Structure et dynamique de protéines isolées : approches statistiques

Pierre Poulain

N° d'ordre : 88–2006 Année 2006

THÈSE

présentée

devant l'UNIVERSITÉ CLAUDE BERNARD - LYON I

pour l'obtention

du DIPLÔME de DOCTORAT

(arrêté du 25 avril 2002)

présentée et soutenue publiquement le

3 juillet 2006

par

Pierre POULAIN

Structure et dynamique de protéines isolées : approches statistiques

Directeur de thèse : Philippe DUGOURD

Jury : M. Jean-Louis BARRAT, président
M. Florent CALVO, examinateur
M. Philippe DUGOURD, directeur de thèse
M. Christophe JOUVET, rapporteur
M. Richard LAVERY, rapporteur
M. Yves-Henri SANEJOUAND, examinateur

UNIVERSITÉ CLAUDE BERNARD - LYON I

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J.F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur R. GARRONE
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur G. ANNAT
Secrétaire Général	M. J.P. BONHOTAL

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Odontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur L. COLLET
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : M. le Professeur J.P. SCHARFF
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M.C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J.C. AUGROS

Remerciements

> *On n'en prend jamais. C'est trop monstrueux, presque fade à force d'opulence sucreuse. Mais voilà. On a trop fait ces derniers temps dans le camaïeu raffiné, l'amertume ton sur ton. On a poussé jusqu'à l'île flottante le léger vaporeux, l'insaisissable, et jusqu'à la coupelle aux quatre fruits rouges la luxuriance estivale mesurée. Alors, pour une fois, on ne saute pas sur le menu la ligne réservée au banana-split.*
>
> Philippe Delerm, *La première gorgée de bière.*

En guise de remerciements, je vous propose la recette de ma thèse. Elle n'a rien d'universel mais reflète pourtant bien les rencontres réalisées au cours de ces trois années au laboratoire.

Dans un premier temps, le choix du sujet est essentiel. Prenez-le plutôt large, de préférence inédit, voire même légèrement ambitieux. Le directeur de thèse, généralement livré avec le sujet, doit être compétent, attentif et surtout patient. Un chef de type Philippe Dugourd convient tout à fait. Assurez-vous enfin que le patron du restaurant, Christian Bordas, soit prêt à vous accueillir dans son établissement.

Commencez à mélanger les ingrédients, en mettant en premier la disponibilité et les compétences des divers membres de l'équipe (Philippe, Rodolphe, Michel, Driss, Isabelle, Mohamad, Francis, Anne, Cosmin, Thibault et Laure). S'il reste encore trop de grumeaux, n'hésitez pas à ajouter quelques conseils extérieurs, notamment ceux d'Allouche. Laissez reposer au moins deux jours par semaine, voire même quelques soirées de plus si nécessaire, en compagnie d'amis comme les Damoiseau, les lyonnais, les IFPiens, les souris et les angevins. Une fois la pâte levée, l'ajout d'épices telles que Florent relèvera particulièrement la saveur de la préparation.

Versez alors l'appareil dans un moule de type 150–200 pages, interligne 1,5 et cuisez immédiatement à four très chaud. Patientez 4 mois. Pour éviter que la thèse ne brûle, détendez-vous avec quelques collègues : Francisco, Sad, Xavier, Aurélie, Marc B., Estelle, Isabelle et Franck. Les discussions et les pauses avec les autres marmitons sont également fortement appréciées : Mohamad, Rola, Pierre B., Francesca, p'tit Bruno, Gaëlle, Antoine, Gulabi...

Une fois la thèse cuite, démoulez aussitôt et laissez refroidir auprès des rapporteurs, Richard Lavery et Christophe Jouvet. Patientez à nouveau un bon mois et profitez-en pour refaire du sport avec votre coach (Marc B.) et vos collègues (Francisco).

Le jour de la soutenance, présentez enfin votre travail de thèse devant un jury ne craignant pas l'indigestion (Jean-Louis Barrat, Richard Lavery, Christophe Jouvet, Florent Calvo, Yves-Henri Sanejouand et Philippe Dugourd) ainsi que devant un public de gourmands. N'oubliez pas de remercier tous vos collègues pour leur disponibilité et leur gentillesse, votre famille pour son affection et vos amis pour les bulles d'oxygène.

Enfin, souvenez-vous que, comme l'existence, une thèse a besoin de beaucoup d'amour pour être savoureuse. Celui de Bénédicte a été essentiel au quotidien, celui de l'apprenti cuisinier que j'ai été, nécessaire pour ce travail.

Bon appétit !

Table des matières

Introduction

Ordinateur [...], si tu n'ouvres pas à l'instant la porte de ce sas, je m'en vais illico voir ta mémoire centrale et te la reprogrammer avec une grosse hache, vu ?

Douglas Adams, *Le guide galactique.*

Le terme protéine vient du grec *proteios* qui signifie « premier en importance » [1]. Ce terme reflète l'omniprésence de ces molécules dans les êtres vivants. Que ce soient les bactéries, les plantes ou les animaux, tous sont essentiellement constitués d'eau et de protéines. Ces dernières interviennent à tous les stades du fonctionnement de notre organisme : l'hémoglobine transporte l'oxygène, l'insuline régule le taux de sucre, les anticorps combattent les infections, la myosine permet à nos muscles de se contracter et le collagène constitue nos tendons et nos ligaments [2].

Les protéines sont produites dans des usines biochimiques appelées ribosomes. À partir de l'information génétique codée dans l'ARN, ces usines construisent une chaîne linéaire d'acides aminés qui adopte alors une structure tridimensionnelle. Les protéines sont en fait des biopolymères naturels avec un squelette très simple, mais la diversité physicochimique de la chaîne latérale des 20 acides aminés naturels permet aux protéines d'adopter une large variété de structures. Cette forme est très importante puisqu'elle porte la fonction, donc l'activité de la protéine ainsi façonnée.

Pauling, Corey et Branson ont été parmi les premiers à lever le voile sur la structure des protéines [3, 4, 5]. En 1951, ils ont ainsi révélé l'existence de deux motifs structuraux, les hélices α et les feuillets β, et ont par la même occasion démontré l'importance de la liaison hydrogène dans la stabilisation de ces motifs. Une liaison hydrogène est une liaison électrostatique entre d'une part un atome d'hydrogène lié à un atome très électronégatif de type azote ou oxygène et, d'autre part, un atome possédant un doublet non liant comme le soufre, l'azote ou l'oxygène. Ces liaisons hydrogènes sont récurrentes dans les phénomènes biologiques. Plus généralement, elles sont directement liées à la vie sur Terre puisque sans elles l'eau liquide n'existerait pas à température ambiante. La découverte fondamentale du groupe de Pauling a été suivie, 3 ans plus tard, du travail d'Anfinsen [6] qui montra que toute l'information nécessaire pour qu'une protéine se replie dans sa structure fonctionnelle est *a priori* contenue dans la séquence. Cette hypothèse forte n'empêche pas pour autant

que certaines protéines subissent une étape de maturation post-traductionnelle et que le repliement puisse être assisté par une autre protéine appelée chaperonne.

Tout est donc mis en œuvre pour qu'une protéine puisse adopter une conformation bien définie, dite native, qui lui permettra d'assurer sa fonction dans l'organisme. Mais, l'apparition de pathologies comme la maladie d'Alzheimer et de Creutzfeldt-Jacob chez l'homme, ou de l'encéphalopathie spongiforme chez la vache montre que tout n'est pas aussi simple. En effet, pour toutes ces maladies souvent mortelles, une protéine saine et soluble se déplie en une forme pathogène et insoluble qui s'accumule en agrégats appelés fibres amyloïdes [7]. Bien évidemment, les scientifiques n'ont pas attendu l'apparition de ces maladies pour s'intéresser à la structure des protéines et à leurs mécanismes de repliement. Depuis Pauling, des efforts toujours croissants, tant expérimentaux que théoriques, sont réalisés pour déterminer ces structures. La demande reste pourtant énorme puisque sur les 100 000 protéines identifiées dans le génome humain [8], seule la structure d'un dixième d'entre elles a pu être résolue. Ainsi, et depuis maintenant 30 ans, les expérimentateurs tentent de déterminer ces géométries par des mesures de diffraction de rayons X sur une protéine cristallisée puis, plus récemment, pour des protéines en solution par des techniques de résonance magnétique nucléaire (RMN). Malgré ces efforts, la grande majorité des structures restent encore inconnues. Parallèlement à l'utilisation de ces techniques avancées, des expériences sont également développées en phase gazeuse afin de déterminer la structure de protéines ou de petits peptides. L'objectif est de mieux comprendre les propriétés intrinsèques impliquées dans les mécanismes de repliement des protéines [9, 10, 11]. On peut notamment citer la fluorescence, la spectroscopie infrarouge, les mesures de mobilité ionique, les mesures de dipôle électrique [12, 13], ou bien encore la spectrométrie de masse associée par exemple à des échanges hydrogène–deutérium.

Les théoriciens participent également à cet effort en proposant des modèles et des méthodes de simulation qui permettent d'accéder aux propriétés et aux comportements des protéines. Une des premières difficultés rencontrée dans la modélisation de protéines a été de comprendre comment, parmi la multitude de conformations possibles, une protéine pouvait adopter son unique forme native dans un temps raisonnable. Une représentation désormais admise repose sur la notion de surface d'énergie potentielle. Cette dernière possède de nombreux minima d'énergie dont le minimum global correspond à la structure à température nulle. La surface d'énergie potentielle est d'autant plus rugueuse que le nombre de degrés de liberté du système est important. Dans les cas les plus simples, ce paysage énergétique prend une forme d'entonnoir dont le fond contient la structure fonctionnelle de la protéine [14]. Pour des surfaces plus complexes, plusieurs structures d'équilibre ou entonnoirs peuvent rentrer en compétition par le jeu de l'entropie. À température finie, la structure native est celle qui minimise l'énergie libre mais qui ne minimise plus nécessairement l'énergie potentielle. L'objectif des simulations sera alors d'explorer ce paysage énergétique sans rester bloqué dans un minimum local. La méthode la plus

intuitive pour simuler des systèmes biologiques est la dynamique moléculaire qui tente de reproduire le comportement d'une protéine tel qu'il s'opére naturellement au cours du temps. Depuis la simulation de l'inhibiteur de la trypsine pancréatique bovine dans le vide (885 atomes) pendant 0,01 ns en 1977, jusqu'à celle d'un canal d'aquaporine dans une membrane lipidique (106 189 atomes) pendant 5 ns en 2002 [1], les progrès réalisés sont considérables, tant du point de vue de la description utilisée, que de l'algorithme de simulation ou bien des moyens de calculs disponibles. Cependant, une trajectoire de dynamique moléculaire de protéine ne peut atteindre au mieux que quelques nanosecondes, ce qui reste encore très insuffisant pour décrire des phénomènes de repliement qui sont de l'ordre de la milliseconde ou de la seconde. Dans une approche purement statistique, pour laquelle on s'intéresse préférentiellement aux propriétés thermodynamiques, on emploiera les méthodes Monte Carlo [15]. Celles-ci reposent sur un échantillonnage stochastique du paysage énergétique. La complexité des protéines nécessite l'emploi d'algorithmes évolués, tels que ceux initialement développés pour des systèmes magnétiques, des agrégats ou des verres [16]. Enfin, la modélisation par homologie tente également de déterminer *in silico* la structure d'une protéine. Cette dernière méthode est basée sur l'alignement des séquences d'acides aminés d'une protéine de structure connue et d'une autre dont on souhaite connaître la géométrie.

Les performances des méthodes théoriques sont régulièrement évaluées par le concours CASP (*Critical Assessment of techniques for protein Structure Prediction*) [17] qui permet de mesurer les progrès réalisés dans chaque domaine de simulation. Même si la prédiction par homologie demeure la plus performante, les méthodes *ab initio* basées sur la seule connaissance de la séquence peptidique s'améliorent graduellement [18].

Cette thèse est une étude théorique des propriétés thermodynamiques de polypeptides en phase gazeuse avec comme objectif une meilleure compréhension des mécanismes fondamentaux impliqués dans le repliement des protéines. Ce travail a été réalisé en forte interaction avec les avancées expérimentales de l'équipe « dipôle électrique, biomolécules et agrégats » du laboratoire de spectrométrie ionique et moléculaire (LASIM), à l'Université Claude Bernard Lyon I.

Dans un premier chapitre, nous allons décrire les sujets de notre étude, à savoir les protéines, ou plus modestement, des petits polypeptides modèles. Nous verrons quelle représentation a été utilisée pour décrire ces systèmes. L'approche par champ de force sera particulièrement expliquée. Enfin, nous discuterons de la complexité des systèmes étudiés et des difficultés inhérentes à la simulation.

Nous nous sommes intéressés essentiellement aux propriétés thermodynamiques des protéines, pour lesquelles nous avons employé une approche statistique basée sur la méthode Monte Carlo, décrite dans le chapitre 2. Cette méthode permet d'échantillonner le paysage énergétique du système étudié mais devient rapidement inefficace à mesure

que la complexité du système augmente. Le recours aux ensembles généralisés, dont nous évoquerons deux classes différentes, permet alors de dépasser ces difficultés. L'algorithme Monte Carlo d'échange a été très largement utilisé dans le présent travail. La méthode Wang-Landau est plus récente et nous proposerons une adaptation pour les systèmes à degrés de liberté continus.

Le chapitre 3 a pour vocation d'illustrer les performances relatives de ces deux méthodes. Nous discuterons en premier lieu du comportement d'un peptide dans un champ électrique intense et inhomogène. Cette étude répondra à une problématique d'abord expérimentale mais aussi plus générale, puisque la réponse d'une protéine à un champ électrique intervient dans de nombreux phénomènes biologiques comme la reconnaissance moléculaire ou l'amarrage d'une protéine avec d'autres molécules. Dans un deuxième temps, nous comparerons notre adaptation de la méthode Wang-Landau à l'algorithme original et au Monte Carlo d'échange.

Le chapitre 4 concernera la simulation de polymères d'alanines. La présence de deux transitions de phase puis leur caractérisation montrera que l'hélice α est la structure d'énergie la plus basse et que le feuillet β est stabilisé entropiquement, comme observé expérimentalement. L'influence d'un champ électrique sur ces transitions de phase sera également quantifiée.

Enfin, nous discuterons dans le dernier chapitre d'une méthode alternative d'analyse basée sur la fragmentation induite par laser de biomolécules en phase gazeuse. Cette technique consiste à exciter localement un peptide puis à analyser par spectrométrie de masse les peptides résultant de la fragmentation. Mon travail a consisté à simuler les systèmes étudiés dans leur état fondamental pour établir un éventuel lien entre leur conformation et la dissociation observée.

Nous terminerons ce manuscrit par une conclusion et les perspectives qui pourraient naturellement prolonger ce travail.

Bibliographie

[1] T. Schlick. *Molecular Modeling and Simulation: an interdisciplinary guide*. Springer-Verlag, New-York, USA, 2002.

[2] W. A. B. Thomasson. Unraveling the Mystery of Protein Folding, 2003. Breakthroughts in Bioscience, Office of Public Affairs from the Federation of American Societies for Experimental Biology, `http://opa.faseb.org/pdf/protfold.pdf`.

[3] L. Pauling, R. B. Corey, and H. R. Branson. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37:205–211, 1951.

[4] L. Pauling and R. B. Corey. Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37:235–240, 1951.

[5] L. Pauling and R. B. Corey. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37:251–256, 1951.

[6] C. B. Anfinsen, R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *The Journal of Biological Chemistry*, 207:201–210, 1954.

[7] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[8] C. M. Dobson and M. Karplus. The fundamentals of protein folding: bringing together theory and experiment. *Current Opinion in Structural Biology*, 9:92–101, 1999.

[9] C. S. Hoaglund-Hyzer, A. E. Counterman, and D. E. Clemmer. Anhydrous Protein Ions. *Chemical Reviews*, 99:3037–3079, 1999.

[10] M. F. Jarrold. Peptides and proteins in the vapor phase. *Annual Review of Physical Chemistry*, 51:179–207, 2000.

[11] E. G. Robertson and J. P. Simons. Getting into shape: Conformational and supramolecular landscapes in small biomolecules and their hydrated clusters. *Physical Chemistry Chemical Physics*, 3:1–18, 2001.

[12] I. Compagnon. *Mesure de dipôle électrique en phase gazeuse : application aux agrégats et aux biomolécules*. PhD thesis, Université Claude Bernard - Lyon I, Lyon, 2003.

[13] M. Abd El Rahim. *Déflexion électrique d'un jet moléculaire : progrès expérimentaux et théoriques*. PhD thesis, Université Claude Bernard - Lyon I, Lyon, 2005.

[14] K. A. Dill. Polymer principles and protein folding. *Protein Science*, 8:1166–1180, 1999.

[15] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, California, USA, 2^{nd} edition, 2002.

[16] D. J. Wales and H. A. Sheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285:1368–1372, 1999.

[17] `http://predictioncenter.org/`.

[18] C. Gibas and P. Jambeck. *Introduction à la bioinformatique*. O'Reilly, Paris, 1^{st} edition, 2002.

Chapitre 1

Protéines et leur modélisation par des champs de force moléculaires

Sommaire

Dans ce chapitre, nous décrivons les systèmes étudiés lors de ce travail de thèse, les peptides, ainsi que leurs propriétés géométriques fondamentales. Nous détaillons ensuite un certain nombre d'observables structurales et électriques, paramètres directement reliés aux expériences réalisées au LASIM. Nous nous penchons également sur l'utilisation des

champs de force qui permettent une description empirique, rapide et robuste de ces systèmes. Enfin, après avoir présenté la problématique liée à la modélisation de biomolécules, nous décrivons rapidement la méthode de dynamique moléculaire.

1.1 Protéines

Les protéines font partie des molécules organiques les plus abondantes chez les êtres vivants et sont essentielles à leur fonctionnement. Elles constituent plus de 50 % du poids sec d'un être vivant. Les protéines ont été découvertes en 1839 par Mulder [1] qui les nomma d'après le grec *proteios* qui signifie premier en importance.

L'intérêt pour les protéines est lié au grand nombre de tâches qu'elles réalisent. Qu'elles soient fonctionnelles comme les enzymes, régulatrices comme l'insuline ou bien encore structurelles comme le collagène, les protéines sont omniprésentes dans l'organisme. Elles se révèlent être des sujets d'étude passionnants au confluent de plusieurs disciplines.

1.1.1 De l'ADN aux protéines

Les protéines sont des chaînes d'acides aminés dont l'assemblage est gouverné par le code génétique. L'acide désoxyribonucléique (ADN) est une molécule, en forme de double hélice, située dans le noyau des cellules procaryotes et qui contient l'information génétique, autrement dit l'ensemble des caractères s'exprimant dans un organisme. Cette information est codée avec 4 bases différentes — adénine (A), cytosine (C), guanine (G) et thymine (T) — regroupées en triplets appelés codons. Le code génétique traduit l'information génétique en protéine via l'acide ribonucléique (ARN).

L'expression d'un gène est ainsi constituée de deux étapes. D'une part, la transcription qui est la copie d'une partie de l'information de l'ADN en ARN, dans lequel la base thymine est remplacée par l'uracile (U). D'autre part, la traduction de l'information génétique qui intervient dans les ribosomes où l'enchaînement de codons de l'ARN est converti en acides aminés.

Parmi les 64 (4^3) codons possibles, seuls 61 codent pour les 20 acides aminés naturels, les trois codons restants étant des codons de fin de traduction. Plusieurs codons différents représentent donc un même acide aminé, on parle de dégénérescence du code génétique.

1.1.2 Polymères d'acides aminés

Synthétisée dans les ribosomes après traduction de l'ARN, une protéine est un assemblage linéaire d'acides aminés. Plus précisément, on parle de polypeptides pour des molécules ayant environ entre 2 et 50 acides aminés. Une protéine peut contenir de 50 à plus de 1000 acides aminés.

Les acides aminés constituent les briques de base des protéines. On compte 20 acides

aminés naturels différents. La grande majorité a une même structure de base (figure 1.1) comprenant une fonction amine (NH_2), une fonction acide carboxylique (COOH), une chaîne latérale (notée R sur la figure 1.1) et un atome d'hydrogène articulés autour d'un atome de carbone asymétrique (C_α) dans sa conformation L. La nature du groupement latéral R différencie les acides aminés (figure 1.2) et est responsable des propriétés associées (acidité, basicité, aromaticité...).

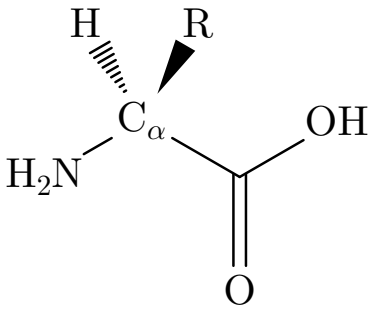


FIG. 1.1 – *Représentation générique des acides aminés (sauf la proline, de forme cyclique).*

La condensation entre deux acides aminés d'une fonction amine et d'une fonction acide carboxylique forme un groupe amide, qui constitue ainsi la liaison peptidique. La structure de cette liaison a été déterminée par Pauling, Corey et Branson par des mesures de diffraction aux rayons X [2]. La résonance entre la double liaison C=O et la liaison simple C-N entraîne que la liaison peptidique est double à 50 % et plane (figure 1.3). De proche en proche, on peut créer un polymère d'acides aminés, autrement dit un polypeptide.

Suite à la condensation, un acide aminé dans un peptide est privé d'un atome d'hydrogène sur sa fonction amine et/ou d'un groupe alcool sur sa fonction acide, on parle alors de résidu (figure 1.4). Le résidu dont le groupe amine n'est pas engagé dans une liaison peptidique est dit « N-terminal » (ou « N-ter »). L'autre extrémité est dite « C-terminale » (ou « C-ter »). Par convention, on écrit la structure d'une protéine de gauche à droite en commençant par le N-terminal (figure 1.4).

Dans les conditions physiologiques (milieu aqueux et pH $\sim$ 7), un peptide adopte spontanément une forme zwitterionique pour laquelle les extrémités N-ter et C-ter sont ionisées, respectivement en $-NH_3^+$ et $-COO^-$. En phase gazeuse, c'est par contre la forme neutre du peptide qui est préférée [3].

1.1.3 Structures des protéines

On distingue quatre niveaux structuraux pour une protéine, respectivement appelés structures primaire, secondaire, tertiaire et quaternaire. Dans la suite, nous illustrerons ces différents niveaux avec la protéine triose phosphate isomérase de la levure (*Saccharomyces cerevisiae*) [4]. La structure de cette protéine est extraite de la *protein data bank* (PDB) [5] où elle porte le code 2YPI.

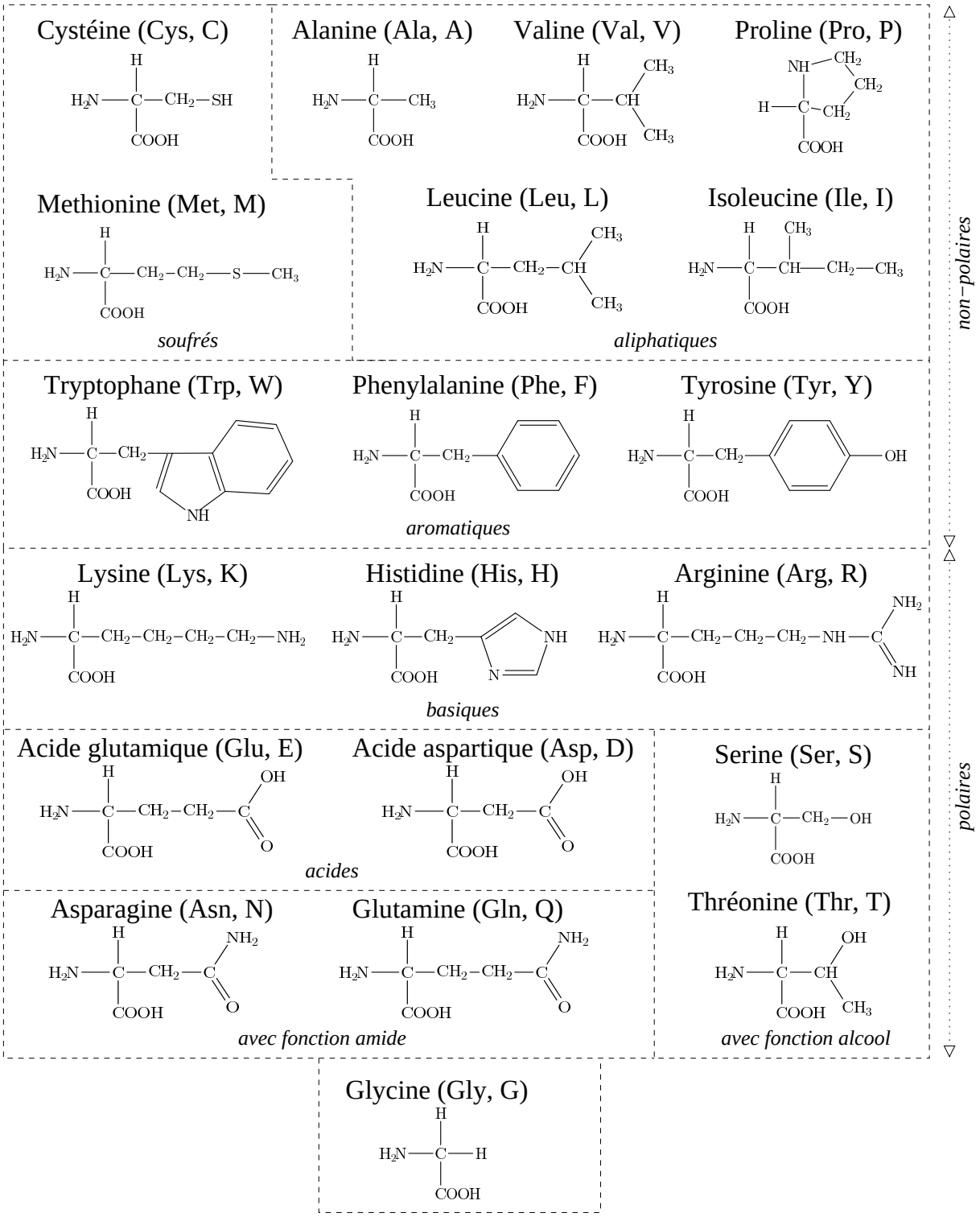


FIG. 1.2 – *Les 20 acides aminés naturels avec pour chacun l'abréviation biochimique en 3 lettres, le code et la formule semi-developpée correspondante.*

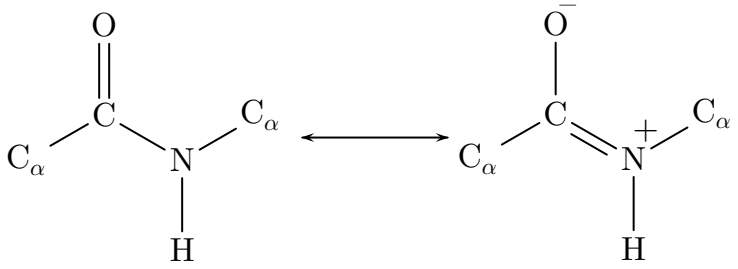


FIG. 1.3 – *Liaison peptidique avec sa forme mésomère.*

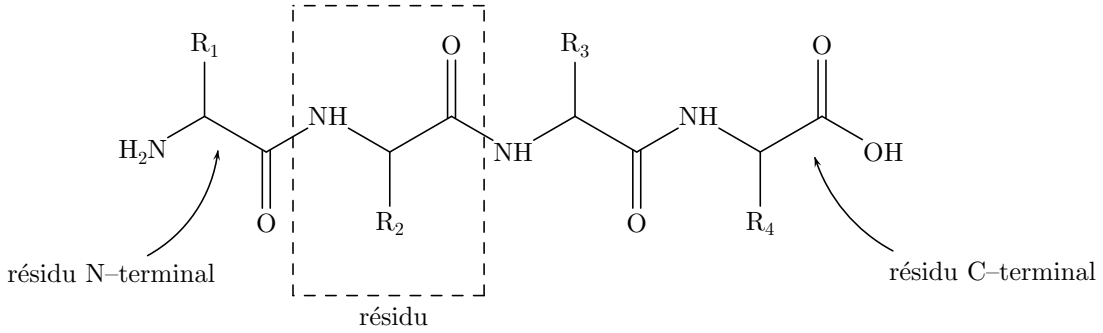


FIG. 1.4 – *Résidus d'un tétrapeptide avec les extrémités N-terminale et C-terminale.*

1.1.3.1 Structure primaire

La structure primaire est l'enchaînement linéaire des acides aminés dans un peptide. Cette structure est aussi appelée séquence. La séquence de la chaîne A de la protéine 2YPI est représentée figure 1.5.

```
ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSVD
QIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEEVKDWTNVV
VAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPEFV
DIINSRN
```

FIG. 1.5 – *Séquence de la chaîne A de la protéine 2YPI.*

1.1.3.2 Structure secondaire

La structure secondaire résulte d'un repliement local de la protéine créé par des interactions stériques et électrostatiques et stabilisée par des liaisons hydrogène. Cette structure a été découverte en 1951 par Pauling, Corey et Branson [2, 6, 7] qui ont déterminé plusieurs motifs structuraux caractéristiques.

On distingue deux angles de torsion principaux dans un résidu (figure 1.6) :

– Φ, angle C-N-C_α-C ;

– Ψ, angle N-C_α-C-N.

La liaison peptidique étant plane, l'angle Ω (C_α-C-N-C_α) est bloqué sur une valeur proche de 180°, ce qui maintient la liaison peptidique dans la conformation trans.

L'angle χ définit la rotation de la chaîne latérale du résidu.

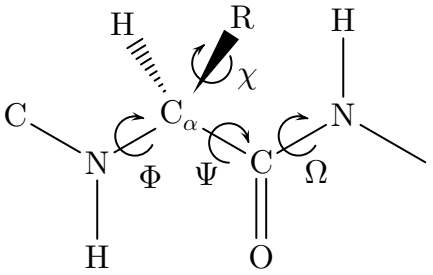


FIG. 1.6 – *Angles de torsion dans un peptide.*

Ainsi, les liaisons N-C_α et C_α-C effectuent librement des mouvements de rotation, mais seules quelques conformations parviennent à minimiser la gêne stérique. On peut répartir ces structures en catégories représentées dans le tableau 1.1.

TAB. 1.1 – *Paramètres angulaires des différentes structures secondaires [8]. La 4^e colonne représente le nombre de résidus nécessaires pour faire un tour, autrement dit l'inclinaison d'un résidu. La dernière colonne permet de connaître la longueur du peptide suivant la structure secondaire considérée et le nombre de résidus.*

Conformation	Φ [°]	Ψ [°]	Résidus/tour	Translation/résidu [Å]
Hélice α droite	-57	-47	3,6	1,50
Hélice α gauche	+57	+47	3,6	1,50
Hélice 3_{10} droite	-49	-26	3,0	2,50
Hélice π droite	-57	-70	4,4	1,15
Hélice gauche du collagène	-51	+153	3,0	3,13
Feuillet plissé β antiparallèle	-139	+135	2,0	3,40
Feuillet plissé β parallèle	-119	+113	2,0	3,20
Chaîne étirée	±180	±180		

Diagramme de Ramachandran

Connaissant les paramètres angulaires (Φ, Ψ), il est possible de dessiner une carte à deux dimensions, appelée diagramme de Ramachandran [9, 10], représentant les couples (Φ, Ψ) préférentiellement adoptés pour chaque résidu dans une protéine (figure 1.7).

Pour une protéine donnée, chaque couple d'angles (Φ, Ψ) tombe dans la zone autorisée [11] qui permet de minimiser la gène stérique. Les seules exceptions sont les résidus proline (acide aminé cyclique) et glycine (acide amine très flexible dont la chaîne latérale se résume à un atome d'hydrogène).

Comme le laisse apparaître la carte de Ramachandran, les deux structures secondaires les plus importantes sont les hélices α (droites) et les feuillets β.

Hélice α

Elle fut la première structure secondaire découverte par Pauling, Corey et Branson [2, 6].

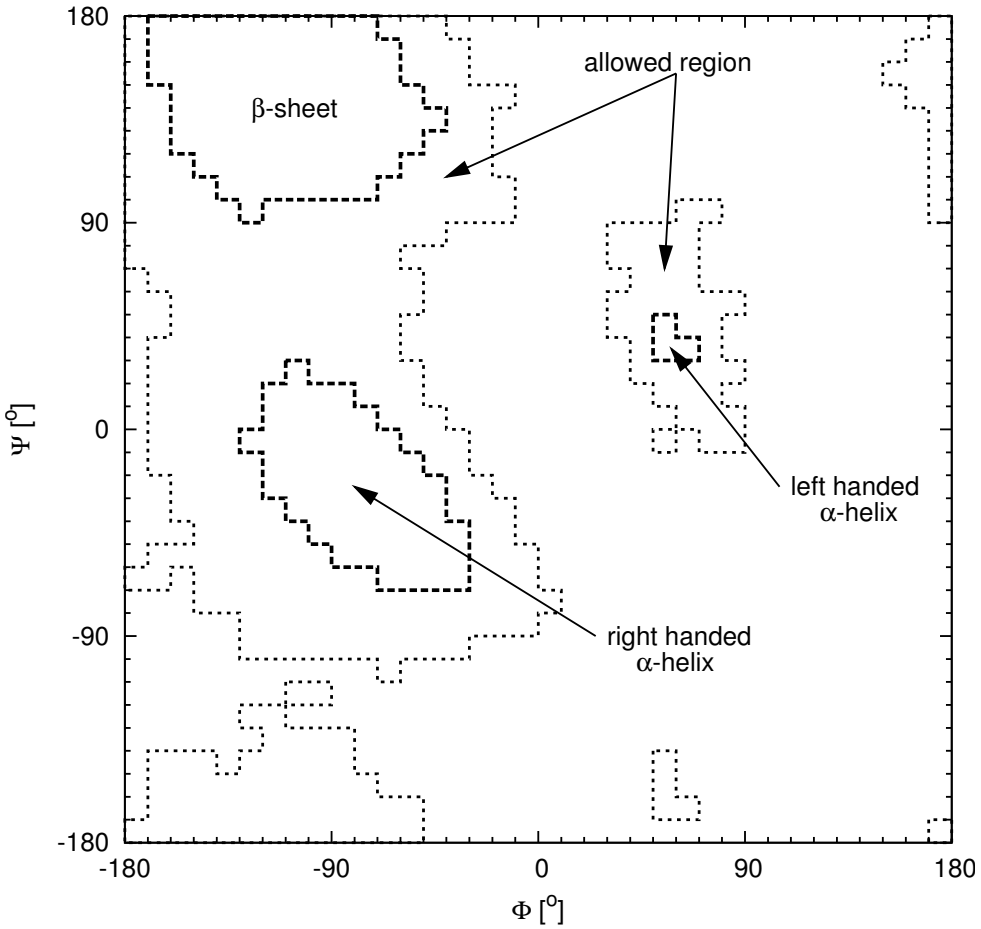


FIG. 1.7 – *Carte de Ramachandran avec les différentes structures secondaires associées.*

À peu près 30 % des résidus des protéines sont dans cette conformation, qui comme son nom l'indique, est hélicoïdale.

Pour chaque résidu, l'angle Φ vaut approximativement -57° et l'angle Ψ -47°. On compte 3,6 résidus dans un tour d'hélice pour une longueur de 5,41 Å [12].

La stabilité de cette structure est due à la liaison hydrogène entre l'atome d'oxygène d'un résidu i et l'hydrogène du groupe NH d'un résidu $i+4$ (figure 1.8). La longueur d'une telle liaison avoisine les 2,86 Å de l'atome d'oxygène à l'atome d'azote [12].

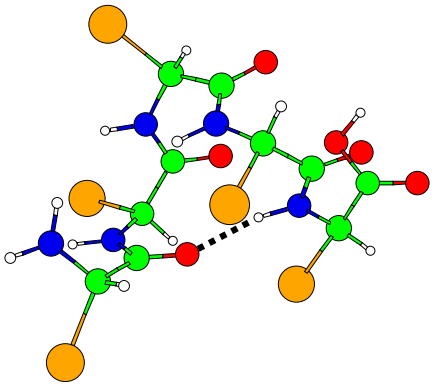


FIG. 1.8 – *Un tour d'hélice α avec la liaison hydrogène entre les résidus i et $i+4$. Les atomes de carbone sont en vert, d'hydrogène en blanc, d'azote en bleu et d'oxygène en rouge. Les chaînes latérales sont représentées en orange. La liaison hydrogène est en pointillés noirs.*

Feuillet β

Cette structure [7] est constituée de chaînes polypeptidiques très étirées (par exemple, Φ = -139° et Ψ = +135° pour le feuillet β antiparallèle). En moyenne, 20 % des résidus dans les protéines sont en feuillet β [8]. Un seul brin β n'est pas particulièrement stable, par contre plusieurs brins le sont grâce aux liaisons hydrogène existantes entre les groupes CO et NH des brins voisins (figure 1.9).

