G_n)$ un chemin de la conformations G_i à la conformation G_j de $\mathcal{G}_P$. Le coût du chemin $P(G_i, G_j)$ est défini comme suit :*

$$\mathcal{C}(P(G_i, G_j)) = \max_{i \in [\![1, n-1]\!]} \omega(G_i, G_{i+1})$$

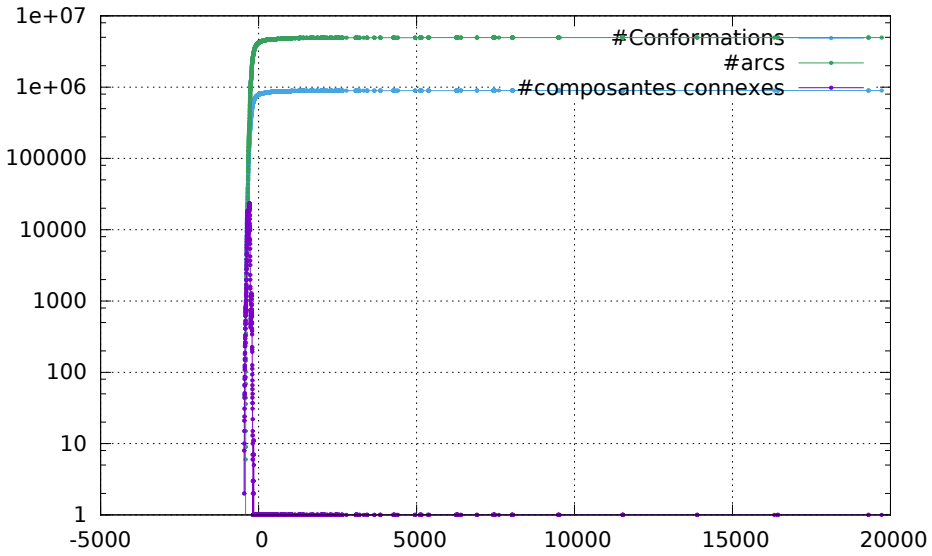

FIGURE 4.5 – Evolution des composantes connexes selon le niveau énergétique. Pour simplifier le dessin, l'axe des ordonnées est présenté en échelle logarithmique. La courbe bleue présente l'évolution des conformations en fonction du niveau énergétique, la courbe verte l'évolution des arcs ajoutés à chaque niveau énergétique et la courbe en violet présente le nombre des composantes connexes en fonction du niveau énergétique.

En d'autres termes, le coût d'un chemin est le maximum des coûts de transitions entre les conformations appartenant à ce chemin.

Soient un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ et $P(G_i, G_j) = (G_1, G_2, ...G_n)$ le chemin de la conformation G_i à G_j de $\mathcal{G}$. Le chemin $P(G_i, G_j)$ est un chemin de coût minimum si pour tout autre chemin $P\prime(G_i, G_j)$ entre ces deux conformations, $\mathcal{C}(P(G_i, G_j)) \leq \mathcal{C}(P\prime(G_i, G_j))$.

Le but est non seulement de trouver un chemin de coût minimum en termes de coût de transitions mais aussi d'essayer de minimiser le nombre de conformations traversées entre les deux conformations en question, donc on cherche un plus court chemin de coût minimum.

D'un point de vue algorithmique, nous utilisons l'algorithme de Dijkstra [52]. Soient un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ ou un sous graphe des possibles (une élimination des conformations en utilisant un niveau énergétique maximal) et deux conformations G_i à G_j de $\mathcal{G}_P$. Le principe de l'algorithme pour trouver un plus court chemin de coût minimum $P(G_i, G_j) = (G_1, G_2, ..., G_n)$ de taille $|P(G_i, G_j)| = n$ ($G_1 = G_i$,$G_n = G_j$) est le suivant : parcourir l'ensemble des conformations de $\mathcal{G}$ de telle sorte pour toute conformation G_k ($2 \leq k \leq n$) visitée, vérifier que $|(P(G_i, G_k)| = \min_{G \in \mathcal{G}_{Nvist}, (G_{k-1}, G) \in \mathcal{A}} (|P(G_i, G)|)$ et $\omega(G_{k-1}, G_k) \leq \mathcal{C}(P(G_i, G_{k-1}))$, $\mathcal{G}_{Nvist}$ étant l'ensemble des conformations non visitées.

Nous avons vu que selon le niveau énergétique choisi, certaines conformations sont accessibles et d'autres non. De même pour les transitions, nous pouvons fixer un coût de transition maximal $\mathcal{C}_{max}$, et seuls les chemins de coût inférieurs à $\mathcal{C}_{max}$ sont pris en compte. Le coût maximal est une estimation de la barrière sur la surface d'énergie potentielle (cf. chapitre 1).

Soit un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$ ou un sous-graphe des possibles (une élimination des conformations en utilisant un niveau énergétique maximal). Trouver un plus court chemin $P(G_i, G_j)$ de coût minimum de la conformation G_i à la conformation G_j, dans notre approche, revient à trouver en premier lieu un plus court chemin en ignorant la barrière ($\mathcal{C}_{max}$ n'est pas fixée). Ensuite le coût de ce chemin sera utilisé comme $\mathcal{C}_{max}$ pour trouver un autre plus court chemin de coût minimum. Nous pouvons résumer les étapes de recherche des chemins avec barrière comme suit :

1. Trouver un plus court chemin $P(G_i, G_j)$ s'il existe entre G_i et G_j avec l'algorithme de Dijkstra (décrit ci-dessus) tel que $\mathcal{C}_{max}$ ignorée,
2. Si $P(G_i, G_j)$ existe, définir une nouvelle barrière $\mathcal{C}_{max} = \mathcal{C}(P(G_i, G_j))$, sinon fin de l'algorithme.
3. Trouver un plus court chemin de coût minimum $P(G_i, G_j)$ s'il existe entre G_i et G_j avec l'algorithme de Dijkstra (décrit ci-dessus) tel que $\mathcal{C}(P(G_i, G_j)) \leq \mathcal{C}_{max}$,
4. Retourner à l'étape (2).

L'idée est d'avoir un plus court chemin sans barrière. Cela permet de définir une barrière maximale qui est le coût du chemin trouvé. Ensuite, baisser au fur et à mesure cette barrière jusqu'à trouver une barrière minimale, en dessous de laquelle il n'existe plus de chemin entre les deux conformations d'entrée (le graphe des possibles est dans ce cas déconnecté). Le chapitre 5 montre un exemple d'évolution des plus courts chemins selon les barrières obtenues.

De même que dans la construction du graphe des possibles, pour réduire les calculs et l'espace mémoire, nous utilisons une table de hachage pour accéder aux conformations. En outre, pour l'ensemble des les conformations non visitées $\mathcal{G}_{Nvist}$, nous rajoutons au fur et à mesure les conformations voisines à la conformation courante. L'ensemble $\mathcal{G}_{Nvist}$ est trié par ordre croissant selon le coût de la transition. Cela permet d'accéder directement à la transition la moins coûteuse.

4.5 Evaluation et performance des algorithmes

L'algorithme a été implémenté en langage C. Le modèle a été conçu pour des peptides isolés. Pour évaluer les performances de la prédiction conformationnelle

nous avons testé la méthode sur des peptides isolés à taille et complexité croissantes.

Le tableau 4.1 résume les résultats numériques trouvés. Il présente le nombre d'atomes (N), le nombre de liaisons hydrogène possibles (M), le nombre de conformations théoriques (C_t), le nombre de conformations possibles trouvées par l'algorithme (C), le nombre maximal de liaisons hydrogène présentes simultanément (M_{max}) et le temps d'exécution (TE)

TABLE 4.1 – Résultats de la prédiction conformationnelle avec l'algorithme proposé.

Formule empirique de la molécule*	N	M	C_t	C	M_{max}	TE
$C_6H_{13}N_2O_3$	24	6	64	17	2	$0.01s$
$C_9H_{18}N_3O_4$	34	12	4096	150	4	$0.05s$
$C_{21}H_{38}N_7O_8$	74	56	7,2 10^{16}	3435798	9	$\approx 219h$
$C_{26}H_{39}N_7O_8$	80	54	2,81 10^{14}	896243	8	$13h\ 63m\ 57s$

Nous remarquons que le nombre de conformations est exponentiel en fonction du nombre de liaisons hydrogène possibles.

La performance de la prédiction proposée apparait dans les grands systèmes moléculaires où on peut voir une grande différence entre le nombre de conformations théorique et le nombre de conformations trouvées par notre méthode. Par exemple, si nous considérons la dernière ligne du tableau qui présente un peptide à 80 atomes, nous avons 54 liaisons hydrogène possibles. Théoriquement, le nombre de conformations possibles dépasse un milliard de conformations alors qu'avec la méthode proposée le nombre de conformations possibles est d'environ 900000 conformations. En outre, nous remarquons que le nombre de liaisons hydrogène présentes simultanément dans chacune des conformations est inférieur au nombre de liaisons hydrogène possibles. Dans le peptide parmi 54 liaisons hydrogène possibles, au plus 8 liaisons hydrogène peuvent être présentes simultanément. Ces résultats confirment que quand le nombre de liaisons hydrogène est élevé la molécule est contrainte et il devient difficile d'ajouter d'autres liaisons hydrogène. D'où la diminution du nombre de conformations à partir d'un certain nombre de liaisons hydrogène jusqu'à ne plus pouvoir ajouter de liaisons hydrogène.

Les figures 4.6, 4.7, 4.8 et 4.9 représentent en histogrammes le nombre de conformations possibles en fonction du nombre de liaisons hydrogène présentes pour les quatre peptides indiqués dans le tableau 4.1. Les différents histogrammes montrent des écarts importants entre le nombre de conformations théoriques et trouvées par notre méthodes selon le nombre de liaisons. Par exemple, pour le $C_{26}H_{39}N_7O_8$ (voir figure 4.9), le nombre de conformations théoriques à six liai-

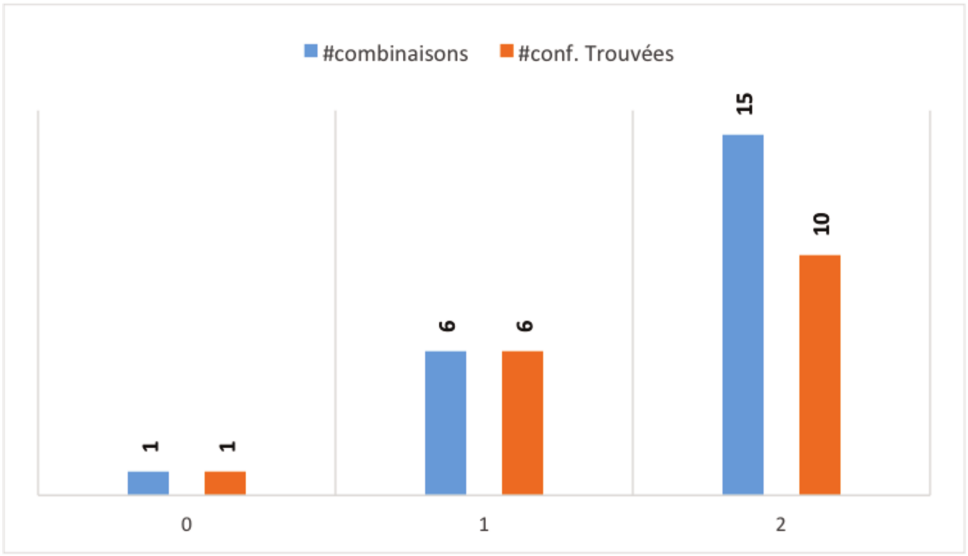

FIGURE 4.6 – Répartition des conformations en fonction du nombre de liaisons hydrogène pour le peptide $C_6H_{13}N_2O_3$. Les histogrammes en orange présentent le nombre de conformations trouvées par l'algorithme et en bleu le nombre de conformations théorique.

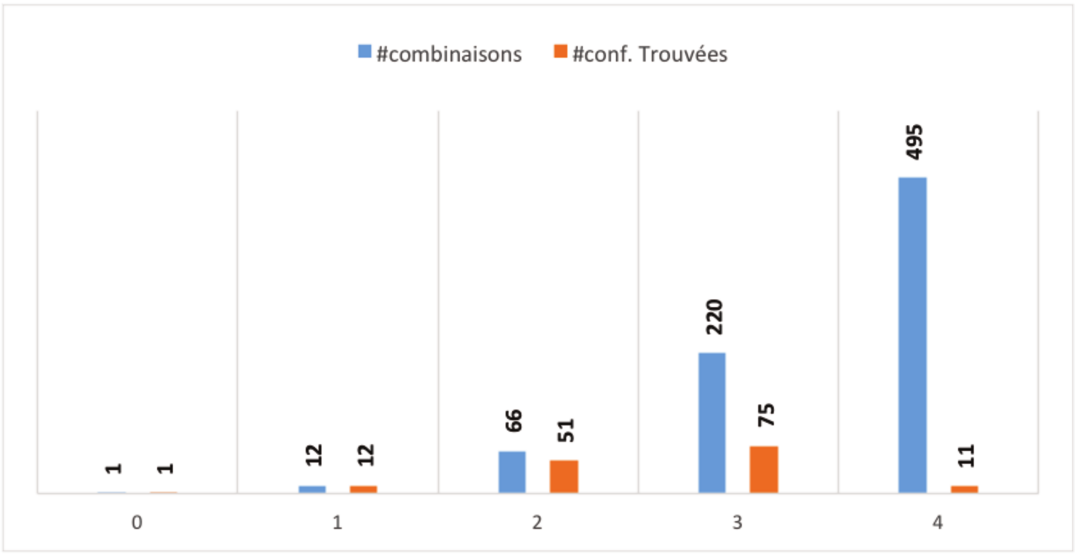

FIGURE 4.7 – Répartition des conformations en fonction du nombre de liaisons hydrogène pour le peptide $C_9H_{18}N_3O_4$. Les histogrammes en orange présentent le nombre de conformations trouvées par l'algorithme et en bleu le nombre de conformations théorique.

sons hydrogène est 32468436 conformations cependant le nombre trouvé par l'algorithme est de 356484 conformations ($\approx 1\%$ du nombre théorique).

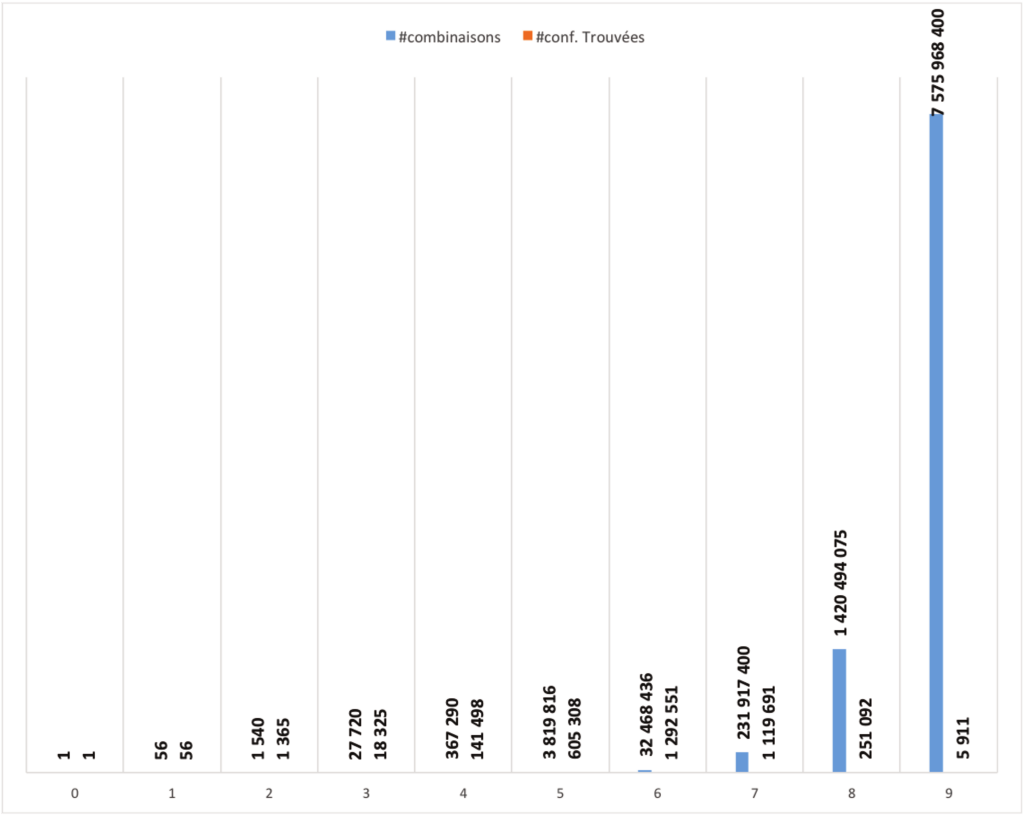

FIGURE 4.8 – Répartition des conformations en fonction du nombre de liaisons hydrogène pour le peptide $C_{21}H_{38}N_7O_8$. Les histogrammes en orange présentent le nombre de conformations trouvées par l'algorithme et en bleu le nombre de conformations théorique.