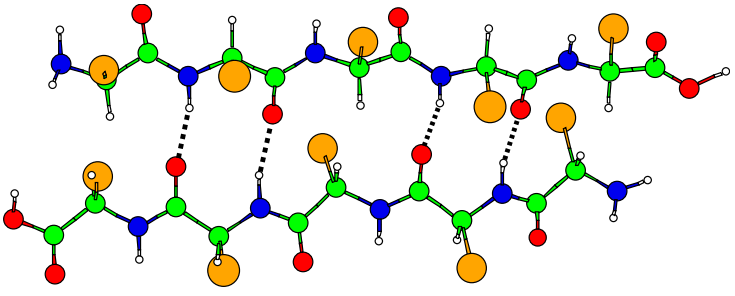
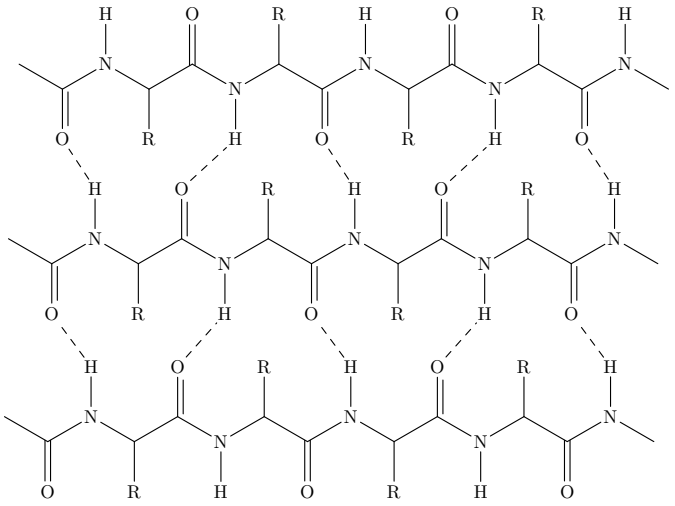


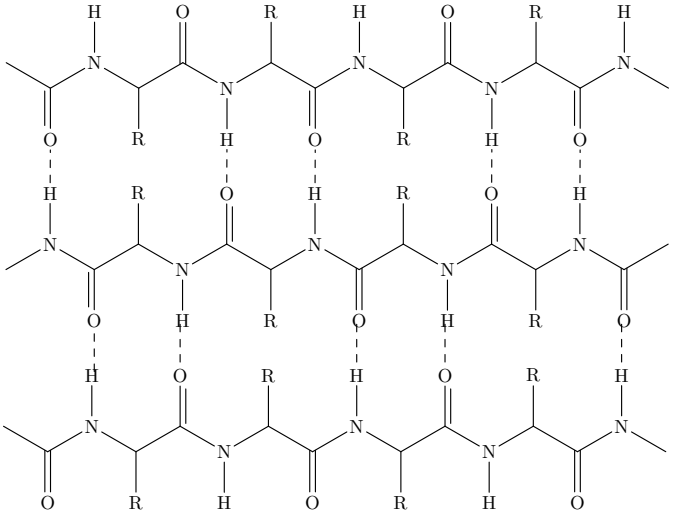
FIG. 1.9 – *Feuillets β antiparallèles. Les atomes de carbone sont en vert, d'hydrogène en blanc, d'azote en bleu et d'oxygène en rouge. Les chaînes latérales sont représentées en orange. Les liaisons hydrogènes sont en pointillés noirs.*

Les feuillets β parallèles et antiparallèles se distinguent par l'alternance des brins au sein du feuillet (figure 1.10). La connexion des différents brins se fait par quelques acides

aminés qui n'ont pas de structure secondaire particulière ; on nomme cependant ces régions tours β (β-*turns*).



(a)



(b)

FIG. 1.10 – *Deux types de feuillets β, (a) parallèles et (b) antiparallèles.*

Les résidus qui se situent entre les différents éléments de structure secondaire et qui n'ont pas eux-mêmes de structure secondaire spécifique ne jouent pas un rôle majeur dans la détermination de la structure et des propriétés d'une protéine [13]. Ces régions sont appelées tours (*turns*) ou boucles (*loops*).

1.1.3.3 Structure tertiaire

Après une succession de repliements locaux, la protéine subit un repliement global qui lui donne sa forme finale et son activité biologique. Ce repliement conduit à l'enfouissement des acides aminés hydrophobes de la protéine. La structure tertiaire d'une protéine peut

contenir plusieurs motifs structuraux secondaires comme des hélices α ou des feuillets β (figure 1.11).

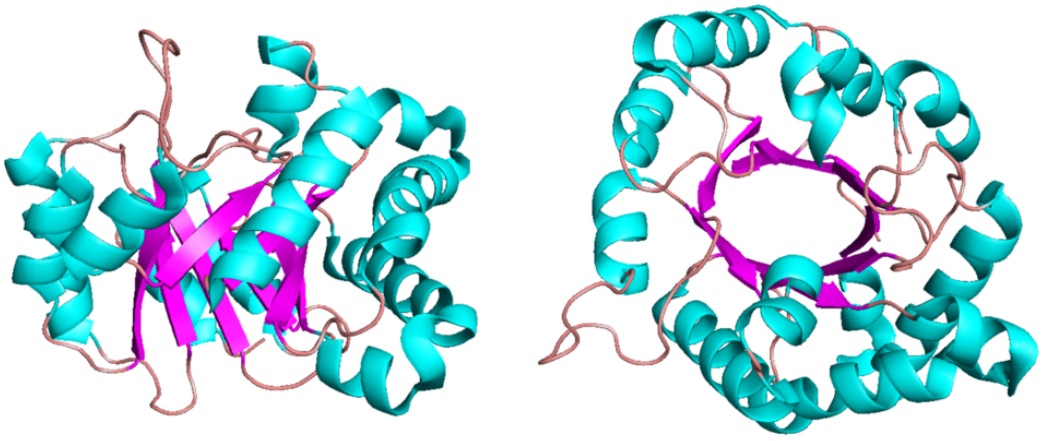


FIG. 1.11 – *Structure tertiaire de la chaîne A de la triose phosphate isomérase de la levure. Les hélices α sont représentées en hélices de couleur bleue, les feuillets β en flèches roses. La structure tertiaire de cette enzyme présente une forme particulière appelée* TIM barrel *[14].*

1.1.3.4 Structure quaternaire

La structure quaternaire concerne un petit nombre de protéines de grande taille issues de l'agrégation de plus petites molécules. En effet, la structure quaternaire est la construction d'une super protéine à partir de peptides en structures tertiaires et parfois même d'autres molécules organiques ou d'atomes métalliques. Ainsi, l'enzyme triose phosphate isomérase de la levure est constitué de deux chaînes A et B et de deux molécules d'acide 2-phosphoglycolique comme représenté sur la figure 1.12.

1.1.4 Relation séquence–structure–fonction

La fonction d'une protéine est intimement liée à sa forme tridimensionnelle. Autrement dit, comprendre le rôle d'une protéine nécessite la connaissance de sa structure.

En 1954, Anfinsen (prix Nobel en 1972) énonça que toute l'information nécessaire au repliement d'une protéine dans sa structure fonctionnelle (native) se trouve dans sa structure primaire, autrement dit dans sa séquence [15, 16]. Cette observation est depuis connue sous le nom de principe d'Anfinsen.

D'après ce principe, il est donc thermodynamiquement possible d'accéder à la structure tertiaire d'une protéine à partir de sa seule séquence. Cette affirmation est satisfaisante pour envisager la modélisation de protéines, la description correcte de la séquence d'acides aminés étant en principe suffisante pour accéder à la structure tertiaire. Cette description est détaillée dans la section 2.

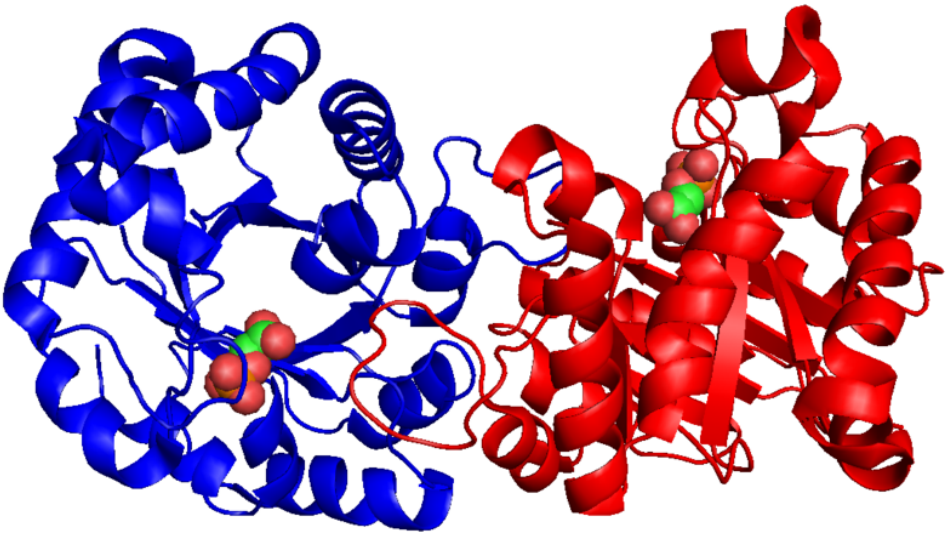


FIG. 1.12 – *Structure quaternaire de l'enzyme triose phosphate isomérase de la levure. La chaîne A est en bleu et la chaîne B en rouge. On aperçoit également deux molécules d'acide 2-phosphoglycolique (*$C_2H_5O_6P$*) [4].*

Considérons maintenant que chaque résidu puisse adopter seulement trois états (Φ, Ψ) possibles. Une petite protéine de 100 acides aminés possède donc 3^{100} soit 5×10^{47} configurations possibles. Si une protéine peut échantillonner 10^{13} structures/seconde (en sachant que la rotation d'une liaison prend 10^{-13} seconde), alors il lui faut 10^{27} années pour trouver la conformation la plus stable [17], ce qui représente plus que l'âge de l'univers ! Pourtant, en milieu biologique, une protéine se replie en des temps allant de la milliseconde à la minute [18]. D'un point de vue cinétique, le repliement d'une protéine aboutit à un paradoxe, dit de Levinthal [19]. La résolution de la structure des protéines est un problème NP complet (*non-polynomical complete*). Nous verrons comment aborder ce problème par la simulation dans la section 3.

L'information séquentielle est, en principe, suffisante pour déterminer la structure d'une protéine. On distingue alors deux stratégies pour prédire la structure tridimensionnelle d'un peptide à partir de sa structure primaire, la modélisation par homologie et la prédiction « *ab initio* » [20]. La première stratégie consiste à aligner la séquence d'une protéine (cible) dont on ne connaît pas la structure, sur la séquence d'une protéine dont la structure est connue (référence). Les algorithmes d'alignement les plus utilisés sont FASTA [21] et BLAST (*basic local alignment search tool*) [22]. Plus la séquence de la protéine cible sera similaire à la séquence de la protéine de référence, plus on pourra raisonnablement penser que les structures des deux protéines sont proches. En deçà de 25 % de similarité, il est par contre difficile d'établir une quelconque correspondance. À l'opposé, la prédiction « *ab initio* » tente de reconstruire la structure d'une protéine à

partir de la seule connaissance de sa séquence et d'un modèle adéquate. Cette deuxième stratégie sera développée dans la section 2.

Parallèlement au calcul, la détermination expérimentale de la structure géométrique d'une protéine fait partie des enjeux majeurs de l'ère post-génomique. En effet, on connaît beaucoup plus de séquences (environ 60 000 000 dans GenBank [23]) que de structures (environ 30 000 dans la PDB). La diffraction par rayons X et les techniques de résonance magnétique nucléaire (RMN) ont permis d'obtenir la structure 3D de protéines cristallisées ou en solution. En revanche, la détermination de la géométrie d'une protéine isolée reste un défi. En s'affranchissant des interactions avec le solvant, elle semble toutefois une approche prometteuse pour comprendre le lien entre séquence et structure tridimensionnelle. Les techniques spectroscopiques associées à des calculs *ab initio* permettent de déterminer la structure d'acides aminés, de dipeptides voire de tripeptides mais ne sont pas encore adaptées à des systèmes de plus grande taille. De nouvelles sondes doivent donc être développées. C'est l'un des axes majeurs de recherche de l'équipe dipôle électrique, biomolécules et agrégats (DEBA) du LASIM, avec notamment l'utilisation du dipôle électrique comme sonde conformationnelle. Cette observable, utilisée pour ce travail de thèse, est décrite ci-dessous ainsi qu'un certain nombre d'autres variables structurales utilisées occasionnellement.

1.1.5 Dipôle électrique, polarisabilité électronique et susceptibilité électrique : des sondes structurales

1.1.5.1 Dipôle électrique, définition et unités

Le moment dipolaire $\vec{\mu}$ d'une molécule caractérise sa distribution de charges. Ainsi, pour N particules chargées ponctuelles, on définit le dipôle électrique par

$$\vec{\mu} = \sum_{i=1}^{N} q_i \vec{r_i}, \tag{1.1}$$

où q_i est la charge de la particule i et $\vec{r_i}$ le vecteur distance de la particule i à l'origine.

Lorsqu'il s'agit d'une distribution de charges continue ρ dans un volume V, on a

$$\vec{\mu} = \iiint \rho(\vec{r})\vec{r}dV. \tag{1.2}$$

Dans le système international, le dipôle est défini en coulomb.mètre (C.m). Plus adapté aux grandeurs rencontrées en chimie, le debye (D) est couramment utilisé. La correspondance avec le système international est [24] :

$$1 \text{ D} = 3{,}336 \times 10^{-30} \text{ C.m}\,. \tag{1.3}$$

Deux particules de charges $+e$ et $-e$ espacées de 1 Å forment un dipôle de 4,18 D, valeur tout à fait adaptée à une utilisation de l'unité debye dans les molécules.

À titre d'exemple, la molécule d'eau possède un dipôle électrique de 1,85 D soit $6{,}0 \times 10^{-30}$ C.m [figure 1.13(a)]. La molécule de chlorométhane n'a, par contre, pas de moment dipolaire en raison de sa symétrie tétraédrique [figure 1.13(b)]. Enfin, les chaînes aliphatiques saturées sont peu polaires car les liaisons C-C sont apolaires et les C-H très peu polaires.

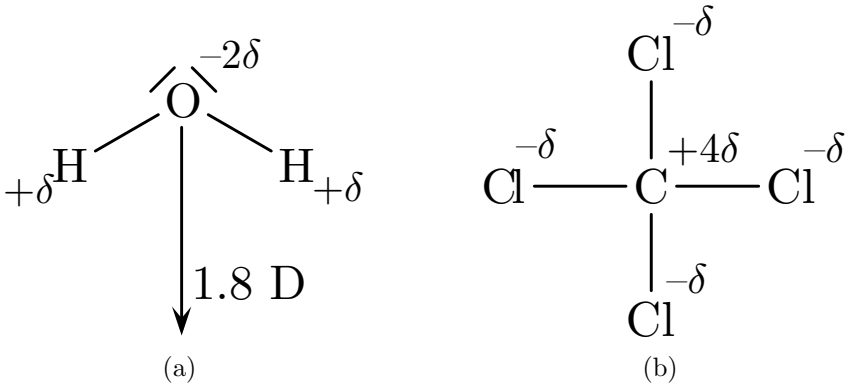


FIG. 1.13 – *Dipôle électrique de la molécule (a) d'eau et (b) de chlorométhane. Le symbole δ indique la présence d'une charge partielle due à la différence d'électronégativité entre les atomes voisins.*

1.1.5.2 Polarisabilité électronique, définition et unités

En absence d'un champ électrique, le dipôle de la molécule est appelé dipôle permanent. Lorsqu'un champ électrique extérieur $\vec{E}$ est appliqué, la distribution du nuage électronique est modifiée ce qui induit un dipôle qui se superpose au dipôle permanent,

$$\vec{\mu} = \vec{\mu_0} + \overleftrightarrow{\alpha} \cdot \vec{E} + \frac{1}{2}\beta \vdots \vec{E}^2 + \frac{1}{6}\gamma \vdots \vec{E}^3 + \dots \tag{1.4}$$

Le premier ordre du dipôle induit est le tenseur de polarisabilité, qui est une matrice 3×3 diagonalisable. Le valeur moyenne α de la polarisabilité statique est

$$\alpha = \frac{1}{3}\mathrm{Tr}(\overleftrightarrow{\alpha}) \tag{1.5}$$

Le terme d'hyperpolarisabilité β ainsi que les termes d'ordre supérieur sont importants pour des champs laser intenses mais négligeables pour les champs électriques statiques utilisés expérimentalement dans le groupe.

La polarisabilité a la dimension d'un volume et s'exprime donc en m^3 dans le système international. Usuellement, on utilise l'unité $Å^3$.

1.1.5.3 Susceptibilité électrique, définition et unités

Sous l'influence d'un champ électrique, l'orientation statistique des moments dipolaires d'une molécule entraîne une polarisation d'orientation. Ce comportement est décrit par la théorie de Langevin-Debye [25, 26]. En présence du champ électrique F, par exemple appliqué suivant l'axe z, l'interaction entre les dipôles électriques et le champ extérieur conduit à une orientation privilégiée des dipôles dans la direction du champ. L'énergie E de la molécule s'écrit

$$E = E_0 - \mu_z F, \tag{1.6}$$

où E_0 est l'énergie de la molécule sans champ électrique et μ_z la coordonnée du dipôle électrique suivant l'axe z. La valeur moyenne du dipôle est

$$\langle \mu_z \rangle = \frac{1}{Z} \sum_{\text{états}} \mu_z e^{-(E_0 - \mu_z F)/k_B T}, \tag{1.7}$$

avec Z la fonction de partition du système,

$$Z = \sum_{\text{états}} e^{-(E_0 - \mu_z F)/k_B T}, \tag{1.8}$$

où k_B est la constante de Boltzmann et T la température. Dans la limite du champ faible, pour laquelle $\mu_z F / k_B T \ll 1$, un développement limité donne la valeur moyenne du dipôle suivant z,

$$\langle \mu_z \rangle_F \simeq \frac{\langle \mu_z^2 \rangle_0}{k_B T} F \simeq \frac{\langle \mu^2 \rangle_0}{3 k_B T} F \tag{1.9}$$

Par symétrie, les moyennes $\langle \mu_x \rangle$ et $\langle \mu_y \rangle$ sont nulles. La susceptibilité électrique est alors définie comme

$$\chi = \frac{\langle \mu_z \rangle_F}{F} = \frac{\langle \mu^2 \rangle_0}{3 k_B T}, \tag{1.10}$$

à laquelle il faut rajouter la polarisabilité électronique,

$$\chi = \frac{\langle \mu^2 \rangle_0}{3 k_B T} + \alpha. \tag{1.11}$$

L'unité de susceptibilité électrique est la même que celle de polarisabilité électronique. Elle s'exprime par conséquent en Å^3. Expérimentalement, c'est cette grandeur qui est mesurée.

1.1.5.4 Dipôle électrique des protéines

Dans les protéines, la liaison peptidique est polarisée et possède un moment dipolaire de 3,5 D parallèle aux liaisons C=O et N-H et orienté de l'oxygène du carbonyle vers l'hydrogène de l'amide [27] (figure 1.14).

Le dipôle d'un peptide est la somme vectorielle des moments dipolaires élémentaires de toutes les liaisons peptidiques à laquelle il faut ajouter la contribution des chaînes latérales polaires et des charges éventuelles. Suivant la structure secondaire adoptée, le

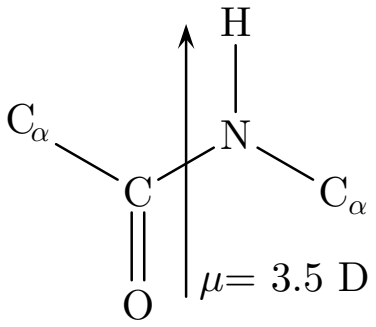


FIG. 1.14 – *Moment dipolaire de la liaison peptidique.*

dipôle total d'un peptide peut varier fortement. Dans la structure en hélice α, les dipôles peptidiques s'ajoutent pour former un macrodipôle [27, 28] aligné sur l'axe de l'hélice [figure 1.15(a)]. Pour un peptide en hélice α constitué de n acides aminés, on a ainsi un dipôle $\mu \sim 3{,}5 \times (n-1)$ D. Par contre, le dipôle d'un feuillet β est très faible, par compensation deux à deux des dipôles électriques individuels [figure 1.15(b)]. Enfin, le dipôle d'une structure non organisée (appelée pelote statistique ou *random coil*) est de l'ordre de $(3{,}5 \times (n-1))^{1/2}$ D [figure 1.15(c)].

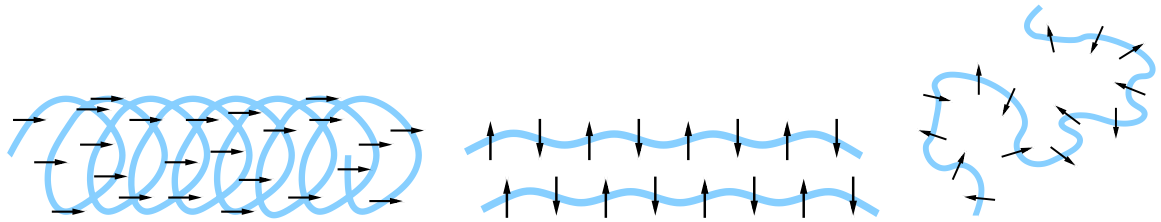


FIG. 1.15 – *Orientations des dipôles élémentaires dans les différentes structures secondaires. (a) Hélice α, (b) feuillet β et (c) pelote statistique.*

Le tableau 1.2 donne le dipôle électrique de quelques protéines riches en hélices α [29]. Le dipôle peut prendre des valeurs très élevées, dépassant plusieurs centaines de debyes.

TAB. 1.2 – *Dipôles électriques de quelques protéines riches en hélices α [29].*

Protéine	Dipôle [D]
Hémoglobine du cheval	480
Myoglobine de l'homme	170
Myoglobine du cheval	163
Albumine de l'œuf	250
Albumine du sérum de cheval	380

Le dipôle électrique d'une protéine peut servir de sonde structurale pour déterminer la structure secondaire adoptée par un peptide. Plus généralement, le moment dipolaire

permet aussi d'identifier des molécules n'ayant pas de structure secondaire particulière (petits peptides). L'équipe de P. Dugourd a montré qu'il est possible expérimentalement de mesurer le dipôle électrique de tels peptides [30, 31]. Il est également prouvé que des peptides isomères en masse peuvent être différenciés par leur dipôle électrique [32]. Le tableau 1.3 rassemble quelques résultats expérimentaux. Dans la partie supérieure, deux couples de peptides isomères en masse, (GW et WG) et (AW et WA), peuvent être distingués par leur dipôle électrique. La partie inférieure du tableau 1.3 donne l'évolution du dipôle pour les peptides WG_n avec $n = 1, 2, 3, 4, 5$. Le dipôle électrique s'avère être également discriminant pour l'identification de tels peptides.

TAB. 1.3 – *Dipôle électrique de quelques peptides extrait de la susceptibilité électrique mesurée expérimentalement [30, 32]. La polarisabilité électronique est calculée à partir d'un modèle additif.*

Molécule	Polarisabilité électronique [Å^3]	Susceptibilité expérimentale [Å^3]	Dipôle [D]
GW	28,4	457 ± 55	7,85 ± 0,42
WG	28,4	241 ± 27	4,70 ± 0,35
AW	30,2	537 ± 107	7,73 ± 0,83
WA	30,2	245 ± 80	4,97 ± 0,96
WG	28,4	214 ± 27	4,70 ± 0,35
WG_2	33,5	289 ± 37	5,50 ± 0,40
WG_3	38,6	289 ± 37	5,45 ± 0,41
WG_4	43,7	375 ± 50	6,27 ± 0,48
WG_5	48,8	391 ± 77	6,34 ± 0,72

1.1.6 Autres observables structurales des protéines

Le dipôle électrique d'une protéine sera l'observable privilégiée dans le cadre de ce travail. Cependant, d'autres grandeurs sont également pertinentes et utilisées pour décrire des systèmes biologiques.

1.1.6.1 Rayon de giration

Le rayon de giration mesure la compacité d'une protéine et est défini [33] comme

$$Rg = \sqrt{\frac{\sum_i m_i (r_i - r_{\mathrm{CM}})^2}{\sum_i m_i}}, \tag{1.12}$$

où r_{CM} est la position du centre de masse, r_i est la position de l'atome i et m_i sa masse.

Pratiquement, prendre toutes les masses égales à 1 permet de simplifier l'expression (1.12) sans affecter la valeur de Rg de manière significative :

$$\tilde{R}g = \sqrt{\frac{\sum_i (r_i - r_{\mathrm{CM}})^2}{N}}, \tag{1.13}$$

où N représente le nombre d'atomes.

Pour les peptides, cette approximation introduit une erreur inférieure à 10 %.

1.1.6.2 Distance bout-à-bout

La distance bout-à-bout permet de mesurer l'étirement d'une chaîne peptidique. Elle peut se définir comme la distance entre l'azote du N-terminal et l'oxygène du C-terminal ou entre les carbones C_α du N-terminal et du C-terminal.

1.1.6.3 Nombre de résidus en hélice ou en feuillet

Pour savoir si un peptide est dans telle ou telle conformation secondaire, il peut être intéressant de calculer le nombre de résidus en hélice α ou en feuillet β (figure 1.7). D'après Peng et Hansmann [34], un résidu est en hélice α si les angles (Φ, Ψ) sont dans l'intervalle (-70° ± 30°, -37° ± 30°). De même, on définit un résidu en feuillet β si les angles (Φ, Ψ) du squelette peptidique sont dans l'intervalle (-140° ± 40°, 140° ± 40°).

1.2 Champs de force utilisés en mécanique moléculaire

La première étape dans la détermination de la structure d'une protéine est le calcul de l'énergie potentielle associée à une conformation donnée. La seconde étape consiste à déterminer la configuration la plus stable, c'est-à-dire la plus basse en énergie. Si l'on désire réaliser des calculs à température finie, on doit en réalité considérer l'énergie libre,

$$\mathcal{F} = U - TS, \tag{1.14}$$

où $\mathcal{F}$ représente l'énergie libre, U l'énergie interne du système, T la température et S l'entropie. La prise en compte du terme entropique peut alors être accomplie avec des simulations de type Monte Carlo ou par dynamique moléculaire.

1.2.1 Modélisation d'une protéine

L'énergie d'une protéine peut être obtenue par différentes approches regroupées sur la figure 1.16. On ne prétend pas ici fournir une description exhaustive de ces méthodes mais seulement une présentation succinte. Pour plus de détail, nous renvoyons le lecteur aux références [35, 36].

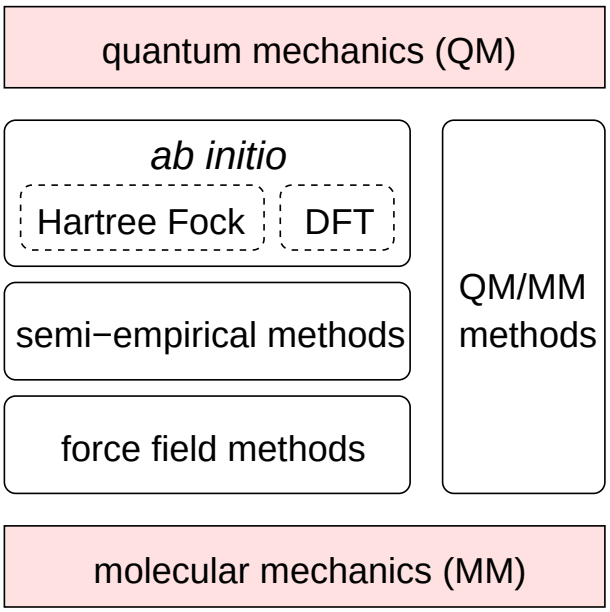


FIG. 1.16 – *Principales méthodes de calcul d'énergie utilisées en simulation moléculaire.*

Les méthodes *ab initio* tentent de résoudre l'équation de Schrödinger dans laquelle on traite les électrons actifs du système. L'objectif de telles techniques est de déterminer l'énergie d'une molécule ainsi que la fonction d'onde ou la densité des électrons. Une méthode très utilisée est la méthode Hartree-Fock, pour laquelle chaque électron perçoit les autres électrons comme un champ moyen. Mais l'omission des interactions de configuration peut avoir des conséquences importantes sur les propriétés des systèmes étudiés [36]. Ces interactions peuvent alors être ajoutées avec différents niveaux d'approximation. Les méthodes *ab initio* sont cependant limitées à des systèmes comprenant peu d'électrons donc peu d'atomes. On assiste depuis une quinzaine d'années au développement des méthodes DFT (*density functional theory*) [37, 38] pour les systèmes biologiques. La théorie de la fonctionnelle de la densité repose sur la détermination de l'état fondamental électronique par sa densité totale au lieu de la fonction d'onde.

Les méthodes semi-empiriques reposent sur les fondements de la mécanique quantique mais supposent un certain nombre d'approximations, notamment sur les calculs des termes non diagonaux dans l'hamiltonien. On peut citer les méthodes de type MNDO (*modified neglect of differential overlap*) [39], AM1 (*Austin model 1*) [40], et PM3 (*parametric model number 3*) [41], toutes les trois basées sur l'approximation NDDO (*neglect of differential diatomic overlap*) [42].

Les méthodes par champs de force ne considèrent, par contre, que les mouvements des noyaux. Le calcul de l'énergie potentielle dépend d'une forme fonctionnelle ainsi que d'un jeu de paramètres empiriques. Les électrons ne sont plus explicitement pris en compte, donc la rupture ou la création de liaisons chimiques n'est plus possible. Le calcul par champ de force est très rapide comparé aux méthodes *ab initio* ou semi-empiriques, et

permet d'aborder les propriétés statistiques ou de traiter des systèmes contenant plusieurs centaines voire plusieurs milliers d'atomes.

Enfin, les méthodes QM/MM [43, 44] sont des méthodes hybrides qui traitent une partie du système quantiquement et l'autre partie par une approche classique. Le traitement de la frontière entre les parties quantique/classique reste cependant délicat et est le sujet de nombreux études et débats.

Dans ce travail, nous avons utilisé l'approche par champs de force qui est adaptée aux systèmes de grande taille et au développement de simulations statistiques.

1.2.2 Modélisation par champs de force

L'utilisation d'un champ de force est pertinente lorsqu'on veut réaliser des simulations de mécanique ou dynamique moléculaire pour des systèmes comprenant plusieurs dizaines voire plusieurs centaines d'atomes.

Un champ de force se définit comme une forme fonctionnelle et un ensemble de paramètres attribué à un type d'atome. Les paramètres du champ de force sont tirés de simulations *ab initio* ou semi-empiriques et/ou de données expérimentales.

Les champs de force les plus couramment utilisés pour les simulations de biomolécules sont AMBER (*assisted model building and energy refinement*) [45], CHARMM (*chemistry at harvard using molecular mechanics*) [46], OPLS (*optimized potentials for liquid simulations*) [47], GROMOS 96 (*Groningen molecular simulation package*) [48] et ECEPP (*empirical conformational energy program for peptides*) [49]. Le tableau 1.4 rassemble les résultats de l'interrogation du moteur de recherche SciFinder Scholar concernant l'utilisation de ces différents champs de force pour la modélisation de protéines. La recherche a porté sur la période 1987–2006. Le champ de force AMBER semble être le plus utilisé.

Tab. 1.4 – *Nombre de publications relatives aux différents champs de force. L'interrogation s'est faite sur le moteur de recherche SciFinder Scholar le 21.01.2005. Les bases de données utilisées sont CAPLUS et MEDLINE. La requête était « X force field for protein or peptide », avec X le nom du champ de force recherché.*

Champ de force	Nombre de références trouvées
AMBER	497
CHARMM	222
OPLS	128
GROMOS	114
ECEPP	91

1.2.3 Champ de force AMBER

Développé depuis le début des années 80 par le groupe de Peter Kollmann, le champ de force AMBER a été décrit pour la première fois en 1984 [45]. Il fait partie de la suite logiciel de mécanique moléculaire du même nom [50].

Le champ de force AMBER dans sa version *ff84* comme son successeur *ff86* [51] ont initialement été mis au point à partir de données expérimentales obtenues en phase gazeuse. Le calcul des charges a été amélioré dans la version *ff94* [52] grâce à l'utilisation de la méthode RESP (*restrained electrostatic potential fit*) [53]. Depuis la version *ff96* [54], les principales améliorations portent sur les paramètres associés aux valeurs des angles Φ et Ψ. En particulier, les champs de force *ff96* et *ff99* [55] corrigent la tendance de *ff94* à favoriser les conformations en hélice α [56].

Ces dernières années, le groupe de García tente également d'améliorer le champ de force AMBER en modifiant les paramètres des angles de torsion [56].

Ce travail de thèse a été réalisé avec le champ de force AMBER dans sa version *ff96*. Ce choix est justifié par des calculs AM1 B3LYP 6-31 G* réalisés par P. Dugourd sur des peptides en conformation hélice α, feuillet β et pelote statistique. Les dipôles calculés pour chaque structure sont en très bon accord avec ceux obtenus à partir de *ff96*, qui désignera par la suite aussi bien l'ensemble des paramètres que le champ de force lui-même.

1.2.4 Types d'atomes utilisés par AMBER *ff96*

Le champ de force *ff96* comprend une forme fonctionnelle et des paramètres associés à des types d'atomes. Un type d'atome comprend :

- l'élément chimique ;
- l'hybridation de l'atome (sp^2 ou sp^3 pour le carbone) ;
- l'environnement autour de l'atome considéré (autres atomes ou groupements particuliers).

Il peut exister, par exemple, plusieurs types pour l'atome de carbone, suivant que ce dernier soit lié à tels ou tels autres atomes. Le tableau 1.5 rassemble les principaux types d'atomes utilisés dans le cadre de ce travail.

1.2.5 Forme fonctionnelle du champ de force AMBER *ff96*

La forme fonctionnelle du champ de force AMBER *ff96* est composée de plusieurs termes regroupés selon le type d'interaction :

- interactions entre atomes liés par 2 ou 3 liaisons chimiques (énergies d'élongation, de flexion, de torsion) ;
- interactions entre atomes non liés, ou séparés par plus de 3 liaisons, (interactions de van der Waals, électrostatique ou liaison hydrogène).

TAB. 1.5 – *Principaux types d'atomes* ff96 *utilisés pour ce travail.*

Type d'atome	Élément chimique et environnement
H	hydrogène dans un groupe amide
HC	hydrogène attaché à un carbone
H2	hydrogène dans un groupe amino (NH_2)
HO	hydrogène dans un groupe alcool
C	carbone sp^2 dans un groupe carbonyle
CT	carbone sp^3 avec 4 constituants (carbone α d'un résidu)
CA	carbone dans un cycle aromatique à six chaînons
CH	carbone unifié sp^3 + 1 hydrogène (groupe CH)
C2	carbone unifié sp^2 + 2 hydrogènes (groupe CH_2)
C3	carbone unifié sp^3 + 3 hydrogènes (groupe CH_3)
N	azote sp^2 dans un groupe amide
NT	azote sp^3 avec 3 substituants
O	oxygène dans un groupe carbonyle
OH	oxygène dans un groupe alcool

La forme fonctionnelle du champ de force *ff96* comprend un terme d'énergie d'élongation, un terme d'energie de flexion, un terme d'énergie de torsion, un terme pour l'énergie d'interaction de van der Waals et un terme d'interactions électrostatiques :

$$\begin{aligned} E_{\text{totale}} = & \sum_{\text{liaisons}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{flexions}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{torsions}} \sum_n \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ & + \sum_{\text{vdW}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] + \sum_{\text{Coulomb}} \frac{q_i q_j}{\epsilon r_{ij}}. \end{aligned} \tag{1.15}$$

Les différentes composantes de cette fonctionnelle sont détaillées ci-après.

1.2.5.1 Énergie d'élongation (*stretching*)

Une liaison covalente existe entre deux atomes qui partagent des électrons. La méthode usuelle pour approximer l'énergie potentielle d'une telle liaison dans des molécules organiques est d'utiliser la loi de Hooke :

$$E_{\text{liaison}} = K_r(r - r_{\text{eq}})^2. \tag{1.16}$$

Ici, r est la longueur de la liaison, r_{eq} est la longueur de la liaison à l'équilibre et K_r est la constante de raideur du ressort qui représente la force de la liaison. Le tableau 1.6 représente quelques valeurs de ces paramètres.

Cette expression de l'énergie de liaison est à la fois simple (une constante de liaison et une distance d'équilibre par type de liaison) et donc rapide à calculer. Cependant, il existe une écriture anharmonique plus précise mais plus coûteuse en temps de calcul de

TAB. 1.6 – *Paramètres r_{eq} et K_r utilisés dans l'énergie de liaison pour le champ de force* ff96. *Les types d'atomes utilisés sont donnés dans le tableau 1.5.*

Liaisons	r_{eq} [Å]	K_r [kcal/(mol.Å^2)]
C-O	1,229	570
C-C2	1,522	317
C-N	1,335	490
C2-N	1,449	337
N-H	1,010	434

ce potentiel, appelée potentiel de Morse. La forme de ce potentiel (inutilisé dans le champ de force *ff96*) est :

$$E_{\text{liaison}} = K_r \left(1 - e^{-a(r-r_{\text{eq}})}\right)^2 . \tag{1.17}$$

Ici le terme supplémentaire a représente l'asymétrie du puits. La comparaison des deux potentiels est représentée sur la figure 1.17.

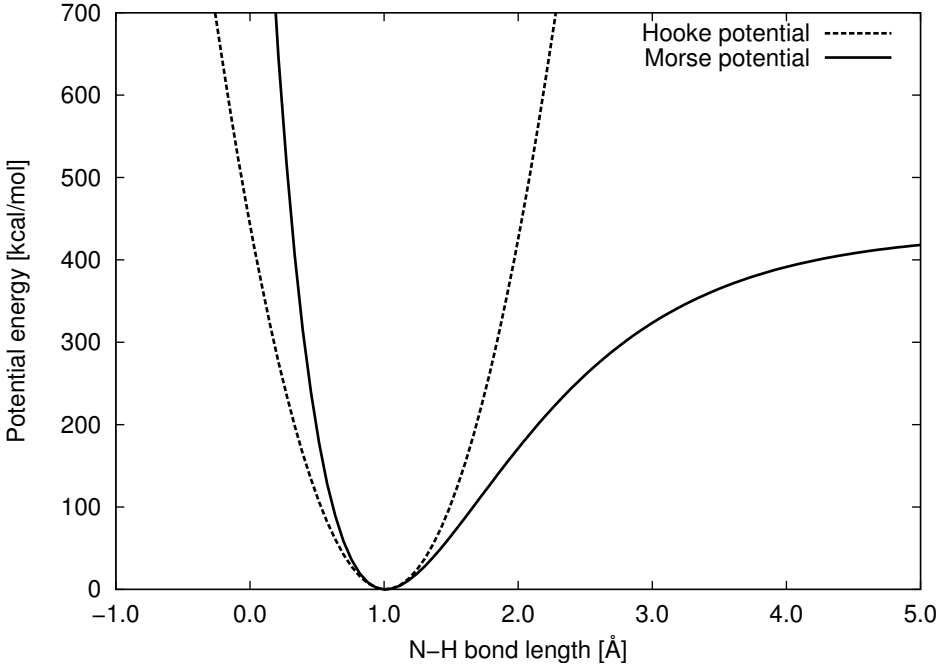


FIG. 1.17 – *Potentiels comparés de Hooke et Morse pour la liaison N-H.*

1.2.5.2 Énergie de flexion (*bending*)

On définit l'angle de flexion $\widehat{ABC}$ par l'angle θ entre les liaisons AB et BC. L'énergie de flexion est représentée par une loi harmonique, similaire à la loi de Hooke,

$$E_{\text{flexion}} = K_\theta(\theta - \theta_{\text{eq}})^2 \tag{1.18}$$

avec θ_{eq} l'angle d'équilibre et K_θ la constante de ressort (tableau 1.7).

TAB. 1.7 – *Paramètres θ_{eq} et K_θ intervenant dans l'énergie de flexion pour le champ de force* ff96. *Les types d'atomes utilisés sont précisés dans le tableau 1.5.*

Angle de flexion	θ_{eq} [degré]	K_θ [kcal/degré^2]
C-N-H	119,8	35,0
C2-N-C	121,9	50,0
C2-N-H	118,4	38,0
C-C2-N	110,3	80,0
C2-C-O	120,4	80,0
C2-C-N	116,6	70,0
O-C-N	122,9	80,0

1.2.5.3 Énergie de torsion

L'angle dièdre ϕ entre quatre atomes A-B-C-D est défini comme l'angle entre les plans ABC et BCD (figure 1.18). Un angle dièdre de 0° correspond à une conformation cis, un angle de 180° à une conformation trans.

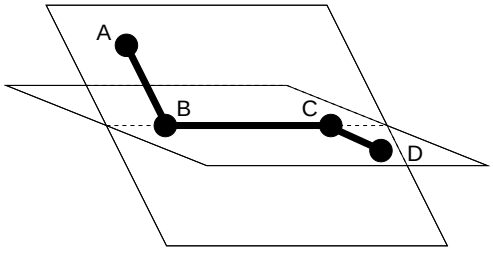


FIG. 1.18 – *Représentation de l'angle de torsion entre les points A-B-C-D.*

Le potentiel de torsion a la forme d'une série de Fourrier (potentiel de Pitzer [57]),

$$E_{\text{torsion}} = \sum_n \frac{V_n}{2}[1 + \cos(n\phi - \gamma)], \tag{1.19}$$

où V_n est la barrière d'énergie de rotation, n est le nombre de minima d'énergie dans une rotation complète et γ est le décalage angulaire (tableau 1.8).

Pour l'angle H-N-C-O, l'énergie de torsion contient deux termes dont la combinaison est présentée figure 1.19.

1.2.5.4 Énergie d'interaction de van der Waals

L'énergie d'interaction de van der Waals correspond à l'interaction entre deux atomes non liés, séparés par au moins 3 liaisons. L'expression de cette interaction se fait sous la forme d'un potentiel de Lennard-Jones :

$$E_{\text{vdW}} = \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right]. \tag{1.20}$$

TAB. 1.8 – *Paramètres n, $V_n/2$ et γ intervenant dans l'énergie de torsion pour le champ de force* ff96. *Les types d'atomes utilisés sont précisés dans le tableau 1.5. La série de Fourrier ne compte en général qu'un ou deux termes, les autres étant nuls par défaut.*

Angle de torsion	n	$V_n/2$ [kcal/mol]	γ [degré]
N-CT-C-O	3	0,067	180
OH-CT-CT-OH	2	0,500	0
H-N-C-O	1	0,650	0
H-N-C-O	2	2,500	180
O-C-CH-N	3	0,100	180

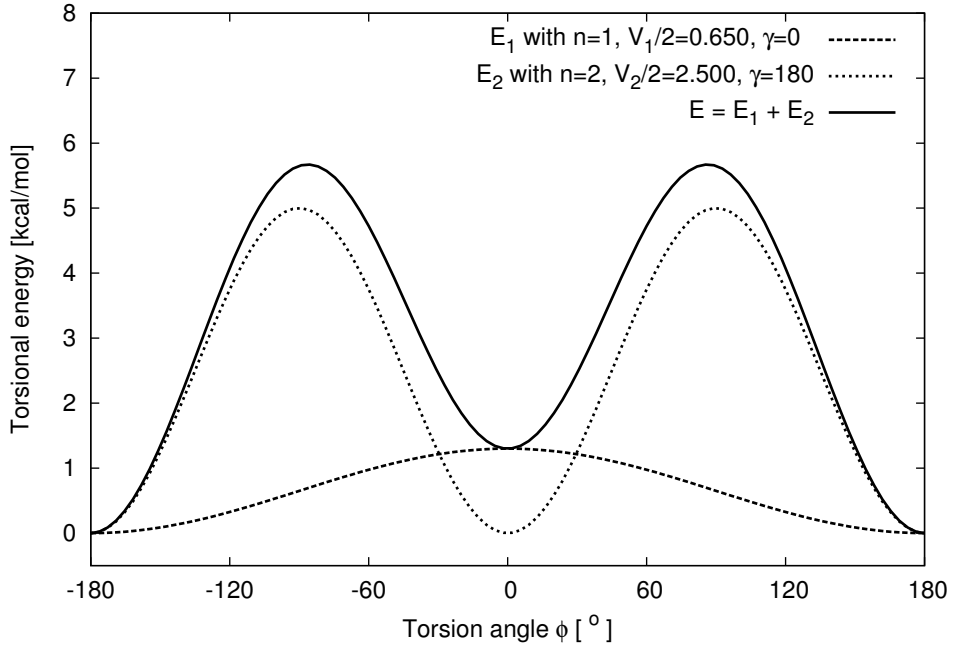


FIG. 1.19 – *Représentation de la fonction du potentiel de torsion pour l'angle H-N-C-O. La série de Fourrier associée comporte deux termes.*

Le champ de force AMBER fournit pour chaque atome i, les paramètres Lennard-Jones ϵ_i et σ_i (tableau 1.9). On définit alors A_{ij} et B_{ij} comme

$$A_{ij} = \epsilon_{ij}(\sigma_{ij})^{12} \text{ et } B_{ij} = 2\epsilon_{ij}(\sigma_{ij})^{6}, \tag{1.21}$$

en utilisant les règles de combinaison de Lorentz-Berthelot,

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \text{ et } \sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}. \tag{1.22}$$

On obtient alors

$$E_{\text{vdW}} = \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]. \tag{1.23}$$

TAB. 1.9 – *Paramètres σ et ϵ utilisés dans l'énergie d'interaction de van der Waals pour le champ de force* ff96. *Les types d'atomes sont précisés dans le tableau 1.5.*

Type d'atome	σ [Å]	ϵ [kcal/mol]
C	1,85	0,120
CT	1,80	0,060
HC	1,54	1,010
HO	1,00	0,020
N	1,75	0,160
NT	1,85	0,120

Dans les versions du champ de force AMBER antérieures à *ff86*, les interactions dues aux liaisons hydrogènes étaient représentées par un potentiel similaire avec une puissance 10 pour le terme attractif au lieu d'une puissance 6 dans l'expression de Lennard-Jones. Depuis la version *ff86*, l'énergie des interactions des liaisons hydrogènes est incluse dans l'énergie d'interaction de van der Waals.

Pour les atomes séparés par exactement trois liaisons chimiques (atomes 1–4), l'énergie de van der Waals associée est divisée par 2,0.

À titre d'exemple, le potentiel de van der Waals est représenté pour les atomes N et C dans la figure 1.20.

1.2.5.5 Énergie d'interaction électrostatique

L'interaction coulombienne entre charges ponctuelles s'applique, comme précédemment, à l'interaction entre deux atomes non liés séparés par au moins trois liaisons, portant tous deux une charge. Cette interaction est représentée par la loi de Coulomb

$$E_{\text{Coulomb}} = \frac{q_i q_j}{\epsilon r_{ij}}, \tag{1.24}$$

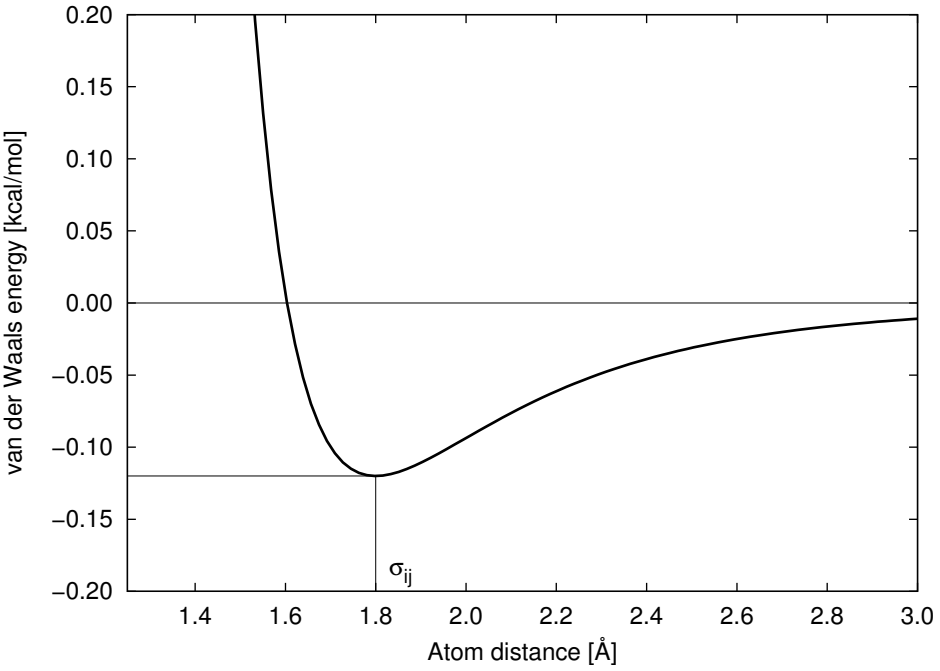


FIG. 1.20 – *Énergie de van der Waals pour les atomes N et C. Le potentiel est répulsif à courte distance (terme en 1/r_{ij}^{12}) et attractif à longue distance (terme en -1/r_{ij}^{6})*

où q_i et q_j sont les charges électriques des atomes i et j, r_{ij} est la distance entre les atomes i et j et ϵ est la constante diélectrique du milieu. Cette dernière dépend de l'environnement. La charge électrique d'un atome est fixe dans le champ de force *ff96* et est paramétrée pour chaque acide aminé (figure 1.21). Plus exactement, seuls les atomes de la forme résidu sont documentés. La charge des atomes aux extrémités N-ter (-NH_2) et C-ter (-COOH) ne sont pas connues. Pour les obtenir, la procédure recommandée par les développeurs d'AMBER est la suivante :

1. optimisation de géométrie par mécanique quantique avec la base HF/6-31 G* ;
2. calcul *single point* de la densité électronique avec la base MP2/6-31 G* ;
3. calcul des charges par dérivation du potentiel électrostatique (méthode RESP [53]).

Cette procédure a été utilisée pour déterminer les charges partielles des atomes aux extrémités neutres N-ter et C-ter pour le dipeptide Ala_2.

Par ailleurs, pour les atomes séparés par exactement trois liaisons chimiques, l'énergie électrostatique est modifiée par un facteur 0,833.

Le tableau 1.10 recense quelques valeurs de constantes diélectriques dans des milieux courants. Pour une simulation dans le vide, il faudrait logiquement prendre ϵ_0 comme constante diélectrique ($\epsilon_0 = 1$). Cependant, une protéine est une molécule organique de taille conséquente, les interactions entre atomes ne sont pas celles d'atomes totalement isolés. On prend donc usuellement des valeurs comprises entre 2 et 5 [58]. Ce procédé nous permet aussi de partiellement prendre en compte la polarisabilité des atomes dans les protéines, qui n'est pas décrite explicitement dans le champ de force utilisé.

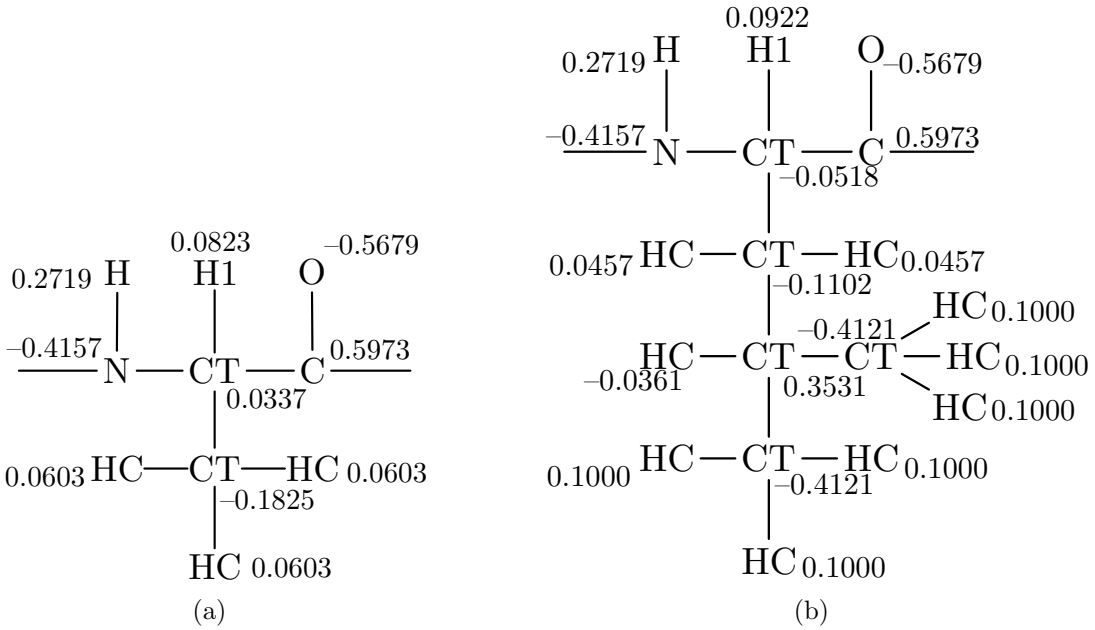


FIG. 1.21 – *Charges fixes paramétrées dans AMBER* 96. *Résidus (a) alanine et (b) leucine.*

TAB. 1.10 – *Constante diélectrique relative ϵ_r de quelques milieux [59]. Elle est définie comme le rapport entre la constante diélectrique ϵ du milieu et la constante diélectrique du vide ϵ_0 ($\epsilon_0 = 8{,}85 \times 10^{-12}$ F/m). Les constantes diélectriques ϵ et ϵ_0 s'expriment en F/m, ϵ_r est donc sans unité.*

Milieu	Constante diélectrique relative ϵ_r
Eau (20 ℃)	80,3
Eau (0 ℃)	87,7
Méthanol (20 ℃)	33,0
Éthanol (20 ℃)	25,3
Cyclohexane (20 ℃)	2,0
Argon liquide (-191 ℃)	1,5
Vide (par définition)	1,0

1.3 Paysage énergétique des biomolécules

1.3.1 Généralités

Des systèmes comme les biomolécules, les verres ou bien encore les agrégats possèdent un nombre très élevé de degrés de liberté. La complexité de la surface d'énergie potentielle évolue exponentiellement avec le nombre de ces degrés de liberté, multipliant ainsi les états stables séparés par des barrières plus ou moins élevées. Pour les protéines, la recherche du minimum global du paysage énergétique est très importante car la structure trouvée sera proche de la structure native, biologiquement active [60].

Dans cet objectif, une exploration exhaustive de tous les minima locaux est impossible. Cependant, seuls les minima les plus bas en énergie ont une influence sur les propriétés du système mais il n'y a, par contre, aucune raison pour que ces minima soient structurellement proches du minimum global. Toute la difficulté dans une simulation sera alors d'explorer largement le paysage énergétique pour recenser les minima locaux les plus stables.

Une simulation Monte Carlo ou de dynamique moléculaire conventionnelle dans l'ensemble canonique peut ainsi rencontrer certaines difficultés à échantillonner le paysage énergétique tourmenté des biomolécules. La figure 1.22 illustre ce problème, certaines barrières ne sont pas franchissables à la température de simulation.

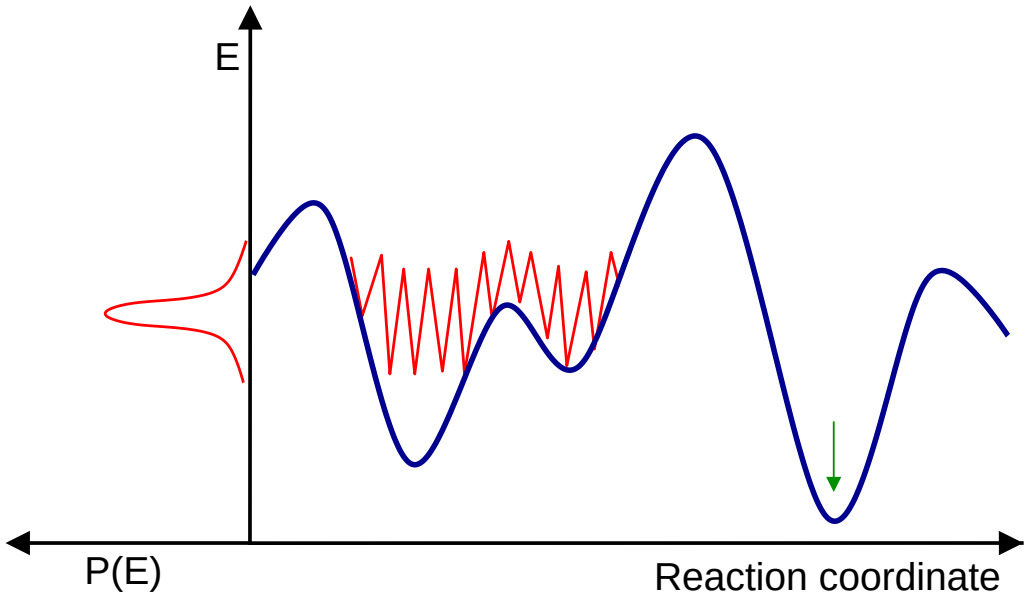


FIG. 1.22 – *Exploration conformationnelle dans l'ensemble canonique. La simulation parcourt un certain nombre de minima locaux mais ne trouve pas systématiquement le minimum global localisé par une flèche. La distribution d'énergie $P(E)$ indique les énergies accessibles à une température donnée.*

1.3.2 Optimisation globale

Les méthodes d'optimisation globale ont pour vocation de répondre au problème d'exploration du paysage énergétique.

1.3.2.1 Recuit simulé

Le recuit simulé (*simulated annealing*) [61] est historiquement une des premières méthodes d'optimisation globale. Sa relative simplicité en fait une méthode encore très utilisée.

Le recuit simulé tente de reproduire le processus naturel de cristallisation. En abaissant graduellement la température d'une substance fondue on peut en principe obtenir un cristal parfait correspondant au minimum global d'énergie [62]. Pratiquement, cette méthode consiste en plusieurs refroidissements et recuits successifs. Partant d'une température suffisamment élevée pour parcourir largement l'espace des phases, on refroidit lentement le système pour trouver les configurations les plus stables. Périodiquement, on chauffe pour permettre au système de quitter les minima locaux dans lesquels il serait piégé. Les étapes de refroidissement et de chauffage sont souvent données par une règle géométrique

$$T_{n+1} = \alpha T_n, \tag{1.25}$$

ou arithmétique

$$T_{n+1} - T_n = \alpha, \tag{1.26}$$

avec $\alpha < 1$ pour un refroidissement et $\alpha > 1$ pour un réchauffement. La figure 1.23 illustre l'évolution de la température au cours d'une simulation par recuit simulé dans le cas d'un comportement géométrique ou arithmétique. La décroissance géométrique de la température est ralentie à mesure que le système se refroidit.

Si la température diminue trop rapidement (phénomène de trempe ou *quenching*), le système peut rester bloqué dans un minimum local d'énergie. Dès lors, pour garantir qu'une simulation par recuit simulé atteigne le minimum global, il faudrait que le recuit soit infiniment long. Bien entendu, ceci est irréalisable en pratique et le recuit simulé ne garantit pas de trouver le minimum global. Cependant, si après plusieurs simulations la même configuration est obtenue, on considérera alors que cette configuration correspond au minimum global [62] ou au moins à une conformation très stable.

1.3.2.2 Autres méthodes d'optimisation

D'autres méthodes d'optimisation globale, autres que le recuit simulé, existent [63]. Nous présentons quelques unes d'entre elles ci-après.

La déformation d'hypersurface [64] consiste à atténuer la rugosité du paysage énergétique en appliquant un potentiel adapté. Le nombre de minima locaux est ainsi réduit et le minimum global plus facilement repéré. Une transformation inverse doit retrouver la

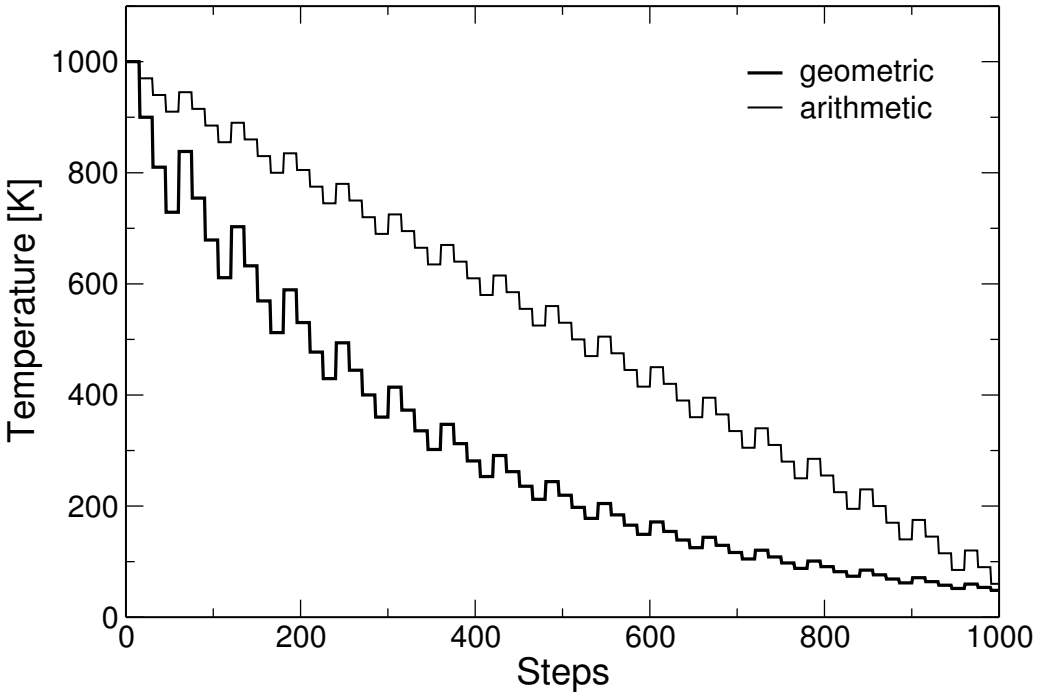


FIG. 1.23 – *Évolution de la température au cours d'une simulation en recuit simulé. Une progression géométrique et arithmétique de la température sont représentées.*

surface d'énergie originale pour que son minimum global puisse être identifié au minimum global de la surface déformée.

La méthode de saut de bassins (*basin-hopping*) [60] déforme l'espace des phases en le discrétisant en un certain nombre de bassins d'attraction. L'énergie d'une structure est associée à l'énergie du minimum local le plus proche, ce qui élimine les barrières. L'exploration de la surface de potentiel se trouve ainsi améliorée.

Les algorithmes génétiques [65, 66] consistent à faire évoluer une population de structures tirées de minima locaux (chromosomes) vers la structure du minimum global. À chaque génération, les énergies des structures optimisées sont calculées. Elles sont ensuite aléatoirement sélectionnées en fonction de leur énergie puis recombinées voire modifiées (mutées), produisant alors une nouvelle génération.

Toutes ces méthodes d'optimisation globale se soucient d'une vaste exploration de l'espace des phases mais elles ne permettent pas l'accumulation de statistiques pour le calcul de moyennes canoniques d'observables. Nous verrons dans le chapitre 2 que l'utilisation d'ensembles généralisés (*generalized ensembles*) [67] permettra de lever en partie ces difficultés.

1.4 Simulation de biomolécules par dynamique moléculaire

La dynamique moléculaire est la méthode la plus intuitive et la plus naturelle pour explorer la surface d'énergie potentielle d'une biomolécule. Historiquement, la première molécule d'intérêt biologique — l'inhibiteur de la trypsine pancréatique bovine (BPTI) — a été modélisée par dynamique moléculaire il y moins de 30 ans [68] et cette technique de simulation est toujours largement utilisée actuellement [69].

Le dynamique moléculaire cherche à obtenir une exploration de la surface d'énergie potentielle, l'accumulation de statistiques et bien évidemment, la construction d'une dynamique réelle, par exemple, de repliement [69].

La section suivante propose une rapide introduction à cette technique. Une description plus complète peut être trouvée dans les ouvrages de Leach [35] et de Frenkel et Smit [62].

1.4.1 Principe et intégrations finies

La dynamique moléculaire est basée sur la résolution numérique des équations du mouvement de Newton :

$$\frac{d^2x_i}{dt^2} = \frac{f_{x_i}}{m_i} = -\frac{1}{m_i}\frac{\partial V}{\partial x_i}, \tag{1.27}$$

où m_i est la masse de la particule i, x_i est une coordonnée et f_{x_i} est la force appliquée à la particule i suivant cette coordonnée, qui dérive de l'énergie potentielle V.

Les vitesses initiales sont déterminées par une distribution de Maxwell-Boltzmann. La probabilité qu'un atome i ait la vitesse v_x à la température T est :

$$P(v_{i,x}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left(-\frac{m_i v_{i,x}^2}{2k_B T}\right). \tag{1.28}$$

L'intégration des équations de Newton se fait numériquement sur des temps (δt) finis et petits. On utilise pour cela les développements en série de Taylor :

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t.\vec{v}(t) + \frac{1}{2}\delta t^2.\vec{a}(t) + \frac{1}{6}\delta t^3.\vec{b}(t) + \cdots \tag{1.29}$$

$$\vec{v}(t+\delta t) = \vec{v}(t) + \delta t.\vec{v}(t) + \frac{1}{2}\delta t^2.\vec{b}(t) + \cdots \tag{1.30}$$

$$\vec{a}(t+\delta t) = \vec{a}(t) + \delta t.\vec{b}(t) + \cdots \tag{1.31}$$

L'algorithme de Verlet aux positions s'exprime par

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t.\vec{v}(t) + \frac{1}{2}\delta t^2.\frac{\vec{f}(t)}{m}; \tag{1.32}$$

$$\vec{r}(t-\delta t) = \vec{r}(t) - \delta t.\vec{v}(t) + \frac{1}{2}\delta t^2.\frac{\vec{f}(t)}{m}; \tag{1.33}$$

pour donner finalement :

$$\vec{r}(t+\delta t) = 2\vec{r}(t) - \vec{r}(t-\delta t) + \delta t^2.\frac{\vec{f}(t)}{m}. \tag{1.34}$$

L'intérêt de cet algorithme réside dans le faible coût de stockage qu'il offre. La position à $t+\delta t$ est en effet définie par $\vec{r}(t)$, $\vec{r}(t-\delta t)$ et $\vec{a}(t)$. Cependant, une perte de précision est due à l'addition d'un terme petit $[\delta t^2.\vec{a}(t)]$ à la différence de deux termes plus grands $[2\vec{r}(t)$ et $\vec{r}(t-\delta t)]$. Enfin, la vitesse ne peut être calculée à la même précision qu'en connaissant la position suivante $[\vec{r}(t+\delta t)]$.

Plusieurs autres algorithmes comme l'algorithme « saute-mouton » (*leap-frog*), Verlet aux vitesses ou prédicteur–correcteur permettent également d'intégrer les équations de Newton.

1.4.2 Principales échelles de temps rencontrées en dynamique moléculaire

L'intégration temporelle se fait de manière discrète en dynamique moléculaire. Il convient donc de définir avec soin le pas de temps utilisé. Le tableau 1.11 relate quelques différents temps caractéristiques des systèmes étudiés. Avec un pas de temps trop court, la simulation ne peut pas converger et explorer correctement l'espace des phases. Réciproquement, un pas trop grand aboutit à une forte instabilité du système [35].

TAB. 1.11 – *Principales échelles de temps rencontrées en dynamique moléculaire pour des systèmes biologiques. [35, 70]*

Mouvement	temps [s]
élongation d'une liaison	10^{-14}
flexion d'une liaison	10^{-14}
rotation des chaînes latérales de surface	10^{-11}–10^{-10}
rotation des chaînes latérales intérieures	10^{-4}–1
repliement de protéines	10^{-6}–10^{2}

Les échelles de temps rencontrées dans les systèmes biologiques peuvent couvrir jusqu'à 16 ordres de grandeurs. Le défi des simulations de biomolécules est de modéliser des phénomènes lents comme le repliement d'une protéine tout en tenant compte de mouvements rapides comme la vibration d'une liaison.

Plusieurs approches tentent de résoudre ce problème. Les algorithmes de dynamique avec contraintes, SHAKE [71] et RATTLE [72], modifient les équations du mouvement de façon à geler les vibrations les plus rapides. Dans la méthode à pas de temps multiples [73], les forces à longue distance sont évaluées moins souvent que celles à courte distance.

Avec ces stratégies et une augmentation significative de la puissance de calcul, la dynamique moléculaire peut actuellement simuler des biomolécules sur des temps de l'ordre de 10 ns ou des systèmes allant jusqu'à 10^4–10^6 atomes [69]. Ces performances sont remarquables mais restent cependant insuffisantes pour simuler le repliement d'une protéine qui peut se dérouler en 10 μs.

Dans le contexte de l'échantillonnage, la méthode Monte Carlo est une stratégie alternative qui permet à la fois une large exploration de la surface d'énergie potentielle et une accumulation de statistiques. Cette méthode a été utilisée pour ce travail de thèse et sera détaillée dans le chapitre 2.

1.5 Conclusion

Les protéines, et plus généralement les biomolécules, sont des systèmes très complexes pouvant adopter une large variété de conformations. Parmi ces nombreuses structures, on distingue deux motifs très ordonnés : les hélices α et les feuillets β, respectivement associés à un dipôle électrique élevé et très faible. Dugourd *et al.* ont montré que, plus généralement, le dipôle électrique peut servir de sonde conformationnelle.

La modélisation de protéines, systèmes pouvant contenir jusqu'à plusieurs milliers d'atomes, n'est pas envisageable par une approche quantique. Moyennant quelques approximations, la mécanique moléculaire peut, par contre, fournir une description relativement fidèle des protéines.

Pour répondre à des préoccupations d'ordre statistique, nous avons utilisé la méthode Monte Carlo pour échantillonner la surface d'énergie potentielle des biomolécules étudiées. Cette surface est tourmentée et présente de nombreux minima locaux. Déterminer le minimum global reste un vrai défi et fait appel à des techniques de simulation évoluées. Le développement de telles techniques Monte Carlo est l'objet du prochain chapitre.

Bibliographie

[1] G. J. Mulder. On the composition of some animal substances. *Journal für praktische Chemie*, 16:129, 1839.

[2] L. Pauling, R. B. Corey, and H. R. Branson. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37:205–211, 1951.

[3] M. F. Jarrold. Peptides and proteins in the vapor phase. *Annual Review of Physical Chemistry*, 51:179–207, 2000.

[4] E. Lolis and G. A. Petsko. Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5 Å resolution: implications for catalysis. *Biochemistry*, 29:6619–6625, 1990.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] L. Pauling and R. B. Corey. Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37:235–240, 1951.

[7] L. Pauling and R. B. Corey. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37:251–256, 1951.

[8] S. Weinman and P. Méhul. *Biochimie - Structure et fonction des protéines*. Dunod, Paris, 2000.

[9] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

[10] G. N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, 23:283–438, 1968.

[11] S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by C-alpha geometry: phi, psi and C-beta deviation. *Proteins: Structure, Function, and Genetics*, 3:437–450, 2003.

[12] T. E. Creighton. *Proteins : structures and molecular properties*. Freeman, 6^{th} edition, 2002.

[13] M. Munson, S. Balasubramanian, K. G. Fleming, A. D. Nagi, R. O'Brien, J. M. Sturtevant, and L. Regan. What makes protein a protein? Hydrophobic core designs that specify and structural properties. *Protein Science*, 5:1584–1593, 1996.

[14] M. E. M. Noble, R. K. Wierenga, A.-M. Lambeir, F. R. Opperdoes, W. H. Thunnissen, K. H. Kalk, H. Groendijk, and W. G. Hol. The Adaptability of the Active Site of Trypanosomal Triosephospahte Isomerase as Observed in the Crystal Structures of Three Different Complexes. *Proteins: Structure, Function, and Genetics*, 10:50–69, 1991.

[15] C. B. Anfinsen, R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *The Journal of Biological Chemistry*, 207:201–210, 1954.

[16] C. B. Anfinsen. Structure of a protein is determined solely by the amino acids sequence info. *Science*, 181:223–230, 1973.

[17] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89:20–22, 1992.