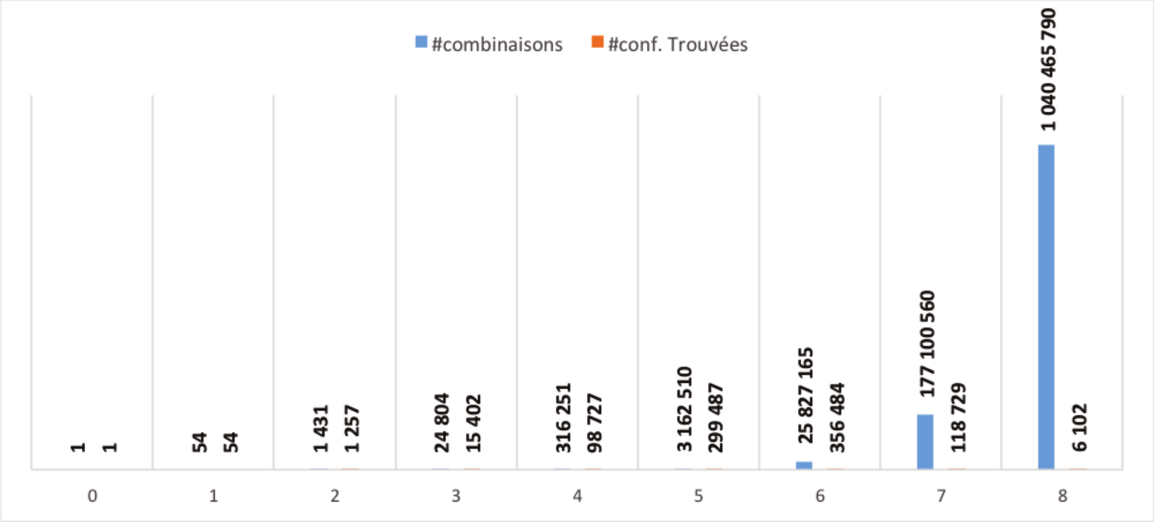

FIGURE 4.9 – Répartition des conformations en fonction du nombre de liaisons hydrogène pour le peptide $C_{26}H_{39}N_7O_8$. Les histogrammes en orange présentent le nombre de conformations trouvées par l'algorithme et en bleu le nombre de conformations théorique.

4.6 Conclusion

Dans ce chapitre, nous avons vu une méthode alternative qui permet de prédire les conformations qu'un peptide isolé peut prendre. Cela indépendamment des simulations et donc du temps et de la géométrie du système. Il s'agit de la construction du graphe des possibles en utilisant des connaissances physiques/chimiques et des notions de la théorie des graphes. Les sommets du graphe représentent les conformations possibles du système et les arcs les transitions possibles entre ces conformations.

D'un part, trouver les conformations les plus stables (minima sur la surface d'énergie potentielle) revient à faire une classification des conformations possibles. Pour cela, nous avons défini un modèle pour calculer les niveaux énergétiques des conformations. Le modèle est basé sur l'ensemble des liaisons hydrogène présentes dans les conformations.

D'autre part, les états de transition représentent dans notre approche les coûts énergétiques qu'il faut pour passer d'une conformation à une autre. Ces coûts sont principalement définis par les contraintes structurelles que subissent les atomes. Les résultats numériques montrent que la méthode proposée donne de bons résultats en termes de nombre de conformations et le chevauchement dans la classification entre les conformations avec une nombre de liaisons hydrogène différents confirment les hypothèses qu'avoir un grand nombre de liaisons hydrogène présente n'implique pas une stabilité meilleure au système moléculaire.

L'objectif de la prédiction conformationnelle est de trouver une présélection de familles de conformations pertinentes. Ces résultats seront présentés aux groupes de recherche de la chimie théorique et computationnelle, afin de les utiliser dans leurs recherches et études, en particulier pour prédire les formes tridimensionnelles des conformations pertinentes ou de faire des simulations intelligentes.

Plus de détails sur résultats de la prédiction conformationnelle seront présentés dans le chapitre suivant.

Chapitre 5

Application des méthodes d'analyse et de prédiction sur les systèmes moléculaires

Dans les chapitres 3 et 4, nous avons présenté deux algorithmes : un algorithme pour analyser la dynamique conformationnelle sur les trajectoires de simulations de dynamique moléculaire et un autre algorithme pour prédire les conformations possibles d'un système moléculaire ainsi que les transitions possibles entre elles et ce indépendamment des simulations.

Dans ce chapitre, nous illustrons les résultats de nos algorithmes sur des systèmes moléculaires en phase gazeuse. Nous avons choisi trois types de trajectoires pour analyser la dynamique conformationnelle : trajectoires de peptides isolés de taille et de complexité croissantes, trajectoires contenant une collision d'un atome d'argon avec un peptide et trajectoires de clusters (cf. chapitre 1). L'objectif est de montrer le fonctionnement et la flexibilité de l'algorithme ainsi que d'expliquer ses différents résultats.

Le modèle proposé pour la prédiction conformationnelle repose sur la dynamique des liaisons hydrogène. A cet effet, les systèmes moléculaires testés sont les peptides isolés. Nous présentons un exemple du graphe des possibles d'un tripeptide composé de 34 atomes. Enfin, nous comparons quelques chemins analysés dans des trajectoires de simulations de ce même peptide avec les chemins théoriques trouvés.

5.1 Analyse des trajectoires de peptides isolés

Dans ce système deux peptides ont été choisis un dipeptide $C_6H_{13}N_2O_3$ et un heptapeptide $C_{21}H_{38}N_7O_8$. Ces peptides ont été déjà analysés et publiés [45, 47]. Les conformations des peptides isolés sont déterminées à partir de la dynamique des liaisons hydrogène. Les liaisons covalentes sont fixes tout au long de la trajectoire et les interactions intermoléculaires électrostatiques n'existent pas dans ce genre de systèmes. Dans l'analyse de la dynamique conformationnelle de cette trajectoire, nous fixons l'ensemble des liaisons covalentes comme ils ont été calculées à la première image, puis nous observons seulement l'évolution des liaisons hydrogène au fil du temps.

La trajectoire de $C_6H_{13}N_2O_3$ a été générée par un simulateur de dynamique moléculaire ab initio, elle contient 10200 images (*i.e.* $4ps$), où chaque image est com-

posée de 24 atomes.

Sur cette trajectoire **quatre conformations stables** ont été identifiées par l'algorithme d'analyse de trajectoire (cf. chapitre 1). La figure 5.1 présente la représentation 3D de ces conformations (le haut de la figure 5.1) ainsi que les graphes mixtes associés (le bas de la figure 5.1). Le nombre de conformations explorées sur cette trajectoire montre que le calcul des ensembles de liaisons et des orbites est effectué très peu de fois. De même, le test d'isomorphisme entre la conformation courante et les conformations déjà identifiées est appliqué au plus pour 4 conformations. Ces résultats affirment qu'en pratique la complexité théorique n'est jamais atteinte.

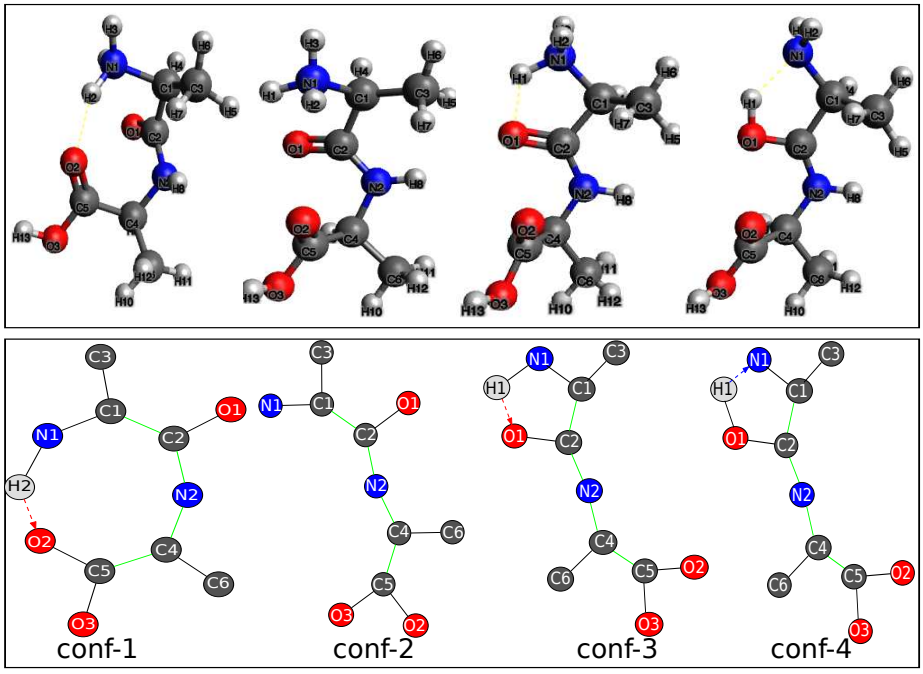

FIGURE 5.1 – Représentation schématique des conformations stables du $C_6H_{13}N_2O_3$ identifiées le long de la trajectoire par l'algorithme. Le haut de la figure présente les représentations 3D des conformations et le bas de la figure présente les graphes mixtes associés. Dans les graphes mixtes, les liaisons covalentes sont présentées en arêtes noires, les axes de rotations conformationnelles en arêtes vertes et les liaisons hydrogène en arcs rouges ou arcs bleus si transfert de proton. La couleur des sommets dépend du type d'atome : les atomes d'oxygène sont présentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique combiné à son numéro d'apparition dans l'image selon son type.

Dans la première conformation (conf-1), une seule liaison hydrogène est formée entre l'atome N1 et l'atome O2 du peptide. Cette liaison hydrogène disparait dans la deuxième conformation (conf-2) et une nouvelle liaison hydrogène apparait dans la troisième conformation (conf-3) entre l'atome N1 et l'atome O1 suite aux

mouvements rotationnels autour de l'axe de rotation conformationnel $C1 - C2$. La quatrième conformation (conf-4) est obtenue à partir de la troisième conformation par un transfert de proton dans la liaison hydrogène $(N1, O1)$.

En analysant le temps de résidence de chaque conformation, la conformation la plus stable en termes de temps est la troisième conformation (conf-3). Ce résultat est confirmé par l'analyse d'évolution en temps des liaisons hydrogène. La figure 5.2 présente cette évolution. L'axe des abscisses présente le temps en pico-seconde (ps) et l'axe des ordonnées présente les liaisons hydrogène. Chaque liaison hydrogène est présentée par une ligne horizontale. Cette ligne apparait en *rouge* si la liaison hydrogène est formée, en *gris* si la liaison hydrogène n'est pas formée mais que l'atome donneur et l'atome accepteur sont suffisamment proches pour former une liaison hydrogène, en *bleu* s'il s'agit d'un transfert de proton et disparait sinon.

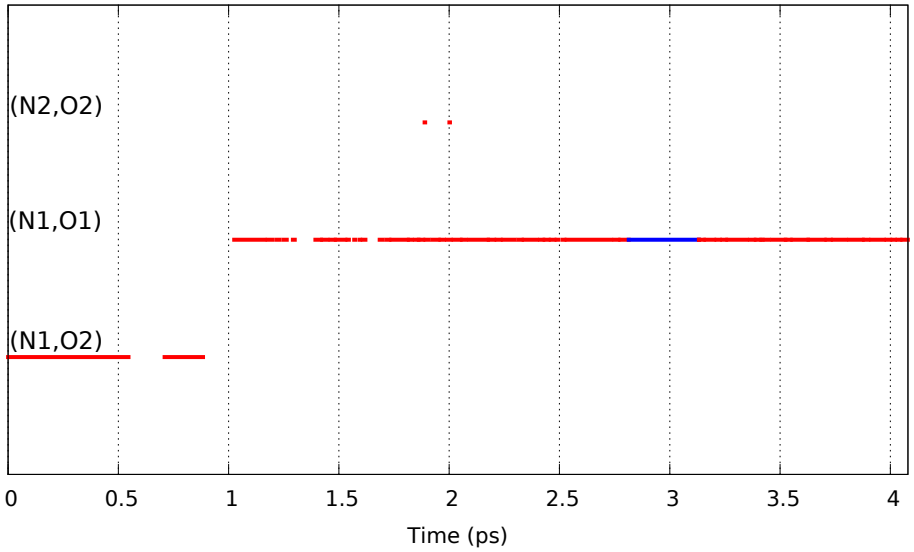

FIGURE 5.2 – Evolution en temps des liaisons hydrogène formées au long de la trajectoire de $C_6H_{13}N_2O_3$. Chaque ligne présente une liaison hydrogène. La ligne est en rouge si une liaison hydrogène est formée et en bleu s'il y a un transfert de proton. Au dessus de chaque ligne est indiqué le couple d'atomes (donneur,accepteur) de la liaison hydrogène.

La première ligne du bas de la figure 5.2 montre l'évolution de la liaison hydrogène $(N1, O2)$. Cette ligne est rouge au début puis elle disparait avant $1ps$. Après $1ps$ une ligne rouge au milieu apparait. Elle représente la liaison hydrogène $(N1, O1)$. On remarque sur cette ligne une partie bleue entre $2.75ps$ et $3.25ps$. Cette période présente le transfert de proton qui a eu lieu entre l'atome d'azote $N1$ et l'atome d'oxygène $O1$. Nous remarquons une troisième liaison hydrogène entre l'atome $N2$ et l'atome $O2$ (ligne rouge en haut de la figure) mais cette liaison apparait très peu de temps. Elle est formée suite au mouvement rotationnel autour de l'axe de rotation conformationnel $C4 - C5$. L'apparition de cette liaison engendre deux états

transitoires comme les montre la figure 5.3.

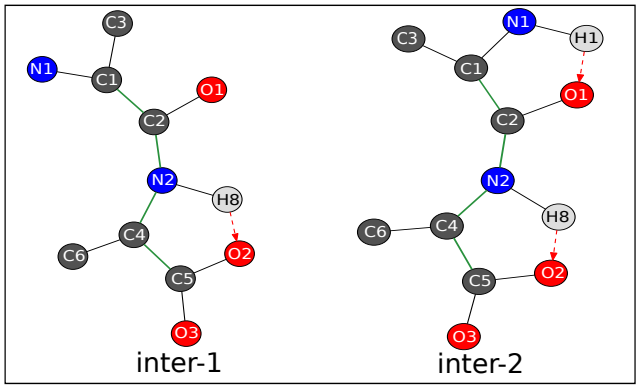

FIGURE 5.3 – Les états transitoires de la trajectoire de $C_6H_{13}N_2O_3$ identifiées par l'algorithme. La figure présente les graphes mixtes associés aux états transitoires identifiées. Les liaisons covalentes sont présentées en arêtes noires, les axes de rotations conformationnelles en arêtes vertes et les liaisons hydrogène en arcs rouges ou arcs bleus si transfert de proton.

La figure 5.4 présente l'évolution de la distance (en Å) de la liaison hydrogène créée entre le couple d'atomes $(\mathrm{N1}, \mathrm{O1})$ (liaison hydrogène qui apparait dans conf-3 et conf-4). La distance de la liaison covalente entre l'atome d'hydrogène et l'atome d'azote (le donneur) est présentée en rouge et la distance de la liaison hydrogène entre l'atome d'hydrogène et l'atome d'oxygène (l'accepteur) est présentée en bleu. Cette illustration a pour but de montrer plus en détails le transfert de proton qui apparait dans l'intervalle $[2.81, 3.13]ps$.

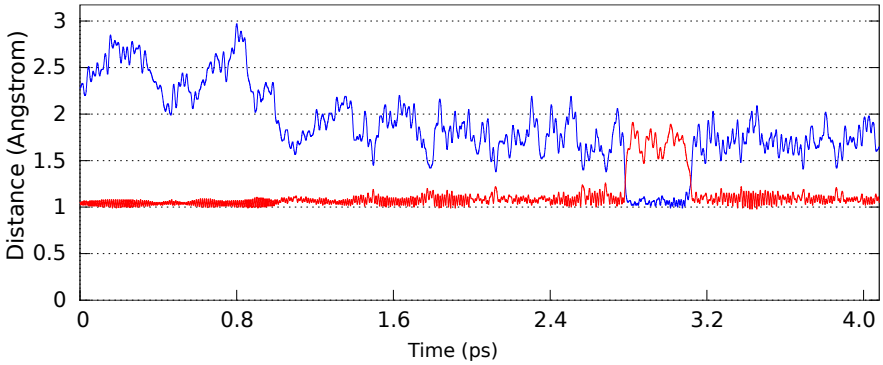

FIGURE 5.4 – Courbe d'évolution de la distance de la liaison hydrogène entre le couple $(\mathrm{N1}, \mathrm{O1})$. La distance de la liaison covalente entre l'atome d'hydrogène et l'atome d'azote (le donneur) est présentée en rouge et la distance de la liaison hydrogène entre l'atome d'hydrogène et l'atome d'oxygène (l'accepteur) est présentée en bleu.

La figure 5.5 présente le graphe des transitions observées le long de la trajectoire $C_6H_{13}N_2O_3$. Chaque arc entre deux conformations G_i et G_j de ce graphe sont étiquetés par deux types d'informations :

- La fréquence de transition de la conformation G_i à la conformation G_j .
- Le type de changement entre la conformation G_i et la conformation G_j (cf. section 3.3.2).

Le graphe de transition à gauche de la figure 5.5 représente le graphe de transitions avec les fréquences de chacune. Nous remarquons une grande fréquence (60) entre les conformations conf-2 et conf-3. Cela est expliqué par les faibles variations de distance autour de la distance de seuil (cf. chapitre 1). Cependant pour le reste des transitions il n y a pas de dynamique (fréquence 1).