[18] A. R. Dinner, A. Sali, and M. Karplus. The folding mechanism of larger model proteins: Role of native structure. *Proceedings of the National Academy of Sciences of the United States of America*, 93:8356–8361, 1996.

[19] C. Levinthal. Are there pathways for protein folding ? *Journal of Chemical Physics*, 65:44–45, 1968.

[20] C. Gibas and P. Jambeck. *Introduction à la bioinformatique*. O'Reilly, Paris, 1^{st} edition, 2002.

[21] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85:2444–2448, 1988.

[22] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[23] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 33:34–38, 2005.

[24] P. Atkins and J. de Paula. *Atkins' physical chemistry*. Oxford University Press, Oxford, 7^{th} edition, 2002.

[25] P. Debye. *Polar molecules*. Dover, New York, 1929.

[26] J. H. Van Vleck. On Dielectric Constants and Magnetic Susceptibilities in the New Quantum Mechanics. Part II Application to Dielectric Constants. *Physical Review*, 30:31–54, 1927.

[27] A. Wada. The alpha-helix as an electric macro-dipole. *Advances in Biophysis*, 9:1–63, 1976.

[28] W. G. J. Hol. The role of the alpha-helix dipole in protein function and structure. *Progress in Biophysics and Molecular Biology*, 45:149–195, 1985.

[29] R. Pethig. *Dielectric and Electronic Properties of Biological Materials*. Wiley, New-York, 1979.

[30] R. Antoine, I. Compagnon, D. Rayane, M. Broyer, P. Dugourd, G. Breaux, F. C. Hagemeister, D. Pippen, R. R. Hudgins, and M. F. Jarrold. Electric susceptibility of unsolvated glycine-based peptides. *Journal of American Chemical Society*, 124:6737–6741, 2002.

[31] R. Antoine, I. Compagnon, D. Rayane, M. Broyer, Ph. Dugourd, G. Breaux, F.C. Hagemeister, D. Pippen, R. R. Hudgins, and M. F. Jarrold. Electric dipole moments and conformations of isolated peptides. *The European Physical Journal D*, 20:583–587, 2002.

[32] R. Antoine, I. Compagnon, D. Rayane, M. Broyer, P. Dugourd, N. Sommerer, M. Rossignol, D. Pippen, F. C. Hagemeister, and M. F. Jarrold. Application of molecular beam deflection time-of-flight mass spectrometry to peptide analysis. *Analytical Chemistry*, 75:5512–5516, 2003.

[33] B. J. Berne and R. Pecora. *Dynamic Light Scattering*. Wiley-Interscience, New-York, 1976.

[34] Y. Peng and U. H. E. Hansmann. Helix versus sheet formation in a small peptide. *Physical Review E*, 68:041911, 2003.

[35] A. R. Leach. *Molecular modelling: principles and applications*. Pearson Education, Harlow, 2^{nd} edition, 2001.

[36] C. J. Cramer. *Essentials of Computational Chemistry : theories and models*. Wiley, Chichester, England, 2^{nd} edition, 2004.

[37] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review B*, 136:864–887, 1964.

[38] W. Kohn and L. J. Sham. Self-consistent Equations Including Exchange and Correlation Effects. *Physical Review A*, 140:1133–1138, 1965.

[39] M. J. S. Dewar and W. Thiel. Ground States of Molecules, 38. The MNDO Method. Approximations and Parameters. *Journal of American Chemical Society*, 99:4899–4907, 1977.

[40] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. AM1: A New General Purpose Quantum Mechanical Model. *Journal of American Chemical Society*, 107:3902–3909, 1985.

[41] J. J. P. Stewart. Optimization of Parameters for Semi-Empirical Methods. I-Method. *Journal of Computational Chemistry*, 10:209–220, 1989.

[42] J.A. Pople, D. L. Beveridge, and P. A. Dobosh. Approximate Self-consistent Molecular Orbital Theory V. Intermediate Neglect of Differential Overlap. *Journal of Chemical Physics*, 47:2026, 1967.

[43] J. Gao. Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 7, pages 119–185, 1996.

[44] P. Amara and M. Field. Combined Quantum Mechanics and Molecular Mechanics Potentials. In P. V. R. Schleyer, editor, *Encyclopedia of Computational Chemistry*, volume 1, pages 431–437. Wiley & sons, 1998.

[45] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, Jr. S. Profeta, and P. K. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of American Chemical Society*, 106:765–784, 1984.

[46] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[47] W. L. Jorgensen and J. Tirado-Rives. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *Journal of American Chemical Society*, 110:1657–1666, 1988.

[48] W. F. van Gunsteren and H. J. C. Berendsen. Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry. *Angewandte Chemie International Edition in English*, 29:992–1023, 1990.

[49] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy Parameters in Polypeptides VII. Geometric Parameters, Partial Charges, Non-bonded Interactions, Hydrogen Bond Interactions and Intrensic Torsional Potentials for Naturally Occurring Amino Acids. *Journal of Physical Chemistry*, 79:2361–2381, 1975.

[50] P. K. Weiner and P. A. Kollman. AMBER: Assisted Model Building with Energy Refinement. A General Program for Modeling Molecules and Their Interactions. *Journal of Computational Chemistry*, 2:287–303, 1981.

[51] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *Journal of Computational Chemistry*, 7:230–252, 1986.

[52] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Jr. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of American Chemical Society*, 117:5179–5197, 1995.

[53] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *Journal of Physical Chemistry*, 97:10269–10280, 1993.

[54] P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille. The Development Application of a 'Minimalist' Organic Biochemical Molecular Mechanic Force Field using a Combination of ab Initio Calculations and Experimental Data. In W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors, *Computer Simulations of Biomolecular Systems*, volume 3, pages 83–96, Dordrecht, The Netherlands, 1997. Kluwer Academic.

[55] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21:1049–1074, 2000.

[56] A. E. García and K. Y. Sanbonmatsu. alpha Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 99:2782–2787, 2002.

[57] K. S. Pitzer. Potential energies for rotation about single bonds. *Discussions of the Faraday Society*, 10:66–73, 1951.

[58] M. K. Gilson and H. H. Honig. The dielectric constant of a folded protein. *Biopolymers*, 25:2097–2119, 1986.

[59] D. R. Lide, editor. *CRC Handbook of Chemistry and Physics*. CRC Press, Boca Raton, USA, 85^{th} edition, 2004.

[60] D. J. Wales and J. P. K Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *Journal of Physical Chemistry A*, 101:5111–5116, 1997.

[61] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[62] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, California, USA, 2^{nd} edition, 2002.

[63] D. J. Wales and H. A. Sheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285:1368–1372, 1999.

[64] S. Schelstraete and H. Verschelde. Finding Minimum-Energy Configurations of Lennard-Jones Clusters Using an Effective Potential. *Journal of Physical Chemistry A*, 101:310–315, 1997.

[65] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan, USA, 1975.

[66] D. E. Goldberg, editor. *Genetic Algorithms in Search, Optimization and Machine Learning*. Kluwer Academic, Boston, USA, 1989.

[67] Y. Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling*, 22:425–439, 2004.

[68] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[69] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646–652, 2002.

[70] T. Schlick. *Molecular Modeling and Simulation: an interdisciplinary guide*. Springer-Verlag, New-York, USA, 2002.

[71] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Chemical Physics*, 23:327–341, 1977.

[72] H. C. Andersen. Rattle: a 'velocity' version of the SHAKE algorithm for molecular dynamics calculations. *Journal of Chemical Physics*, 52:24–34, 1983.

[73] W. B. Streett, D. J. Tildesley, and G. Saville. Multiple time step methods in molecular dynamics. *Molecular Physics*, 35:639–648, 1978.

Chapitre 2

Méthodes Monte Carlo dans les ensembles généralisés

Sommaire

Comme nous l'avons discuté dans le chapitre précédent, l'utilisation d'un champ de force définit l'énergie d'une protéine dans une configuration donnée. Pour déterminer les propriétés structurales de ces molécules, il nous faut connaître les conformations qui correspondent à un minimum d'énergie. Les propriétés thermodynamiques sont quant à elles données en mécanique statistique par la probabilité de trouver une configuration particulière.

Les méthodes Monte Carlo sont parfaitement adaptées pour l'exploration de l'espace conformationnel et l'accumulation de statistiques. Dans un premier temps, nous aborderons le principe général de ces algorithmes. Puis nous traiterons leurs extensions dans les ensembles généralisés, que ce soit des méthodes de trajectoires multiples ou d'échantillonnage non-boltzmannien. Pour chacune d'entre elles, les traitements statistiques qui permettent d'accéder aux moyennes canoniques des observables étudiées sont détaillés.

2.1 Méthode Monte Carlo Metropolis

La méthode Monte Carlo originale a été introduite au début des années 40. Elle doit son nom à l'utilisation des nombres aléatoires, en référence aux casinos situés à Monte Carlo. Au XVIII$^{\text{e}}$ siècle, Buffon avait déjà eu recours aux tirages de nombres aléatoires dans son problème des aiguilles. La valeur de π était estimée par un jet répété d'une aiguille sur des droites parallèles. Pour une aiguille de longueur ℓ et des droites espacées de cette même distance, la probabilité que l'aiguille coupe l'une des droites s'avère être $2/\pi$.

2.1.1 Principe général

Le calcul de la moyenne statistique d'une observable $\mathcal{A}$ en intégrant tout l'espace des phases $\{X\}$ est impossible lorsque le nombre de degrés de liberté D est grand. Ainsi, pour D degrés de liberté et seulement 10 points par dimension, il faudrait calculer 10^D fois l'observable $\mathcal{A}(X)$,

$$\langle \mathcal{A} \rangle = \int \mathcal{A}(X) d^D X. \tag{2.1}$$

Plutôt que d'évaluer $\mathcal{A}(X)$ sur N points régulièrement répartis (méthode des trapèzes, de Gauss...), on peut également évaluer l'observable considérée sur N points disposés aléatoirement (*simple* ou *random sampling*). On a alors la somme discrète

$$\langle \mathcal{A} \rangle \simeq \frac{1}{N} \sum_i \mathcal{A}(X_i). \tag{2.2}$$

Cependant, un système moléculaire comporte beaucoup plus de conformations à haute énergie qu'à basse énergie. Les propriétés intéressantes sont le plus souvent recherchées à des températures où les conformations de basses énergies jouent un rôle significatif dans le calcul de la fonction de partition. Une autre méthode pour calculer une moyenne statistique consiste alors à échantillonner l'espace des phases avec un ensemble de points X_i répartis suivant la distribution de l'ensemble canonique $\rho(X)$. On parle alors d'échantillonnage d'importance ou *importance sampling*. Si le nombre de points est suffisant et représentatif de l'ensemble des configurations possibles, alors la moyenne canonique de l'observable $\mathcal{A}$

est telle que

$$\langle \mathcal{A} \rangle \simeq \frac{\sum_i \mathcal{A}(X_i)\rho(X_i)}{Z}, \tag{2.3}$$

où ρ est la distribution de l'ensemble canonique ($\rho(X_i) = \exp[-\beta E(X_i)]$) avec la température inverse $\beta = 1/k_B T$, k_B la constante de Boltzmann, T la température et $E(X_i)$ l'énergie de l'état X_i. La fonction de partition Z à la température T est définie comme

$$Z = \sum_i e^{-\beta E(X_i)}. \tag{2.4}$$

Lorsqu'une telle série de points est créée suivant la distribution de Boltzmann, la moyenne de l'observable $\mathcal{A}$ revient à l'équation (2.2). Ces configurations peuvent être engendrées par une chaîne de Markov.

2.1.2 Chaîne de Markov

Une chaîne de Markov $\{X_i\}$ est une suite de points X_i qui, pour un nombre infini de points, tend vers la distribution $\rho(X)$:

$$X_0 \to X_1 \to \cdots \to X_i \to X_{i+1} \to \cdots \tag{2.5}$$

Quels que soient deux états X et Y, on peut passer de l'un à l'autre aléatoirement. On définit alors une probabilité de transition $\pi(X \to Y)$ qui ne dépend pas du temps mais uniquement des états X et Y. En particulier, elle ne dépend pas des états antérieurs occupés par le système. Enfin, on peut toujours atteindre un état Y à partir d'un état X de l'espace des états. La probabilité π satisfait donc

$$\sum_Y \pi(X \to Y) = 1 \quad \text{pour tout } X. \tag{2.6}$$

La chaîne de Markov atteint la distribution ρ voulue lorsqu'on a ergodicité et que le bilan détaillé est respecté.

2.1.2.1 Ergodicité

Dans une chaîne de Markov, l'ergodicité stipule qu'il est possible d'atteindre n'importe quel point Y à partir de n'importe quel point X en un nombre *fini* d'étapes.

2.1.2.2 Bilan détaillé

La réversibilité microscopique donne à la limite asymptotique

$$\sum_Y \rho(Y)\pi(Y \to X) = \rho(X) \quad \text{pour tout } X. \tag{2.7}$$

La condition du bilan détaillé suffit à assurer cette égalité :

$$\frac{\pi(X \to Y)}{\pi(Y \to X)} = \frac{\rho(X)}{\rho(Y)} \quad \text{pour tous } X \text{ et } Y. \tag{2.8}$$

Cette dernière équation signifie que le flux des mouvements de X vers Y est compensé par le flux des mouvements de Y vers X.

2.1.3 Algorithme de Metropolis

Dans un premier temps, on passe aléatoirement de l'actuel point X_i vers le nouveau point X. Ensuite, on détermine si ce nouvel état est accepté comme point de la chaîne de Markov. Les mouvements sont donc essayés avec une probabilité $\alpha(X_i \to X)$ puis acceptés avec une probabilité $\mathrm{acc}(X_i \to X)$,

$$\pi(X_i \to X) = \alpha(X_i \to X) \times \mathrm{acc}(X_i \to X). \tag{2.9}$$

Le choix des mouvements se fait aléatoirement, donc $\alpha(X_i \to X) = \alpha(X \to X_i)$. Les équations (2.8) et (2.9) donnent

$$\frac{\mathrm{acc}(X_i \to X)}{\mathrm{acc}(X \to X_i)} = \frac{\rho(X)}{\rho(X_i)}. \tag{2.10}$$

Dans le cas d'une distribution canonique, $\rho(X) \propto e^{-\beta E(X)}$, on a

$$\frac{\mathrm{acc}(X_i \to X)}{\mathrm{acc}(X \to X_i)} = \frac{\rho(X)}{\rho(X_i)} = e^{-\beta(E(X)-E(X_i))}. \tag{2.11}$$

Metropolis *et al.* proposèrent en 1953 [1] :

$$\mathrm{acc}(X_i \to X) = \begin{cases} e^{-\beta[E(X)-E(X_i)]}, & \text{si } E(X_i) < E(X); \\ 1, & \text{si } E(X_i) \geq E(X). \end{cases} \tag{2.12}$$

En pratique l'algorithme Monte Carlo de Metropolis est le suivant :

1. on crée aléatoirement une conformation X à partir de la conformation X_i ;
2. on calcule la différence d'énergie $\Delta E = E(X) - E(X_i)$ entre ces deux conformations ;
3. – si $\Delta E \leq 0$, alors le mouvement est accepté ;
   – si $\Delta E > 0$, on lui attribue la probabilité $p = \exp(-\beta \Delta E)$. Le mouvement est accepté si p est supérieur à un nombre aléatoire tiré entre 0 et 1, sinon il est rejeté ;
4. – si le mouvement est accepté, alors X devient la nouvelle conformation X_{i+1} ;
   – si le mouvement est rejeté, alors X_i reste la conformation actuelle et est *recopiée* dans X_{i+1}.

Enfin, il faut remarquer que l'accumulation de données statistiques doit se faire lorsque

le mouvement est accepté comme rejeté, le rejet étant tout aussi important que l'acceptation d'une nouvelle conformation.

2.1.4 Pas de déplacement

La nouvelle conformation X est créée à partir de l'ancienne conformation X_i en la modifiant aléatoirement, par exemple en déplaçant un atome n de coordonnées initiales (x_n, y_n, z_n) :

$$\begin{aligned} x_n(X) &= x_n(X_i) + \Delta(\text{rand} - 1/2), \\ y_n(X) &= y_n(X_i) + \Delta(\text{rand} - 1/2), \\ z_n(X) &= z_n(X_i) + \Delta(\text{rand} - 1/2), \end{aligned} \tag{2.13}$$

où rand représente un nombre aléatoire dans l'intervalle [0,1[et Δ est le pas de déplacement qui doit être ajusté suivant la température de simulation. En effet, un pas de déplacement trop petit entraîne de faibles modifications de structure et donc de faibles variations d'énergie. La plupart des mouvements seront acceptés et l'espace conformationnel faiblement exploré. Au contraire, un pas de déplacement trop grand conduit à des modifications importantes de structure et donc à des mouvement presque tous refusés. La convergence sera là encore très lente. Pour qu'elle soit optimale, il faut que le nombre de mouvements acceptés soit comparable au nombre de mouvements rejetés, d'où la nécessité d'ajuster Δ pour atteindre cet équilibre. Cette adaptation peut se faire régulièrement au cours de la simulation, pas trop souvent cependant pour conserver la réversibilité microscopique de la chaîne de Markov [2]. On peut également fixer Δ dès le début de la simulation en connaissant la température de simulation.

On définit le taux d'acceptance comme le rapport entre le nombre de mouvements acceptés sur le nombre de mouvements tentés. En pratique, le pas de déplacement Δ doit être adapté de façon à ce que le taux d'acceptance reste dans l'intervalle 30–70 % [3, 4].

2.1.5 Générateur de nombres aléatoires

La génération d'une chaîne de Markov ainsi que l'algorithme de Metropolis reposent sur l'utilisation constante de nombres aléatoires. Un générateur de nombres aléatoires idéal étant très coûteux et peu pratique (car basé sur la désintégration de noyaux radioactifs), il est nécessaire d'utiliser un générateur de nombres pseudo-aléatoires. Autrement dit, il faut utiliser un algorithme capable de produire une suite de nombres dont la distribution se rapproche le plus possible d'une distribution purement aléatoire.

Le générateur de nombres pseudo-aléatoires utilisé dans le cadre de ce travail est l'algorithme *Mersenne Twister* [5] développé par Matsumoto et Nishimura en 1996. Ce générateur a été choisi pour sa rapidité et sa très grande période ($2^{19937} - 1 \sim 10^{6000}$), à comparer avec la période ($\sim 10^9$) [6] de l'algorithme utilisé par la bibliothèque ANSI

C accompagnant le compilateur *gcc*. L'implémentation de cet algorithme en C++ a été réalisée par Zubin Dittia [7].

2.1.6 Simulation dans l'ensemble canonique

Considérons un système composé de N degrés de liberté dans l'ensemble canonique à la température T. Chaque état X du système est pondéré par le facteur de Boltzmann

$$W_B(X,T) = \frac{e^{-\beta E(X)}}{Z(T)}. \tag{2.14}$$

La probabilité de distribution canonique $P_B(E,T)$ de l'énergie potentielle s'écrit alors comme le produit de la densité d'états microcanonique du système $\Omega(E)$ par le facteur de Boltzmann $W_B(X,T)$:

$$P_B(E,T) \propto \Omega(E) W_B(X,T). \tag{2.15}$$

Lorsque l'énergie E augmente, $\Omega(E)$ croît rapidement et W_B décroît exponentiellement, ce qui produit une distribution de l'ensemble canonique piquée autour de l'énergie moyenne à la température T.

La densité d'états $\Omega(E_0)$ représente le nombre d'états X ayant une énergie E_0. Elle est définie comme

$$\Omega(E_0) = \int \delta[E(X) - E_0] dX. \tag{2.16}$$

La connaissance de la densité d'états suffit en principe pour connaître la fonction de partition par transformée de Laplace

$$Z(T) = \int e^{-\beta E} dX = \int \Omega(E) e^{-\beta E} dE. \tag{2.17}$$

Par suite, la moyenne canonique d'une observable $\mathcal{A}$ dépendant uniquement de E est donnée par

$$\langle \mathcal{A} \rangle_T = \frac{\int \mathcal{A}(E) P_B(E,T) dE}{Z(T)} = \frac{\int \mathcal{A}(E) \Omega(E) e^{-\beta E} dE}{\int \Omega(E) e^{-\beta E} dE}, \tag{2.18}$$

avec $\mathcal{A}(E)$ l'histogramme de $\mathcal{A}$ en fonction de E.

On peut également déterminer l'ensemble des propriétés thermodynamiques comme l'énergie libre

$$F(T) = -\frac{1}{\beta} \ln[Z(T)] = -\frac{1}{\beta} \ln \int \Omega(E) e^{-\beta E} dE, \tag{2.19}$$

l'énergie interne,

$$U(T) = \langle E \rangle_T = \frac{\int E \Omega(E) e^{-\beta E} dE}{Z(T)} = \frac{\int E \Omega(E) e^{-\beta E} dE}{\int \Omega(E) e^{-\beta E} dE}, \tag{2.20}$$

ou encore la chaleur spécifique,

$$C_V(T) = \frac{\partial U}{\partial T} = \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{k_B T^2}. \tag{2.21}$$

2.1.7 Simulations dans les ensembles généralisés

Comme nous l'avons vu dans le chapitre 1, une simulation conventionnelle dans l'ensemble canonique peut rencontrer certaines difficultés à échantillonner un paysage énergétique tourmenté et rester bloquée dans l'un des nombreux minima locaux existants.

Des simulations dans les ensembles généralisés [8, 9] peuvent contourner ces difficultés. Elles effectuent une marche aléatoire dans l'espace des phases qui peut franchir les barrières d'énergie et explorer plus largement la surface de potentiel. De telles simulations permettent d'atteindre le minimum global d'énergie et de calculer les moyennes canoniques à une température donnée par différentes techniques de repondération [10].

Pour échantillonner une distribution dont le poids statistique de chaque état X est $W[E(X)]$, le critère de Metropolis est

$$\mathrm{acc}(X_i \to X) = \min\left[1, \frac{W[E(X)]}{W[E(X_i)]}\right], \tag{2.22}$$

et $W(E)$ conduit à la probabilité d'énergie

$$P(E) \propto \Omega(E) W(E). \tag{2.23}$$

Dans le cas général, le poids statistique $W(E)$ est donné par la nature de l'ensemble considéré. Ainsi dans l'ensemble canonique,

$$W(E) = e^{-\beta E}. \tag{2.24}$$

Dans l'ensemble microcanonique à l'énergie totale E_0, ce poids devient

$$W(E) = (E_0 - E) e^{\frac{3N-8}{2}} \Theta(E_0 - E), \tag{2.25}$$

avec N le nombre d'atomes et Θ la fonction d'échelon de Heaviside. Dans l'ensemble de Tsallis [11], on souhaite obtenir une distribution qui favorise les énergies élevées

$$W(E) = \left[1 + \beta \frac{E - E_0}{\gamma}\right]^{-\gamma}, \tag{2.26}$$

avec γ un paramètre de la distribution qui pour une valeur infinie redonne la distribution de Boltzmann.

D'autres méthodes peuvent être également employées, comme le pavage du paysage énergétique proposé par Hansmann *et al.* [12]. Un échantillonnage idéal serait obtenu pour

une probabilité P uniforme, ce qui conduit à

$$W(E) = \frac{1}{\Omega(E)}. \tag{2.27}$$

Contrairement au poids statistique dans l'ensemble canonique, le poids $W(E)$ n'est ici pas connu car dépendant de la densité d'états $\Omega(E)$ qu'il faut déterminer. On cherchera alors à l'estimer par différentes méthodes dans les ensembles généralisés.

Dans les sections suivantes, nous détaillons deux grandes classes d'algorithmes. La première est basée sur des trajectoires multiples et la deuxième sur un échantillonnage non-boltzmannien de la surface d'énergie potentielle.

2.2 Méthode des trajectoires multiples : Monte Carlo d'échange

2.2.1 *Simulated tempering*

La méthode *simulated tempering* [13, 14] est la première méthode à avoir introduit la notion de trajectoires multiples et pour laquelle la température devient une variable dynamique.

Au cours d'une simulation, la configuration et la température sont modifiées par le poids de trempe simulée

$$W_{\mathrm{ST}}(E,T) = e^{-\beta E + a(T)}, \tag{2.28}$$

où $a(T)$ est choisie de façon à ce que la distribution de probabilité de la température soit uniforme :

$$P_{\mathrm{ST}}(T) \propto \Omega(E) W_{\mathrm{ST}}(E,T) \equiv \text{const.} \tag{2.29}$$

Le poids de trempe simulée P_{ST} est obtenu par itérations. La température T est discrétisée en M températures et plusieurs simulations sont réalisées à T_m, $m = 1 \ldots M$. Des échanges de conformations sont tentés entre les températures adjacentes.

Une fois que W_{ST} est déterminé, on peut calculer la densité d'états du système par la technique des histogrammes multiples que nous traiterons ultérieurement, et ainsi calculer les moyennes canoniques des observables considérées aux températures voulues.

Une évolution naturelle de cette méthode a été la méthode Monte Carlo d'échange qui s'affranchit du calcul de la fonction $a(T)$.

2.2.2 Monte Carlo d'échange

La méthode Monte Carlo d'échange [13, 14, 15, 16, 17, 18] est aussi connue sous le nom de *replica exchange method (REM)*, *parallel tempering* ou bien encore *multiple markov chain method*.

Initialement conçue pour les verres de spin [17], cette technique a été introduite pour les biomolécules par Hansmann [19] et est très utilisée pour ces systèmes depuis [20, 21, 22, 23, 24, 25, 26]. Cette méthode peut convenir pour répondre aux problématiques de ce travail.

2.2.2.1 Algorithme

La méthode Monte Carlo d'échange consiste à faire évoluer M copies indépendantes (répliques) du système à M températures fixées ($T_1 < \ldots < T_m < \ldots < T_M$).

À une température T_m donnée, une nouvelle configuration (n) est créée à partir de l'ancienne (o). La probabilité d'accepter la nouvelle configuration est donnée par

$$p = \min\left(1, e^{-\beta_m \Delta E}\right), \tag{2.30}$$

avec la température inverse $\beta_m = 1/k_B T_m$ et $\Delta E = E_n - E_o$ la différence d'énergie entre la nouvelle et l'ancienne configuration.

Chaque réplique progresse ainsi de façon indépendante pendant un certain nombre de pas Monte Carlo (en général, quelques dizaines ou centaines de pas) puis l'échange des répliques X_m et X_{m+1}, d'énergies respectives E_m et E_{m+1}, aux températures voisines m et $m+1$ est accepté avec la probabilité :

$$\mathrm{acc}\left[(X_m, \beta_m) \to (X_m, \beta_{m+1})\right] = \min\left[1, e^{-E_m(\beta_{m+1}-\beta_m)}\right]; \tag{2.31}$$

$$\mathrm{acc}\left[(X_{m+1}, \beta_{m+1}) \to (X_{m+1}, \beta_m)\right] = \min\left[1, e^{-E_{m+1}(\beta_m-\beta_{m+1})}\right]. \tag{2.32}$$

Il vient alors que

$$\mathrm{acc}(X_m \rightleftharpoons X_{m+1}) = \min\left(1, e^{\Delta\beta\Delta E}\right), \tag{2.33}$$

avec $\Delta\beta = \beta_{m+1} - \beta_m$ et $\Delta E = E_{m+1} - E_m$.

Le principe de fonctionnement de la méthode Monte Carlo d'échange est résumé dans la figure 2.1. Cette méthode peut être vue comme M simulations Monte Carlo (trajectoires) à M températures distinctes, avec des échanges occasionnels entre elles.

Au cours de la simulation, on enregistre à chaque température T_m l'histogramme des énergies $H_m(E)$ obtenues au cours de la simulation et celui ou ceux des observables considérées $H_m(\mathcal{A})$. En pratique, on enregistre ces histogrammes avec deux dimensions $H_m(E, \mathcal{A})$. Les histogrammes sont incrémentés tous les N degrés de liberté pour limiter les corrélations.

Une simulation canonique à basse température explore les voisinages des minima d'énergie mais peut difficilement sauter les barrières de potentiel [figure 2.2(a)]. À température élevée, les barrières sont facilement franchies mais la simulation ne reste pas suffisamment longtemps dans un minimum pour permettre de l'identifier [figure 2.2(b)]. La méthode Monte Carlo d'échange réalise par contre un va-et-vient entre les basses et les

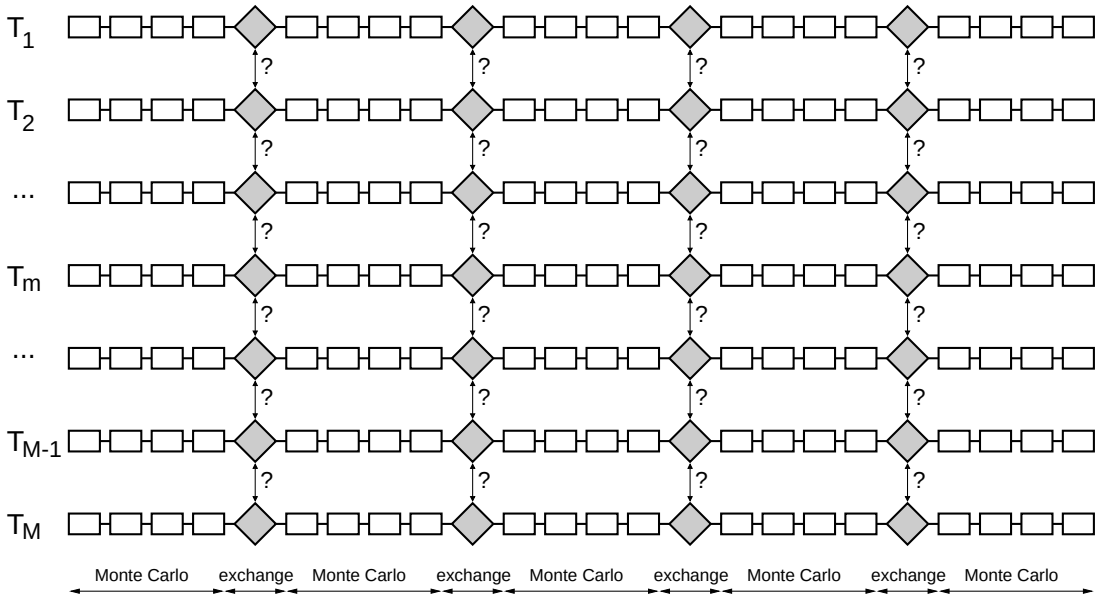


FIG. 2.1 – *Principe de la méthode Monte Carlo d'échange.*

hautes températures via l'échange de répliques. L'espace des phases est ainsi largement parcouru et les minima d'énergie sont identifiables par les trajectoires froides [figure 2.2(c)].

2.2.2.2 Affectation des températures

En Monte Carlo standard, l'efficacité d'un mouvement dépend de la taille du pas de déplacement. En Monte Carlo d'échange, l'efficacité des échanges dépend de l'écart entre les températures de la simulation et peut être quantifiée par le recouvrement entre les distributions d'énergie potentielle.

Comme illustré dans la figure 2.3(a), l'échange des répliques entre deux températures ne se produit pas si les distributions d'énergie ne se recouvrent pas. La simulation se résume alors à M simulations Monte Carlo totalement isolées. Si les recouvrements sont au contraire trop importants [figure 2.3(c)], les échanges tentés sont acceptés trop souvent. Suivant les températures considérées, la simulation reste bloquée dans un minimum local ou n'explore pas les minima locaux. Enfin pour que l'écart entre les températures soit optimal, on considère que la probabilité d'accepter l'échange de répliques entre deux températures adjacentes doit être proche de 50 % permettant ainsi une exploration large et détaillée du paysage énergétique.

Les températures d'une simulation par Monte Carlo d'échange sont souvent distribuées géométriquement [10], de façon à rapprocher les températures les plus basses et ainsi conserver un taux d'échange à peu près constant sur tout l'intervalle de températures considéré [27],

$$T_m = T_{\max} \times \left(\frac{T_{\min}}{T_{\max}}\right)^{\frac{m-1}{M-1}} \qquad \text{avec } 1 \leq m \leq M. \tag{2.34}$$

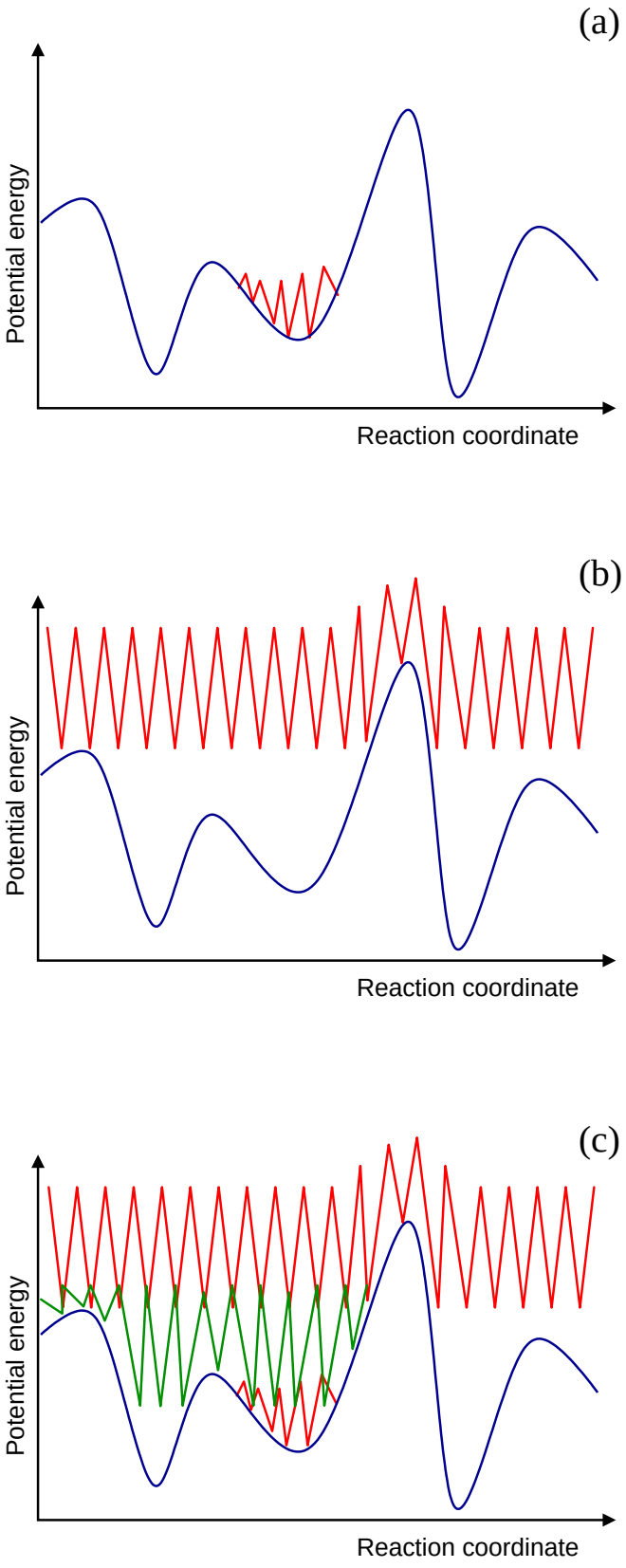


FIG. 2.2 – *Exploration de la surface d'énergie potentielle, (a) à basse et (b) haute température. (c) Avec la méthode Monte Carlo d'échange.*

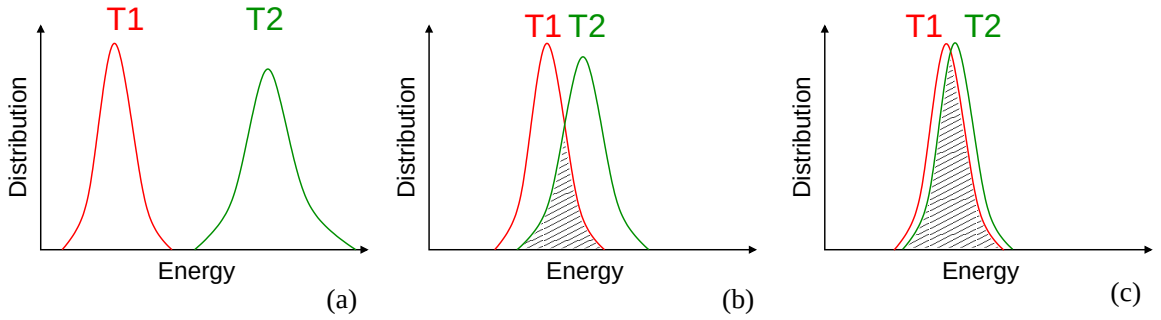


FIG. 2.3 – *Distribution de la probabilité d'énergie pour deux répliques aux températures voisines T_1 et T_2. (a) Recouvrement nul, (b) optimal et (c) trop important.*

Avec T_m la température m, $T_{\min}$ et $T_{\max}$ respectivement les températures minimale et maximale de la simulation.