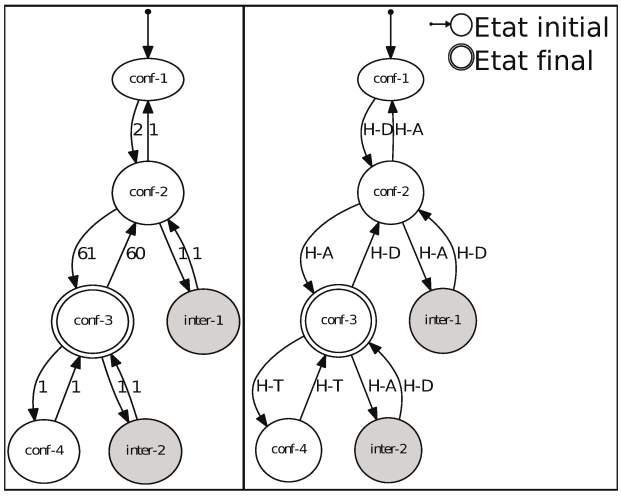

FIGURE 5.5 – Graphe des transitions observées le long de la trajectoire du $C_6H_{13}N_2O_3$. La partie gauche de la figure présente le graphe avec les fréquences de transitions et la partie à droite présente les types de changements entre ces transitions. Les conformations stables sont présentées en cercles blancs tandis que les états transitoires sont présentés en cercles gris.

Le graphe à droite de la figure 5.5 indique les différents type de changements entre les conformations. Par exemple, la transition de la conformation conf-1 à la conformation conf-2 est due à la Disparition d'une liaison Hydrogène (H-D). Ce graphe montre aussi la conformation finale que le système moléculaire a adapté. Dans cette trajectoire il s'agit de la conformation conf-3.

La deuxième trajectoire choisie est celle d'un heptapeptide $C_{21}H_{38}N_7O_8$. Cette trajectoire a été également générée par un simulateur de dynamique moléculaire ab initio, elle contient 26601 images (*i.e.* $10,6ps$), où chaque image est composée de 74 atomes. Cette trajectoire est plus longue et le peptide traité est plus gros et plus complexe que le premier peptide présenté.

Durant l'analyse de cette trajectoire, nous avons remarqué plus de dynamique conformationnelle. L'algorithme a identifié 9 conformations stables et 35 états transitoires (cf. section 3.6). La figure 5.6 montre un exemple d'une conformation identifiée le long de la trajectoire.

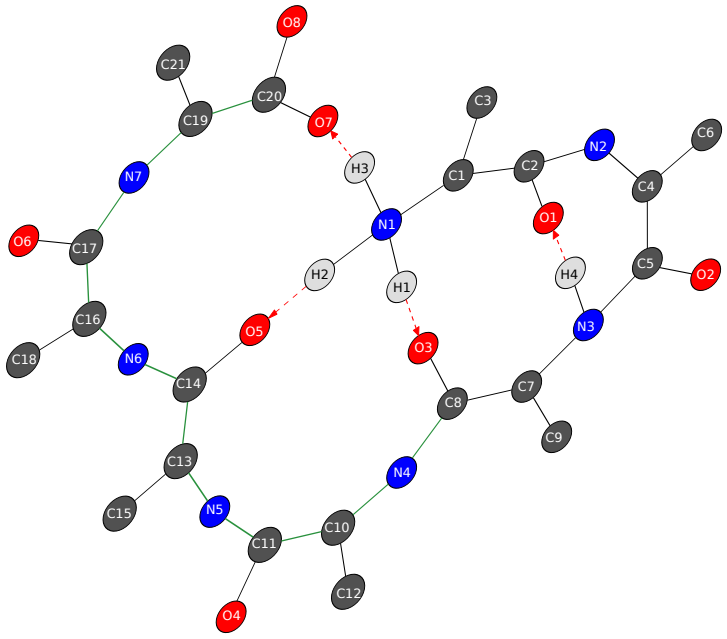

FIGURE 5.6 – Un graphe d'une conformation explorée durant la trajectoire du $C_{21}H_{38}N_7O_8$. Les liaisons covalentes sont présentées en arêtes noires, les axes de rotations conformationnelles en arêtes vertes et les liaisons hydrogène en arcs rouges ou arcs bleus si transfert de proton. La couleur des sommets dépend du type d'atome : les atomes d'oxygène sont présentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique combiné à son numéro d'apparition dans l'image selon son type.

Le nombre total des liaisons hydrogène identifiées est de 8 liaisons. Dans le total des conformations explorées, ce nombre varie entre 1 et 7 liaisons hydrogène formées simultanément. L'évolution de ces liaisons hydrogène est montrée dans la figure 5.7.

La figure montre une grande dynamique dans les liaisons hydrogène $(\mathrm{N3}, \mathrm{O1})$, $(\mathrm{N7}, \mathrm{O4})$ et $(\mathrm{N5}, \mathrm{O2})$ (enchainement d'apparition et de disparition de liaisons traduit par les lignes rouges qui apparaissent et disparaissent en continu, suite à des faibles variations de distance autour de la distance de seuil) contrairement aux liaisons hydrogène $(\mathrm{N1}, \mathrm{O5})$ et $(\mathrm{N1}, \mathrm{O7})$ qui sont plus stables. Nous remarquons aussi quelques liaisons hydrogène qui apparaissent rarement comme la liaison hydrogène $(\mathrm{N6}, \mathrm{O4})$. Cette évolution explique le nombre élevé d'états transitoires identifiés.

5.2 Analyse des trajectoires de dissociation de peptides induite par collision

Dans le but d'illustrer la capacité de l'algorithme à reconnaitre les changements de conformations, en particulier le changement des liaisons covalentes, nous avons

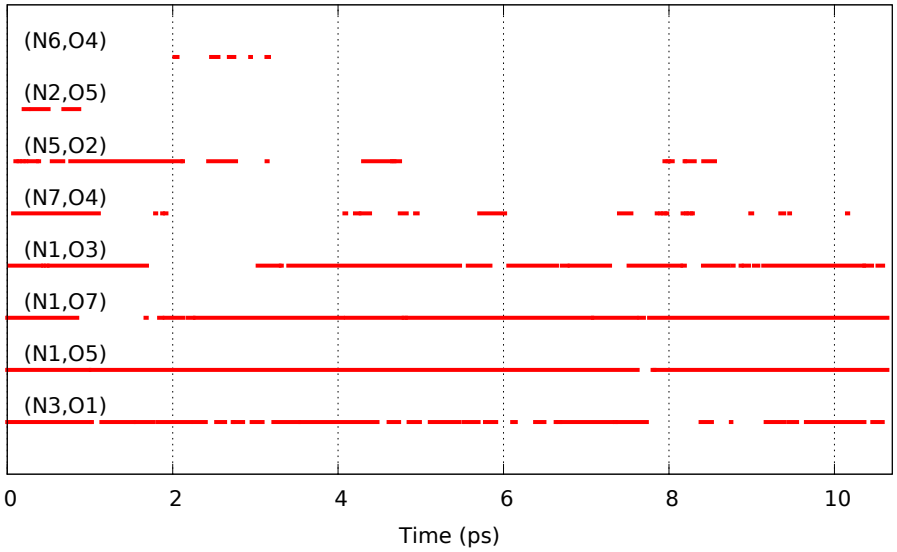

FIGURE 5.7 – Evolution en temps des liaisons hydrogène formées au long de la trajectoire de $C_{21}H_{38}N_7O_8$. Chaque ligne présente une liaison hydrogène. La ligne est en rouge si une liaison hydrogène est formée. Au dessus de chaque ligne est indiqué le couple d'atomes (donneur,accepteur) de la liaison hydrogène.

testé l'algorithme sur un autre type de système. Nous avons choisi une trajectoire contenant collision entre un atome d'argon et un dipeptide $C_4H_{10}N_3O_2Ar$ [48]. La trajectoire contient 51 images, où chaque image est composée de 20 atomes. Dans ce genre de système, les trajectoires générées sont relativement courtes, comparées aux trajectoires des peptides isolés. L'objectif de l'analyse de ce système est d'examiner les fragments engendrés par la collision entre un gaz inerte et un peptide et suivre leur évolution.

Les conformations sont déterminées exclusivement à partir de la dynamique des liaisons covalentes. **Trois conformations** ont été identifiées le long de la trajectoire (voir figure 5.8) : la première conformation (conf-1) présente la structure initiale du peptide (un seul fragment). La deuxième conformation (conf-2) est obtenue de la première conformation après collision du peptide avec l'atome d'argon. La collision casse la liaison covalente $[N2, C3]$ et engendre deux fragments pour le peptide comme le montre la figure 5.8 (au milieu). Dans la troisième conformation (conf-3), une nouvelle liaison covalente est créée entre l'atome $C2$ et l'atome $O1$ dans un des fragments de la deuxième conformation sans que le nombre de fragments ne change.

Comme pour les peptides isolés, l'algorithme génère le graphe des transitions observées au cours de la trajectoire. Ce graphe est présenté dans la figure 5.9. Nous remarquons une seule transition de la conformation conf-1 à la conformation conf-2. Cette transition est due à une liaison covalente qui disparait. Ensuite, une

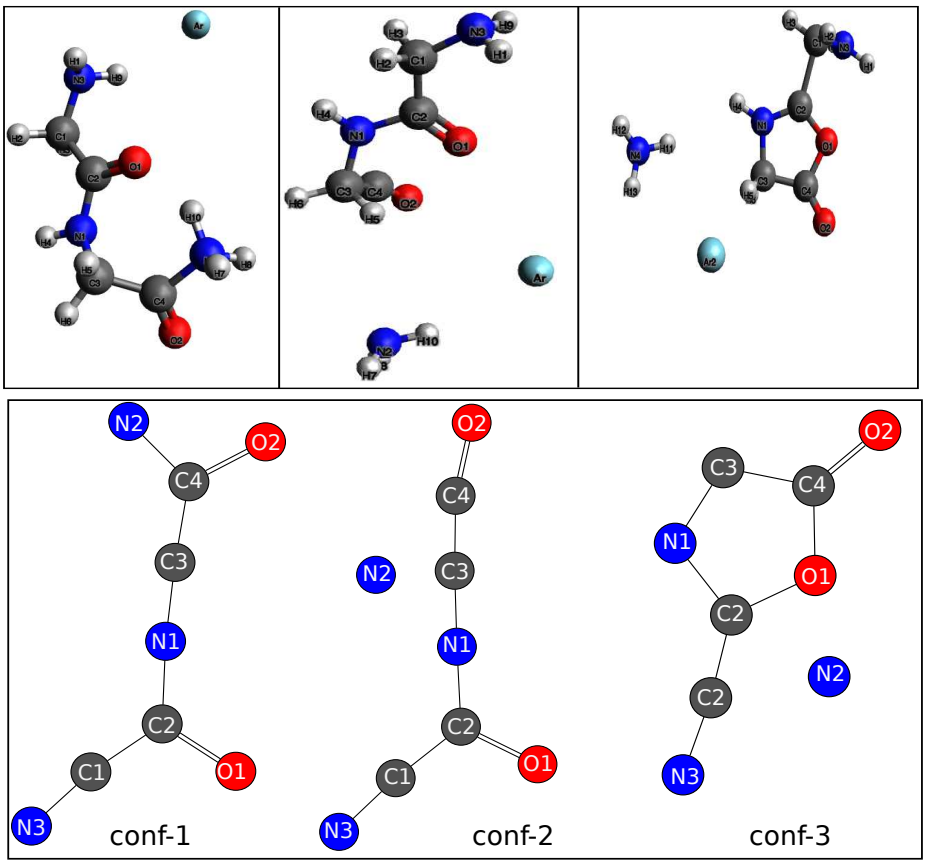

FIGURE 5.8 – Représentation schématique des conformations stables du $C_4H_{10}N_3O_2Ar$ identifiées le long de la trajectoire par l'algorithme. Le haut de la figure présente les représentations 3D des conformations et le bas de la figure présente les graphes mixtes associés. Dans les graphes mixtes, les liaisons covalentes sont présentées en arêtes noires et les liaisons hydrogène en arcs rouges ou arcs bleus si transfert de proton. La couleur des sommets dépend du type d'atome : les atomes d'oxygène sont présentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique combiné à son numéro d'apparition dans l'image selon son type.

transition de la conformation conf-2 à la troisième conformation (conf-3) avec une apparition d'une nouvelle liaison covalente.

5.3 Analyse des trajectoires de clusters

Le troisième type de systèmes que nous avons testé avec notre algorithme (cf. chapitre 3) sont les clusters. Dans ce type, les conformations sont déterminées principalement de la dynamique des interactions intermoléculaires électrostatiques. Nous avons choisi un cluster type eau/ion ($Li^+(H_2O)_4$). Trois trajectoires ont été choisies pour le même cluster mais générées avec trois températures : $50K$, $300K$ et $400K$. Chaque trajectoire contient 50001 images ($20ps$), où chaque image est

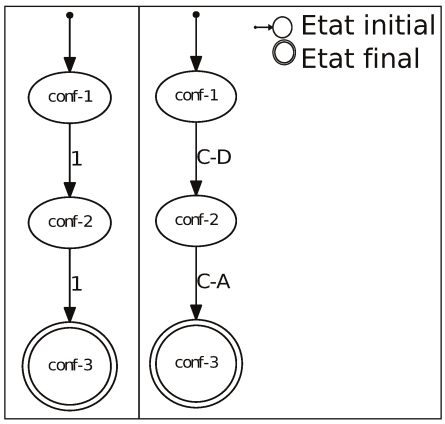

FIGURE 5.9 – Graphe des transitions observées le long de la trajectoire du $C_4H_{10}N_3O_2Ar$. La partie gauche de la figure présente le graphe avec les fréquences de transitions et la partie à droite présente les types de changements entre ces transitions. Les conformations stables sont présentées en cercles blancs tandis que les états transitoires sont présentés en cercles gris.

composée de 13 atomes (4 molécules d'eau $(H_2O)_4$ et un ion de type lithium Li^+). Le temps d'exécution pour l'analyse de ces trajectoires est $6,72$ secondes.

TABLE 5.1 – Résultats d'analyse de la dynamique conformationnelle des trajectoires du $Li^+(H_2O)_4$.

T °	#Conf.	#E.T.	#I.B.	#H.B.	Type de dynamique
$50K$	1	0	2	3	Aucune
$300K$	1	2	3	0-2	Dynamique de liaisons hydrogène
$400K$	2	3	3-4	0-2	Dynamique de liaisons hydrogène et interactions intermoléculaires électrostatiques

Le tableau 5.1 résume les résultats de l'analyse de la dynamique conformationnelle de ces trajectoires. Nous présentons le nombre de conformations stables (#conf.), le nombre d'états transitoires (#E.T.), le nombre de interactions intermoléculaires électrostatiques (#I.B.), le nombre de liaisons hydrogène (#H.B.) et le type de dynamique en termes de liaisons qui a eu lieu le long de chaque trajectoire.

Les résultats du tableau 5.1 montrent que la température influence la dynamique conformationnelle du système moléculaire. A basse température ($50K$), il n y a aucune dynamique, cependant, plus la température est élevée, plus il y a de dynamique dans les liaisons et les interactions ($400K$).

La figure 5.10 présente les conformations identifiées le long de la trajectoire. A $50K$, une seule conformation à été identifiée avec deux interactions intermoléculaires électrostatiques et trois liaisons hydrogène stables. A $300K$, une conformation a été identifiée mais avec plus de dynamique en comparaison de la première tra-

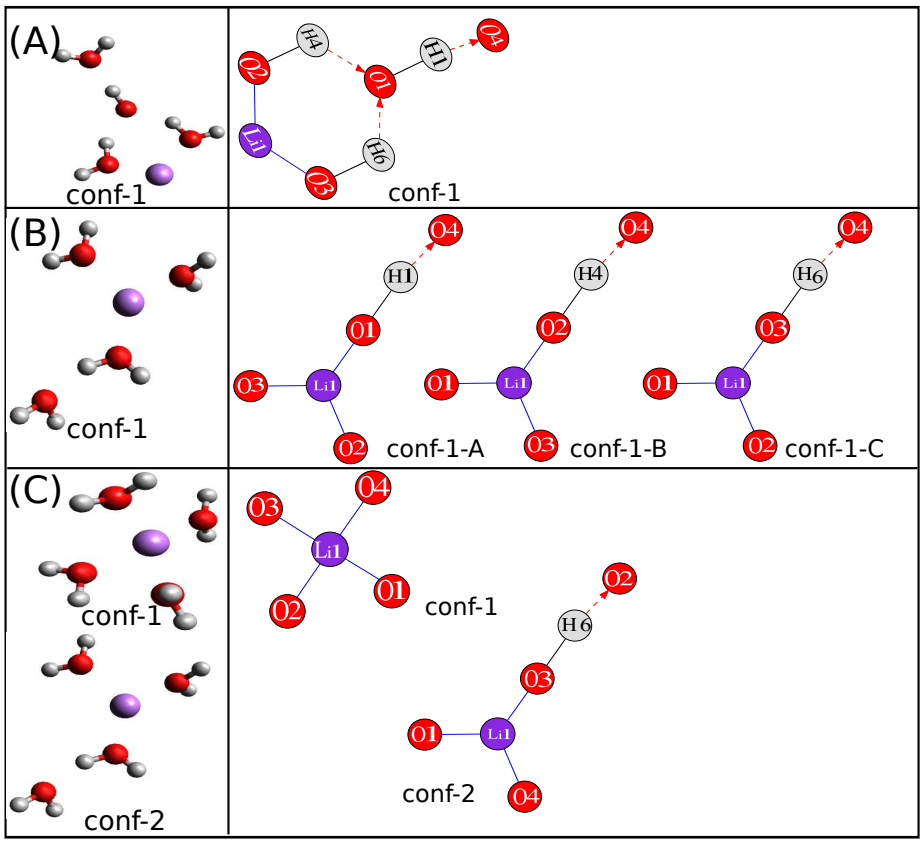

FIGURE 5.10 – Représentation schématique des conformations stables du cluster $Li^+(H_2O)_4$. La partie gauche de la figure présente les représentations 3D des conformations et la partie droite présente les graphes mixtes associés. Dans les graphes mixtes, les liaisons covalentes sont présentées en arêtes noires, les liaisons hydrogène en arcs rouges et les interactions intermoléculaires électrostatiques en arêtes bleu. Les figures (A), (B) et (C) représentent les conformations identifiées aux températures $50K$, $300K$ et $400K$ respectivement.

jectoire. Trois interactions intermoléculaires électrostatiques stables sont formées avec trois molécules d'eau et la quatrième molécule d'eau tourne autour de ces molécules et forment des liaisons hydrogène. Enfin, à $400K$, il y a eu une dynamique des interactions intermoléculaires électrostatiques et de liaisons hydrogène. La première conformation, identifiée dans la trajectoire, contenant 4 interactions intermoléculaires électrostatiques, puis une molécule d'eau s'éloigne jusqu'à la disparition d'une interaction intermoléculaire électrostatique. Cette molécule d'eau forme des liaisons hydrogène avec les trois molécules d'eau attachées au lithium. Pour mieux voir la dynamique conformationnelle du cluster $Li^+(H_2O)_4$ en fonction de la température, la figure 5.11 présente les graphes des transitions de chaque trajectoire. Nous remarquons d'après ces graphes, que plus la température est élevée, plus le nombre de transitions est élevé.