Cependant, une répartition linéaire des températures peut tout à fait être envisageable et donner également des résultats corrects pour certains systèmes [21].

La figure 2.4 représente les différents modes de distribution des températures, géométrique, linéaire ou hybride. Pour ce travail, nous avons utilisé une distribution à 85 % géométrique et 15 % linéaire, comme utilisé par Dugourd *et al.* dans [28],

$$\begin{aligned} T_m = 0.85 \times T_{\max} \left(\frac{T_{\min}}{T_{\max}}\right)^{\frac{m-1}{M-1}} \\ + 0.15 \left[T_{\max} - \frac{m-1}{M-1}(T_{\max} - T_{\min})\right] \qquad \text{avec } 1 \leq m \leq M. \end{aligned} \tag{2.35}$$

La figure 2.5 rassemble, à titre d'exemple, les distributions de probabilité de l'énergie à toutes les températures d'une simulation Monte Carlo d'échange du peptide Ala_4. Aux températures les plus faibles pour lesquelles les distributions d'énergie sont les plus piquées, les températures sont plus resserrées pour que les recouvrements de distribution soient à peu près constants.

Chaque réplique du système évolue à une température particulière. Le Monte Carlo d'échange est naturellement parallélisable en affectant une température à chaque nœud de calcul [19]. En pratique, et pour un nombre de processeurs disponibles limités, on cherche à optimiser la grille de températures. Pour cela, il peut être intéressant d'attribuer plusieurs températures par processeur [29].

Toutes nos simulations Monte Carlo d'échange ont été effectuées avec un programme C++ développé dans l'équipe. La librairie MPICH/MPI [30, 31] a été employée pour paralléliser le code.

Deux stratégies existent pour les échanges. Soit on échange les répliques à température fixée, soit l'inverse. Dans la mesure où, pour chaque température, on stocke également les statistiques des propriétés physiques du système, il s'avère plus économique d'échanger les configurations entre les trajectoires.

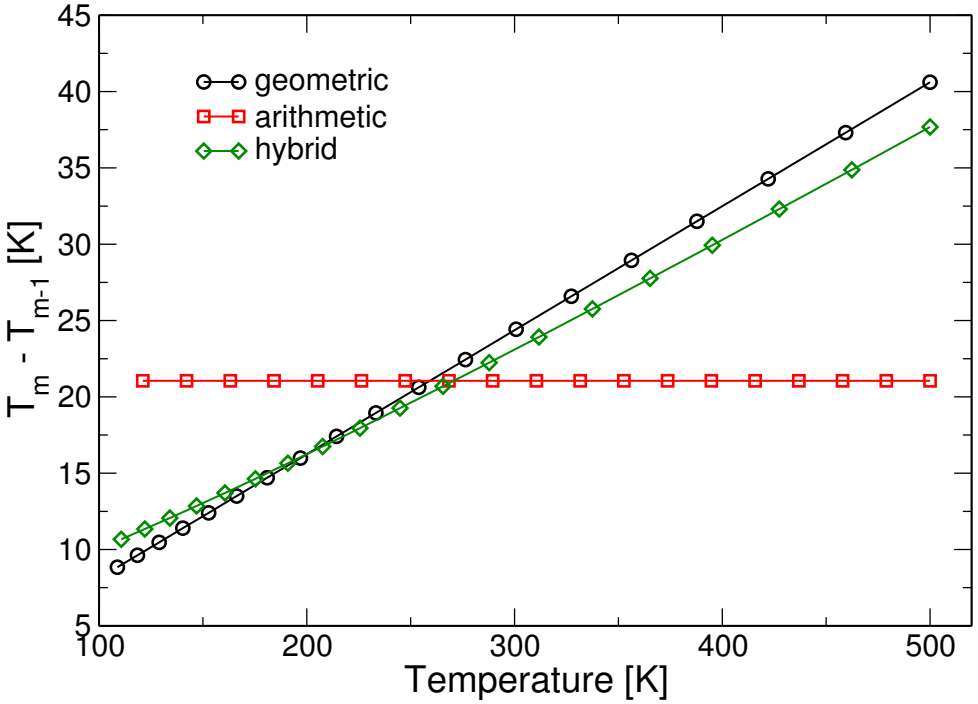


FIG. 2.4 – *Distribution de la température pour 20 trajectoires entre 100 et 500 K suivant une progression linéaire, géométrique ou hybride. La progression hybride utilisée pour ce travail est à 85 % géométrique et 15 % linéaire.*

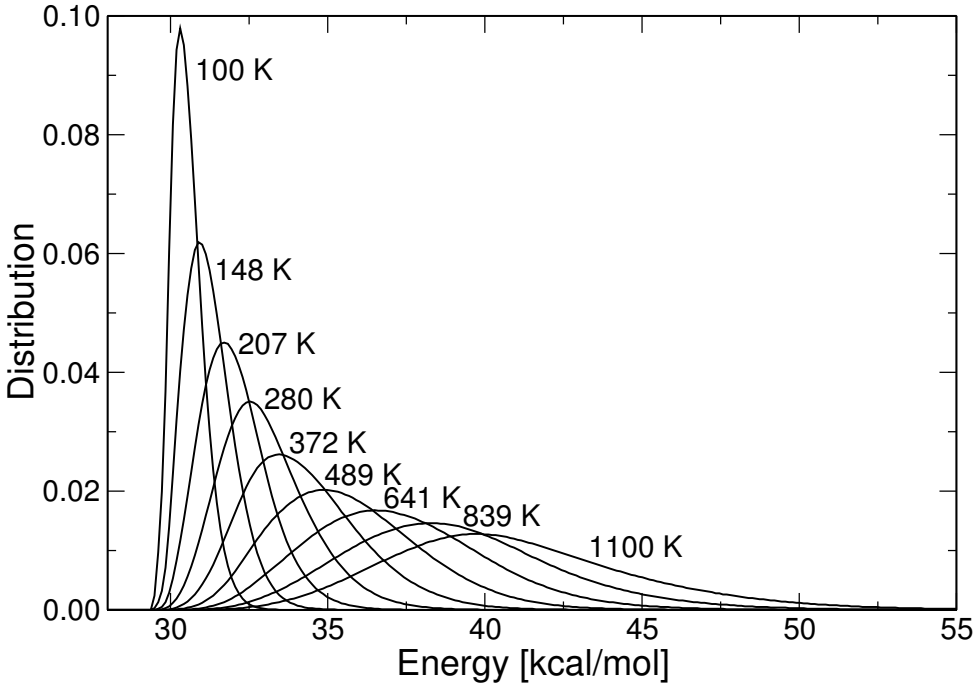


FIG. 2.5 – *Distribution de la probabilité d'énergie pour une simulation Monte Carlo d'échange du peptide Ala_4 comptant 9 températures entre 100 et 1100 K.*

Le nombre de températures à affecter à la simulation pour que le recouvrement des distributions d'énergie soit à peu près constant est proportionnel à la racine carrée du nombre de degrés de liberté. Cette évolution a d'abord été observée par Hukushima et Nemoto [17] puis par Predescu *et al.* [32]. Si la simulation Monte Carlo d'échange est relativement simple à mettre en œuvre pour des systèmes de taille réduite, elle peut devenir problématique pour de grands systèmes. On doit en effet allouer un nombre important de températures à la simulation ce qui nécessite d'importants moyens de calcul parallélisés.

Enfin, une allocation optimale des températures et donc un bon recouvrement des histogrammes d'énergie ne suffit parfois pas à garantir la convergence d'une simulation Monte Carlo d'échange. Il est alors instructif de suivre l'évolution des répliques échangées entre les températures.

2.2.2.3 Stratégie d'échange

Les échanges de répliques entre les différentes températures sont cruciaux dans la méthode de Monte Carlo d'échange. La probabilité d'accepter l'échange entre deux répliques à deux températures différentes est donnée par l'équation (2.33). Cette probabilité décroît exponentiellement avec la différence de températures inverses [22], ce qui fait qu'en pratique, on tente des échanges entre les répliques dont les températures sont voisines [27]. Ainsi, chaque réplique évolue indépendamment pendant un certain nombre de pas Monte Carlo puis un échange est essayé. Une température est ensuite choisie au hasard puis on tente l'échange de répliques avec une des températures adjacentes. Une stratégie alternative d'échange a été proposée par Calvo [33]. Elle consiste à envisager tous les échanges possibles entre toutes les conformations puis d'en sélectionner une en fonction de sa probabilité de succès.

2.2.2.4 Performance et convergence de la méthode

Taux d'acceptance Monte Carlo

Dans une simulation Monte Carlo d'échange, la majorité du temps de calcul est utilisée pour propager chaque réplique à sa température. Il est donc important de s'intéresser en premier lieu aux taux d'acceptance Monte Carlo pour chaque température. Le tableau 2.1 représente par exemple les taux d'acceptance obtenus pour une simulation du dipeptide WG. Dans le cas présent, les taux d'acceptance sont supérieurs à 60 %, les pas de déplacement des mouvements Monte Carlo sont donc un peu trop petits.

Taux d'acceptance des échanges entre températures

Un critère de performance souvent considéré est le taux d'acceptance des échanges entre les températures adjacentes. Mitsutake *et al.* [10] recommandent qu'il soit uniforme et suffisamment grand, c'est-à-dire supérieur à 20 % [34]. Pour un système modèle, Predescu

TAB. 2.1 – *Taux d'acceptance Monte Carlo à chaque température. Cas du dipeptide WG avec 9 températures comprises entre 100 K et 1100 K.*

Température [K]	Taux d'acceptance [%]
100	78
148	76
207	74
280	72
372	71
489	69
641	68
839	66
1100	64

et al. [35] montrent que le taux d'acceptance optimal est proche de 39 % et qu'au-delà, les performances ne sont plus améliorées.

En pratique, lorsqu'on obtient un taux d'échange entre répliques trop faible, on augmente alors le nombre de répliques pour le même intervalle de températures. Si les ressources de calcul ne le permettent pas, on réduit alors l'intervalle de températures.

Le tableau 2.2 donne les taux d'acceptance des échanges dans le cas d'une simulation Monte Carlo d'échange avec 9 températures. Les valeurs obtenues sont bien supérieures à 20 % et relativement uniformes. Les échanges sont plus nombreux entre les températures élevées.

TAB. 2.2 – *Taux d'acceptance des échanges entre températures adjacentes. Simulation réalisée pour le dipeptide WG avec 9 températures comprises entre 100 K et 1100 K [36].*

Paires de températures [K]	Taux d'acceptance [%]
$100 \leftrightarrow 148$	47
$148 \leftrightarrow 207$	50
$207 \leftrightarrow 280$	56
$280 \leftrightarrow 372$	60
$372 \leftrightarrow 489$	62
$489 \leftrightarrow 641$	65
$641 \leftrightarrow 839$	68
$839 \leftrightarrow 1100$	70

Des taux d'acceptance élevés sont nécessaires pour assurer que les statistiques accumulées soient correctes, mais ils ne sont cependant pas toujours suffisants pour garantir la convergence globale de la simulation.

Évolution temporelle des échanges

Une simulation par Monte Carlo d'échange doit permettre aux conformations à haute température de communiquer avec des températures plus basses et inversement. Afin de s'en assurer, on peut suivre l'évolution des répliques pour une température donnée (figure 2.6). Réciproquement, on peut aussi suivre l'évolution de la température adoptée par une réplique particulière (figure 2.7). Dans les exemples présentés, l'échantillonnage a été efficace. Une réplique précise passe bien par toutes les températures de la simulation, ceci afin que les conformations les plus stables puissent être explorées et que le franchissement de barrières soit possible. Pour être rigoureux, il faut vérifier que toutes les températures rencontrent bien toutes les répliques.

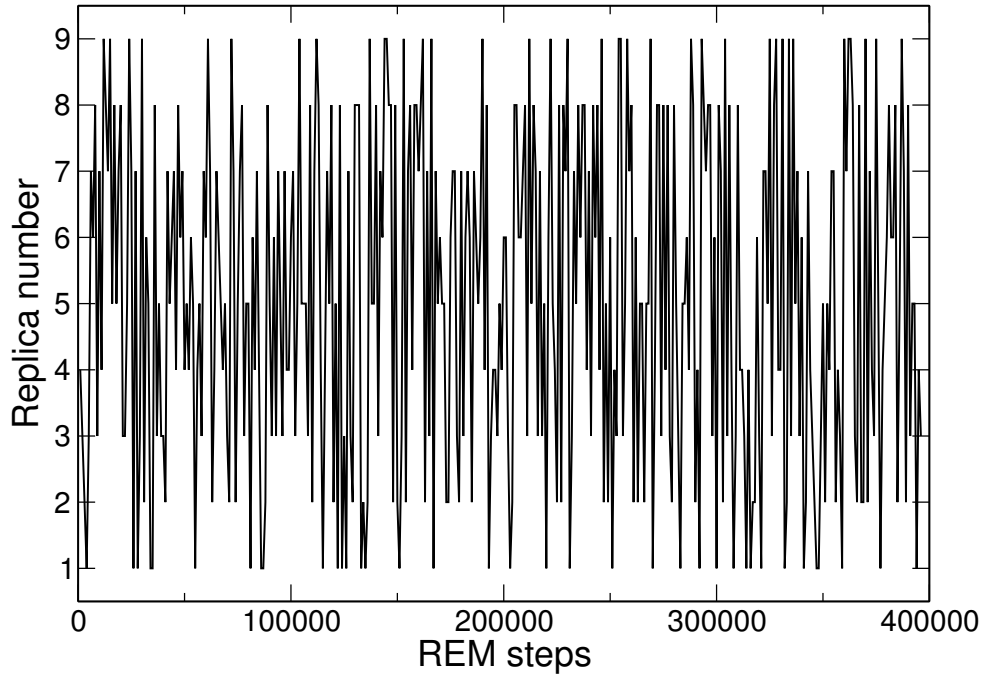


FIG. 2.6 – *Simulation Monte Carlo d'échange du dipeptide WG. Évolution des répliques rencontrées à la température de 280 K.*

Enfin, l'évolution de grandeurs physiques à une température apporte des informations sur la convergence de la simulation. Pour les biomolécules, on peut s'intéresser à la moyenne de l'énergie (figure 2.8) et à celle du dipôle électrique (figure 2.9). Les moyennes présentées se stabilisent rapidement et convergent vers la valeur d'équilibre. Dans le cas présent, on peut en déduire que l'échantillonnage est efficace et que le temps de la simulation est suffisant.

2.2.3 Calcul des moyennes canoniques

Le calcul de la moyenne canonique d'une observable à partir d'une simulation Monte Carlo d'échange se fait par simple accumulation d'une moyenne si la température considérée est une des températures de la simulation. Or, on peut être intéressé par des proprié-

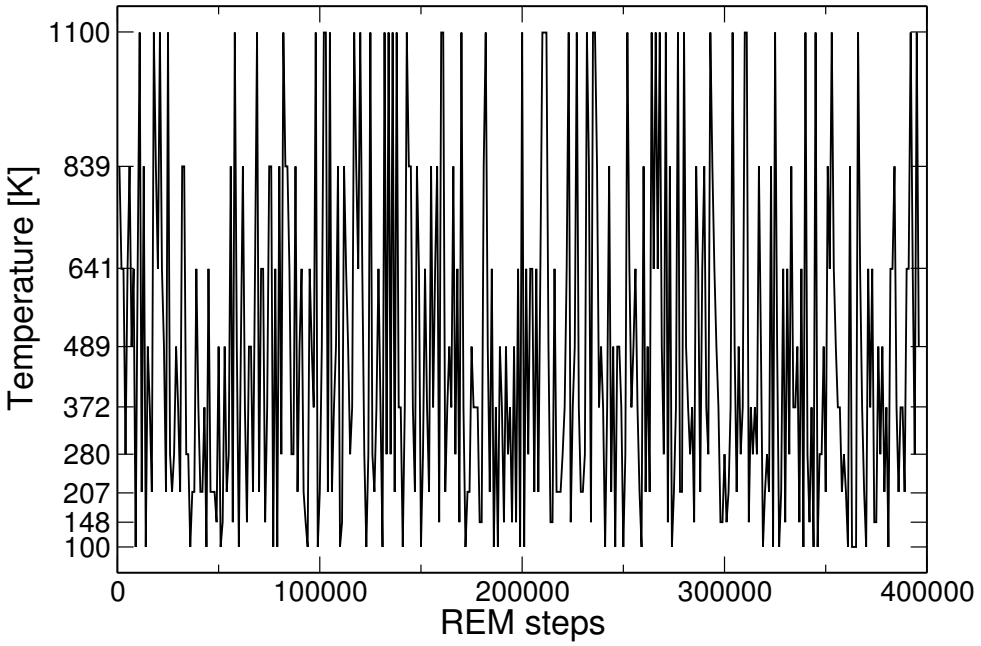


FIG. 2.7 – *Simulation Monte Carlo d'échange du dipeptide WG. Évolution des températures rencontrées par la réplique 1.*

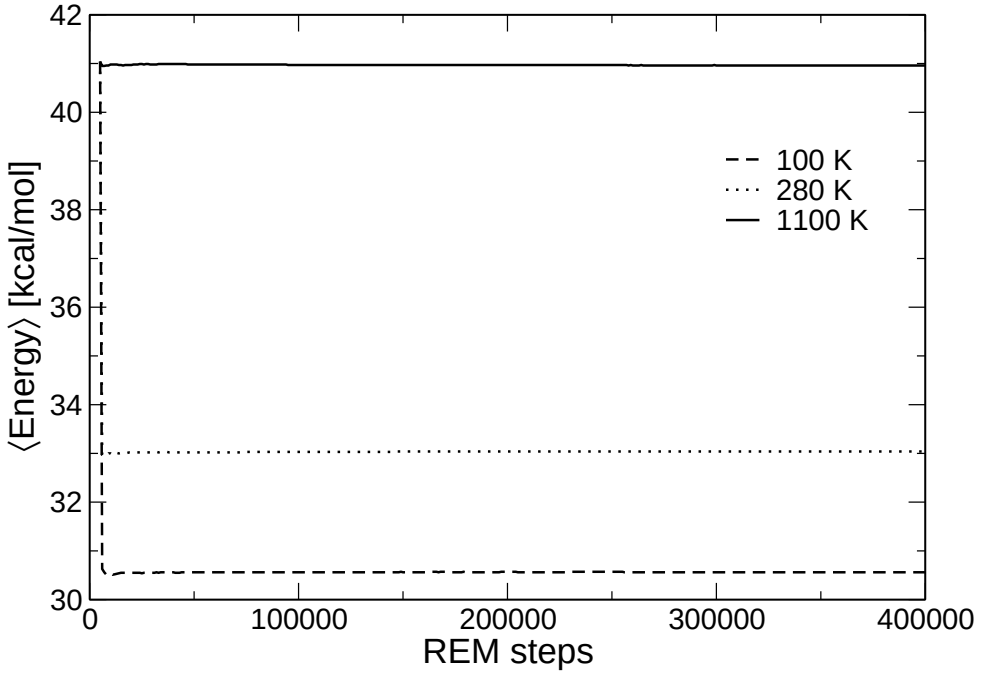


FIG. 2.8 – *Simulation Monte Carlo d'échange du dipeptide WG avec 9 températures comprises entre 100 K et 1100 K. Évolution de l'énergie moyenne aux températures de 100 K, 280 K et 1100 K.*

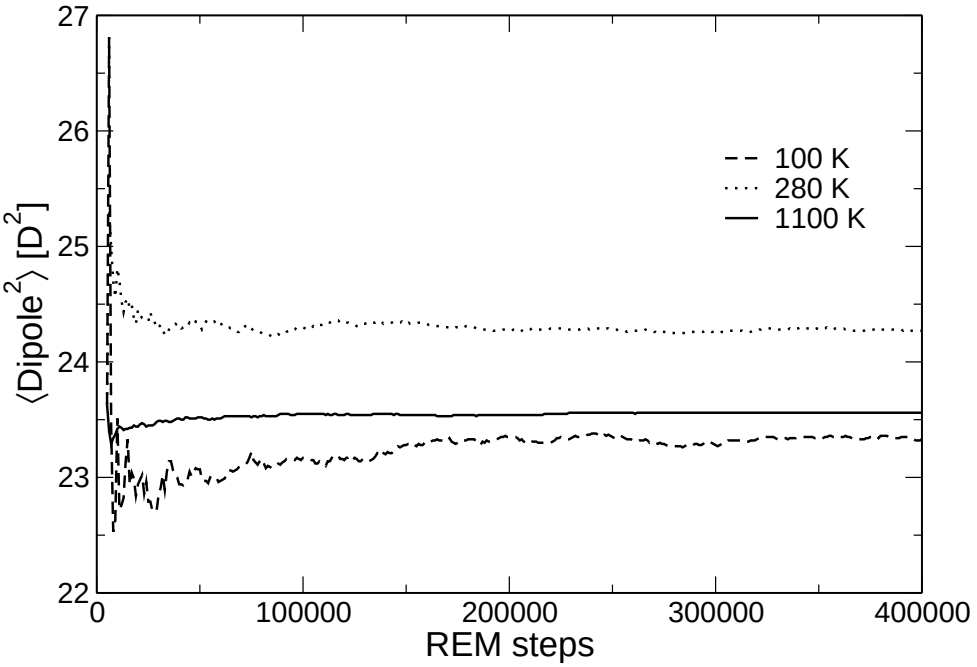


FIG. 2.9 – *Simulation Monte Carlo d'échange du dipeptide WG avec 9 températures comprises entre 100 K et 1100 K. Évolution de la moyenne du dipôle électrique aux températures de 100 K, 280 K et 1100 K.*

tés statistiques à des températures intermédiaires. La méthode des histogrammes multiples [37, 38] permet précisément d'interpoler des moyennes d'observables à partir de plusieurs simulations réalisées à différentes températures. Elle fournit également une estimation de la densité d'états.

2.2.3.1 Moyennes canoniques à une température de simulation

La moyenne d'une observable $\mathcal{A}$ à la température de simulation T_m est calculée directement à partir de N mesures de $\mathcal{A}$:

$$\langle \mathcal{A} \rangle_{T_m} = \frac{1}{N} \sum_i \mathcal{A}_i \tag{2.36}$$

2.2.3.2 Méthode des histogrammes multiples

Si on souhaite calculer la moyenne d'une observable à une température différente des températures de simulation Monte Carlo d'échange, on peut utiliser la méthode des histogrammes multiples ou *weighted histogram analysis method* (WHAM) [37, 38] qui a été développée par Ferrenberg et Swendsen [37]. Pour une simulation Monte Carlo d'échange à M températures différentes, on calcule par exemple la moyenne de l'énergie à n'importe

quelle température inverse β, où β n'est pas nécessairement égal à l'un des β_m, par

$$\langle E \rangle = \frac{\sum_E E P(E,\beta)}{\sum_E P(E,\beta)}, \tag{2.37}$$

avec $P(E,\beta)$ qui représente la probabilité d'occuper l'énergie E à la température inverse $\beta = 1/k_B T$. La probabilité $P(E,\beta)$ peut être déterminée par la technique des histogrammes multiples qui consiste à résoudre de manière auto-itérative le système

$$P(E,\beta) = \frac{\sum_m k_m^{-1} H_m(E) e^{-\beta E}}{\sum_m h_m k_m^{-1} e^{f_m - \beta_m E}}; \tag{2.38}$$

$$e^{-f_m} = \sum_E P(E,\beta_m), \tag{2.39}$$

avec f_m qui représente l'énergie libre du système à la température T_m et $k_m = 1$ si les configurations successives rencontrées lors de la simulation Monte Carlo sont indépendantes [9, 37]. $H_m(E)$ est l'histogramme des énergies pour la température m et h_m est le nombre de mesures effectuées à chaque température, autrement dit, le nombre de points contenus dans l'histogramme précédent :

$$h_m = \sum_E H_m(E). \tag{2.40}$$

Par itération, on calcule $P(E,\beta)$ et f_m en initialisant f_m à 0 [38]. La convergence s'obtient en général rapidement en quelques dizaines d'itérations. En outre,

$$P(E,\beta) \propto \Omega(E) e^{-\beta E}. \tag{2.41}$$

$\Omega(E)$ est donc la densité d'états obtenue directement en utilisant l'équation (2.38) et avec $k_m = 1$:

$$\Omega(E) = \frac{\sum_m H_m(E)}{\sum_m h_m e^{f_m - \beta_m E}}. \tag{2.42}$$

La moyenne canonique d'une observable autre que l'énergie à la température inverse β s'exprime par

$$\langle \mathcal{A} \rangle = \frac{\sum_{\mathcal{A},E} \mathcal{A} P(E,\mathcal{A},\beta)}{\sum_{\mathcal{A},E} P(E,\mathcal{A},\beta)}. \tag{2.43}$$

où $P(E, \mathcal{A}, \beta)$ est donnée par

$$P(E, \mathcal{A}, \beta) = \frac{\sum_m H_m(E, \mathcal{A}) e^{-\beta E}}{\sum_m h_m e^{f_m - \beta_m E}}. \tag{2.44}$$

Notons que la connaissance de $H_m(E, \mathcal{A})$ donne directement accès aux moyennes canoniques des températures T_m de simulation. La moyenne canonique de l'observable $\mathcal{A}$ est alors

$$\langle \mathcal{A} \rangle_{T_m} = \frac{\sum_E \sum_{\mathcal{A}} \mathcal{A} H_m(E, \mathcal{A})}{\sum_E \sum_{\mathcal{A}} H_m(E, \mathcal{A})}. \tag{2.45}$$

2.2.3.3 Illustration de la repondération par histogrammes multiples

Pour illustrer la méthode de calcul de moyennes canoniques par les histogrammes multiples, nous avons enregistré à chaque température d'une simulation Monte Carlo d'échange un histogramme à deux dimensions, énergie et carré du dipôle électrique, $H_m(E, \mu^2)$. Cette observable joue un rôle important dans la compréhension de la structure des peptides qui nous intéressent. La figure 2.10 représente la moyenne du carré du dipôle électrique ($\mathcal{A} = \mu^2$) en fonction de la température. Les valeurs aux températures de simulation sont obtenues directement avec l'équation (2.45). Les moyennes aux températures intermédiaires sont calculées par la méthode des histogrammes multiples. Les moyennes obtenues par interpolation se superposent parfaitement à celles déterminées pour chaque température de simulation.

2.3 Échantillonnage non-boltzmannien : cas de la méthode Wang-Landau

Une seconde approche pour explorer efficacement le paysage énergétique des biomolécules est l'emploi d'un échantillonnage non-boltzmannien. En particulier, les algorithmes multicanoniques utilisent un tel échantillonnage pour obtenir une probabilité $P(E)$ constante et ainsi effectuer une marche uniforme de la surface des énergies potentielles.

2.3.1 Ensemble multicanonique

La méthode multicanonique a longtemps été la méthode la plus utilisée dans les ensembles généralisés. Elle est aussi appelée *entropic sampling*. L'ensemble multicanonique est largement utilisé pour les verres de spin [39] et les protéines [40].

Contrairement à l'ensemble canonique, une simulation dans l'ensemble multicanonique pondère chaque état du système par un facteur $W_{\mathrm{mu}}(E)$ différent du poids de Boltzmann

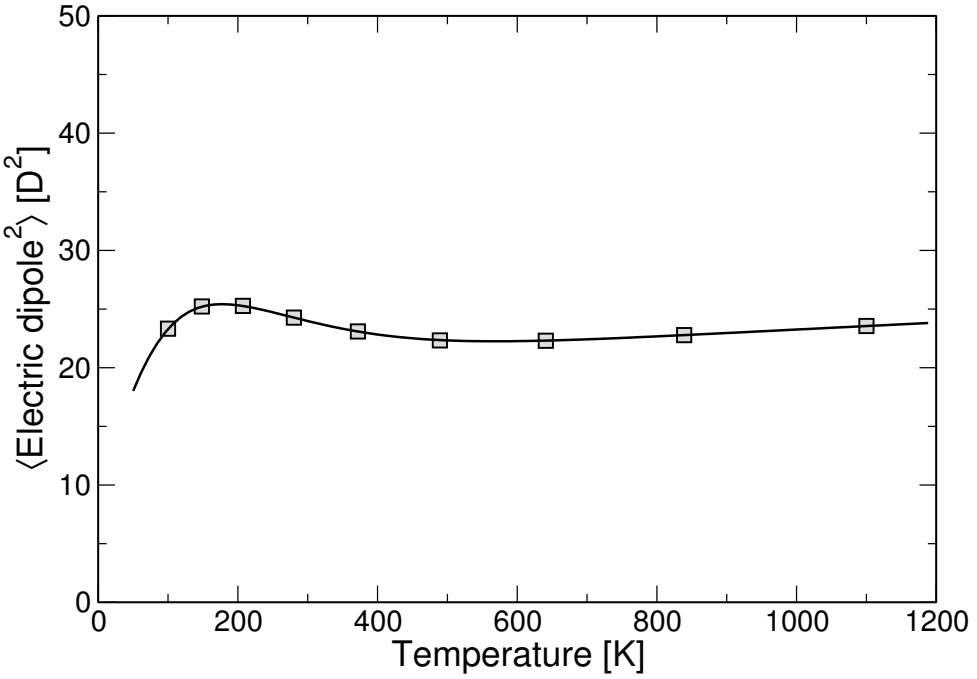


FIG. 2.10 – *Évolution de la moyenne thermique du carré du dipôle pour le dipeptide WG. Les carrés représentent les moyennes obtenues aux températures de la simulation Monte Carlo d'échange. Le trait plein correspond aux moyennes obtenues par la méthode des histogrammes multiples.*

de telle façon que la distribution en énergie soit équiprobable,

$$P_{\mathrm{mu}}(E) \propto \Omega(E) W_{\mathrm{mu}}(E) = \mathrm{const.} \tag{2.46}$$

Comme illustré sur la figure 2.11, cette distribution uniforme assure une marche aléatoire sur la surface d'énergie potentielle. La densité d'états $\Omega(E)$ du système n'est généralement pas connue à l'avance. Le poids multicanonique $W_{\mathrm{mu}}(E)$ doit donc être calculé numériquement et par itérations [39, 41]. Pour ce faire, on initialise le poids multicanonique avec une distribution de Boltzmann à une température élevée (β_0 faible),

$$W^0_{\mathrm{mu}}(E) = e^{-\beta_0 E}. \tag{2.47}$$

La première itération produit une distribution d'énergie $P^1_{\mathrm{mu}}(E)$. La nouvelle densité d'états $\Omega^1(E)$ est calculée comme

$$\Omega^1(E) = \frac{P^1_{\mathrm{mu}}(E)}{W^0_{\mathrm{mu}}(E)}, \tag{2.48}$$

et le nouveau poids multicanonique devient alors

$$W^1_{\mathrm{mu}}(E) = \frac{1}{\Omega^1(E)} = \frac{W^0_{\mathrm{mu}}(E)}{P^1_{\mathrm{mu}}(E)}. \tag{2.49}$$

On procède ainsi par itérations, jusqu'à obtenir $W_{\mathrm{mu}}(E)$ avec la précision voulue. On

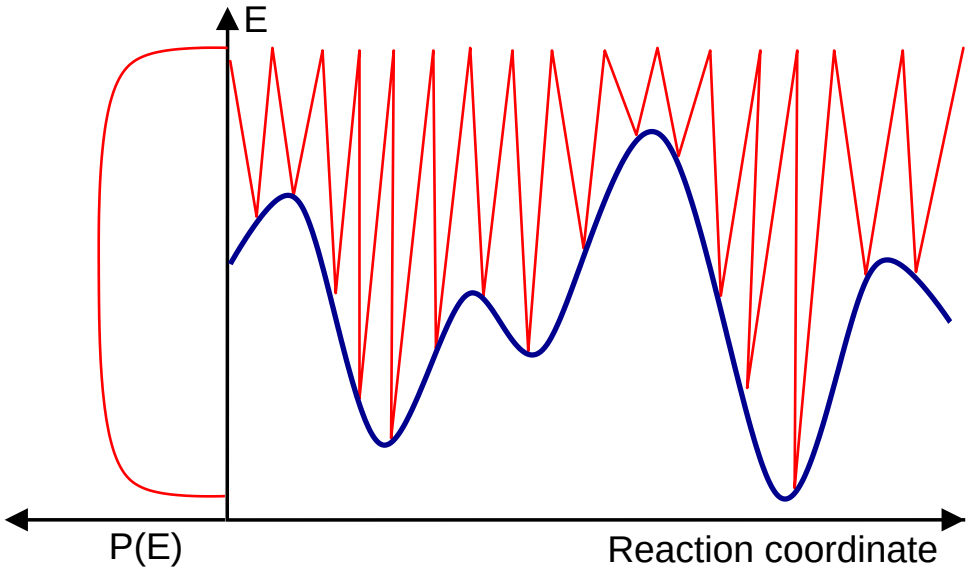


FIG. 2.11 – *Marche aléatoire sur la surface d'énergie potentielle, assurée par la distribution d'énergie équiprobable.*

effectue ensuite une dernière simulation multicanonique pour accumuler suffisamment de statistiques, notamment l'histogramme de la distribution d'énergie potentielle $H_{\mathrm{mu}}(E)$ et la distribution de l'observable voulue $\mathcal{A}$ en fonction de l'énergie, $\mathcal{A}$(E). La densité d'états $\Omega(E)$ est obtenue par une repondération et l'application de l'équation (2.46) :

$$\Omega(E) = \frac{H_{\mathrm{mu}}(E)}{W_{\mathrm{mu}}(E)}. \tag{2.50}$$

La moyenne canonique de $\mathcal{A}$ à la température T est alors obtenue par application de l'équation (2.18), soit

$$\langle A \rangle_T = \frac{\sum_E \mathcal{A}(E) \frac{H_{mu}(E)}{W_{mu}(E)} e^{-\beta E}}{\sum_E \frac{H_{mu}(E)}{W_{mu}(E)} e^{-\beta E}}. \tag{2.51}$$

Comme nous l'avons vu, la méthode multicanonique impose le calcul d'un facteur de pondération par itérations, ce qui est souvent long et fastidieux. Une méthode plus récente comme la méthode Wang-Landau permet un calcul plus aisé des poids multicanoniques.

2.3.2 Méthode Wang-Landau

La méthode Wang-Landau est similaire à la méthode multicanonique décrite plus haut, mais la manière itérative dont les poids sont déterminés diffère.

Elle a été développée par Wang et Landau en 2001 [42, 43] et est en constante amélioration depuis. L'élégance et l'apparente simplicité de cette méthode ont conduit à un large spectre d'utilisations, comme les systèmes magnétiques [42, 43, 44, 45, 46], les cristaux

liquides [47], les fluides [48, 49], les polymères [50, 51, 52], les biomolécules [53, 54, 55], les agrégats atomiques [56, 57] ou les verres structuraux [58] et verres de spins [59, 60].

2.3.2.1 Généralités

L'algorithme Wang-Landau tente d'estimer avec précision la densité d'états microcanonique $\Omega(E)$ en réalisant une marche aléatoire sur l'espace des énergies potentielles et en pénalisant les états au fur et à mesure qu'ils sont visités. La densité d'états $g(E)$ estimée est discrétisée sur N énergies comprises entre $E_{\min}$ et $E_{\max}$. Au début de la simulation, la densité d'états n'est pas connue et elle est arbitrairement initialisée à 1 pour toutes les énergies. On désigne par $H(E)$ l'histogramme des énergies visitées au cours d'une itération. Un critère d'uniformité (discuté ultérieurement) permet de déterminer si cet histogramme est plat et donc si la simulation a convergé. Une fois ce critère vérifié, on procède à une autre itération pour affiner la construction de la densité d'états.

Une itération Wang-Landau est constituée d'un certain nombre de pas Monte Carlo. La probabilité d'accepter un mouvement Monte Carlo entre l'ancienne configuration (o) et la nouvelle (n) d'énergies respectives E_o et E_n est donnée par

$$\mathrm{acc}(o \to n) = \min\left(1, \frac{g(E_o)}{g(E_n)}\right) = \min\left(1, e^{s(E_o)-s(E_n)}\right). \tag{2.52}$$

Pratiquement, on utilise $s(E) = \ln g(E)$, grandeur similaire à une entropie, pour éviter de manipuler des nombres trop importants.