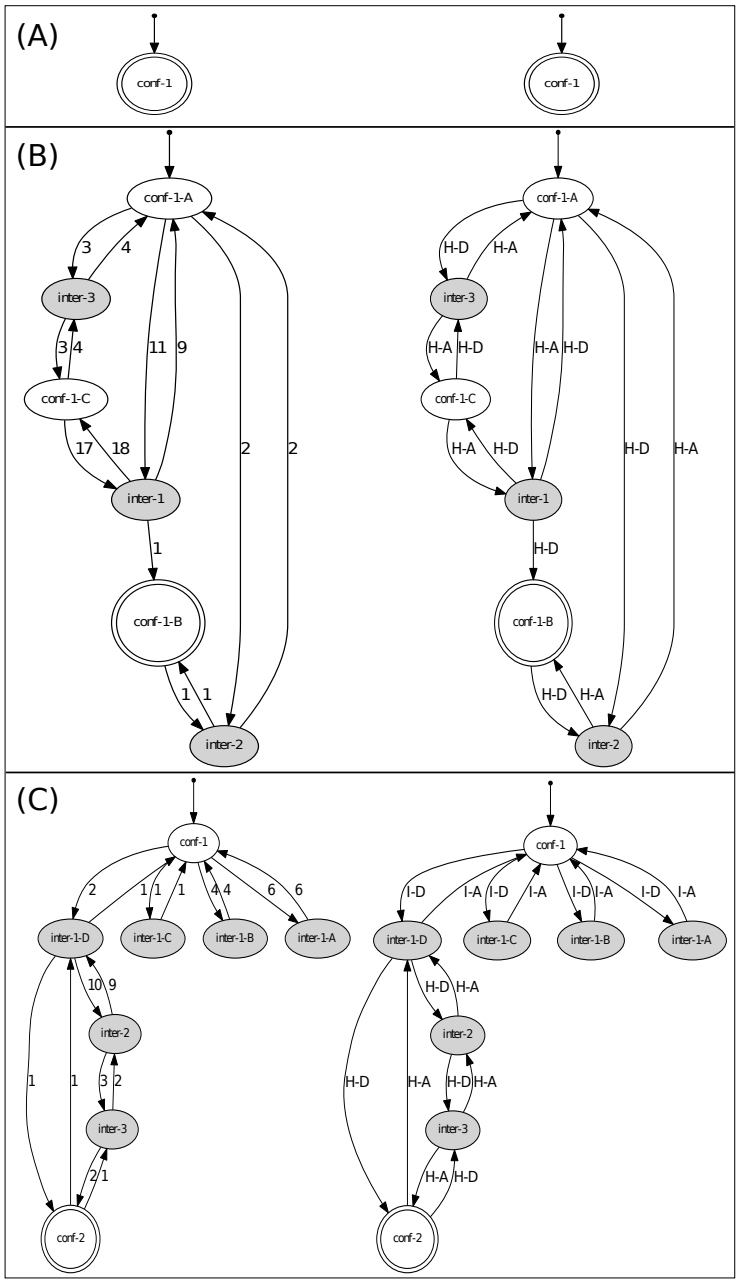

FIGURE 5.11 – Graphes des transitions observées le long des trajectoires du cluster $Li^+(H_2O)_4$. La partie gauche de la figure présente les graphes avec les fréquences de transitions et la partie à droite présente les types de changements entre ces transitions. Les conformations stables sont présentées en cercles blancs tandis que les états transitoires sont présentés en cercles gris. Les figures (A), (B) et (C) présentent les conformations identifiées aux températures $50K$, $300K$ et $400K$ respectivement.

5.4 Prédiction conformationnelle pour un peptide isolé

Pour illustrer la prédiction conformationnelle, nous avons choisi le tripeptide $C_9H_{18}N_3O_4$. La figure 5.12 présente la conformation initiale G_0^0 de ce tripeptide.

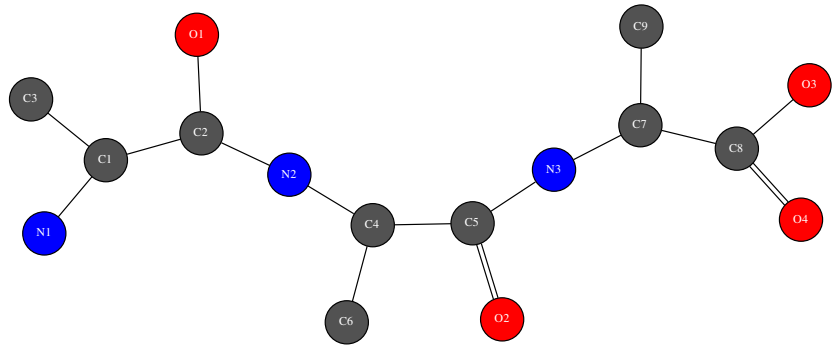

FIGURE 5.12 – Graphe mixte à la conformation initiale du tripeptide $C_9H_{18}N_3O_4$. La conformation initiale du système comporte que les liaisons covalentes (présentées en arêtes noires).

A partir de l'ensemble d'atomes et de l'ensemble de liaisons covalentes, l'ensemble des liaisons hydrogène possibles est :

$$Hydro = \{(\mathrm{N3}, \mathrm{O4}), (\mathrm{N3}, \mathrm{O3}), (\mathrm{N2}, \mathrm{O2}), (\mathrm{N1}, \mathrm{O1}), (\mathrm{O3}, \mathrm{O2}),$$

$$(\mathrm{N3}, \mathrm{O1}), (\mathrm{N2}, \mathrm{O4}), (\mathrm{N2}, \mathrm{O3}), (\mathrm{N1}, \mathrm{O2}), (\mathrm{O3}, \mathrm{O1}), (\mathrm{N1}, \mathrm{O3}), (\mathrm{N1}, \mathrm{O4})\}$$

En tenant compte des règles données en chapitre 2 et chapitre 4, 12 liaisons hydrogène sont possibles pour ce tripeptide. Après la prédiction conformationnelle, avec l'algorithme présenté en chapitre 4, seulement 4 liaisons peuvent exister simultanément, et ce du principalement à la règle de domination entre les liaisons hydrogène. Au dela de 4 liaisons hydrogène les atomes sont contraints et il n y a plus d'axes de rotation pour pouvoir former d'autres liaisons hydrogène (cf. sections 3.6.2 et 4.1.1) Le graphe des possibles construit contient 150 conformations possibles et 371 transitions possibles. La figure 5.13 montre l'histogramme du nombre de conformations possibles en fonction du nombre de liaisons hydrogène présentes. Nous remarquons que les conformations avec 3 liaisons hydrogène présentes sont les plus nombreuses.

La figure 5.14 montre un exemple de graphe des possibles avec une barrière (cf. chapitre 1) sur le niveau d'énergie des conformations à -52. Nous remarquons que le graphe comporte 115 conformations avec 8 composantes connexes : une grande composante connexe avec 108 conformations et 7 autres composantes connexes unitaires contenant chacune une seule conformation. Nous remarquons que les com-

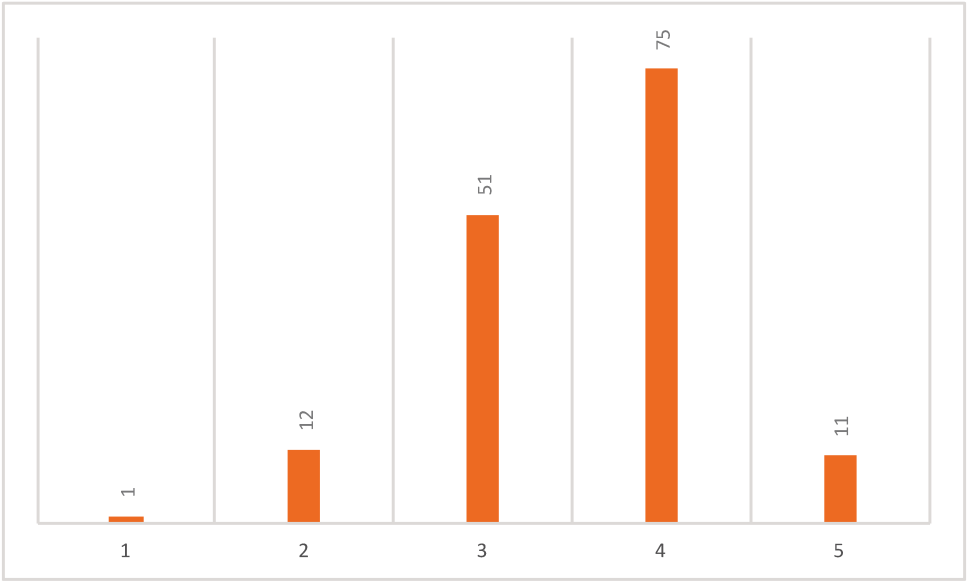

FIGURE 5.13 – Histogramme du nombre de conformations possibles en fonction du nombre de liaisons hydrogène pour le tripeptide $C_9H_{18}N_3O_4$.

posantes avec une seule conformation comportent toutes des conformations avec 3 liaisons hydrogène, ce qui signifie qu'il faut lever plus la barrière pour avoir des conformations avec moins de liaisons hydrogène et donc pouvoir les connecter. Le graphe des possibles total est un graphe connexe.

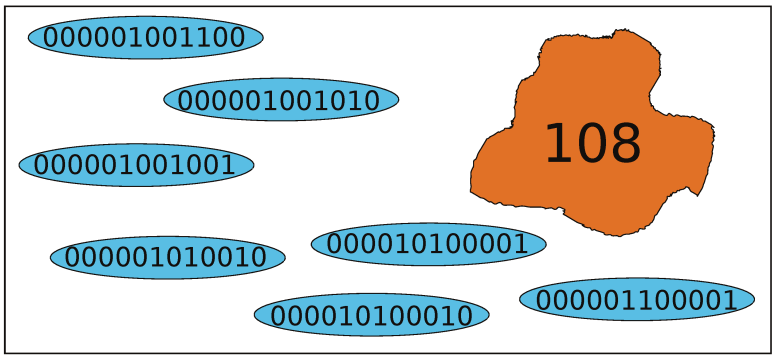

FIGURE 5.14 – Les composantes connexes du graphe des possibles avec niveau d'énergie maximal de (-52). La composante en orange représente la plus grande composante du graphe des possibles elle contient 108 conformations, et le reste des composantes connexes (en bleu) sont de même taille. Chacune contient une seule conformation qui est représentée sur la figure par une séquence binaire (voir le texte pour plus de détails sur cette représentation binaire).

Pour faciliter l'analyse, les conformations sont présentées par une séquence binaire dont la taille est le nombre de liaisons hydrogène. Par exemple pour ce tripeptide, une conformation est une séquence binaire de taille 12. Les liaisons hydrogène sont numérotées de 1 à 12. De gauche à droite, le premier bit de la séquence présente la liaison hydrogène 1 $(\mathrm{N3}, \mathrm{O4})$, le deuxième bit présente la liaison hydrogène

2 (N3, O3) et ainsi de suite, jusqu'au dernier bit, il présente la liaison hydrogène 12 (N1, O4). Chaque bit est à 1 si la liaison hydrogène est présente et 0 dans le cas contraire. La conformation 001000001100 désigne une conformations avec les trois liaisons hydrogène : (N2, O2), (N1, O2) et (O3, O1).

Parmi les objectifs du graphe des possibles est de trouver des plus courts chemins de coût minimum d'une conformation à une autre. Soient les deux conformations possibles les 000001010010 et 001001000010 (voir figure 5.15) dont les deux liaisons hydrogène (N3, O1) et (N1, O4) sont en commun.

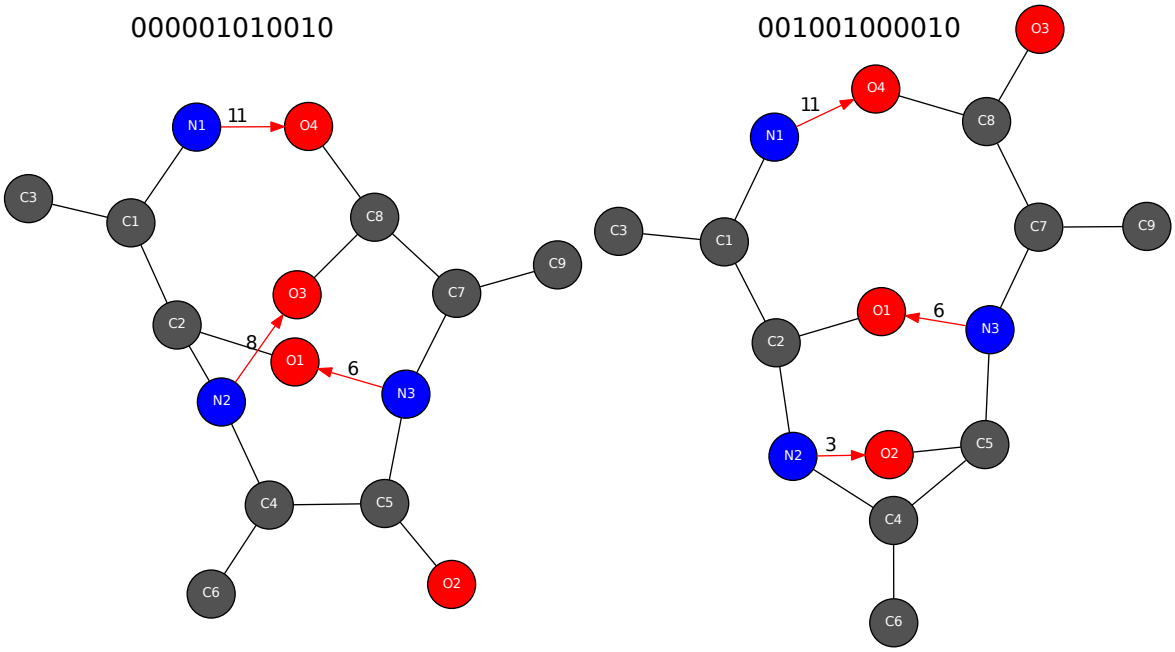

FIGURE 5.15 – Graphes mixtes des conformations 000001010010 et 001001000010. Les liaisons covalentes sont présentées en arêtes noires et les liaisons hydrogène en arcs rouges. La couleur des sommets dépend du type d'atome : les atomes d'oxygène sont présentés en rouge, les atomes de carbone en gris foncé, les atomes d'azote en bleu et les atomes d'hydrogène en gris clair. L'étiquette attribuée à chaque sommet est son type chimique combiné à son numéro d'apparition selon son type. Les labels sur les liaisons hydrogène présentent les numéros de ces liaisons.

Pour trouver des plus courts chemins de coût minimum de la conformation 000001010010 à la conformation 001001000010, l'algorithme cherche un plus court chemin entre ces deux conformations, en ignorant la barrière. Le coût du chemin trouvé devient la nouvelle barrière et ainsi de suite, jusqu'à que l'algorithme ne trouve plus de chemin au dessous de la barrière fixée.

L'analyse a donné 3 chemins de la conformation 000001010010 à la conformation 001001000010. Le tableau 5.2 résume les trois chemins trouvés :

Chaque ligne du tableau 5.2 indique la barrière utilisée, la taille du chemin trouvé (nous rappelons qu'il s'agit d'un plus court chemin en termes de nombre

TABLE 5.2 – Meilleurs chemins de la conformation 000001010010 à la conformation 001001000010.

N°	Barrière	Taille	Coût	chemin
1	-	3	83.33	000001010010 000001000010 (0.00) 001001000010 (83.33)
2	83.33	5	70.00	000001010010 000001000010 (0.00) 000000000010 (0.00) 001000000010 (70.00) 001001000010 (58.33)
3	70.00	7	58.33	000001010010 000001000010 (0.00) 000000000010 (0.00) 000000000000 (0.00) 001000000000 (25.00) 001000000010 (45.00) 001001000010 (58.33)

de conformations traversées y compris les conformations de départ et d'arrivée), le coût du chemin (cf. section 4.4) et l'ensemble des conformations traversées (les conformations sont affichées par ordre de visite). Pour chaque conformation visitée, le tableau indique entre parenthèse le coût de la transition pour arriver à cette conformation.

Nous remarquons que plus la barrière est basse plus le chemin est long. Le chemin le moins coûteux est le troisième chemin trouvé, avec un coût de 58.33. Il n'existe pas de chemin entre les conformations 000001010010 et 001001000010 et dont le coût est inférieur à 58.33. Cependant, en termes de nombre de conformations il est plus long avec 7 conformations (y compris les conformations 000001010010 et 001001000010) que les deux autres chemins trouvés. En analysant la dernière colonne du tableau qui indique les conformations traversées, nous constatons que le troisième chemin consiste en la disparition de toutes les liaisons hydrogène présentes puis à les former ou reformer par ordre croissant de la taille de la liaison (taille du cycle engendré par la liaison hydrogène).

Si nous reprenons le chemins dans le sens inverse, c'est-à-dire de la conformation 001001000010 à la conformation 000001010010, l'algorithme à trouvé que deux chemins. Contrairement au chemin de 001001000010 à la conformation 000001010010, le chemin le moins coûteux ne consiste pas la disparition de toutes les liaisons hydrogène présentes dans la conformation 001001000010 puis la formation ou la re-formation des liaisons hydrogène présentes dans la conformation 000001010010, ce qui est rend le modèle proposé intéressant d'un point de vu chi-

mique. Le tableau 5.3 résume les résultats trouvés.

TABLE 5.3 – Meilleurs chemins entre les conformations 001001000010 et 000001010010.

N°	Barrière	Taille	Coût	chemin
1	-	3	119.05	001001000010 000001000010 (0.00) 000001010010 (119.05)
2	119.05	5	73.81	001001000010 001001000000 (0.00) 000001000000 (0.00) 000001010000 (45.24) 000001010010 (73.81)

Tous les chemins sont donnés en annexe D.

5.5 Evaluation de la prédiction conformationnelle vis à vis l'analyse des trajectoires

Nous reprenons l'exemple du tripeptide $C_9H_{18}N_3O_4$, dont les trajectoires ont été analysés par l'algorithme présenté en chapitre 3 et le graphe des possibles a été présenté dans la section 5.4. La conformation sans liaisons hydrogène est présentée dans la figure 5.12.

Dans la prédiction, 12 liaisons hydrogène possibles ont été identifiées. Sur les trajectoires de ce peptide qui ont été analysées avec l'algorithme présenté en chapitre 3, seules 5 liaisons hydrogène ont été trouvées : $\{(N1, O3), (N1, O1), (N3, O1), (N3, O4), (N2, O2)\}$. Parmi ces liaisons hydrogène, au plus 3 liaisons hydrogène apparaissent simultanées dans des conformations.

La figure 5.16 présente un exemple du graphe de transitions trouvé pour une de ces trajectoires.

Nous remarquons qu'entre la conformations 000100000001 et la conformation 000101000001 il y a une liaisons hydrogène en plus qui est la liaison $(N3, O1)$. Le graphe de transition montre qu'il n y a pas de chemin direct de la conformation 000100000001 à la conformation 000101000001. Il y a deux conformations intermédiaires : la conformation 000000000001 dont la liaison hydrogène $(N1, O1)$ disparait et la conformation 000101000000 où la liaison hydrogène $(N3, O1)$ apparaît. La figure 5.17 montre les graphes mixtes des conformations traversées dans le chemin observé.

En appliquant l'algorithme de la prédiction conformationnelle, deux chemins ont été identifiés. Le plus court chemin de coût minimum de la conformation 000100000001 à la conformation 000101000001 diffère du chemin observé. Dans le

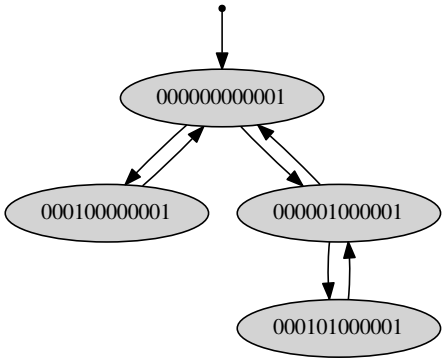

FIGURE 5.16 – Graphe de transition d'une trajectoire. Les étiquettes sur les sommets présentent la séquence binaire qui représente chaque conformation explorée.

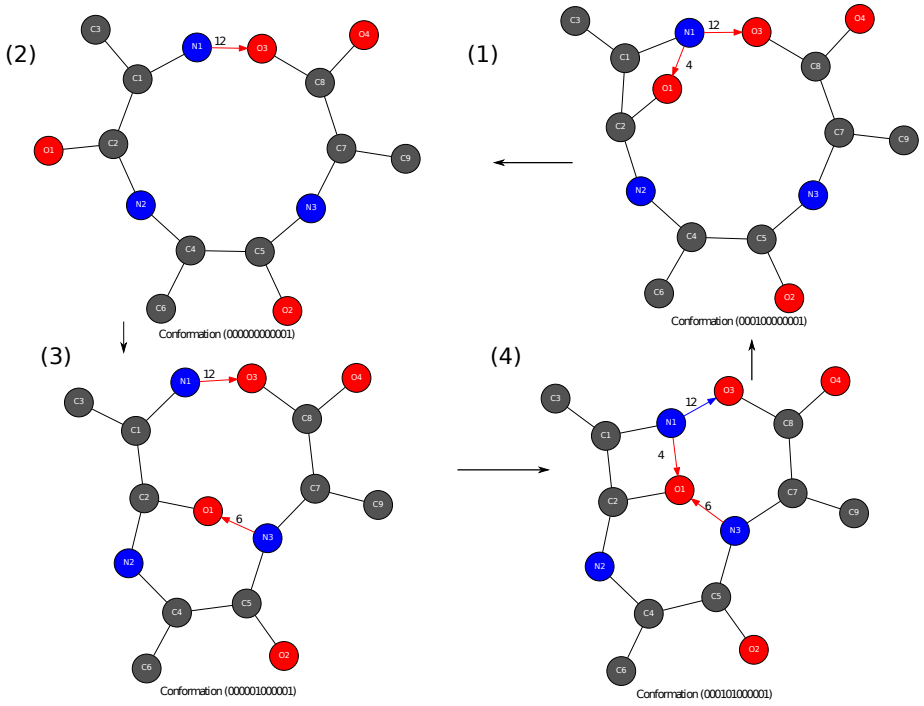

FIGURE 5.17 – Le chemin observé entre la conformation 000100000001 et la conformation 000101000001. Les graphes mixtes des conformations sont données par ordre dans le chemin.

chemin théorique, nous explorons d'abord la conformation 000100000000 où la liaison hydrogène $(\mathrm{N1}, \mathrm{O3})$ n'est plus présente, ensuite la conformation 000101000000 avec l'apparition de la liaison hydrogène $(\mathrm{N3}, \mathrm{O1})$ avec un coût de transition de 58.33 et enfin arriver à la conformation 000101000001 avec un coût de transition de 91.67 en créant la conformation 000101000001. La figure 5.18 montre les graphes mixtes des conformations traversées dans ce chemin.

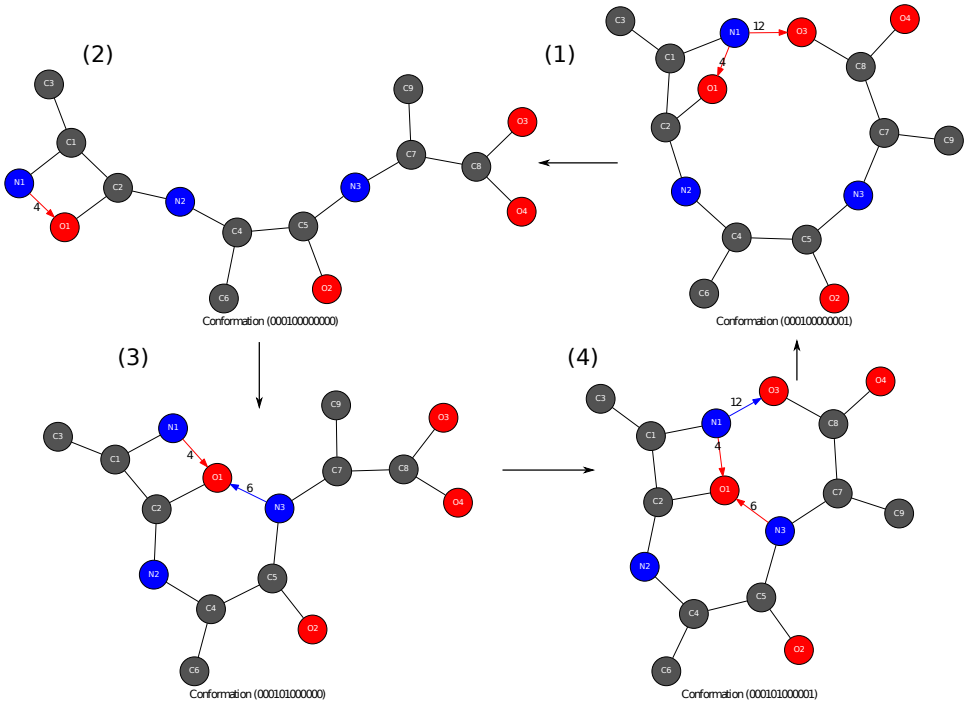

FIGURE 5.18 – Un plus court chemin de coût minimum de la conformation 000100000001 à la conformation 000101000001. Les graphes mixtes des conformations sont données par ordre dans le chemin.

Le coût de ce chemin est donc 91.67. L'analyse du chemin réel (chemin observé dans la trajectoire) avec l'algorithme de prédiction a indiqué que le coût de ce chemin est de 116.00. La transition la plus coûteuse dans le chemin réel est le passage de la conformation 000001000001 à la conformation 000101000001, cela est expliqué, selon notre modèle, par le fait que tous les atomes impliqués dans la liaison $(\mathrm{N1}, \mathrm{O1})$ sont inclus dans un cycle, ce qui contraint les axes de rotation. Nous avons également constaté que le chemin directe de la conformation 000100000001 à la conformation 000101000001 a un coût relativement proche au coût du chemin réel, avec un coût de 105.00 qui représente le coût d'apparition de la liaison $(\mathrm{N3}, \mathrm{O1})$. La différence entre les deux chemins est que les atomes de la liaison $(\mathrm{N1}, \mathrm{O1})$ sont dans un cycle de taille 8 et les atomes de la liaison $(\mathrm{N3}, \mathrm{O1})$ sont dans un cycle de taille 10, en autres termes, les atomes de la liaison $(\mathrm{N1}, \mathrm{O1})$ dans le chemin réel sont plus contraints que les atomes de la liaison $(\mathrm{N3}, \mathrm{O1})$ dans le chemin théorique. Bien que notre approche trouve des chemins différents sur les chemins observés, ces chemins trouvés par notre approche restent cohérents et réalistes d'un point de vu chimique.

5.6 Conclusion

Dans ce chapitre, nous avons illustré, l'application de l'algorithme présenté dans le chapitre 3 sur des trajectoires de différents systèmes moléculaires : trajectoires

de peptides isolés, trajectoires de collision conduisant à une fragmentation du peptide et clusters d'eau. La résultats obtenus montrent la flexibilité d'algorithme et sa capacité d'identifier les conformations sur différentes trajectoires.

L'inconvénient de cet algorithme est précédemment vu quand il y a une grande dynamique de liaisons comme nous l'avons vu pour le peptide $C_{21}H_{38}N_7O_8$. En effet, il y a une difficulté dans la séparation des conformations stables des états transitoires. La méthode proposée en utilisant des pourcentages sur les durées d'apparition de la conformation (cd. section 3.6.1) devient insuffisante dans ces cas.

Le deuxième algorithme proposé, au cours de cette thèse, permet de prédire les conformations possibles d'un système moléculaire en utilisant des mesures ah doc, sans connaissance explicite de la géométrie de la molécule et donc sans connaissance de la surface d'énergie potentielle. Les résultats montrent que l'algorithme donne une classification pertinente des conformations possibles ainsi que les chemins à coûts minimum entre ces conformations sont cohérents avec les hypothèses fixées dans l'approche. La difficulté de cette étape figure dans la comparaison entre les résultats de prédiction et les résultats d'analyse de trajectoires (comparaison des chemins sur le graphe des possibles avec les graphes expérimentaux obtenus par l'analyse des trajectoires). Cette difficulté revient du fait que les trajectoires de simulation de dynamique moléculaire contiennent peu de changements conformationnels. En outre, quand le nombre d'atomes augmente le nombre de conformations possibles augmente. Cependant, nous avons constaté que le nombre de conformations dans le graphe des possibles est faible en comparaison avec le nombre de combinaisons possibles théoriquement des liaisons hydrogène (cf. section 4.5). D'un point de vue chimique, bien que le nombre de conformations possibles obtenu reste trop important, la recherche des chemins de coût minimum entre les conformations peut donner des chemins cohérents et réalistes.

Chapitre 6

Conclusion et Perspectives

Au cours de cette thèse, nous avons présenté deux algorithmes complémentaires dont l'objectif est de comprendre et prédire l'évolution d'un système moléculaire en termes de conformations. Le premier algorithme présenté dans le chapitre 3 a pour objectif d'analyser la dynamique conformationnelle de trajectoires de simulations de dynamique moléculaire. Nous avons utilisé dans cet algorithme les graphes mixtes pour modéliser les conformations, où les sommets représentent les atomes du système moléculaire et les arcs/arêtes représentent les liaisons et les interactions entre ces atomes. Définir la conformation comme un graphe mixte permet à l'algorithme d'être adapté et appliqué facilement à n'importe quel système moléculaire. Une trajectoire de simulation de dynamique moléculaire n'est qu'une évolution des positions des atomes du systèmes à des intervalles de temps réguliers. A chaque intervalle de temps, nous considérons les coordonnées cartésiennes des atomes du système, ce qu'on a appelé une **image**. L'algorithme utilise des règles de géométrie (distances, angles, etc.) sur ces coordonnées pour trouver les liaisons (liaisons covalentes, liaisons hydrogène et interactions électrostatiques) formées entre les atomes et donc avoir les graphes mixtes des conformations. Nous avons remarqué pour les trajectoires, que nous avons analysées, que d'une image à la suivante les atomes bougent peu et donc sur une trajectoire il y a un sous-ensemble d'images appelées **images de référence** qui permettent de couvrir toutes les conformations explorées dans la trajectoire. Sur ces images de référence, nous avons défini des **orbites** qui représentent pour un atome donné les atomes les plus proches de lui en termes de distance géométrique. L'idée est, plutôt que de prendre tous les atomes du système moléculaire à chaque fois, de ne prendre en compte qu'un ensemble d'atomes qui peuvent potentiellement former une liaison chimique pour un atome donné. Bien que l'utilisation des images de référence et des orbites ne réduit pas la complexité théorique de l'algorithme, elle permet, en pratique, d'éviter un calcul systématique à chaque image de la trajectoire et donc de réduire le nombre de calculs effectués.

L'algorithme proposé a été testé sur trois types de systèmes : trajectoires de peptides isolés de taille et de complexité croissantes, trajectoires de dissociation de peptides induite par collision et des trajectoires de clusters. L'objectif en considérant des systèmes différents était de montrer le fonctionnement et la flexibilité de l'algorithme ainsi qu'expliquer ses différents résultats.

Lors de l'analyse des trajectoires, nous avons constaté que les simulations ex-

plorent très peu de conformations (moins d'une dizaine de conformations pour les trajectoires analysées et même avec des trajectoires de 50000 images). En effet, même une simulation extrêmement longue en temps ne garantit pas l'exploration de tous les bassins conformationnels possibles d'un système moléculaire, ce qui signifie que nous n'avons qu'une partie de la surface d'énergie potentielle. De plus, en termes de temps de calculs, la génération d'une trajectoire de dynamique moléculaire ab initio prend en général quelques jours (trajectoires des molécules isolées) voire quelques semaines (trajectoires d'interfaces solide-liquide). Ces observations nous ont amenés à developper un algorithme qui permet de prédire les conformations possibles que peut adopter un système moléculaire. Cet algorithme consiste à construire un graphe appelé **graphe des possibles**. Il s'agit d'un graphe orienté pondéré où les sommets représentent les conformations possibles et les arcs représentent les transitions entre ces conformations. L'idée est de s'affranchir des calculs moléculaires, en particulier du calcul de l'énergie potentielle, pour trouver ces conformations et les transitions entre elles. Le graphe des possibles est construit en utilisant des mesures *ah doc* et des concepts de théorie des graphes qu'on applique aux graphes mixtes des conformations. Cet algorithme permet non seulement de prédire les parties non observées de la surface d'énergie potentielle, mais il représente également une méthode alternative pour les groupes de recherche en chimie théorique et computationnelle pour trouver les conformations les plus stables et les points de passages entre elles (états de transitions sur la surface d'énergie potentielle). En effet, cet algorithme permet de définir une pondération liée au niveau énergétique des conformations du graphe des possibles, ce qui permet d'avoir une classification de ces conformations et donc de pouvoir choisir les plus stables. Pour les états de transition, nous avons proposé un modèle qui permet d'estimer le coût énergétique pour aller d'une conformation à une autre.