Si le mouvement est accepté, on garde, comme en Monte Carlo Metropolis usuel, la nouvelle configuration. Si le mouvement est rejeté, l'ancienne configuration devient la nouvelle. Après chaque pas Monte Carlo, la densité d'états de la nouvelle configuration est multipliée par un facteur de modification f supérieur à 1

$$g(E) \times f \to g(E), \tag{2.53}$$

et donc l'entropie est augmentée de $\ln f$,

$$s(E) + \ln f \to s(E). \tag{2.54}$$

L'histogramme des énergies visitées est quant à lui incrémenté de 1,

$$H(E) + 1 \to H(E). \tag{2.55}$$

Le facteur de modification f est un des éléments cruciaux d'une simulation Wang-Landau car il participe à la construction progressive de g. Au début de la simulation, f prend une valeur arbitraire, généralement e^1, de façon à atteindre rapidement toutes les énergies physiquement accessibles [43]. Régulièrement, un critère d'uniformité est testé et,

s'il est vérifié, f est diminué exponentiellement, en prenant par exemple

$$\sqrt{f} \to f. \tag{2.56}$$

On recommence alors une itération avec la nouvelle valeur de f en partant de la densité d'états obtenue à l'itération précédente. Ce principe de fonctionnement est résumé dans la figure 2.12.

Au tout début de la simulation Wang-Landau, le bilan détaillé n'est pas rigoureusement respecté : la probabilité de passer de l'état o à l'état n n'est pas la même que celle de passer de n à o. Cependant, au cours de la simulation, f diminue exponentiellement et les modifications de $g(E)$ sont de plus en plus petites, jusqu'à être négligeables en regard de $g(E)$. Le bilan détaillé est alors à nouveau respecté [61].

La simulation se termine lorsque $f-1$ (ou $\ln f$) atteint une valeur suffisamment petite (par exemple, $f-1 < 10^{-7}$). La densité d'états obtenue est alors connue avec une précision de l'ordre de f.

À titre d'exemple, la figure 2.13 représente l'évolution du logarithme de la densité d'états au cours d'une simulation Wang-Landau pour le peptide Ala_4. La construction de la densité d'états se fait par incrémentations successives de $\ln g$. L'exploration de la surface d'énergie se fait progressivement au cours de la simulation, les énergies les plus élevées étant manifestement visitées en premier. La région pour laquelle $\ln g$ est nul correspond aux énergies non encore visitées ou aux énergies physiquement non accessibles pour le système étudié.

L'histogramme complet de $\ln g(E)$ après la dernière itération est présenté dans la figure 2.14. La densité d'états (comme son logarithme) croît naturellement avec l'énergie. Elle peut s'étaler jusqu'à une vingtaine d'ordres de grandeur. Les variations d'entropie sont particulièrement sensibles à basse énergie.

2.3.2.2 Critères de convergence

La convergence de g entre deux itérations successives est suivie en vérifiant périodiquement l'uniformité de l'histogramme des énergies visitées $H(E)$. Dans l'algorithme original [42, 43], l'histogramme H est considéré comme uniforme si toutes les cases des énergies accessibles sont peuplées par au moins 95 % du nombre moyen de visites. Ce critère apparaît cependant trop restrictif. Shell *et al.* [61] ont alors proposé qu'un nombre minimal de visites soit atteint dans chaque case d'énergie, pour garantir que chaque point de la densité d'états ait été visité. Enfin, on peut également se fixer, en plus de ce critère, un nombre de pas maximal par itération au-delà duquel l'itération en cours est arrêtée.

La figure 2.15 représente les histogrammes des énergies visitées (a, c et e) et du logarithme de la densité d'états (b, d et f) pour, respectivement, la 1re, la 10e et la dernière (19e) itération d'une simulation Wang-Landau. Un nombre maximal global de pas Monte

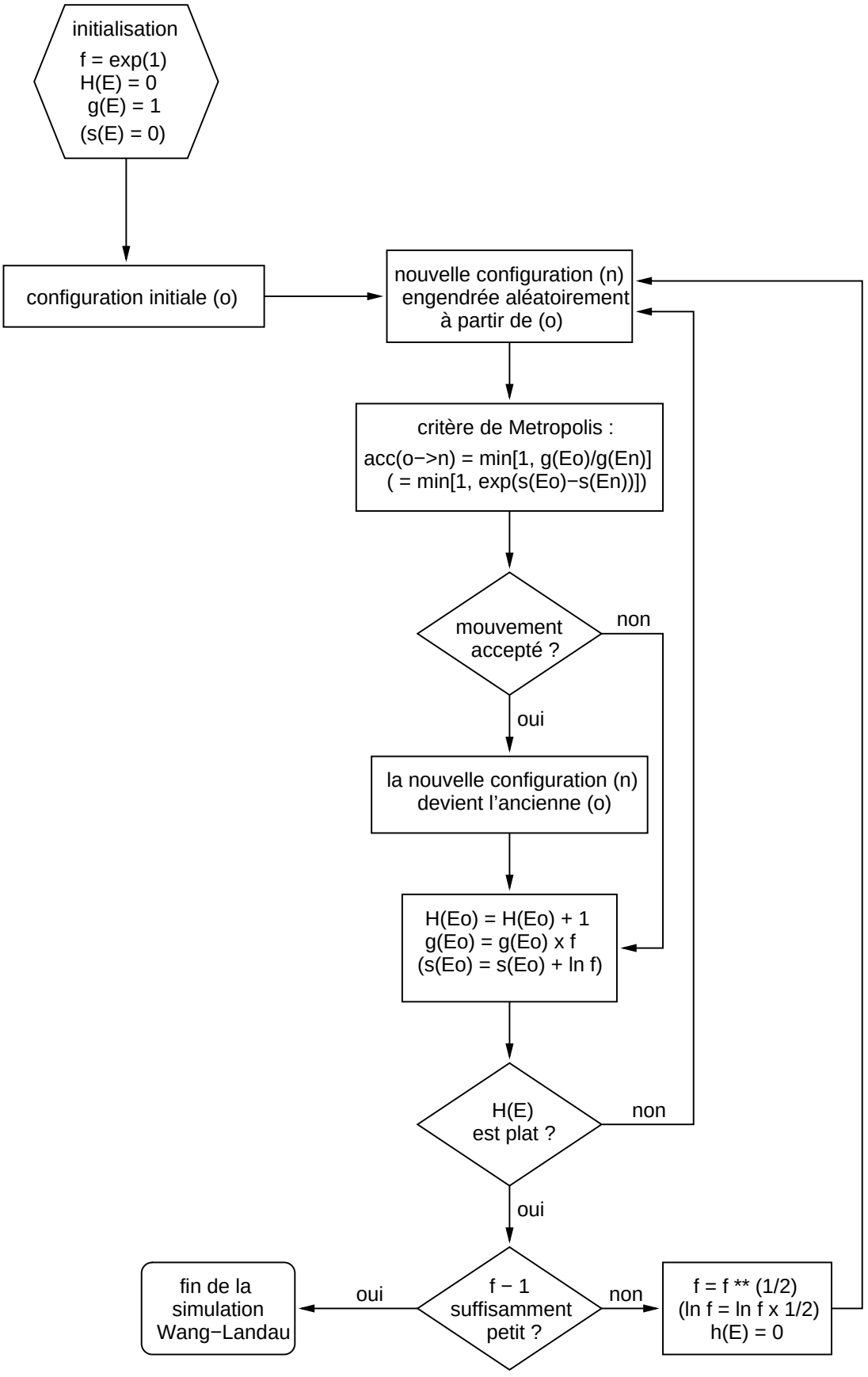


FIG. 2.12 – *Principe de l'algorithme Wang-Landau.*

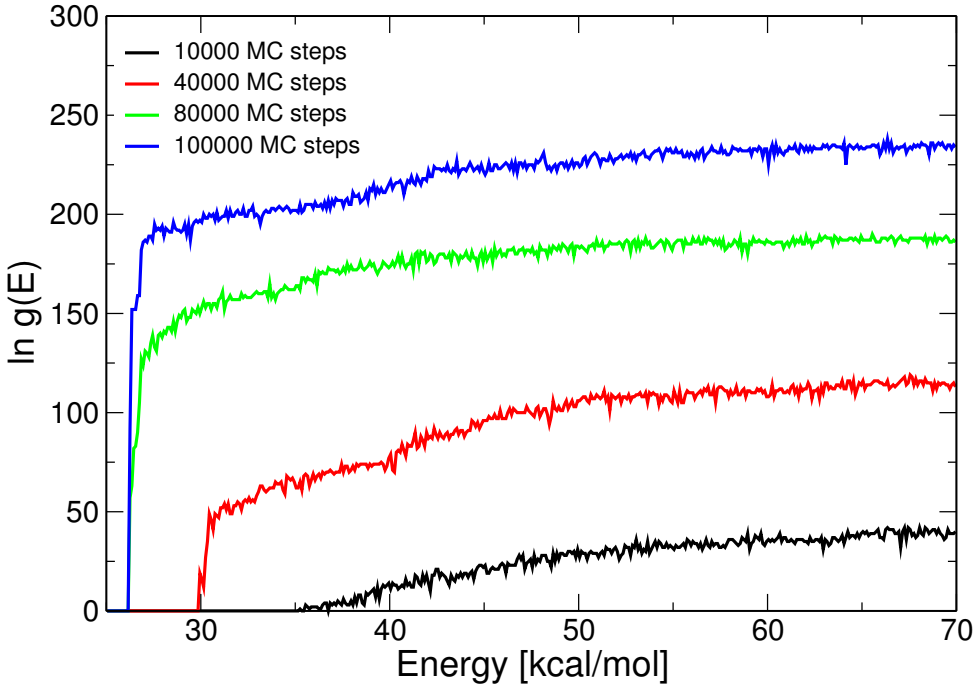


FIG. 2.13 – *Évolution du logarithme de la densité d'états du peptide Ala_4 au cours d'une simulation Wang-Landau.*

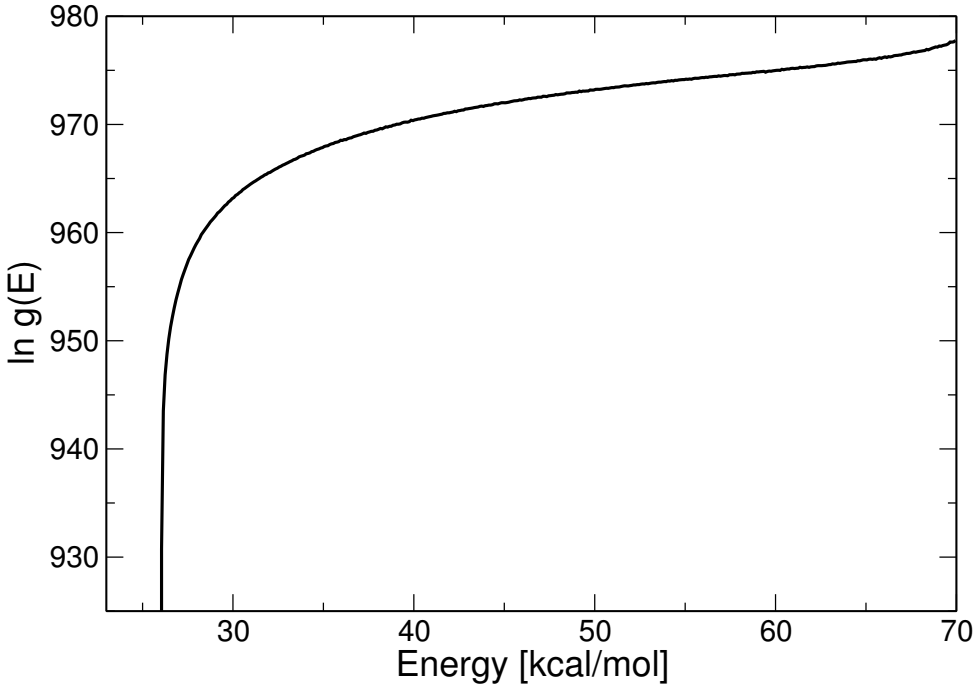


FIG. 2.14 – *Logarithme de la densité d'états finale du peptide Ala_4 après une simulation Wang-Landau.*

Carlo pour chaque itération permet d'avoir des histogrammes suffisamment plats pour que la simulation converge. À mesure que le facteur de modification f diminue, de légères fluctuations peuvent apparaître à la frontière entre les zones physiques et les zones non-physiques. La construction de la densité d'états se fait grossièrement lors des premières itérations [figure 2.15 (b)], et plus finement avec les itérations suivantes [figure 2.15 (d et f)].

2.3.2.3 Stratégies alternatives et améliorations possibles

Afin de tirer parti des ressources de calcul parallélisées, une des premières améliorations proposées [43, 61, 62] pour l'algorithme Wang-Landau a consisté à diviser l'intervalle d'énergie en plusieurs sous-domaines puis à réaliser une simulation Wang-Landau sur chaque fenêtre d'énergie. La densité d'états globale du système est obtenue en ajustant *a posteriori* les densités d'états calculées pour chaque domaine d'énergie. Cette stratégie convient effectivement bien au calcul parallèle mais ne s'avère pas pertinente si d'importantes barrières énergétiques dans une fenêtre empêchent le système d'atteindre l'équilibre dans les fenêtres adjacentes [61].

Une deuxième amélioration prend en compte les mouvements Monte Carlo dont l'énergie tombe en dehors de l'intervalle considéré [63]. Schultz *et al.* [63] proposent de modifier les histogrammes g et H correspondant à la configuration rejetée. Cette modification permet d'éviter certains effets de bord observables aux limites de l'intervalle d'énergie.

Enfin, une amélioration plus qualitative a été proposée par Zhou et Bhatt (ZB) [64]. Ces auteurs ont montré que la convergence de g vers la vraie densité d'états $\Omega(E)$ se fait avec un certaine erreur statistique qui, en pratique, peut être réduite en réalisant au minimum K pas Monte Carlo pour chaque case, avec

$$K \propto \frac{1}{\sqrt{\ln f}}. \tag{2.57}$$

2.3.3 Simulation de systèmes à énergie continue et phénomènes d'accumulation

La simulation de systèmes dont les degrés de liberté évoluent continuement peut être complexe, surtout lorsque le paysage énergétique est rugueux [65]. Certaines régions de l'espace des configurations sont très étroites et l'algorithme Wang-Landau peut alors rencontrer des difficultés à explorer ces régions. Si ces domaines d'énergie sont découverts tardivement au cours de la simulation, leur densité d'états sera alors très petite comparée à celles des autres régions de la surface d'énergie. L'équation (2.52) montre que les mouvements tentés pour sortir de ces trous entropiques seront presque toujours tous rejetés. Le système reste alors bloqué dans ces régions jusqu'à ce que la densité d'états associée soit comparable à la densité d'états des cases voisines de la surface d'énergie. Il se produit une accumulation.

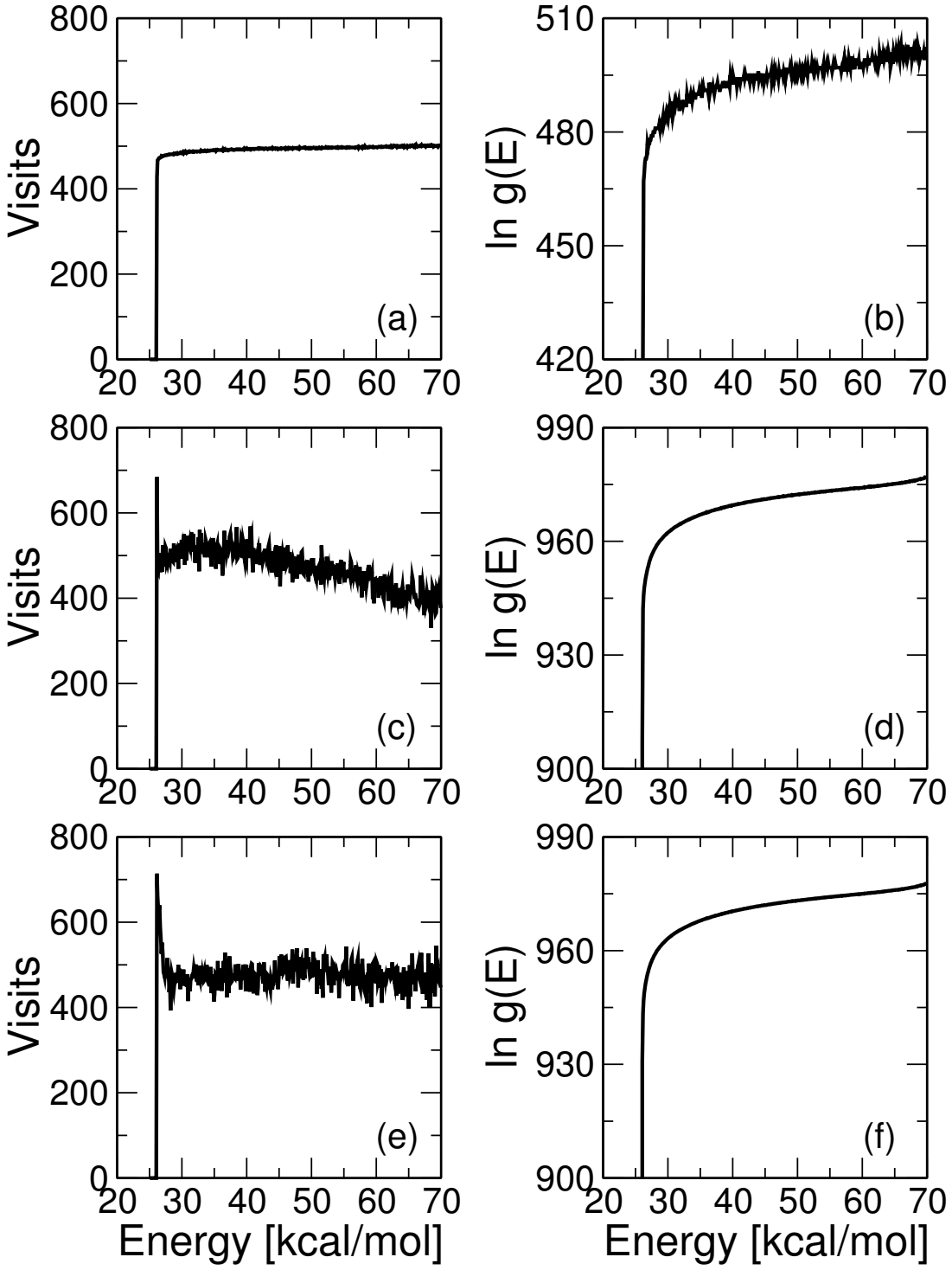


FIG. 2.15 – *(a, c et e) Histogrammes des énergies visitées et (b, d et f) de l'entropie microcanonique pour une simulation Wang-Landau du peptide Ala_4. (a et b) 1re (c et d) 10^{e} et (e et f) 19^{e} itérations.*

La figure 2.16 illustre ce phénomène d'accumulation pour le peptide Ala_4. Lors de l'itération 4 [figure 2.16(a)], la simulation atteint une région d'énergie qui n'avait pas été explorée jusqu'alors. À ce stade, $\ln f = 6{,}25 \times 10^{-2}$ et il faut donc un nombre important de pas Monte Carlo pour rejoindre l'entropie (~ 900) des cases adjacentes, comme le montre l'histogramme des visites de la figure 2.16(b). Les phénomènes d'accumulation sont d'autant plus fréquents et nocifs (jusqu'à plusieurs millions de pas Monte Carlo) que la surface d'énergie potentielle du système simulé présente des bassins en compétition. L'apparition d'accumulations au cours d'une simulation Wang-Landau est handicapante puisqu'une proportion non négligeable du temps de calcul est alors utilisée pour combler ces déficits entropiques.

Face à ce problème d'accumulation, un critère de convergence basé sur un nombre minimum de visites proportionnel à $1/\sqrt{\ln f}$ pour chaque case de l'histogramme $H(E)$ est trop contraignant. En effet, prise dans une accumulation, une simulation ne peut plus explorer la surface d'énergie, empêchant sa convergence avec un tel critère. On comprend alors mieux pourquoi il est pertinent de convertir cette condition locale en un nombre maximal global de pas Monte Carlo proportionnel à $1/\sqrt{\ln f}$ pour chaque itération.

La solution adoptée durant ce travail de thèse est développée ci-après.

2.3.3.1 Méthode Wang-Landau pour les systèmes continus

Dans le cadre d'une collaboration avec F. Calvo du laboratoire de chimie physique quantique de Toulouse, nous avons entrepris l'étude de systèmes à degrés de liberté continus par diverses implémentations de la méthode Wang-Landau [66] afin d'en évaluer les performances. Nous avons considéré deux catégories de systèmes modèles : les peptides et les agrégats Lennard-Jones. Dans cette section, je ne détaillerai que le premier type de systèmes.

Diverses stratégies ont été développées pour résoudre les problèmes d'accumulation. Récemment, Zhou *et al.* [67] ont implémenté une étape de modification globale dans laquelle la densité d'états est partiellement augmentée dans la région où elle est déjà supérieure à un seuil donné, de façon à repousser le système dans ses confins les moins explorés. Ces auteurs ont également développé une procédure spécifique adaptée au caractère continu des degrés de liberté. Pour cela, ils introduisent une fonction de modification continue $k(E)$ qui s'étale sur plusieurs cases [67] au lieu d'une seule dans l'algorithme original [42, 43]. Ainsi, si l'énergie E_0 est visitée, alors toutes les entropies $s(E)$ sont modifiées par

$$s(E) + \gamma k\left(\frac{E - E_0}{\delta}\right) \rightarrow s(E), \tag{2.58}$$

avec γ et δ des constantes ajustables pour chaque système. La fonction k peut être par exemple de type gaussienne :

$$k(x) = \exp(-x^2). \tag{2.59}$$

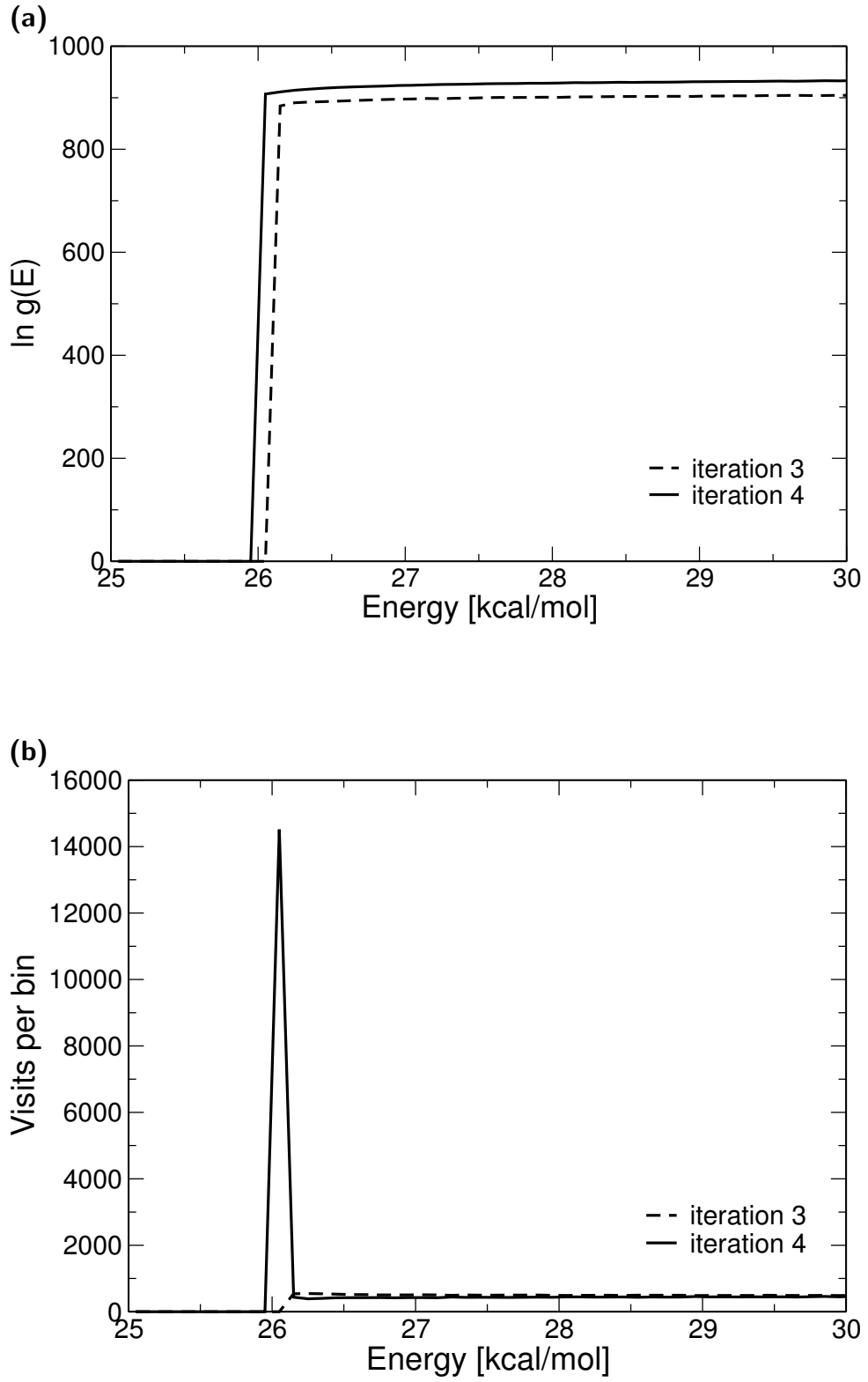


FIG. 2.16 – *Accumulation observée avec l'algorithme Wang-Landau pour le peptide* Ala_4. *(a) La comparaison des entropies après les itérations 3 et 4 montre que la simulation explore une nouvelle région d'énergie physiquement accessible pendant l'itération 4. (b) La simulation tente de combler la densité d'états correspondante pour atteindre un niveau comparable avec les densités d'états des autres régions voisines de l'histogramme. Pour cela, le système passe l'essentiel du temps de calcul dans la nouvelle case de l'histogramme, occasionnant un phénomène d'accumulation.*

Les variations d'entropie dues à l'incrémentation par un facteur de modification continu sont schématisées sur la figure 2.17. L'entropie est la plus affectée sur la case correspondante à E_0 mais le caractère continu de f permet également de modifier les entropies des cases proches de E_0. En pratique, on modifie $s(E)$ sur quelques cases voisines de E_0.

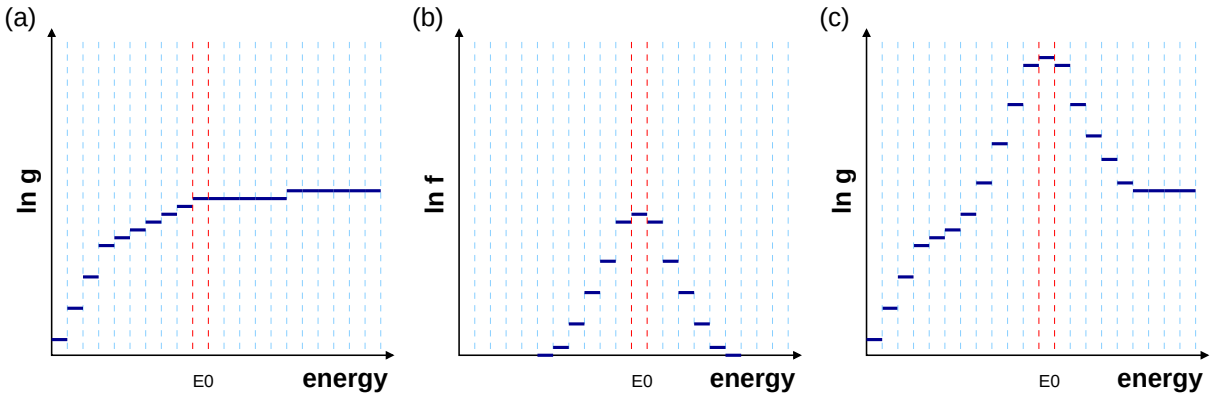


FIG. 2.17 – *Incrémentation de l'entropie par un facteur de modification continu d'après [67]. (a) Entropie avant la modification, (b) facteur de modification continu, (c) entropie après la modification.*

Cet algorithme est difficile à paramétrer à cause des nombreuses variables à ajuster, tant pour la modification globale que pour la modification continue de la densité d'états. De plus, notre expérience montre qu'il n'apporte pas d'améliorations majeures pour les systèmes qui nous intéressent (peptides et agrégats Lennard-Jones).

Dans des travaux portant sur des cristaux liquides sur réseau, Jayasri, Sastry et Murthy (JSM) [47] ont développé un algorithme dans lequel chaque itération Wang-Landau de facteur f est elle-même subdivisée en plusieurs sous-étapes. Le facteur f est légèrement réduit après chaque sous-étape (typiquement, $f^{0,9} \rightarrow f$) et reprend sa valeur initiale plus élevée une fois qu'un certain nombre de sous-étapes sont effectuées. Après quelques itérations, la valeur initiale de f est réduite pour obtenir une meilleure estimation de la densité d'états. L'originalité de cette méthode réside dans le fait que f y varie de façon non monotone au cours de la simulation. En conséquence, les déficits entropiques peuvent être comblés plus rapidement et l'accumulation évitée avec les augmentations ponctuelles de f.

En réinitialisant f à la valeur initiale de la précédente itération, on aide le système à sortir des trous entropiques, mais on détruit aussi la précision avec laquelle la densité d'états était jusqu'alors connue. Ainsi, dans l'implémentation originale proposée par Jayasri *et al.* [47], la simulation comprend 50 itérations Wang-Landau, chacune divisée en 150 sous-étapes. Les 40 premières itérations débutent avec $f = 100$, les 9 suivantes avec $f = 10$ et la dernière avec $f = e^1$. Avec une telle évolution du facteur de modification, la densité d'états ne peut être précise qu'aux toutes dernières itérations.

2.3.3.2 Méthode Wang-Landau recuit

Notre implémentation de la méthode Wang-Landau vise à simuler des polypeptides et plus généralement, des systèmes ayant des degrés de liberté continus et pouvant potentiellement poser des problèmes d'accumulation tels que décrits précédemment. Pour cela, nous combinons la démarche de JSM avec la méthode de Zhou et Bhatt. Nous gardons l'idée de JSM d'augmenter occasionnellement le facteur de modification f mais sans pour autant égaler ou dépasser sa valeur précédente. Nous imposons à f une évolution proche des variations de température obtenues lors d'une simulation en recuit simulé (voir chapitre 1). Cette méthode sera qualifiée par la suite de Wang-Landau recuit (WL-recuit ou *WL-annealing*).

La simulation Wang-Landau recuit est organisée en M itérations, chacune d'entre elles est composée de N sous-étapes. Le facteur de modification initial est réduit exponentiellement après chaque sous-étape

$$f^{\alpha} \rightarrow f \quad \text{avec } \alpha < 1. \tag{2.60}$$

Au début de l'itération suivante, f est augmenté de

$$f^{1/\alpha^{N-1}} \rightarrow f. \tag{2.61}$$

Si on note f_0 le facteur de modification initial au début de la première itération, alors la valeur de f à la dernière sous-étape de la dernière itération sera ${f_0}^{\alpha^{M+N-2}}$.

De façon à également prendre en compte la demande croissante en statistiques à mesure que f diminue, on suit également la stratégie de Zhou et Bhatt d'adapter la longueur de chaque sous-étape en fonction de $1/\sqrt{\ln f}$. On passe donc de plus en plus de temps à construire la densité d'états à mesure que que la simulation progresse, tout en obtenant une estimation de plus en plus précise de celle-ci.

Notre implémentation diffère de l'algorithme Wang-Landau original, comme de celui de Zhou et Bhatt ou Jayasri *et al.* dans la manière dont f varie au cours de la simulation. Ces différentes variations sont présentées dans la figure 2.18. Avec la méthode WL-recuit, f est modifié beaucoup plus souvent qu'avec, par exemple, l'algorithme original.

2.3.4 Calcul des moyennes canoniques suite à une simulation Wang-Landau

Une fois que f atteint la valeur désirée, la densité d'états obtenue est alors connue avec la précision f. La connaissance de $g(E)$ permet en principe de calculer, à n'importe quelle température T, toutes les grandeurs thermodynamiques du système.

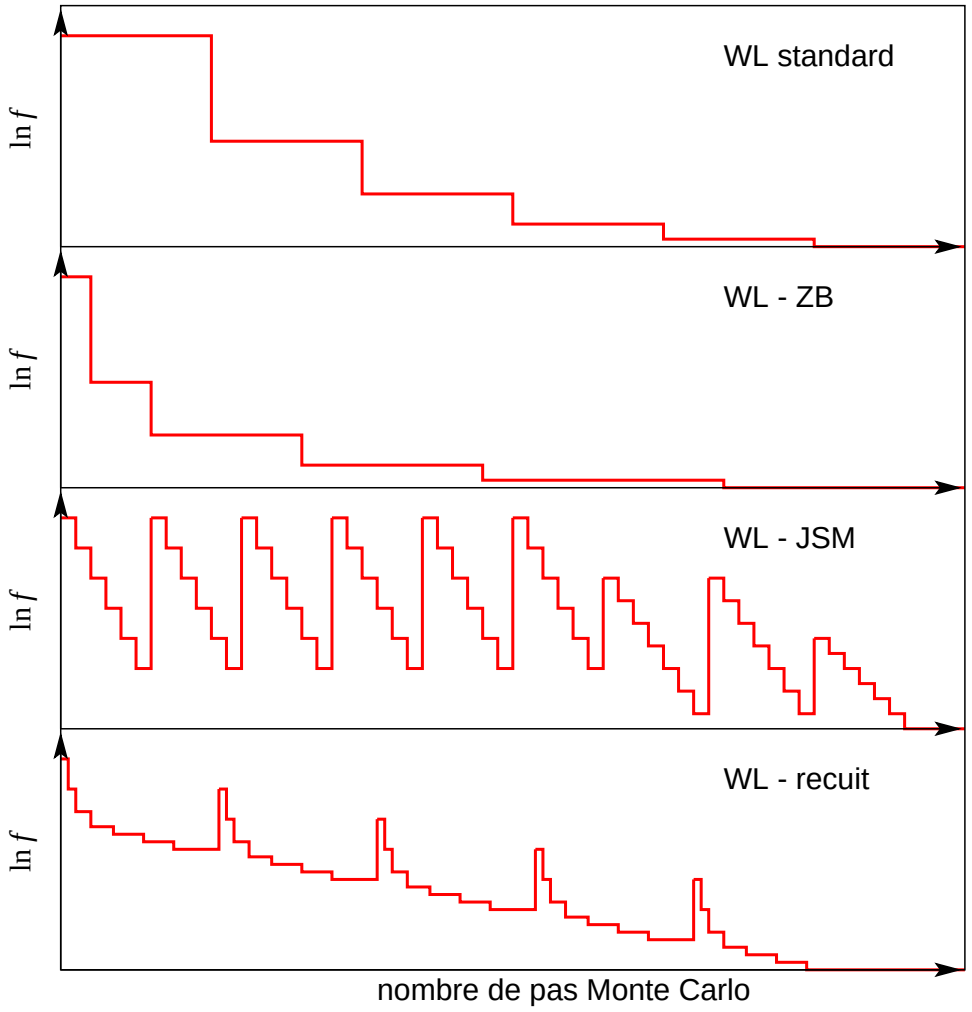


FIG. 2.18 – *Variations schématiques du facteur de modification f au cours de différentes implémentations de simulations Wang-Landau. De haut en bas, algorithme Wang-Landau original (WL standard), Wang-Landau avec la condition de Zhou et Bhatt (WL-ZB), Wang-Landau avec les variations de Jayasri, Sastry et Murthy (WL-JSM) et notre implémentation de la méthode Wang-Landau (WL-recuit).*

2.3.4.1 Simulation Wang-Landau à une dimension avec une coordonnée de réaction

La moyenne canonique d'une observable physique peut être obtenue directement à partir d'une simulation Wang-Landau en modifiant légèrement l'algorithme original [42, 43]. Pour cela, on indexe les histogrammes de densité d'états g et de visites H, non plus avec l'énergie, mais avec une coordonnée de réaction λ qui correspond à l'observable considérée. On cherchera alors à avoir un histogramme H plat sur la variable λ. Le critère de Metropolis de l'équation (2.52) devient alors

$$\mathrm{acc}(o \to n) = \min\left(1, \frac{g(\lambda_o)}{g(\lambda_n)} e^{-\beta \Delta E}\right), \tag{2.62}$$

avec $g(\lambda_o)$ la densité d'états de l'ancienne configuration (o), $g(\lambda_n)$ la densité d'états de la nouvelle configuration (n) et $\Delta E = E_n - E_o$, la différence d'énergie entre la nouvelle et l'ancienne configuration.