Lors du développement du deuxième algorithme, nous ne nous sommes intéressés qu'aux liaisons hydrogène, c'est-à-dire, que les changements de conformations pris en compte sont seulement les changements dans l'ensemble des liaisons hydrogène. Par conséquent, l'algorithme a été testé que sur des systèmes contenant uniquement des changements de liaisons hydrogène. Nous avons testé l'algorithme sur des peptides de taille et de complexité croissantes. Nous avons remarqué que le nombre de conformations possibles obtenues par l'algorithme est faible en comparaison avec le nombre de conformations théoriques (le nombre total de combinaisons). D'un point de vue chimique, bien que le nombre de conformations possibles obtenues reste trop important, la recherche des chemins de coût minimum entre les conformations peut donner des chemins cohérents et réalistes. Nous avons également vu que pour aller d'une conformation à une autre, il peut y avoir plusieurs chemins. Les chemins observés dans les trajectoires sont inclus dans le graphe des

possibles. Il était donc important d'étudier la différence entre le coût énergétique des chemins réels (résultats d'analyse des trajectoires) et le coût énergétique des chemins théoriques (graphe des possibles). Les résultats obtenus ont montré que les chemins de coût minimum restent cohérents d'un point de vue chimique. Parfois nous trouvons que le chemin observé est le chemin a coût minimum et d'autres fois le plus court chemin à coût minimum est différent que le chemin observé. Dans ce dernier cas de différence, nous avons remarqué que le chemin obtenu par l'algorithme de prédiction reste cohérent avec le modèle proposé et réaliste d'un point de vue chimique.

En ce qui concerne les perspectives, pour l'algorithme d'analyse de trajectoires, nous visons à adapter le modèle à d'autres systèmes moléculaires, à savoir les interfaces solide-liquide. Nous avons commencé à tester l'algorithme sur des clusters, qui ont donné des résultats encourageant pour étendre l'analyse à des systèmes plus complexes. Nous avons remarqué dans certaines trajectoires qu'il y a des faibles variations répétitives des distances autour du seuil, ce qui rend la méthode proposée pour séparer les conformations stables des états transitoires peu efficace. Donc il faut chercher une autre méthode pour bien capturer ce type de dynamique.

Par ailleurs, la contribution concernant la prédiction conformationnelle est une première étude exploratoire. L'objectif à présent serait d'ajouter d'autres critères pour diminuer le nombre de conformations possibles, à savoir introduire des propriétés sur les angles entre les liaisons tout en oubliant la géométrie exacte de la molécule. Le graphe d'une conformation peut être vu comme un polygone irrégulier. L'ajout d'une liaison hydrogène revient à ajouter un cycle, donc faire un repliement d'une partie de la molécule. L'idée est de trouver une distribution d'angles sur cette partie qui donne un repliement formant une liaison hydrogène, sachant qu'en chimie chaque type chimique d'atome possède un intervalle bien défini d'angles à respecter. S'il y a une convergence donc la liaison hydrogène peut se former et donc le cas contraire cette liaison hydrogène ne peut pas être ajoutée. Cela permettra de préparer le terrain pour les groupes de recherche afin de prédire les formes tridimensionnelles des conformations possibles et de faire des simulations plus intelligentes.

Pour les coûts énergétiques des transitions, nous supposons que la disparition d'une liaison hydrogène n'a pas de coût. Cela n'est pas totalement vrai, car si nous prenons l'exemple de deux liaisons hydrogène (a_1, b_1) et (a_2, b_2) telles que le cycle obtenu par la liaison hydrogène (a_1, b_1) est totalement inclus dans le cycle obtenu par la liaison hydrogène (a_2, b_2), la suppression de la liaison hydrogène (a_1, b_1) puis la liaison (a_2, b_2) n'a pas en réalité le même coût que la suppression de la liaison (a_2, b_2) puis la liaison (a_1, b_1). A cet effet, pour raffiner le modèle proposé, il serait intéressant d'ajouter d'autres paramètres dans le calcul du coût énergétique de

transition qui tiennent compte de ce genre de situations. De plus, une idée serait d'ajouter les chaines de Markov et donc les probabilités pour trouver les transitions possibles entre les conformations.

Enfin, une interface web a été développée pour pouvoir tester l'algorithme. Des exemples de trajectoires sont disponibles sur cette interface ainsi qu'un guide qui permet d'expliquer l'utilisation de cette interface et de l'algorithme. L'interface web est accessible sur le lien `http://hydrochronographe.prism.uvsq.fr`. Nous travaillons également sur les codes des deux algorithmes et leurs manuels d'utilisation pour les rendre en libre accès et pour qu'ils puissent être utilisés par les différents groupes de recherche de la chimie théorique et computationnelle.

Annexes

Chapitre A

Définitions et notation

A.1 Graphes

Définition 34 (Graphe)**.** *Un graphe est défini par un couple $G = (V, E)$ tel que*

- *V est un ensemble fini de sommets.*
- *E est un ensemble de couples de sommets de V.*

Un graphe peut être orienté ou non :

- *Dans un **graphe orienté** : les couples de E sont orientés, c'est-à-dire (a, b) et (b, a) sont deux couples différents dans le graphe. Le couple (a, b) est appelé **arc**.*
- *Dans un **graphe non orienté** : les couples de E ne sont pas orientés, c'est-à-dire (a, b) est équivalent à (b, a). Le couple (a, b) est appelé **arête**. Dans une graphe non orienté, on parle d'une **paire** de sommets $[a, b]$ au lieu de couple (a, b).*

*Un graphe contenant à la fois des arêtes et des arcs est appelé **graphe mixte**.*

Définition 35 (Graphe pondéré)**.** *Un graphe orienté (resp. non orienté) pondéré est défini par un triplet $G = (V, E, \phi)$ où chaque arc (resp. arête) est affecté d'un nombre réel, appelé poids de cet arc (resp. arête). La fonction ϕ définit la pondération :*

$$\begin{array}{rccc} \phi & : & E & \rightarrow & \mathbb{R} \\ & & (a,b) & \mapsto & \phi(a,b) \end{array}$$

Définition 36 (Chemin/chaine)**.** *Dans un graphe orienté (resp. non orienté) $G = (V, E)$, un **chemin** d'un sommet a à un sommet b (resp. une **chaine** entre deux sommets a et b) est une séquence $(a_1, a_2, ..a_n)$ où $a_1 = a$, $a_n = b$ et chaque deux sommets consécutifs a_i et a_{i+1} sont reliés par un arc (resp. une arête) de G. Ce chemin (resp. une **chaine**) est dit **élémentaire**, si tous les sommets sont tous distincts. La **longueur** du chemin (resp. une **chaine**) est la taille de la séquence parcourue, c'est-à-dire le nombre de sommets dans le chemin.*

Définition 37 (Circuit/cycle)**.** *Dans un graphe orienté (resp. non orienté) $G = (V, E)$, un chemin (resp. une chaine) $(a_1, a_2, ..a_n)$ forme un **circuit** (resp. **cycle**) si $a_1 = a_n$ et le chemin (resp. une chaine) comporte au moins un arc (resp. une arête). Le cycle est élémentaire si les sommets $a_2, a_2, ..a_n$ sont tous distincts.*

Définition 38 (Composante connexe/composante fortement connexe)**.** *Soit $G = (V, E)$ un graphe non orienté et $a \in V$. On définit la* ***composante connexe*** *de G contenant a par :*

$$C_a = \{b \in V | \textit{il existe une chaine de a à b}\}$$

Autrement dit, la composante connexe contenant le sommet a est l'ensemble des sommets qu'il est possible d'atteindre par une chaîne à partir de a.

Dans un graphe orienté, On définit la ***composante fortement connexe*** *de G contenant a par :*

$$C_a = \{b \in V | \textit{il existe un chemin de a à b}\}$$

Définition 39 (Isthme)**.** *Soient un graphe non orienté $G = (V, E)$ et une arête $[a, b]$ de E. On dit que $[a, b]$ est un isthme si sa suppression de E augmente le nombre de composantes connexes de G.*

Chapitre B

Choix des paramètres

B.1 Paramètres de calcul

TABLE B.1 – Caractéristiques des atomes utilisés.

Atome	Symbole	Nombre max de liaisons	Rayon de covalence
Carbone	C	4	0.76Å
Oxygène	O	2	0.71Å
Hydrogène	H	1	0.31Å
Azote	N	3	0.66Å
Lithium	Li	0	1.33Å
Argon	Ar	0	0.96Å

TABLE B.2 – Les valeurs par défaut des paramètres utilisés.

Paramètre	Unité	Valeur
D_C	Å	1.22
D_H	Å	2.3
D_I	Å	2.5
α_H	*degrés*	120
T_r	%	1
$\alpha_{\mathcal{O}_H}$	-	3
$\alpha_{\mathcal{O}_C}$	-	2
$\alpha_{\mathcal{O}_I}$	-	2

Chapitre C

Algorithmes

Données : un ensemble de conformations $\mathcal{G} = \{G_1, G_2, ..., G_C\}$ dont l'ensemble des liaisons covalentes E_C est le même.
Résultat : $A_r \subset E_C$ l'ensemble des axes de rotation.

$A_r \leftarrow \emptyset$;
pour tous les *paires* $[a, b]$ *dans* E_C **faire**
  pour tous les G_i *dans* $\mathcal{G}$ **faire**
    $E_C \leftarrow E_C \setminus \{[a, b]\}$;
    Trouver un chemin T entre $[a, b]$ par parcours en largeur de G_i ;
    si T *est vide et* $[a, b] \notin A_r$ **alors**
      $A_r \leftarrow A_r \cup \{[a, b]\}$;
    $E_C \leftarrow E_C \cup \{[a, b]\}$;

Algorithme 1 : Algorithme pour l'identification des mouvements rotationnels.

Données : une séquence de trajectoires $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n$
Résultat : $\mathcal{G} = \{G_1, G_2, ..., G_C\}$, ensemble des conformations non isomorphes, $\mathcal{G}_\mathcal{I} = (\mathcal{G}, \mathcal{A}) = \bigcup_{i \in [\![1,n]\!]} \mathcal{G}_{\mathcal{I}_i}$ le graphe de transition associé à toutes les trajectoires.

$\mathcal{G} \leftarrow \emptyset$;
$\mathcal{A} \leftarrow \emptyset$;
pour i *de* 2 **a** S **faire**
  Construire $\mathcal{G}_{\mathcal{I}_i} = (\mathcal{G}_i, \mathcal{A}_i)$ de la trajectoire $\mathcal{I}_i$ avec l'algorithme **??** et en utilisant $\mathcal{G}$;
  $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{G}_i$;
  $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{A}_i$;

Algorithme 2 : Algorithme d'analyse de plusieurs trajectoires simultanément.

Données : une séquence d'images $I_1, I_2, ..., I_S$ d'une trajectoire $\mathcal{I}$.
Résultat : $\mathcal{G} = \{G_1, G_2, ..., G_C\}$, $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$, ensemble des conformations non isomorphes, $\mathcal{G}_\mathcal{I} = (\mathcal{G}, \mathcal{A}, \rho, \tau)$ le graphe de transition associé et les $\mathcal{I}_{G_i} = \{I^1_{G_i}, I^2_{G_i}, ..., I^k_{G_i}; 1 \leq k \leq S\}$, les instants d'apparition de la conformation G_i

1 Calculer les orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$ et initialiser l'image de référence à I_1 pour toutes les orbites ;
2 Construire la conformation G_1 à partir de l'image I_1;
3 Initialiser $\mathcal{I}_{G_1}$ à I_1;
4 $\mathcal{G} \leftarrow G_1$;
5 $\mathcal{A} \leftarrow \emptyset$;
6 Calculer la signature de G_1 avec l'algorithme de Mckay (signature(G_1));
7 **pour** i *de* 2 **a** S **faire**
8 **si** $max_1 + max_2 > (\alpha_{\mathcal{O}_C} - 1) \times D_C$ **alors**
9 Calculer les orbites $\mathcal{O}_C$, $\mathcal{O}_H$ et $\mathcal{O}_I$ et changer l'image de référence à I_i pour toutes les orbites ;
10 **sinon**
 si $max_1 + max_2 > (\alpha_{\mathcal{O}_H} - 1) \times D_H$ **alors**
11 Calculer les orbites $\mathcal{O}_H$ et changer l'image de référence à I_i pour ces orbites ;
 si $max_1 + max_2 > (\alpha_{\mathcal{O}_I} - 1) \times D_I$ **alors**
12 Calculer les orbites $\mathcal{O}_I$ et changer l'image de référence à I_i pour ces orbites ;
13 Construire G_i en calculant les liaisons autour des orbites ;
14 **si** $A_{H_i} = A_{H_{i-1}}$, E_{C_i}=$E_{C_{i-1}}$ *et* E_{I_i}=$E_{I_{i-1}}$ *par matrice d'adjacence* **alors**
 G_i et G_{i-1} sont isomorphes
15 **sinon**
16 Calculer la signature de G_i ;
17 **si** $\exists$ G_j *dans* $\mathcal{G}$ *isomorphe à* G_i *(signature(G_j) = signature(G_i))* **alors**
18 $\mathcal{G} \leftarrow \mathcal{G} \cup \{G_j\}$;
19 $\mathcal{I}_{G_j} \leftarrow \mathcal{I}_{G_j} \cup \{I_i\}$;
20 $\mathcal{A} \leftarrow \mathcal{A} \cup \{(G_{i-1}, G_j)\}$;
21 **sinon**
22 $\mathcal{G} \leftarrow \mathcal{G} \cup \{G_i\}$;
23 $\mathcal{I}_{G_i} \leftarrow \{I_i\}$;
24 $\mathcal{A} \leftarrow \mathcal{A} \cup \{(G_{i-1}, G_i)\}$;

Algorithme 3 : Algorithme général pour l'identification des conformations sur une trajectoire.

Données : une conformation initiale $G = (V, E_C, A_H, E_I)$.
Résultat : $\mathcal{G}_P = (\mathcal{G}, \mathcal{A})$ le graphe des possibles.
G_i^{k-1}

Calculer $Hydro$ à partir de V ;
Construire G^0 à partir de G ($G^0 = (V, E_C, \emptyset, \emptyset)$) ;
$\mathcal{G}^0 \leftarrow \{G^0\}$;
$\mathcal{G} \leftarrow \emptyset$;
$\mathcal{A} \leftarrow \emptyset$;
pour i *de* 1 **a** $|Hydro|$ **faire**
    $\mathcal{G}^i \leftarrow \emptyset$;
    pour tous les $G^{i-1} = (V_i, E_{C_i}, A_{H_i}, E_{I_i}) \in \mathcal{G}^{i-1}$ **faire**
        pour tous les $(a, b) \in Hydro$ *et* $(a, b) \notin A_{H_i}$ **faire**
            Construire G^i à partir de G^{i-1} ($G^i = (V_i, E_{C_i}, A_{H_i} \cup \{(a, b)\}, E_{I_i})$) ;
            si $G^i \notin \mathcal{G}^i$ **alors**
                $\mathcal{G}^i \leftarrow \mathcal{G}^i \cup \{G^i\}$;
                $\mathcal{A} \leftarrow \mathcal{A} \cup (G^{i-1}, G^i)$;
$\mathcal{G} = \bigcup_{i \in [\![0, |Hydro|]\!]} \mathcal{G}^i$;

Algorithme 4 : Algorithme pour construire le graphe des possibles.

Données : un graphe des possibles $\mathcal{G}_P = (\mathcal{G}, \mathcal{A}, \Omega, \omega)$, deux conformations $G_D = (V_D, E_{C_D}, A_{H_D}, E_{I_D})$ et $G_F = (V_D, E_{C_D}, A_{H_D}, E_{I_D})$ de $\mathcal{G}_P$ et une barrière $\mathcal{C}_{max}$
Résultat : $P(G_D, G_F)$ un plus court chemin de coût minimum de G_D à G_F

Initialiser $\mathcal{G}_{vist}$ à G_D ;
Initialiser $\mathcal{G}_{Nvist}$ à $\mathcal{G} \setminus \{G_D\}$;
tant que $\mathcal{G}_{Nvist} \neq \emptyset$ **faire**
    pour tous les (G_k, G_l), $G_k \in \mathcal{G}_{vist}$, $G_l \in \mathcal{G}_{Nvist}$ *et* $(G_k, G_l) \in \mathcal{A}$ **faire**
        Choisir G_l tel que $|(P(G_D, G_l)| = \min_{G \in \mathcal{G}_{Nvist}, (G_k, G) \in \mathcal{A}} (|P(G_D, G)|)$ et $\omega(G_k, G_l) \leq \mathcal{C}_{max}$;
        $\mathcal{G}_{Nvist} \leftarrow \mathcal{G}_{Nvist} \setminus \{G_l\}$;
        $\mathcal{G}_{vist} \leftarrow \mathcal{G}_{vist} \cup \{G_l\}$;
        Sauvegarder que G_l est arrivé par G_k dans $P(G_D, G_F)$;
        si G_l *est isomorphe à* G_F **alors**
            il existe un plus court chemin de coût minimum de G_D à G_F
        si $\mathcal{G}_{Nvist} = \emptyset$ **alors**
            il n'existe pas de chemin entre G_D et G_F

Algorithme 5 : Algorithme pour trouver un plus court chemin de coût minimum entre deux conformations.

Chapitre D

Prédiction des conformations pour un peptide

D.1 Liste des conformations trouvées pour le tripeptide $C_6H_{13}N_2O_3$

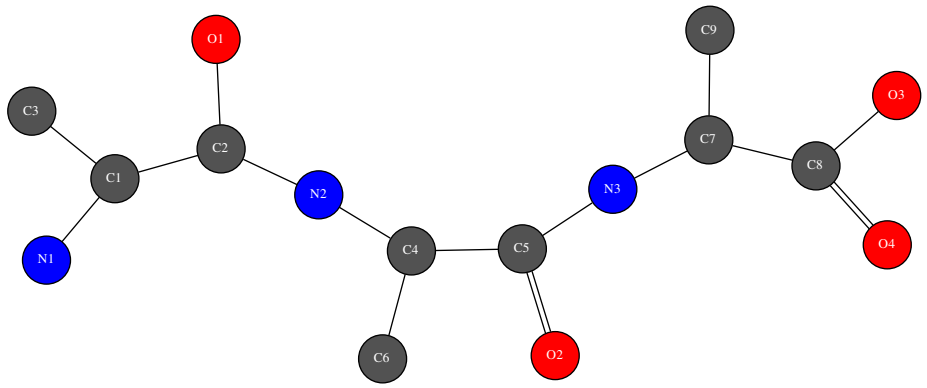

FIGURE D.1 – Graphe mixte à la conformation initiale du tripeptide $C_9H_{18}N_3O_4$. La conformation initiale du système comporte que les liaisons covalentes (présentées en arêtes noires).

TABLE D.1 – Liste des liaisons hydrogène possibles pour le tripeptide $C_6H_{13}N_2O_3$

N°	Liaison hydrogène	Taille du cycle
1	(N3, O4)	5
2	(N3, O3)	5
3	(N2, O2)	5
4	(N1, O1)	5
5	(O3, O2)	7
6	(N3, O1)	7
7	(N2, O4)	8
8	(N2, O3)	8
9	(N1, O2)	8
10	(O3, O1)	10
11	(N1, O3)	11
12	(N1, O4)	11

TABLE D.2 – Liste des conformations possibles du tripeptide $C_6H_{13}N_2O_3$ ordonné par niveau d'énergie.

N°	conformation	Nombre de liaisons hydrogène	Energie
1	001011000010	4	-108
2	101000001100	4	-106
3	100110000001	4	-106
4	100110000010	4	-106
5	100110001000	4	-103
6	101100000100	4	-103
7	101100000001	4	-99
8	011100000010	4	-99
9	001001000110	4	-95
10	011100000000	3	-86
11	101100000000	3	-86
12	001110000000	3	-86
13	100110000000	3	-85
14	010100100000	3	-84
15	010100001000	3	-84
16	100100010000	3	-84
17	100100001000	3	-84
18	011000001000	3	-83
19	101000001000	3	-83
20	100100100000	3	-83
21	001011000000	3	-83
22	000111000000	3	-82
23	000110100000	3	-82
24	001100000100	3	-80
25	100100000100	3	-80
26	000110001000	3	-80
27	100010001000	3	-80
28	000011001000	3	-80
29	100010000010	3	-78
30	100010000001	3	-78
31	001001000010	3	-78

32	001001000001	3	-78
33	000001101000	3	-78
34	000001011000	3	-78
35	001000001100	3	-78
36	101000000100	3	-77
37	010100000100	3	-77
38	000011000010	3	-77
39	000101100000	3	-77
40	000101010000	3	-77
41	001001000100	3	-77
42	011000000010	3	-76
43	101000000001	3	-76
44	001100000010	3	-76
45	001100000001	3	-76
46	010100000010	3	-76
47	100100000001	3	-76
48	001010000010	3	-76
49	000010101000	3	-76
50	000100101000	3	-76
51	010000101000	3	-76
52	000100011000	3	-76
53	100000011000	3	-76
54	000100001100	3	-76
55	000100001010	3	-76
56	000100001001	3	-76
57	010000100010	3	-75
58	100000100100	3	-75
59	100000100001	3	-75
60	100000010010	3	-75
61	001000001010	3	-75
62	001000001001	3	-75
63	000000101100	3	-74
64	001000000110	3	-74
65	011000000100	3	-73
66	101000000010	3	-73

67	100100000010	3	-73
68	001010000001	3	-73
69	000101000010	3	-73
70	000101000001	3	-73
71	100000101000	3	-73
72	000110000010	3	-72
73	100000001100	3	-72
74	010000000110	3	-72
75	000110000001	3	-68
76	010000001100	3	-68
77	000001001100	3	-66
78	000001100001	3	-64
79	000010100010	3	-64
80	000010100001	3	-64
81	000001010010	3	-64
82	000001001010	3	-64
83	000001001001	3	-64
84	011000000110	4	-64
85	100000101100	4	-64
86	100100000000	2	-56
87	010100000000	2	-56
88	001100000000	2	-56
89	101000000000	2	-56
90	011000000000	2	-56
91	100000001000	2	-55
92	010000001000	2	-55
93	000100010000	2	-55
94	000100100000	2	-55
95	000101000000	2	-55
96	001010000000	2	-55
97	000110000000	2	-55
98	000100000100	2	-54
99	000010001000	2	-54
100	001001000000	2	-54
101	000011000000	2	-54

102	100010000000	2	-54
103	000100001000	2	-53
104	100000010000	2	-53
105	010000100000	2	-53
106	100000000001	2	-52
107	001000000001	2	-52
108	000100000001	2	-52
109	010000000010	2	-52
110	001000000010	2	-52
111	000100000010	2	-52
112	100000000100	2	-52
113	001000000100	2	-52
114	000000101000	2	-52
115	000000011000	2	-52
116	001000001000	2	-51
117	100000100000	2	-51
118	010000000100	2	-50
119	000001001000	2	-50
120	000001010000	2	-50
121	000010100000	2	-50
122	000001100000	2	-50
123	100000000010	2	-49
124	000010000010	2	-48
125	000000001100	2	-48
126	000010000001	2	-45
127	000001000001	2	-45
128	000001000010	2	-45
129	000001000110	3	-44
130	000001000100	2	-41
131	000000100001	2	-40
132	000000001001	2	-40
133	000000010010	2	-40
134	000000001010	2	-40
135	000000100100	2	-40
136	000000100010	2	-34

137	000000000110	2	-33
138	100000000000	1	-26
139	010000000000	1	-26
140	001000000000	1	-26
141	000100000000	1	-26
142	000010000000	1	-24
143	000001000000	1	-24
144	000000100000	1	-23
145	000000010000	1	-23
146	000000001000	1	-23
147	000000000100	1	-21
148	000000000010	1	-20
149	000000000001	1	-20

D.2 Chemin entre les conformations 000001010010 et 001001000010

Les figures D.2, D.3 et D.4 montrent les trois chemins trouvés entre les conformations 000001010010 et 001001000010 à différentes barrières d'énergie.

D.3 Chemin entre 001001000010 et 000001010010

Les figures D.5 et D.6 montrent les deux chemins trouvés entre les conformations 001001000010 et 000001010010 à différentes barrières d'énergie.

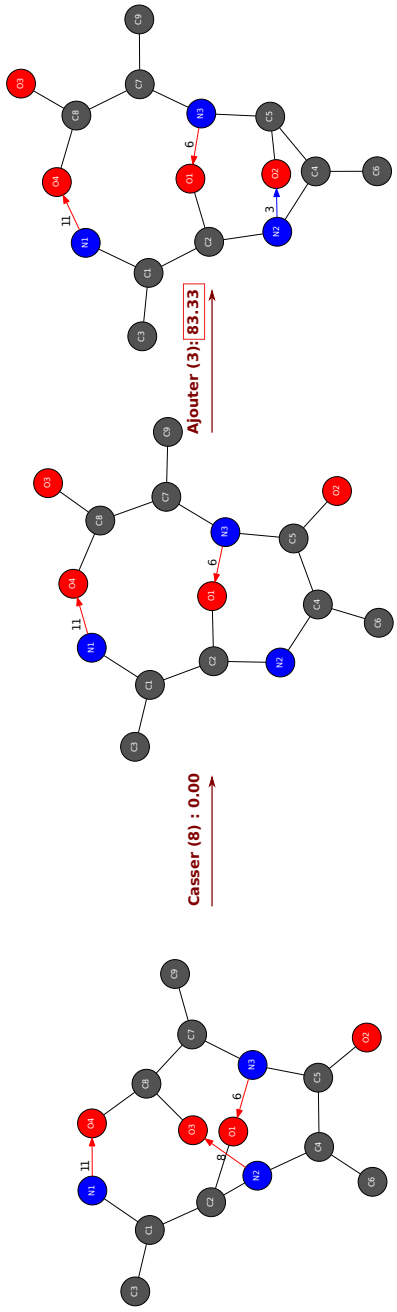

FIGURE D.2 – Chemin entre la conformation 000001010010 et la conformation 001001000010 sans barrière. Les graphes mixtes des conformations sont données par ordre dans le chemin. Les labels sur les arcs présentent le type de transition (ajout ou cassure d'une liaison hydrogène), le numéro de la liaison hydrogène concernée (entre parenthèses) et le coût de la transition.

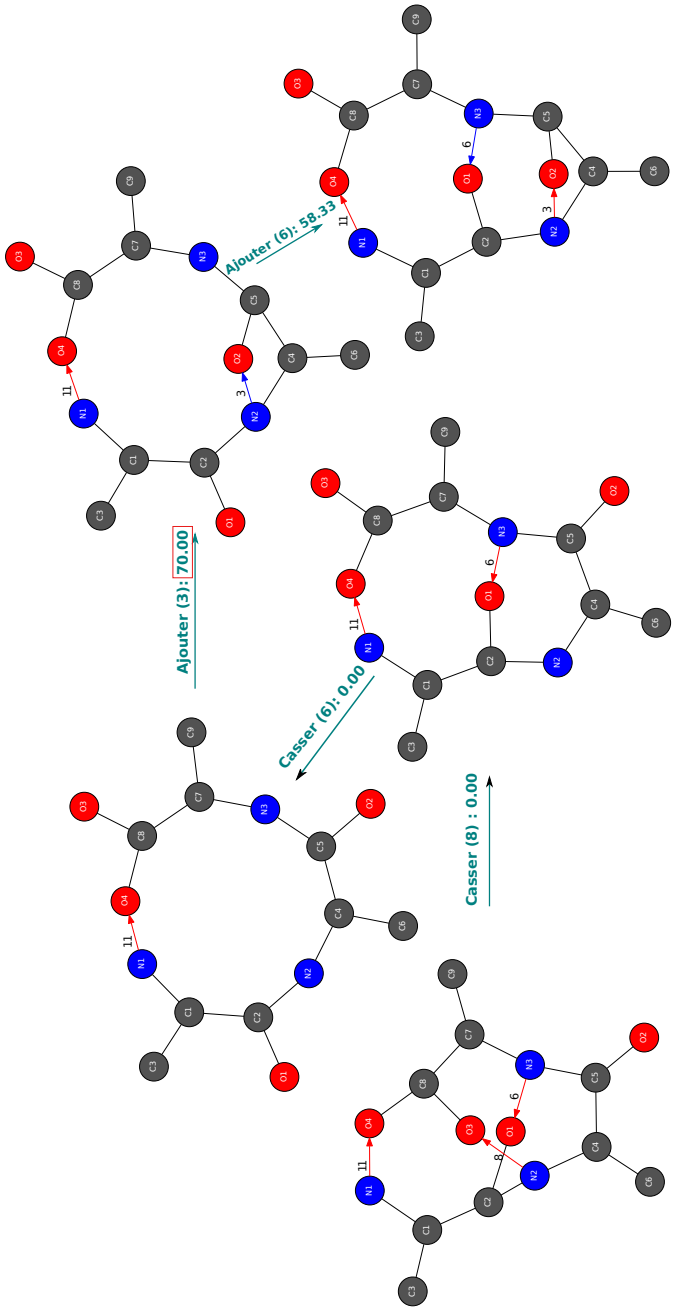

FIGURE D.3 – Chemin entre la conformation 000001010010 et la conformation 001001000010 sous la barrière 83.33. Les graphes mixtes des conformations sont données par ordre dans le chemin. Les labels sur les arcs présentent le type de transition (ajout ou cassure d'une liaison hydrogène), le numéro de la liaison hydrogène concernée (entre parenthèses) et le coût de la transition.

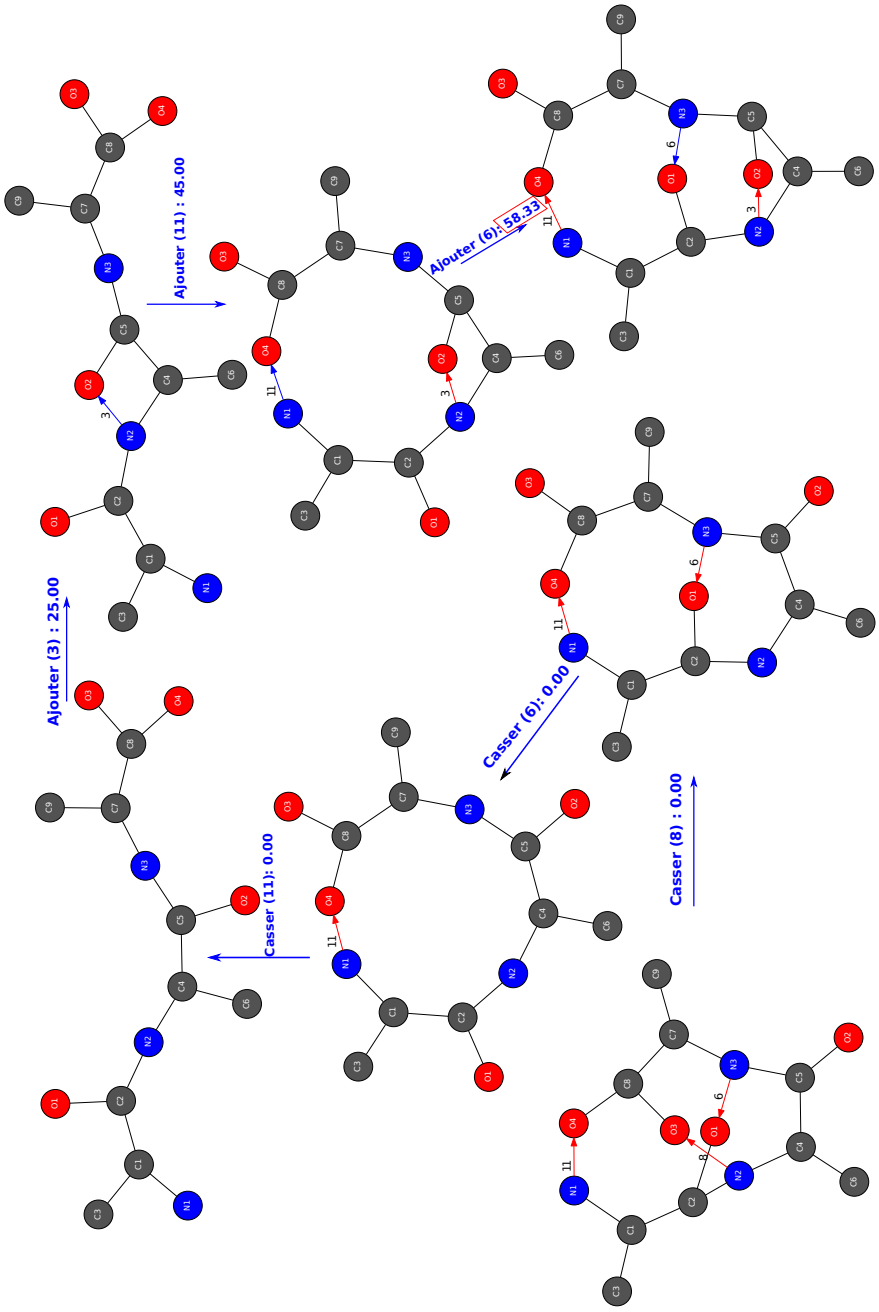

FIGURE D.4 – Chemin entre la conformation 000001010010 et la conformation 001001000010 avec la barrière 70.00. Les graphes mixtes des conformations sont données par ordre dans le chemin. Les labels sur les arcs présentent le type de transition (ajout ou cassure d'une liaison hydrogène), le numéro de la liaison hydrogène concernée (entre parenthèses) et le coût de la transition.

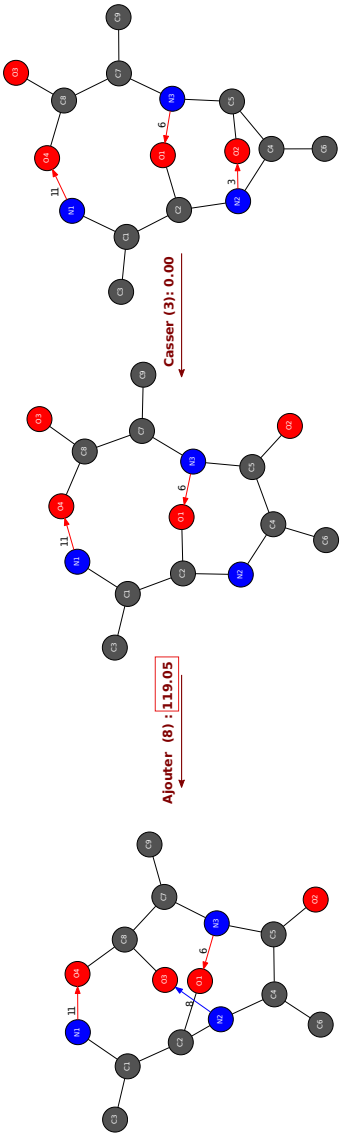

FIGURE D.5 – Chemin entre la conformation 000001010010 et la conformation 001001000010 sans barrière. Les graphes mixtes des conformations sont données par ordre dans le chemin. Les labels sur les arcs présentent le type de transition (ajout ou cassure d'une liaison hydrogène), le numéro de la liaison hydrogène concernée (entre parenthèses) et le coût de la transition.