La coordonnée de réaction peut être n'importe quelle observable physique ayant un intérêt pour la simulation, comme un angle dièdre [56], une coordonnée cartésienne [56] ou la distance bout-à-bout [55].

Une telle simulation donne accès à la moyenne canonique de l'observable considérée à la seule température de simulation T. On perd une des forces de l'algorithme original qui permet *a posteriori* de calculer les moyennes canoniques à n'importe quelle température.

2.3.4.2 Exploration multicanonique à partir d'une simulation Wang-Landau

La simulation multicanonique, comme la méthode Wang-Landau, repose sur la construction d'histogrammes plats et vise à réaliser une marche aléatoire sur la surface d'énergie potentielle. Dans une première étape, on calcule les poids multicanoniques par itérations successives. Ensuite, on effectue une simulation avec les poids multicanoniques obtenus au cours de laquelle on accumule des statistiques pour l'observable $\mathcal{A}$ étudiée. Une dernière étape de repondération permet de calculer les moyennes canoniques aux températures voulues.

La détermination des poids multicanoniques étant assez difficile, nous pouvons adopter la stratégie suivante :

1. réaliser une simulation Wang-Landau pour calculer la densité d'états $g(E)$;
2. utiliser la densité d'états obtenue à la dernière itération de la simulation Wang-Landau dans une simulation multicanonique où les statistiques sont accumulées sur E et $\mathcal{A}$.

Dans cette dernière étape, on engendre un certain nombre de configurations par un processus stochastique. Les mouvements sont acceptés avec la probabilité donnée dans l'équation (2.52). Que le mouvement soit accepté ou rejeté, on enregistre la valeur de

l'observable $\mathcal{A}$ dans un histogramme à deux dimensions, énergie et observable, $H(E, \mathcal{A})$. La qualité de la simulation peut être contrôlée en vérifiant bien que l'histogramme accumulé est plat sur la dimension d'énergie.

La moyenne de l'observable $\mathcal{A}$ à la température T est alors calculée en repondérant l'histogramme $H(E, \mathcal{A})$ par transformée de Laplace

$$\langle \mathcal{A} \rangle_T = \frac{\sum_{\mathcal{A}} \sum_{E} \mathcal{A} H(E, \mathcal{A}) g(E) e^{-\beta E}}{\sum_{\mathcal{A}} \sum_{E} H(E, \mathcal{A}) g(E) e^{-\beta E}}. \tag{2.63}$$

2.3.4.3 Simulation Wang-Landau à deux dimensions, énergie et paramètre d'ordre

L'utilisation du paramètre d'ordre $\mathcal{A}$ [50, 65, 68], en complément de l'énergie, peut aider le système à dépasser certaines barrières et explorer plus largement l'espace des phases. Une simulation Wang-Landau à deux dimensions consiste alors à construire la densité d'états du système à la fois en énergie et en paramètre d'ordre $\mathcal{A}$, $g(E, \mathcal{A})$. Cette méthode permet de garder la corrélation entre ces deux grandeurs. Tout comme l'énergie, le paramètre d'ordre $\mathcal{A}$ est discrétisé. Un histogramme des énergies et des valeurs du paramètre d'ordre visitées $H(E, \mathcal{A})$ est également construit et on espère qu'il sera uniforme à la fin de chaque itération.

La moyenne canonique de l'observable $\mathcal{A}$ à la température T est alors donnée par

$$\langle \mathcal{A} \rangle_T = \frac{\sum_{\mathcal{A}} \sum_{E} \mathcal{A} g(E, \mathcal{A}) e^{-\beta E}}{\sum_{\mathcal{A}} \sum_{E} g(E, \mathcal{A}) e^{-\beta E}}. \tag{2.64}$$

2.4 Conclusion

La méthode Monte Carlo est une méthode d'échantillonnage permettant l'accumulation de statistiques. Les systèmes étudiés ici possèdent de nombreux degrés de liberté et une simulation canonique rencontre beaucoup de difficultés à quitter les minima locaux à basse température.

Les simulations dans les ensembles généralisés, qu'elles utilisent des algorithmes à trajectoires multiples comme en Monte Carlo d'échange ou à échantillonnage non-boltzmannien comme la méthode Wang-Landau, franchissent plus facilement les barrières de potentiel.

La méthode Monte Carlo d'échange repose sur la simulation Monte Carlo de M répliques à M températures. L'échange de répliques est tenté périodiquement. La répartition des températures est une étape cruciale qui peut cependant être aisément résolue en adoptant une distribution géométrique ou à la fois géométrique et linéaire. La convergence de la méthode Monte Carlo d'échange est assurée pour des taux d'échange satisfaisants et

des échanges de répliques les plus variés possibles. Les moyennes canoniques des observables étudiées sont obtenues par une repondération des statistiques accumulées et une interpolation via la méthode des histogrammes multiples.

La méthode Wang-Landau tente d'estimer avec précision la densité d'états microcanonique en réalisant une marche aléatoire sur le paysage énergétique en pénalisant les états au fur et à mesure qu'ils sont visités. La simulation de systèmes continus engendre des phénomènes d'accumulation nocifs à la convergence. En utilisant plusieurs améliorations déjà existantes, nous avons proposé un algorithme de Wang-Landau « recuit » adapté aux systèmes continus. Enfin, l'utilisation d'un paramètre d'ordre, en plus de l'énergie, dans la construction de la densité d'états peut s'avérer intéressant pour décrire avec précision les propriétés à basse température. Bien évidemment, le paramètre d'ordre doit être pertinent vis à vis du système étudié. Pour les peptides, le dipôle électrique ou une distance atomique particulière conviennent à cet effet. Les moyennes canoniques sont déterminées en repondérant la densité d'états calculée.

Malgré l'apparente simplicité de la méthode Wang-Landau originale, notre expérience montre que l'ajustement des nombreux paramètres optimaux est une étape importante de la simulation. Hormis la répartition du nombre de pas Monte Carlo avec le facteur de modification f, les valeurs initiale et finale de f, le taux de décroissance α après chaque itération, les nombres d'itérations M et de sous-étapes N sont dépendants du système considéré. Comparativement, la méthode Monte Carlo d'échange nécessite nettement moins d'ajustements, principalement une bonne répartition des températures.

Dans le chapitre suivant, nous illustrerons ces algorithmes d'échantillonnage sur plusieurs peptides. Nous comparerons également différentes variantes de la méthode Wang-Landau avec la méthode Monte Carlo d'échange.

Bibliographie

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[2] M. A. Miller, L. M. Amon, and W. P. Reinhardt. Should one ajust the maximum step size in a Metropolis Monte Carlo simulation? *Chemical Physics Letters*, 331:278–284, 2000.

[3] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, California, USA, 2^{nd} edition, 2002.

[4] A. R. Leach. *Molecular modelling: principles and applications*. Pearson Education, Harlow, 2^{nd} edition, 2001.

[5] M. Matsumoto and T. Nishimura. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8:3–30, 1998.

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 2^{nd} edition, 2002.

[7] `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/VERSIONS/C-LANG/c-lang.html`.

[8] Y. Iba. Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, 12:623–656, 2001.

[9] Y. Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling*, 22:425–439, 2004.

[10] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. *Biopolymers*, 60:96–123, 2001.

[11] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[12] U. H. E. Hansmann and L. Wille. Global optimization by energy landscape paving. *Physical Review Letters*, 88:068105, 2002.

[13] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *Journal of Chemical Physics*, 96:1776–1783, 1992.

[14] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19:451–458, 1992.

[15] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57:2607–2609, 1986.

[16] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. K. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Fairfax Station, 1991. Interface Foundation.

[17] K. Hukushima and K. Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan*, 65:1604–1608, 1996.

[18] E. Marinari, G. Parisi, and J. Ruiz-Lorenzo. Numerical Simulations of Spin Glass Systems. In A. P. Young, editor, *Spin Glasses and Random fields*, volume 12 of *Directions in Condensed Matter Physics*, Singapore, 1998. World Scientific.

[19] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281:140–150, 1997.

[20] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314:141–151, 1999.

[21] D. Gront, A. Kolinski, and J. Skolnick. Comparison of three Monte Carlo conformational search strategies for a protein-like polymer models: Identyfication of low energy structures and folding thermodynamics. *Journal of Chemical Physics*, 113:5065–5071, 2000.

[22] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *Journal of Chemical Physics*, 113:6042–6051, 2000.

[23] A. E. García and K. Y. Sanbonmatsu. alpha Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 99:2782–2787, 2002.

[24] B. S. Kinnear, M. F. Jarrold, and U. H. E. Hansmann. All-atom generalized-ensemble simulations of small proteins. *Journal of Molecular Graphics and Modelling*, 22:397–403, 2004.

[25] A. Schug, T. Herges, and W. Wenzel. All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method. *Proteins: Structure, Function, and Bioinformatics*, 57:792–798, 2004.

[26] H. Kokubo and Y. Okamoto. Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: Case of two helices. *Journal of Chemical Physics*, 120:10837–10847, 2004.

[27] K. Y. Sanbonmatsu and A. E. García. Structure of Met-Enkephalin in Explicit Aqueous Solution Using Replica Exchange Molecular Dynamics. *Proteins: Structure, Function, and Genetics*, 46:225–234, 2002.

[28] P. Dugourd, R. Antoine, G. Breaux, M. Broyer, and M. F. Jarrold. Entropic Stabilization of Isolated β-Sheets. *Journal of American Chemical Society*, 127:4675–4679, 2005.

[29] D. J. Earl and M. W. Deem. Optimal Allocation of Replicas to Processors on Parallel Tempering Simulations. *Journal of Physical Chemistry B*, 108:6844–6849, 2004.

[30] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, 22:789–828, 1996.

[31] W. D. Gropp and E. Lusk. *User's Guide for* `mpich`*, a Portable Implementation of MPI.* Mathematics and Computer Science Division, Argonne National Laboratory, 1996. ANL-96/6.

[32] C. Predescu, M. Predescu, and C. V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *Journal of Chemical Physics*, 120:4119–4128, 2004.

[33] F. Calvo and P. Labastie. Numerical simulations of the shape and 'phase transitions' in finite systems. *European Journal of Physics*, 26:S23–S30, 2005.

[34] N. Rathore, M. Chopra, and J. J. de Pablo. Optimal allocation of replicas in parallel tempering simulations. *Journal of Chemical Physics*, 122:024111, 2005.

[35] C. Predescu, M. Predescu, and C. V. Ciobanu. On the Efficiency of Exchange in Parallel Tempering Monte Carlo Simulations. *Journal of Physical Chemistry B*, 109:4189–4196, 2005.

[36] P. Poulain, R. Antoine, M. Broyer, and P. Dugourd. Monte Carlo simulations of flexible molecules in a static electric field: electric dipole and conformation. *Chemical Physics Letters*, 401:1–6, 2005.

[37] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo Data Analysis. *Physical Review Letters*, 63:1195–1198, 1989.

[38] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13:1011–1021, 1992.

[39] B. A. Berg and T. Celik. New approach to spin-glass simulations. *Physical Review Letters*, 69:2292–2295, 1992.

[40] U. H. E. Hansmann and Y. Okamoto. New Monte Carlo algorithms for protein folding. *Current Opinion in Structural Biology*, 9:177–183, 1999.

[41] Y. Okamoto and U. H. E. Hansmann. Thermodynamics of helix-coil transitions studied by multicanonical algorithms. *Journal of Physical Chemistry*, 99:11276–11287, 1995.

[42] F. Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters*, 86:2050–2053, 2001.

[43] F. Wang and D. P. Landau. Determining the density of states for classical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64:056101, 2001.

[44] B. J. Schultz, K. Binder, and M. Müller. Flat Histogram Method of Wang-Landau and N-Fold Way. *International Journal of Modern Physics C*, 13:477–494, 2002.

[45] M. Troyer, S. Wessel, and F. Alet. Flat Histogram Methods for Quantum Systems: Algorithms to Overcome Tunneling Problems and Calculate the Free Energy. *Physical Review Letters*, 90:120201, 2003.

[46] B. J. Schulz, K. Binder, and M. Müller. First-order interface localization-delocalization transition in thin Ising films using Wang-Landau sampling. *Physical Review E*, 71:046705, 2005.

[47] D. Jayasri, V. S. S. Sastry, and K. P. N. Murthy. Wang-Landau Monte Carlo simulation of isotropic-nematic transition in liquid crystals. *Physical Review E*, 72:036702, 2005.

[48] Q. Yan, R. Faller, and J. J. de Pablo. Density-of-states Monte Carlo method for simulation of fluids. *Journal of Chemical Physics*, 116:8745–8749, 2002.

[49] M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos. Flat-Histogram Dynamics and Optimization in Density of States Simulations of Fluids. *Journal of Physical Chemistry B*, 108:19748–19755, 2004.

[50] V. Varshney, T. E. Dirama, T. Z. Sen, and G. A. Carri. A Minimal Model for the Helix-Coil Transition of Wormlike Polymers. Insights from Monte Carlo Simulations and Theoretical Implications. *Macromolecules*, 37:8794–8804, 2004.

[51] G. A. Carri, R. Batman, V. Varshney, and T. E. Dirama. A Monte Carlo simulation study of the mechanical and conformational properties of networks of helical polymers. I. General concepts. *Polymer*, 46:3809–3817, 2005.

[52] V. Varshney and G. A. Carri. Coupling between Helix-Coil and Coil-Globule Transitions in Helical Polymers. *Physical Review Letters*, 95:168304, 2005.

[53] N. Rathore and J. J. de Pablo. Monte Carlo simulation of proteins through a random walk in energy space. *Journal of Chemical Physics*, 116:7225–7230, 2002.

[54] N. Rathore, T. A. Knotts IV, and J. J. de Pablo. Density of states simulations of proteins. *Journal of Chemical Physics*, 118:4285–4290, 2003.

[55] N. Rathore, Q. Yan, and J. J. de Pablo. Molecular simulation of the reversible mechanical unfolding of proteins. *Journal of Chemical Physics*, 120:5781–5788, 2004.

[56] F. Calvo. Sampling along reaction coordinates with the Wang-Landau method. *Molecular Physics*, 100:3421–3427, 2002.

[57] F. Calvo and P. Parneix. Statistical evaporation of rotating clusters. I. Kinetic energy released. *Journal of Chemical Physics*, 119:256–264, 2003.

[58] R. Faller and J. J. de Pablo. Density of states of a binary Lennard-Jones glass. *Journal of Chemical Physics*, 119:4405–4408, 2003.

[59] P. Dayal, S. Trebst, S. Wessel, D. Würtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith. Performance Limitations of Flat-Histogram Methods. *Physical Review Letters*, 92:097201, 2004.

[60] M. D. Costa, J. Viana Lopes, and J. M. B. Lopes dos Santos. Analytical study of tunneling times in flat histogram Monte Carlo. *Europhysics Letters*, 72:802–808, 2005.

[61] M. S. Shell, P. G. Debenetti, and A. Z. Panagiotopoulos. Generalization of the Wang-Landau method for off-lattice simulations. *Physical Review E*, 66:056703, 2002.

[62] M. O. Khan, G. Kennedy, and D. Y. C. Chan. A scalable parallel Monte Carlo method for free energy simulations of molecular systems. *Journal of Computational Chemistry*, 26:72–77, 2004.

[63] B. J. Schulz, K. Binder, M. Müller, and D. P. Landau. Avoiding boundary effects in Wang-Landau sampling. *Physical Review E*, 67:067102, 2003.

[64] C. Zhou and R. N. Bhatt. Understanding and improving the Wang-Landau algorithm. *Physical Review E*, 72:025701, 2005.

[65] A. Tröster and C. Dellago. Wang-Landau Sampling with Self-Adaptive Range. *Physical Review E*, 71:066705, 2005.

[66] P. Poulain, F. Calvo, R. Antoine, M. Broyer, and P. Dugourd. Performances of Wang-Landau algorithms for continuous systems. *Physical Review E*, 73:056704, 2006.

[67] C. Zhou, T. C. Schulthess, S. Torbrügge, and D. P. Landau. Wang-Landau Algorithm for Continuous Models and Joint Density of States. *Physical Review Letters*, 96:120201, 2006.

[68] D. P. Landau, S.-H. Tsai, and M. Exler. A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. *American Journal of Physics*, 72:1294–1302, 2004.

Chapitre 3

Applications et comparaison des méthodes Monte Carlo d'échange et Wang-Landau

Sommaire

Le chapitre précédent a détaillé les fondements théoriques des algorithmes à trajectoires multiples et à échantillonnage non-boltzmannien. Le présent chapitre a pour objectif d'illustrer et comparer les méthodes Monte Carlo d'échange et Wang-Landau sur différents peptides.

La première section présente la simulation en Monte Carlo d'échange du dipeptide tryptophane-glycine (Trp-Gly ou WG) dans un champ électrique statique et intense. La deuxième section a pour objectif de comparer différentes variantes de la méthode Wang-Landau avec la méthode Monte Carlo d'échange. La polyalanine Ala_8 et le pentapeptide chargé $AGWLK^+$ sont utilisés comme systèmes modèles dans cette étude.

3.1 Simulation Monte Carlo de molécules dans un champ électrique statique : dipôle électrique et conformation

Les premières simulations Monte Carlo d'échange effectuées pour ce travail de thèse nous ont permis d'étudier le comportement d'une molécule dans un champ électrique [1]. Plus précisément, l'étude a porté sur le dipeptide WG soumis à un champ électrique uniforme. Or, les forces électrostatiques et l'interaction avec un champ externe jouent un rôle très important dans la structure et les propriétés des biomolécules [2, 3, 4, 5, 6]. Dans la plupart de ces phénomènes électrostatiques, la réponse d'une protéine dans un champ électrique externe peut conduire à une relaxation structurale et d'orientation. Pour une protéine en solution, l'effet du solvant s'ajoute à la réponse avec le champ électrique, ce qui impose de traiter à la fois la protéine et le solvant [7, 8, 9]. Au sein de notre équipe [10], un dispositif expérimental a été développé pour l'étude des biomolécules en phase gazeuse, s'affranchissant ainsi des effets du solvant. Cette expérience consiste en un jet moléculaire plongé dans un champ électrique statique et inhomogène [11, 12, 13, 14]. La figure 3.1 schématise le montage de déflexion électrique d'un jet moléculaire qui provient d'une source à vaporisation laser MALD (*matrix assisted laser desorption*). La détection des molécules est assurée par un spectromètre à temps de vol sensible en position.

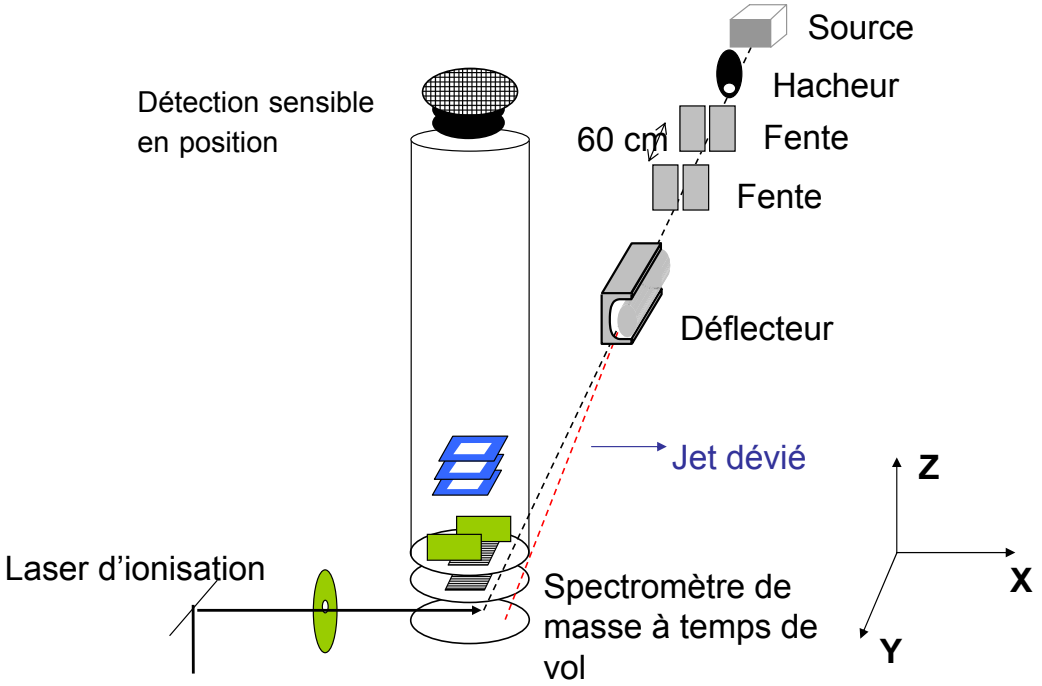


FIG. 3.1 – *Dispositif expérimental de déflexion électrique d'un jet moléculaire.*

La moyenne du dipôle électrique dans le champ externe est déduite de la déflexion du jet moléculaire. Lorsque le champ appliqué est faible et que la réponse linéaire s'applique, la projection du dipôle moyen sur l'axe du champ est proportionnel au champ externe et est donné par la formule de Langevin-Debye [15, 16].

$$\langle \mu_z \rangle = \chi F_z = \left(\frac{\langle \mu^2 \rangle_{0,T}}{3 k_B T} + \alpha_e \right) F_z. \tag{3.1}$$

Dans cette expression, χ est la susceptibilité électrique de la molécule en phase gazeuse, F_z est le champ électrique appliqué (suivant l'axe z), α_e est la polarisabilité électronique statique, $\langle \mu^2 \rangle_{0,T}$ est la moyenne du carré du dipôle électrique de la molécule sans champ appliqué à la température T.

Lorsque le champ électrique est intense, on peut penser qu'une orientation significative de la molécule se produit et que la formule (3.1) n'est plus valide. En particulier, les protéines sont des molécules flexibles à température ambiante [17] et un champ électrique pourrait induire d'importantes déformations structurales.

3.1.1 Simulations

Nous avons réalisé des simulations Monte Carlo d'échange sur le dipeptide WG en phase gazeuse et en présence d'un champ électrique externe uniforme.

Dans un champ électrique homogène, l'énergie d'un peptide est la somme de l'énergie conformationnelle E_0 et de l'interaction avec le champ électrique

$$E = E_0 - \vec{\mu} \cdot \vec{F}, \tag{3.2}$$

où E_0 est l'énergie conformationnelle donnée par le champ de force AMBER *ff96* [18], $\vec{F}$ est le champ électrique appliqué et $\vec{\mu}$ le dipôle de la molécule donné par

$$\vec{\mu} = \sum_i q_i \vec{r_i} + \alpha_e \vec{F}. \tag{3.3}$$

Avec q_i et $\vec{r_i}$, respectivement la charge partielle et la position de l'atome i. Les charges partielles sont définies par le champ de force AMBER dans sa version *ff96* ; elles sont constantes pendant la simulation et ne dépendent pas du champ. La polarisabilité électronique α_e (28,4 Å^3 pour WG) est obtenue par un modèle additif [10]. Elle ne dépend pas de la conformation ni de l'orientation de la molécule. Pour simplifier les calculs, elle ne sera pas considérée par la suite. Ainsi, le dipôle ne va dépendre du champ que via les changements de conformation dues au champ.

Les simulations Monte Carlo d'échange sont constituées de 10^8 pas Monte Carlo avec 5 températures à 200, 269, 354, 463 et 600 K, les 2×10^6 premiers pas Monte Carlo sont utilisés pour la thermalisation et ne sont pas inclus dans les statistiques. Les géométries de départ sont initialisées aléatoirement. Lors de chaque cycle Monte Carlo, les angles

dièdres (Φ, Ψ) de la chaîne peptidique sont modifiés un à un, ainsi que ceux de la chaîne latérale du tryptophane. Cela représente un total de 8 angles de torsion. Un échange de répliques est tenté toutes les 100 cycles Monte Carlo. Les simulations ont été réalisées à 7 différentes valeurs du champ électrique, de 0 à 10^9 V/m. Le champ électrique est disposé suivant l'axe z. Les observables μ^2, μ_x, μ_y, μ_z et E sont suivies et enregistrées dans des histogrammes multidimensionnels après chaque pas Monte Carlo. Les simulations ont été réalisées sur un PC biprocesseur Xeon 1,5 GHz.

Les taux d'acceptance des échanges entre les températures de la simulation pour un champ électrique de 0 et 10^8 V/m sont donnés dans le tableau 3.1. Les valeurs obtenues sont satisfaisantes.

La figure 3.2 représente l'évolution des températures rencontrées par une réplique particulière pendant les simulations avec un champ nul et de 10^8 V/m. Qu'elle que soit la valeur du champ, toutes les températures sont rencontrées régulièrement, preuve d'un échange efficace des répliques.

TAB. 3.1 – *Taux d'acceptance des échanges entre les différentes températures pour les champs électriques de* 0 *et* 10^8 *V/m.*

Paires de températures [K]	Taux d'acceptance [%]	
	$F = 0$	$F = 10^8$ V/m
$200 \leftrightarrow 269$	59	59
$269 \leftrightarrow 354$	61	60
$354 \leftrightarrow 463$	63	63
$463 \leftrightarrow 600$	65	65

La figure 3.3 donne les moyennes de l'énergie et du dipôle électrique pour la température la plus basse (200 K) de la simulation Monte Carlo d'échange avec différentes valeurs du champ électrique. Dans les deux cas, la moyenne converge rapidement vers les valeurs d'équilibre.

La marche aléatoire en température (figure 3.2), les valeurs des taux d'acceptance (tableau 3.1) et l'évolution des moyennes en énergie et en dipôle électrique (figure 3.3) confirment que le temps de simulation a été suffisant et que le calcul a convergé.

3.1.2 Résultats

La figure 3.4 montre l'évolution de $\langle\mu^2\rangle$ en fonction de la température à champ nul. À température élevée, la diminution de cette moyenne est due à des structures étirées ayant des dipôles faibles. Comme attendu, les valeurs moyennes de μ_x, μ_y et μ_z sont nulles sans champ externe. Ceci fournit au passage un critère supplémentaire quant à la bonne convergence des simulations. À 300 K, la moyenne du carré du dipôle obtenue par la méthode des histogrammes multiples vaut 24,08 D^2.

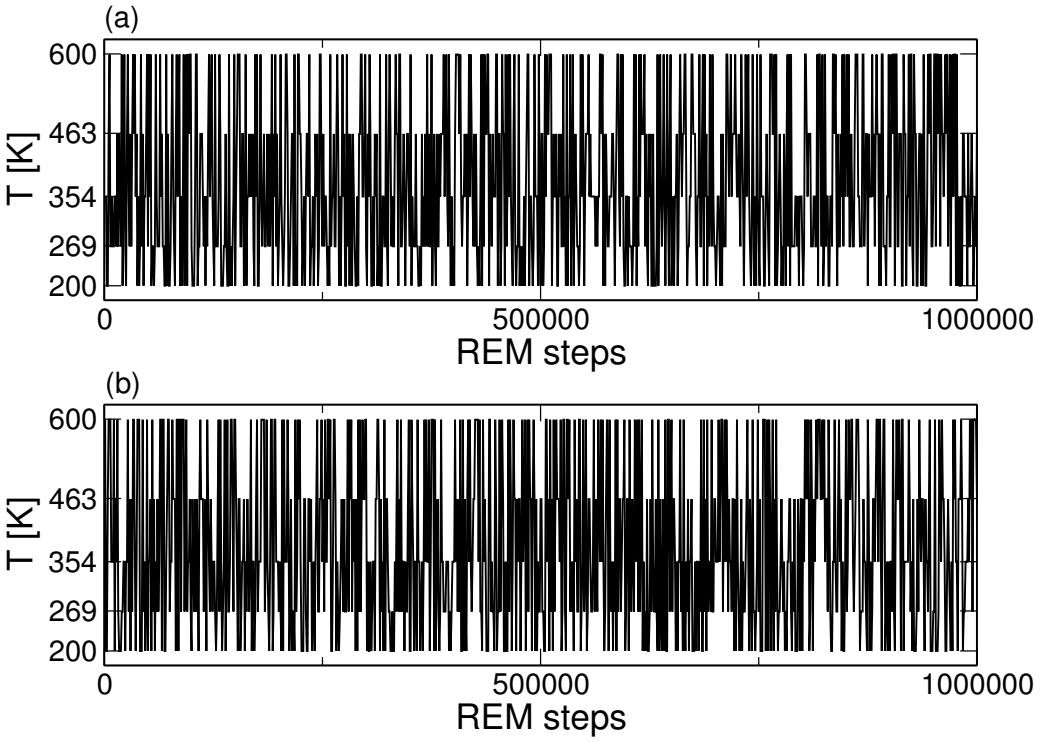


FIG. 3.2 – *Simulation Monte Carlo d'échange de WG avec 5 températures comprises entre 200 K et 600 K. Évolution des températures rencontrées par la réplique 1 pour (a) un champ nul et avec (b) un champ de* 10^8 *V/m.*

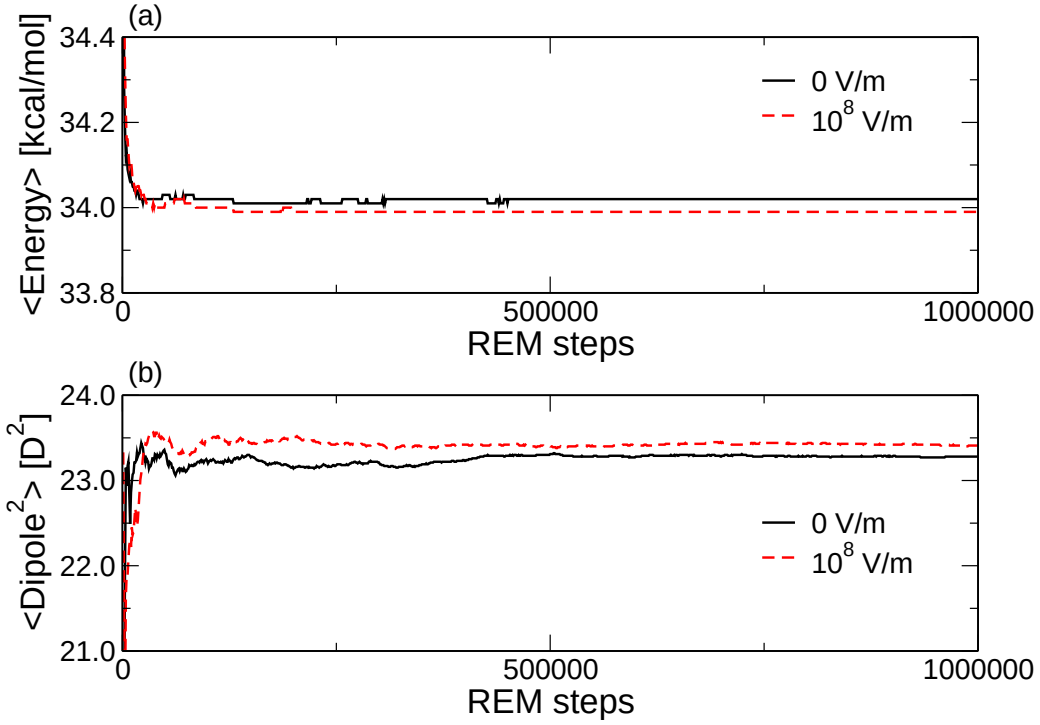


FIG. 3.3 – *Simulation Monte Carlo d'échange de WG avec 5 températures comprises entre 200 K et 600 K. Évolution des moyennes (a) de l'énergie et (b) du dipôle électrique pour la température la plus basse (200 K) à différentes valeurs du champ électrique.*

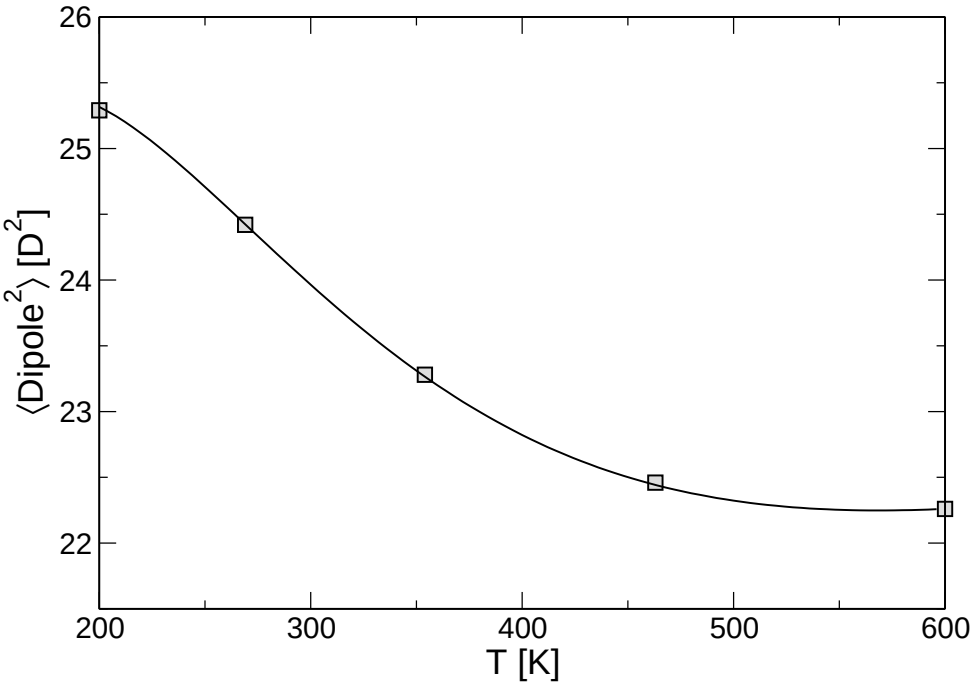


FIG. 3.4 – *Variations de la moyenne du carré du dipôle électrique en fonction de la température, sans champ électrique. Les carrés correspondent aux moyennes obtenues pour les températures de la simulation. La ligne continue correspond aux moyennes calculées par la méthodes des histogrammes multiples.*

Avec une polarisabilité électronique α_e de 28,4 Å^3 et l'équation (3.1), on obtient une susceptibilité de 222 Å^3. Cette valeur est cohérente avec la valeur expérimentale de 214 ± 27 Å^3 mesurée par Antoine *et al.* [10]. La susceptibilité théorique de 240 Å^3 obtenue avec le champ de force CHARMM [10] est également en accord avec cette valeur.

La figure 3.5 représente la moyenne du dipôle électrique suivant l'axe z en fonction du champ électrique externe. Les moyennes du dipôle suivant les axes x et y ont également été calculées mais sont en valeur absolue inférieures à 10^{-2} D. Sont également représentées les moyennes du dipôle prévues par la théorie de la réponse linéaire [formule de Langevin-Debye, équation (3.1)]. Les moyennes du carré du dipôle utilisées sont celles calculées sans champ électrique, à savoir 25,29 D^2 ; 24,42 D^2 ; et 22,26 D^2 à respectivement 200 K, 269 K et 600 K. À 600 K, les simulations Monte Carlo sont en accord avec la théorie de la réponse linéaire. À 200 K et 269 K, deux domaines peuvent être distingués. En présence d'un champ électrique faible, le dipôle calculé est proportionnel au champ électrique et est en accord avec la formule de Langevin-Debye. En effet, cette formule est valide pour $\mu F/k_B T \ll 1$. Avec $\mu^2 \sim 25$ D^2, cela correspond à $F \ll 1{,}7 \times 10^8$ V/m à 200 K et $F \ll 2{,}2 \times 10^8$ V/m à 268 K. Les expériences de déflexion électrique menées dans le groupe [10, 19] sont réalisées avec un champ $F \leq 1{,}5 \times 10^7$ V/m pour lequel l'équation (3.1) est donc valide.

Au-delà de la limite du champ faible, un phénomène de saturation apparaît clairement. La moyenne du dipôle suivant l'axe z n'est plus donnée par la théorie de la réponse linéaire. Si on considère une molécule rigide avec un dipôle μ_0, la moyenne $\langle \mu_z \rangle$ tendrait vers μ_0,

ce qui correspond à l'alignement de la molécule suivant l'axe z du champ électrique. Pour une molécule linéaire, l'évolution de la moyenne du dipôle dans un champ externe est donnée par la fonction de Langevin $\mathcal{L}(x) = \coth(x) - 1/x$ [15]. Pour une molécule flexible, la situation est plus complexe car une orientation ainsi qu'une modification de la conformation peuvent se produire en présence d'un champ électrique.