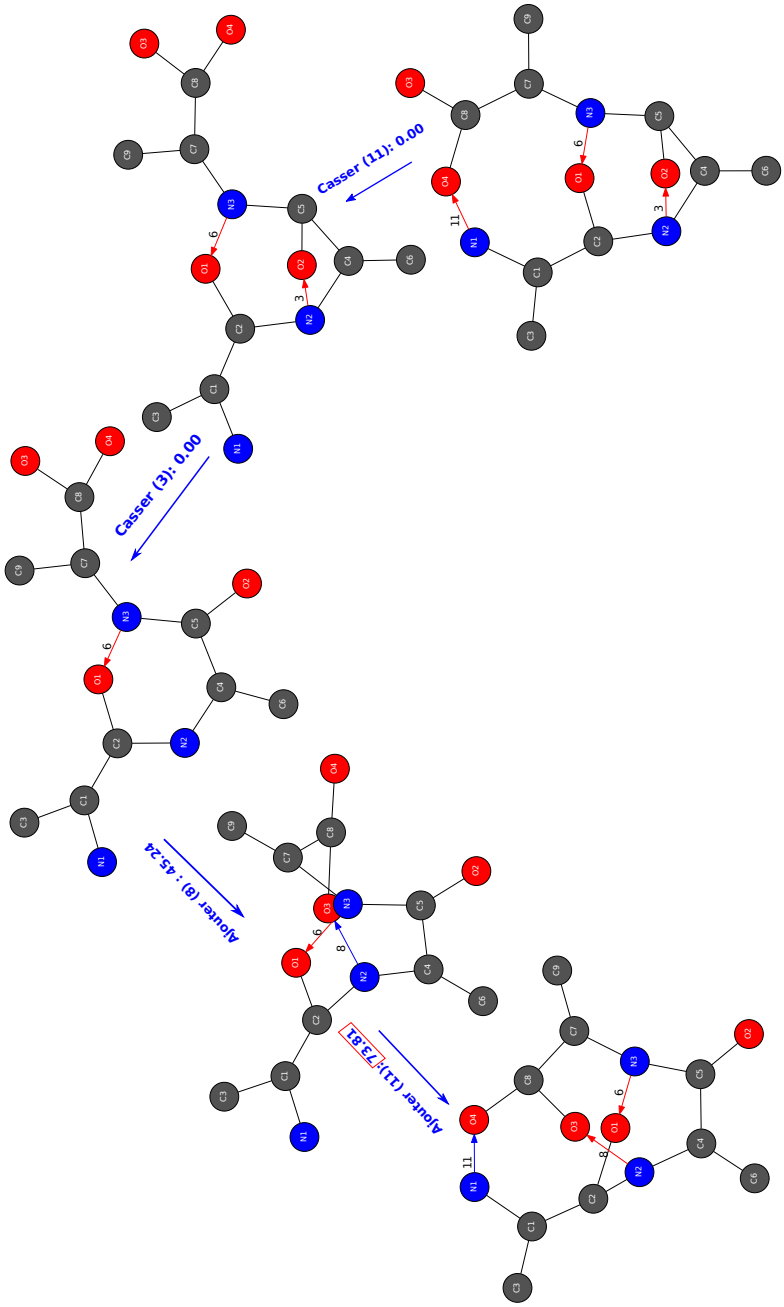

FIGURE D.6 – Chemin entre la conformation 000001010010 et la conformation 001001000010 sous la barrière 119.05. Les graphes mixtes des conformations sont données par ordre dans le chemin. Les labels sur les arcs présentent le type de transition (ajout ou cassure d'une liaison hydrogène), le numéro de la liaison hydrogène concernée (entre parenthèses) et le coût de la transition.

Chapitre E

Interface Web

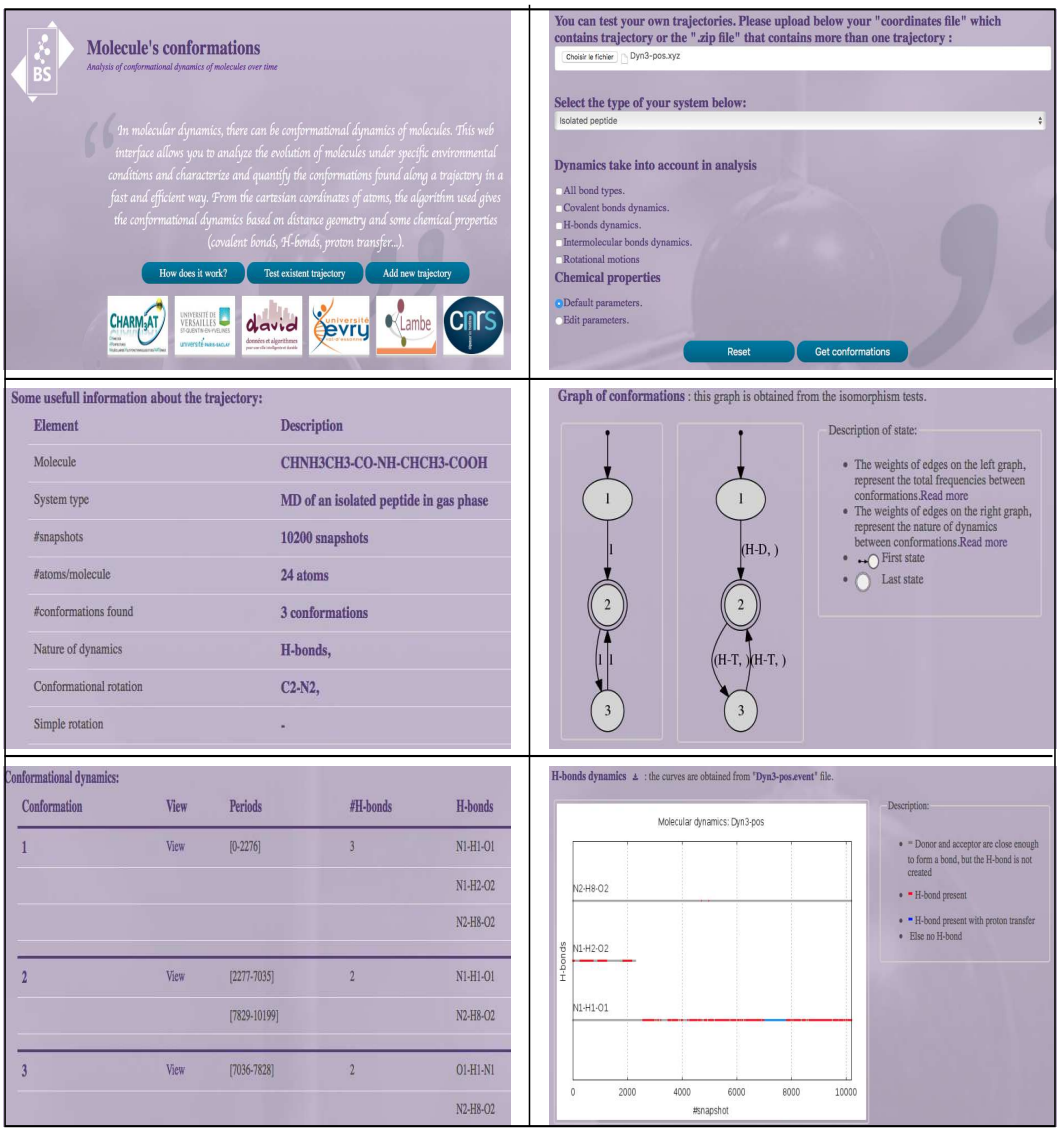

FIGURE E.1 – Captures d'écran des résultats présentés par l'interface web. L'interface web est accessible sur le lien http://hydrochronographe.prism.uvsq.fr

>

Bibliographie

[1] Jonathan CLAYDEN, Nick GREEVES et Stuart WARREN : *Chimie organique : une approche orbitalaire*. De Boeck Supérieur, 2013.

[2] Andrew R LEACH : *Molecular modelling : principles and applications*. Pearson education, 2001.

[3] Carl T KELLEY : *Iterative methods for optimization*. SIAM, 1999.

[4] W. TU et R. W. MAYNE : Studies of multi-start clustering for global optimization. *International Journal for Numerical Methods in Engineering*, 53(9):2239–2252, 2002.

[5] Emilio MARTÍNEZ-NÚÑEZ : An automated method to find transition states using chemical dynamics simulations. *Journal of computational chemistry*, 36(4):222–234, 2015.

[6] James G SPEIGHT *et al.* : *Lange's handbook of chemistry*, volume 1. McGraw-Hill New York, 2005.

[7] Wilfred F van GUNSTEREN et Herman JC BERENDSEN : Computer simulation of molecular dynamics : Methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition*, 29(9):992–1023, 1990.

[8] Daan FRENKEL et Berend SMIT : *Understanding molecular simulation : from algorithms to applications*, volume 1. Academic press, 2001.

[9] M-P GAIGEOT : Unravelling the conformational dynamics of the aqueous alanine dipeptide with first-principle molecular dynamics. *The Journal of Physical Chemistry B*, 113(30):10059–10062, 2009.

[10] Veronika BRÁZDOVÁ et David R BOWLER : *Atomistic computer simulations : a practical guide*. John Wiley & Sons, 2013.

[11] Helmut GRUBMÜLLER, Helmut HELLER, Andreas WINDEMUTH et Klaus SCHULTEN : Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation*, 6(1-3):121–142, 1991.

[12] A CIMAS, TD VADEN, TSJA DE BOER, LC SNOEK et M-P GAIGEOT : Vibrational spectra of small protonated peptides from finite temperature md simulations and irmpd spectroscopy. *Journal of Chemical Theory and Computation*, 5(4): 1068–1078, 2009.

[13] Marcus D HANWELL, Donald Ephraim CURTIS, David C LONIE, Tim VANDERMEERSCH, Eva ZUREK et Geoffrey R HUTCHISON : Avogadro : An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics*, 4:17, 2012.

[14] William HUMPHREY, Andrew DALKE et Klaus SCHULTEN : Vmd : visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[15] Barbara Logan MOONEY, L René CORRALES et Aurora E CLARK : MoleculaRnetworks : an integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. *Journal of computational chemistry*, 33(8):853–860, 2012. PMID : 22278855.

[16] Abdullah OZKANLAR et Aurora E. CLARK : ChemNetworks : a complex network analysis tool for chemical systems. *Journal of Computational Chemistry*, 35(6): 495–505, 2014.

[17] Matthew HUDELSON, Barbara Logan MOONEY et Aurora E. CLARK : Determining polyhedral arrangements of atoms using PageRank. *Journal of Mathematical Chemistry*, 50(9):2342–2350, 2012.

[18] David GOLDBERG : Genetic algorithms in optimization, search and machine learning. *Reading : Addison-Wesley*, 1989.

[19] DB MCGARRAH et Richard S JUDSON : Analysis of the genetic algorithm method of molecular conformation determination. *Journal of Computational Chemistry*, 14(11):1385–1395, 1993.

[20] Dewi Retno Sari SAPUTRO et Purnami WIDYANINGSIH : Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). *In AIP Conference Proceedings*, volume 1868, page 040009. AIP Publishing, 2017.

[21] Margaret H WRIGHT : Direct search methods : Once scorned, now respectable. *Pitman Research Notes in Mathematics Series*, pages 191–208, 1996.

[22] Andrew R CONN, Katya SCHEINBERG et Luis N VICENTE : *Introduction to derivative-free optimization*. SIAM, 2009.

[23] Jun GU, Bin DU et Panos PARDALOS : Multispace search for protein folding. *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS*, 94:47–68, 1997.

[24] H Bernhard SCHLEGEL : Optimization of equilibrium geometries and transition structures. *Journal of Computational Chemistry*, 3(2):214–218, 1982.

[25] Charles J CERJAN et William H MILLER : On finding transition states. *The Journal of chemical physics*, 75(6):2800–2806, 1981.

[26] Johannes KÄSTNER et Paul SHERWOOD : Superlinearly converging dimer method for transition state search. *The Journal of chemical physics*, 128(1): 014106, 2008.

[27] Yihan SHAO, Laszlo Fusti MOLNAR, Yousung JUNG, Jörg KUSSMANN, Christian OCHSENFELD, Shawn T BROWN, Andrew TB GILBERT, Lyudmila V SLIPCHENKO, Sergey V LEVCHENKO, Darragh P O'NEILL *et al.* : Advances in methods and algorithms in a modern quantum chemistry program package. *Physical Chemistry Chemical Physics*, 8(27):3172–3191, 2006.

[28] Graeme HENKELMAN et Hannes JÓNSSON : Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics*, 113(22):9978–9985, 2000.

[29] E WEINAN, Weiqing REN et Eric VANDEN-EIJNDEN : Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *The Journal of chemical physics*, 2007.

[30] Emilio MARTÍNEZ-NÚÑEZ : An automated transition state search using classical trajectories initialized at multiple minima. *Physical Chemistry Chemical Physics*, 17(22):14912–14921, 2015.

[31] D BARTH, S BOUGUEROUA, M-P GAIGEOT, F QUESSETTE, R SPEZIA et S VIAL : A new graph algorithm for the analysis of conformational dynamics of molecules. *In Information Sciences and Systems 2015*, pages 319–326. Springer, 2016.

[32] Eugene M LUKS : Isomorphism of graphs of bounded valence can be tested in polynomial time. *In Foundations of Computer Science, 1980., 21st Annual Symposium on*, pages 42–49. IEEE, 1980.

[33] Brendan D MCKAY *et al.* : *Practical graph isomorphism*. Department of Computer Science, Vanderbilt University Tennessee, US, 1981.

[34] Brendan D. MCKAY et Adolfo PIPERNO : Practical graph isomorphism, {II}. *Journal of Symbolic Computation*, 60(0):94 – 112, 2014.

[35] Stephen G HARTKE et AJ RADCLIFFE : Mckay's canonical graph labeling algorithm. *Communicating mathematics*, 479:99–111, 2009.

[36] Sébastien SORLIN et Christine SOLNON : A new filtering algorithm for the graph isomorphism problem. *In 3rd International Workshop on Constraint Propagation and Implementation, CP2006*, 2006.

[37] José Luis LÓPEZ-PRESA, Antonio Fernández ANTA et Luis Núñez CHIROQUE : Conauto-2.0 : Fast isomorphism testing and automorphism group computation. *arXiv preprint arXiv :1108.1060*, 2011.

[38] Hans L BODLAENDER : Polynomial algorithms for graph isomorphism and chromatic index on partial k-trees. *Journal of Algorithms*, 11(4):631–643, 1990.

[39] Paul T DARGA, Karem A SAKALLAH et Igor L MARKOV : Faster symmetry discovery using sparsity of symmetries. *In Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, pages 149–154. IEEE, 2008.

[40] Tommi JUNTTILA et Petteri KASKI : Engineering an efficient canonical labeling tool for large and sparse graphs. *In 2007 Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 135–149. SIAM, 2007.

[41] Johannes KOBLER, Uwe SCHÖNING et Jacobo TORÁN : *The graph isomorphism problem : its structural complexity*. Springer Science & Business Media, 2012.

[42] László BABAI, Anuj DAWAR, Pascal SCHWEITZER et Jacobo TORÁN : 3.21 complexity classes and the graph isomorphism problem. *The Graph Isomorphism Problem*, page 15, 2016.

[43] László BABAI : Graph isomorphism in quasipolynomial time [extended abstract]. *In Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 684–697. ACM, 2016.

[44] Jonathan L GROSS et Jay YELLEN : *Graph theory and its applications*. CRC press, 2005.

[45] DC MARINICA, G GREGOIRE, C DESFRANCOIS, JP SCHERMANN, D BORGIS et MP GAIGEOT : Ab initio molecular dynamics of protonated dialanine and comparison to infrared multiphoton dissociation experiments. *The Journal of Physical Chemistry A*, 110(28):8802–8810, 2006.

[46] A CIMAS, TD VADEN, TSJA DE BOER, LC SNOEK et M-P GAIGEOT : Vibrational spectra of small protonated peptides from finite temperature md simulations and irmpd spectroscopy. *Journal of chemical theory and computation*, 5(4): 1068–1078, 2009.

[47] Amel SEDIKI, Lavina C SNOEK et Marie-Pierre GAIGEOT : N–h+ vibrational anharmonicities directly revealed from dft-based molecular dynamics simulations on the ala 7 h+ protonated peptide. *International Journal of Mass Spectrometry*, 308(2):281–288, 2011.

[48] Riccardo SPEZIA, Jonathan MARTENS, Jos OOMENS et Kihyung SONG : Collision-induced dissociation pathways of protonated gly 2 nh 2 and gly 3 nh 2 in the short time-scale limit by chemical dynamics and ion spectroscopy. *International Journal of Mass Spectrometry*, 388:40–52, 2015.

[49] Vincent BRITES, Alvaro CIMAS, Riccardo SPEZIA, Nicolas SIEFFERT, James M LISY et Marie-Pierre GAIGEOT : Stalking higher energy conformers on the potential energy surface of charged species. *Journal of Chemical Theory and Computation*, 11(3):871–883, 2015.

[50] Joseph Douglas HORTON : A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing*, 16(2):358–366, 1987.

[51] J ANDVANDERVORST MEIJERINK et Henk A van der VORST : An iterative solution method for linear systems of which the coefficient matrix is a symmetric m-matrix. *Mathematics of computation*, 31(137):148–162, 1977.

[52] S SKIENA : Dijkstra's algorithm. *Implementing Discrete Mathematics : Combinatorics and Graph Theory with Mathematica, Reading, MA : Addison-Wesley*, pages 225–227, 1990.