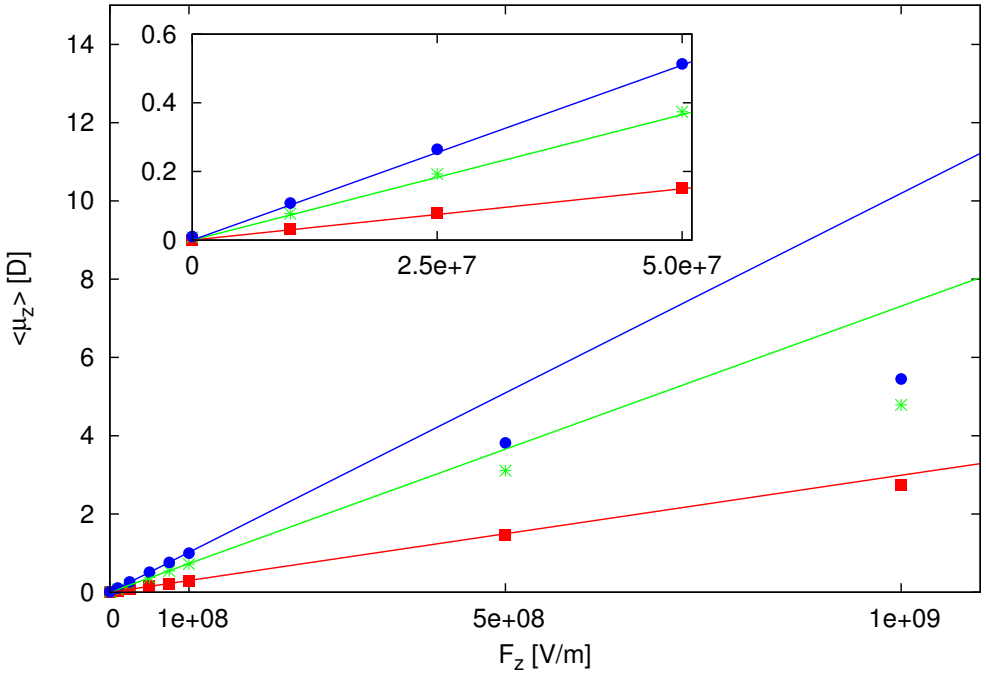


FIG. 3.5 – *Moyenne du dipôle électrique projeté suivant l'axe du champ électrique z ($\langle\mu_z\rangle$) pour différentes températures. Les carrés rouges correspondent à la température de 600 K, les étoiles vertes 269 K et les ronds bleus à 200 K. Les lignes correspondent aux valeurs données par la formule de Langevin-Debye. Un agrandissement entre $F = 0$ et 5×10^7 V/m est mis en insert.*

La figure 3.6 montre la distribution de la moyenne du carré du dipôle obtenue à différentes valeurs du champ électrique et différentes températures. Pour $F = 0$, $T = 200$ K et 269 K, la distribution présente 3 pics distincts centrés à 15 D^2, 27 D^2 et 50 D^2. On dénombre 4 familles de structures qui peuvent être attribuées à ces pics (figure 3.7). Le tout premier pic de la figure 3.6 est dû à deux familles de structures représentées en (a) et (a') sur la figure 3.7. Ces deux structures sont stabilisées par les interactions entre l'atome d'hydrogène du groupe carboxylique et l'atome d'oxygène de la liaison peptidique, ainsi qu'entre l'oxygène non lié du groupe carboxylique et l'hydrogène connecté à l'atome d'azote du groupement indole. La structure (a) est la plus stable à 200 K et à champ nul. Cette structure est dominante dans des simulations effectuées à $T = 50$ K. Les structures (b) et (c) sont stabilisées par l'interaction entre l'indole et le groupe carboxylique et diffèrent par un retournement du groupe carboxylique. La valeur élevée du dipôle électrique de la dernière famille est due à l'addition constructive des dipôles du groupe indole, de

la liaison peptidique et du groupe carboxylique. La simulation explore une large variété de structures et la molécule WG est flexible. Les structures représentées sur la figure 3.7 sont représentatives des structures correspondantes aux pics observés sur la distribution du carré du dipôle électrique de la figure 3.6. À $T = 600$ K, les pics sont beaucoup moins marqués et le poids relatif des structures avec un faible dipôle est augmenté, en accord avec la figure 3.4.

Les distributions à $F = 10^7$ V/m et 10^8 V/m sont similaires à la distribution obtenue pour $F = 0$. De telles intensités de champ électrique ne sont pas suffisantes pour produire des modifications significatives des conformations du dipeptide. Pour $F = 10^9$ V/m et $T = 200$ K et 269 K, on observe une distribution différente du carré du dipôle. Les 3 pics précédents sont présents mais avec des poids relatifs différents. L'interaction avec le champ électrique est plus élevée pour les structures ayant un fort dipôle et un champ électrique intense stabilise clairement ces structures. Pour une molécule donnée, ces changements de conformation peuvent être renforcés soit en augmentant l'intensité du champ électrique, soit en diminuant la température.

Pour observer des modifications significatives de structure avec les champs électriques disponibles dans les expériences de déflexion de jet moléculaire (typiquement entre 10^7 et 10^8 V/m), la meilleure stratégie consiste à choisir une molécule qui présente une compétition entre des structures à faible et fort dipôle. On pourrait par exemple imaginer d'utiliser un peptide possédant une structure globulaire ou en feuillet β, avec un faible dipôle, et une structure en hélice α avec un dipôle élevé. Dans ce cas, le contrôle de structure pourrait se faire avec les intensités de champ électrique disponibles expérimentalement. On pourrait ainsi espérer reproduire les modifications structurales qui ont lieu dans les organismes vivants, sous l'effet des forces électrostatiques.

3.2 Comparaison des algorithmes Wang-Landau avec la méthode Monte Carlo d'échange pour des polypeptides

Nous avons évalué les performances de quelques algorithmes Wang-Landau introduits dans le chapitre 2. Il s'agit de la méthode Wang-Landau originale (WL standard), Wang-Landau avec la condition Zhou et Bhatt (WL-ZB) sur le nombre de pas Monte Carlo, Wang-Landau avec le schéma de température non monotone de Jayasri *et al.* et notre propre algorithme (WL-recuit). Les systèmes étudiés sont deux peptides, la polyalanine Ala_8 et le pentaptide chargé [Ala-Gly-Trp-Leu-Lys + H]$^+$ (AGWLK$^+$). La molécule de polyalanine est un système modèle pertinent pour l'étude de la transition hélice α – pelote statistique [21]. AGWLK$^+$ est un peptide modèle expérimental construit pour étudier la dissociation induite par collision ou par laser [22, 23]. Cette molécule présente différents comportements de fragmentation suivant son état de charge +1 ou +2. Ces différences

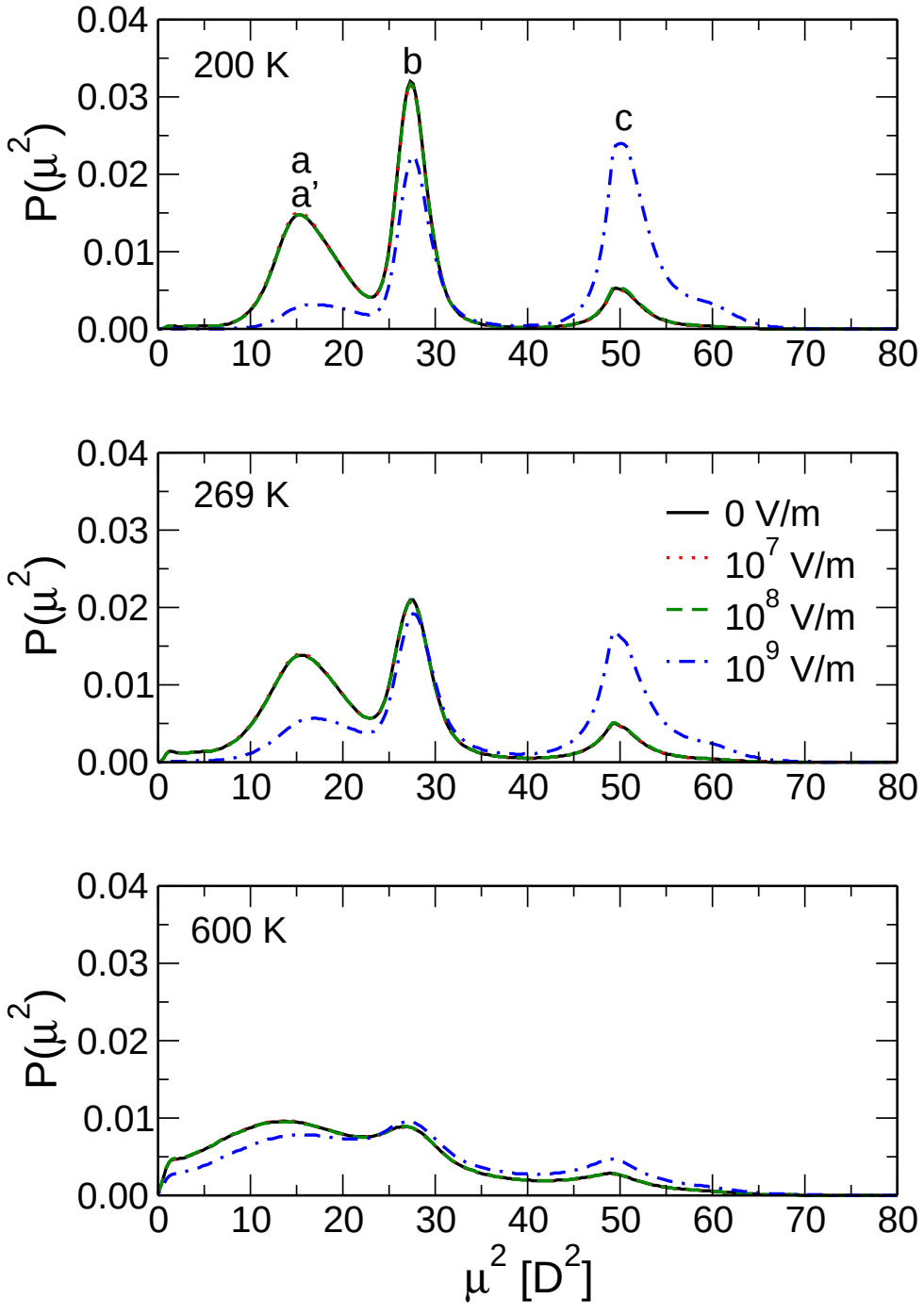


FIG. 3.6 – *Distribution de la probabilité du carré du dipôle électrique pour différentes valeurs du champ électrique à T = 200 K, 269 K et 600 K. Les symboles a, a', b et c correspondent aux valeurs du carré du dipôle des structures représentées figure 3.7.*

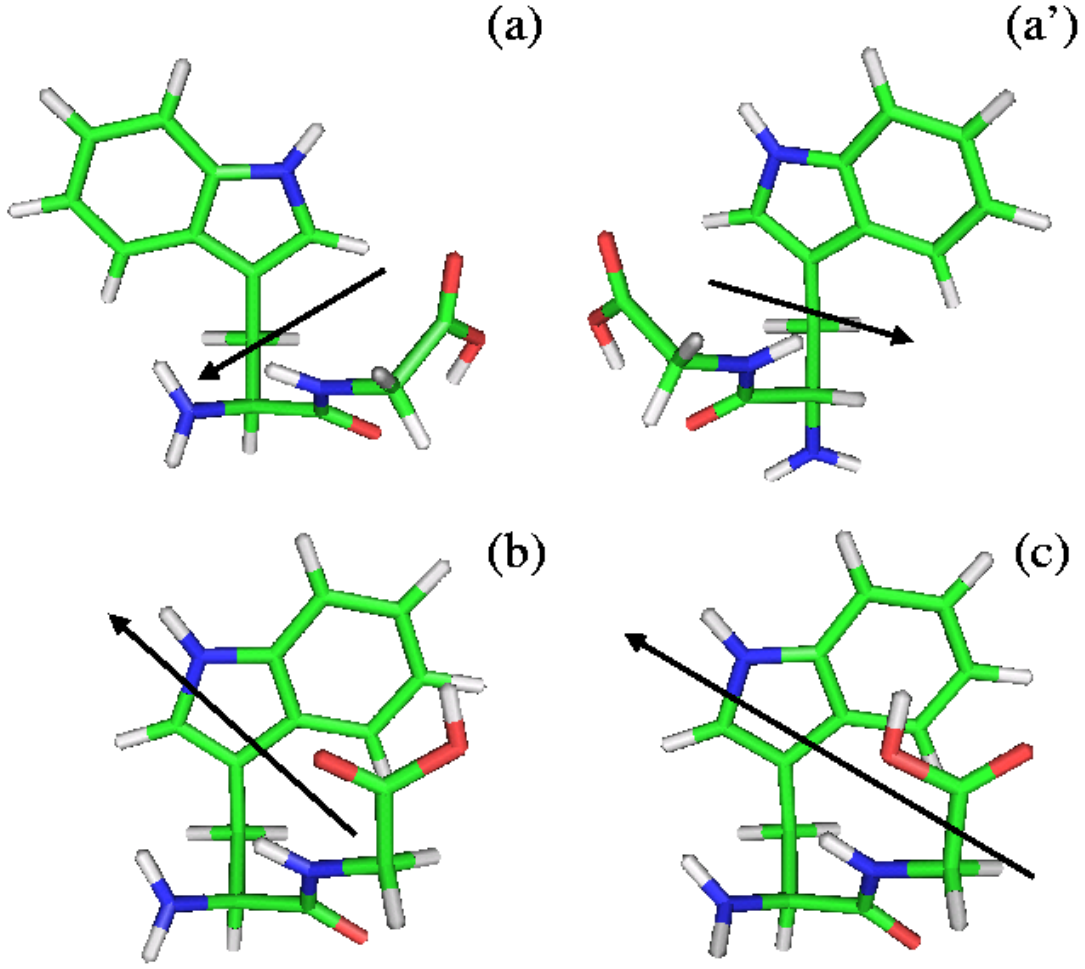


FIG. 3.7 – *Structures représentatives obtenues à 269 K lors de la simulation et représentées avec PyMOL [20]. Les atomes de carbone sont représentés en vert, les azotes en bleu, les oxygènes en rouge et les atomes d'hydrogène en gris. (a) E = 29,56 kcal/mol et* μ^2 *= 15,18 D*^{2}*. (a') E = 29,95 kcal/mol et* μ^2 *= 14,66 D*^{2}*. (b) E = 29,79 kcal/mol et* μ^2 *= 27,12 D*^{2}*. (c) E = 30,43 kcal/mol et* μ^2 *= 51,52 D*^{2}*. Les flèches noires montrent l'orientation et l'ampleur du dipôle permanent des molécules.*

ont été en partie attribuées à des modifications de conformation [23] sur lesquelles nous reviendrons dans le chapitre 4. Les deux peptides ont été modélisés en phase gazeuse par le champ de force AMBER [24] avec les paramètres *ff96* [18] et une constante diélectrique $\epsilon = 2$.

Pour ces deux systèmes, des simulations Monte Carlo d'échange avec des statistiques élevées sont utilisées comme référence. La densité d'états et les propriétés thermiques associées sont obtenues par la méthode des histogrammes multiples. Les simulations Monte Carlo d'échange ont été effectuées avec 19 répliques entre 100 K et 1000 K pour Ala_8 et avec 9 répliques entre 180 K et 1100 K pour $AGWLK^+$. Les températures des différentes répliques suivent une progression à 85 % géométrique [25, 26, 27, 28]. Un cycle Monte Carlo consiste à déplacer aléatoirement chaque angle dièdre du squelette peptidique (en incluant les angles Φ et Ψ) et des chaînes latérales une fois chacun. Les longueurs des liaisons et les angles de flexion sont maintenus constants. Les géométries de départ sont initialisées aléatoirement. La simulation pour Ala_8 comprend un total de $6{,}2 \times 10^9$ pas Monte Carlo dont les $6{,}2 \times 10^8$ premiers pas sont utilisés pour la thermalisation. La simulation pour $AGWLK^+$ est composée de $6{,}8 \times 10^8$ pas Monte Carlo dont $6{,}8 \times 10^7$ pas de thermalisation.

3.2.1 Wang-Landau à une dimension d'énergie

Les différentes variantes de l'algorithme Wang-Landau ont été implémentées en suivant l'évolution du facteur de modification schématisée figure 3.8. Toutes les simulations débutent avec $f_0 = e^1$ et se terminent avec la même précision $\ln f \simeq 2 \times 10^{-6}$. Le coût cumulé en temps de calcul est le même pour toutes les méthodes, à savoir $4{,}2 \times 10^8$ pas Monte Carlo pour Ala_8 et $AGWLK^+$.

Pour les simulations Wang-Landau standard et ZB, on compte 20 itérations, le nombre de pas par itération étant dépendant de f uniquement dans le dernier cas. L'implémentation de JSM est composée de 50 itérations et 85 sous-étapes par itération. Les 40 premières itérations débutent avec $f_0 = e^1$, les 9 suivantes avec $f_0 = e^1/10$ et la dernière avec $f_0 = e^1/100$. Comme dans l'article original [29], le facteur de modification est diminué par $f^{0,9} \to f$ après chaque sous-étape. Notre implémentation du Wang-Landau recuit est constituée de 16 itérations, elles-mêmes divisées en 5 sous-étapes. Le facteur f est modifié par $\sqrt{f} \to f$ comme dans l'algorithme Wang-Landau standard [30, 31]. Ces valeurs de paramètres ont été choisies pour permettre à $\ln f$ de varier d'un ordre de grandeur dans une itération. Les variations réelles de f pour chaque implémentation Wang-Landau sont représentées dans la figure 3.8. Pour les méthodes WL-ZB et WL-recuit, la diminution de f est relativement douce comparativement aux deux autres méthodes. Enfin, l'intervalle d'énergie est discrétisé en 400 cases pour toutes les simulations.

Les capacités calorifiques obtenues à partir de Monte Carlo d'échange et des simulations Wang-Landau sont représentées figure 3.9 pour les deux polypeptides. Pour Ala_8, la courbe de capacité calorifique présente un pic très net proche de 250 K, ce qui est

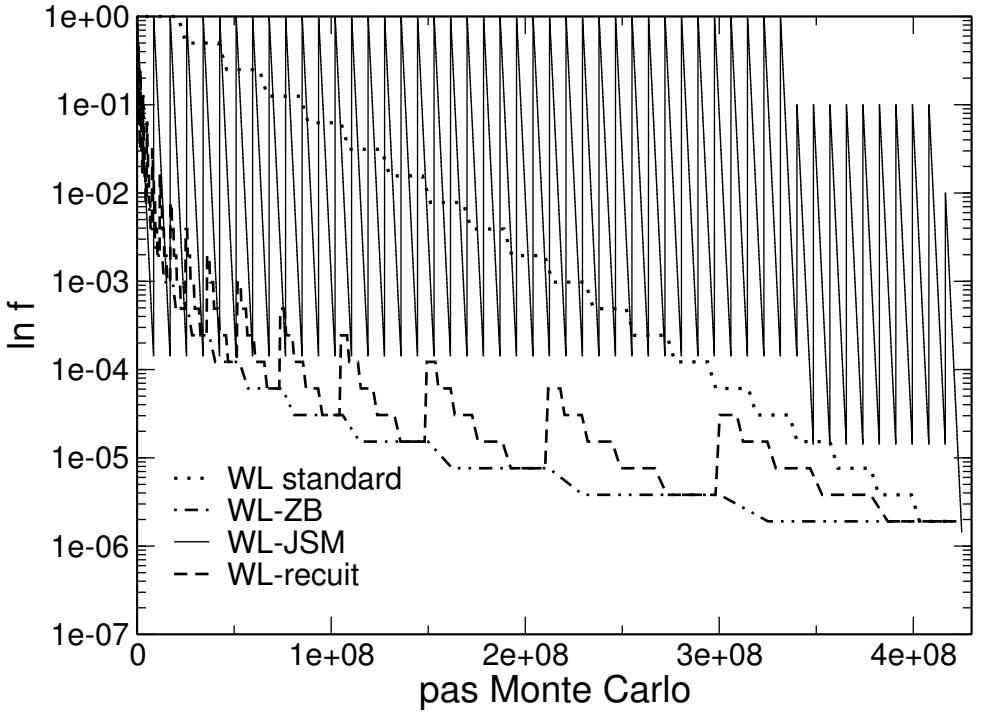


FIG. 3.8 – *Variations du facteur de modification f pour les simulations Wang-Landau des peptides Ala_8 et $AGWLK^+$. Tous les calculs débutent avec $f_0 = e^1$ et se terminent avec $\ln f \simeq 2 \times 10^6$ pour un coût cumulé de $4{,}2 \times 10^8$ pas Monte Carlo.*

comparable aux résultats obtenus pour Ala_{10} par Hansmann *et al.* avec le champ de force ECEPP/2 [32, 33]. Seule la simulation Wang-Landau recuit reproduit correctement la courbe complète et le pic observé, avec moins de pas Monte Carlo que pour les calculs Monte Carlo d'échange. Les méthodes Wang-Landau standard et ZB divergent clairement, respectivement en surestimant et sous-estimant la température de transition. Le désaccord le plus flagrant est obtenu pour l'implémentation JSM, qui ne semble pas converger pour ce système.

La courbe de chaleur spécifique de $AGWLK^+$ ne présente pas un pic net comme c'est le cas pour la polyalanine. Ici, la transition entre l'état fondamental et la pelote statistique se fait de manière plus continue comme on pourrait l'attendre vu la taille plus modeste du pentapeptide. Les résultats de l'implémentation Wang-Landau recuit sont les plus proches des données obtenues par Monte Carlo d'échange, avec un nombre similaire de pas Monte Carlo. Les algorithmes WL-standard et WL-ZB sont en meilleur accord que précédemment mais la méthode WL-JSM reste inefficace en termes de convergence globale de la capacité calorifique.

Nous avons répété les simulations précédentes de façon indépendante, sans obtenir de résultats sensiblement meilleurs. Seule l'implémentation Wang-Landau recuit s'avère reproductible. Il semble donc que l'algorithme WL-recuit avec l'évolution de f proposée soit la plus efficace pour simuler les deux peptides étudiés.

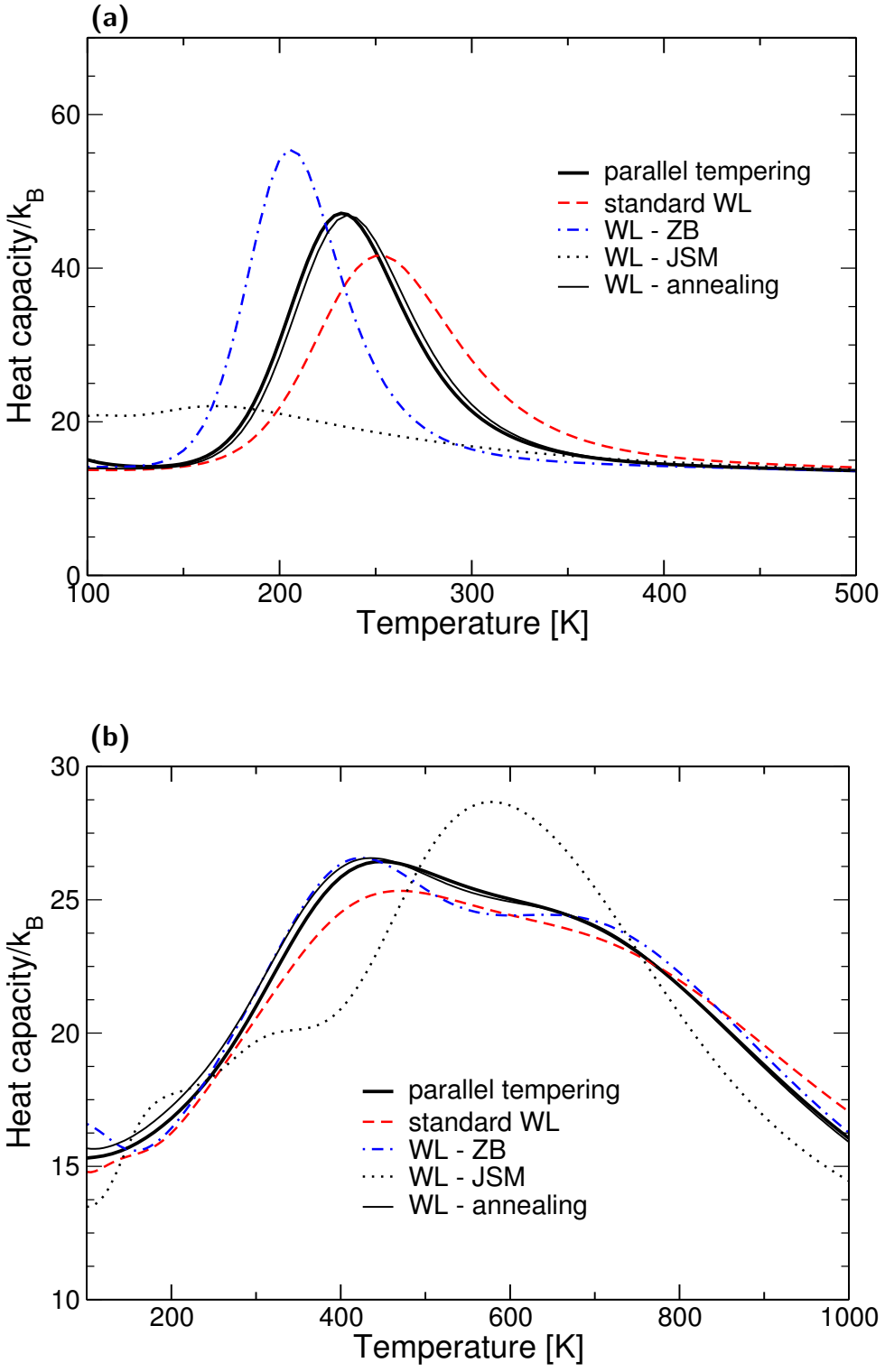


FIG. 3.9 – *Courbes de capacité calorifique pour les polypeptides en phase gazeuse, en fonction de la température. Les résultats des différentes implémentations de la méthode Wang-Landau sont comparés avec les données provenant de Monte Carlo d'échange (*parallel tempering*). (a)* Ala_8*. (b)* $AGWLK^+$*.*

3.2.2 Évolution du paramètre d'ordre

Nous avons déterminé la moyenne thermique $\langle d \rangle$ de la distance entre l'azote du groupe indole du tryptophane et l'azote protoné de la chaîne latérale de la lysine de $AGWLK^+$. Cette distance procure un bon paramètre d'ordre pour distinguer la forme native de la pelote statistique. Pour cela, nous avons mené une simulation multicanonique supplémentaire pour construire l'histogramme à deux dimensions, énergie et distance, $H(E, d)$. Un total de $1{,}5 \times 10^9$ configurations ont été échantillonnées en utilisant la densité d'états en énergie calculée par la méthode Wang-Landau recuit. La figure 3.10 représente les variations de $\langle d \rangle$ avec la température obtenues à partir de Monte Carlo d'échange (avec ou sans repondération par les histogrammes multiples) et de la simulation multicanonique. La moyenne $\langle d \rangle$ présente une augmentation monotone avec la température, dont le taux maximal coïncide à la température du pic de la capacité calorifique. La distance d apparaît effectivement comme une bonne sonde des changements de conformation du peptide $AGWLK^+$. Le très bon accord obtenu entre les deux méthodes de calcul indique que notre implémentation de l'algorithme Wang-Landau est compétitive par rapport au Monte Carlo d'échange pour ce système.

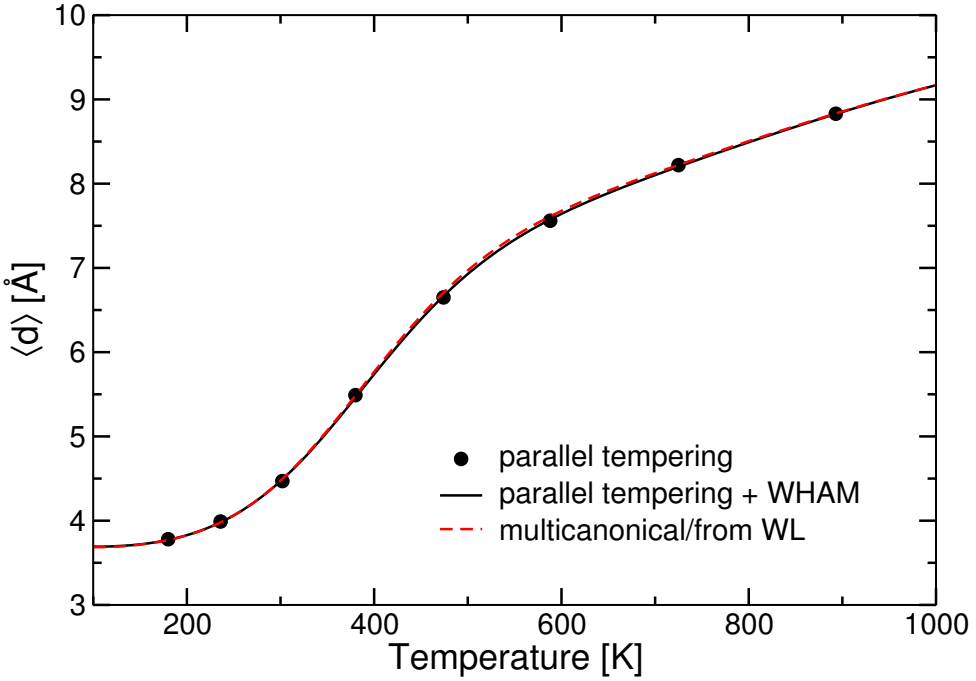


FIG. 3.10 – *Moyenne thermique $\langle d \rangle$ de la distance entre l'azote du groupe indole du tryptophane et l'azote protoné de la chaîne latérale de la lysine de $AGWLK^+$. Données obtenues à partir des simulations Monte Carlo d'échange et multicanonique. La densité d'états utilisée dans le calcul multicanonique provient de la méthode Wang-Landau recuit. Pour le Monte Carlo d'échange, les résultats directs de la simulation Monte Carlo d'échange aux différentes températures et les données continues après analyse par histogrammes multiples sont représentés.*

3.2.3 Wang-Landau à deux dimensions

Nous avons également effectué des simulations Wang-Landau à deux dimensions. Les paramètre d'ordre utilisés sont respectivement le dipôle électrique pour Ala_8, et la distance entre l'azote du groupe indole du tryptophane et l'azote protoné de la chaîne latérale de la lysine, pour $AGWLK^+$. Les variations du facteur de modification f sont identiques à celles de notre implémentation recuit. Chaque simulation est initialisée avec $f_0 = e^1$, se termine lorsque $\ln f \simeq 2 \times 10^{-6}$ et compte 2×10^9 pas Monte Carlo.

Dans toutes les simulations, l'intervalle du paramètre d'ordre est discrétisé en 150 points. Avec 400 cases en énergie, la simulation doit effectuer une marche aléatoire sur au plus 60000 cases, ce qui implique un effort numérique beaucoup plus important que par Wang-Landau à une seule dimension d'énergie.

La figure 3.11 représente le logarithme de la densité d'états obtenu à deux dimensions pour le pentapeptide $AGWLK^+$. On remarque la présence de zones non-physiques qui ne sont pas explorées, à la fois en énergie, comme pour le Wang-Landau à une dimension, mais aussi en coordonnée de réaction.

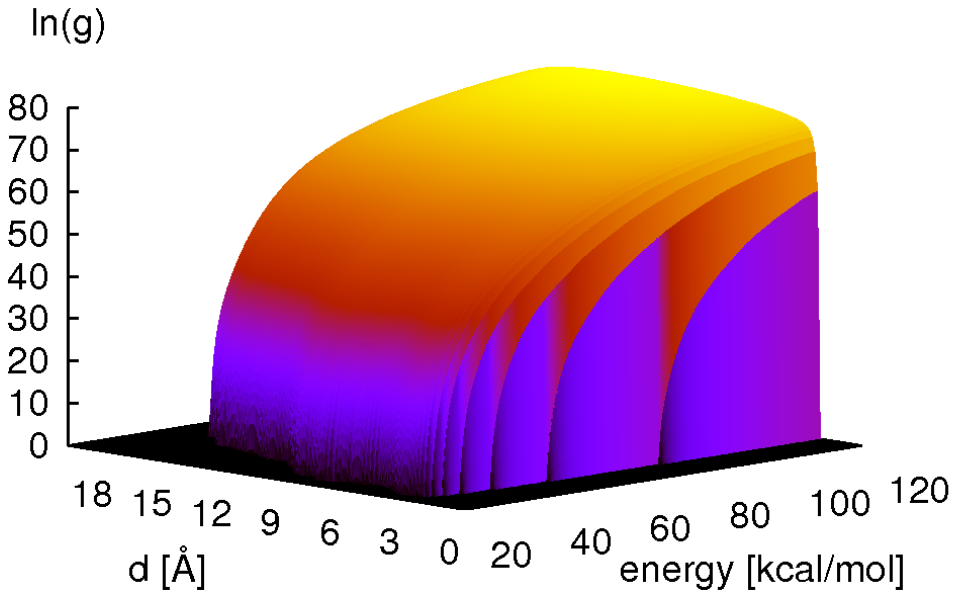


FIG. 3.11 – *Logarithme de la densité d'états pour une simulation à deux dimensions du pentapeptide $AGWLK^+$*

Comme nous l'avons expliqué dans la section précédente, il est possible de calculer la moyenne canonique $\langle \mathcal{A} \rangle$ du paramètre d'ordre à partir de la densité d'états obtenue à partir d'une simulation Wang-Landau à deux dimensions.

La figure 3.12 illustre, pour les deux peptides, les capacités calorifiques déterminées à

partir de Monte Carlo d'échange, de la méthode WL-recuit à une dimension en énergie et de la méthode Wang-Landau à deux dimensions. Un très bon accord est obtenu entre les simulations Monte Carlo d'échange et les simulations Wang-Landau à une seule dimension. À faible température, la construction d'une entropie à deux dimensions semble donner des résultats légèrement supérieurs.

Contrairement à la méthode Wang-Landau à une dimension, le calcul de la moyenne du paramètre d'ordre se fait sans simulation supplémentaire, en utilisant l'équation (2.64).

La figure 3.13 représente la moyenne canonique de d en fonction de la température pour AGWLK$^+$. Les résultats obtenus sont indistinguables de ceux calculés par Monte Carlo d'échange et par simulation multicanonique avec la densité d'états issue de la simulation Wang-Landau recuit à une dimension.

3.2.4 Convergence globale et temps tunnel

Le temps tunnel est défini comme le nombre moyen de pas Monte Carlo nécessaires pour traverser l'intervalle d'énergie considéré et revenir [34]. Le temps tunnel est fortement corrélé avec la topologie de la surface d'énergie explorée [34]. Dayal *et al.* [35] ont récemment étudié ce critère pour évaluer les performances des méthodes comme l'algorithme Wang-Landau à produire des histogrammes plats. Pour des surfaces d'énergie potentielle assez simples, un temps tunnel court est un bon indicateur qui traduit le succès d'un échantillonnage en histogrammes plats [34]. Dans un travail récent, Costa *et al.* [36] ont montré que le temps tunnel nécessaire pour passer d'une région de faible à haute entropie est plus long que celui pour passer d'une région de haute vers basse entropie.

Avec les deux peptides précédents (Ala_8 et AGWLK$^+$), nous avons choisi de suivre la convergence des différentes méthodes Wang-Landau avec une observable thermodynamique comme la capacité calorifique. Nous avons également mesuré le temps tunnel nécessaire pour traverser l'intervalle d'énergie, des basses vers les hautes énergies (L→H) et inversement (H→L). Les 5 premières et les 5 dernières cases du domaine d'énergie accessible sont arbitrairement utilisées pour définir respectivement les états de basse (L) et de haute (H) énergie. Les variations du nombre d'évènements tunnel au cours des simulations Wang-Landau pour les polypeptides Ala_8 et AGWLK$^+$ sont représentées figure 3.14. Le nombre d'évènements tunnels pour les deux systèmes est assez faible, quelques centaines d'évènements seulement par simulation. Une des simulations qui présente le plus d'évènements tunnels est la simulation Wang-Landau JSM qui, pourtant, n'est pas capable de reproduire les capacités calorifiques des deux peptides. L'implémentation Wang-Landau recuit n'offre pas un nombre d'évènements tunnels très élevé.

Les temps tunnels moyens pour chaque simulation sont rassemblés dans le tableau 3.2. La méthode recuit présente des temps moyens parmi les plus élevés. On note également l'absence de corrélation entre le temps tunnel moyen et le nombre d'évènements tunnels.

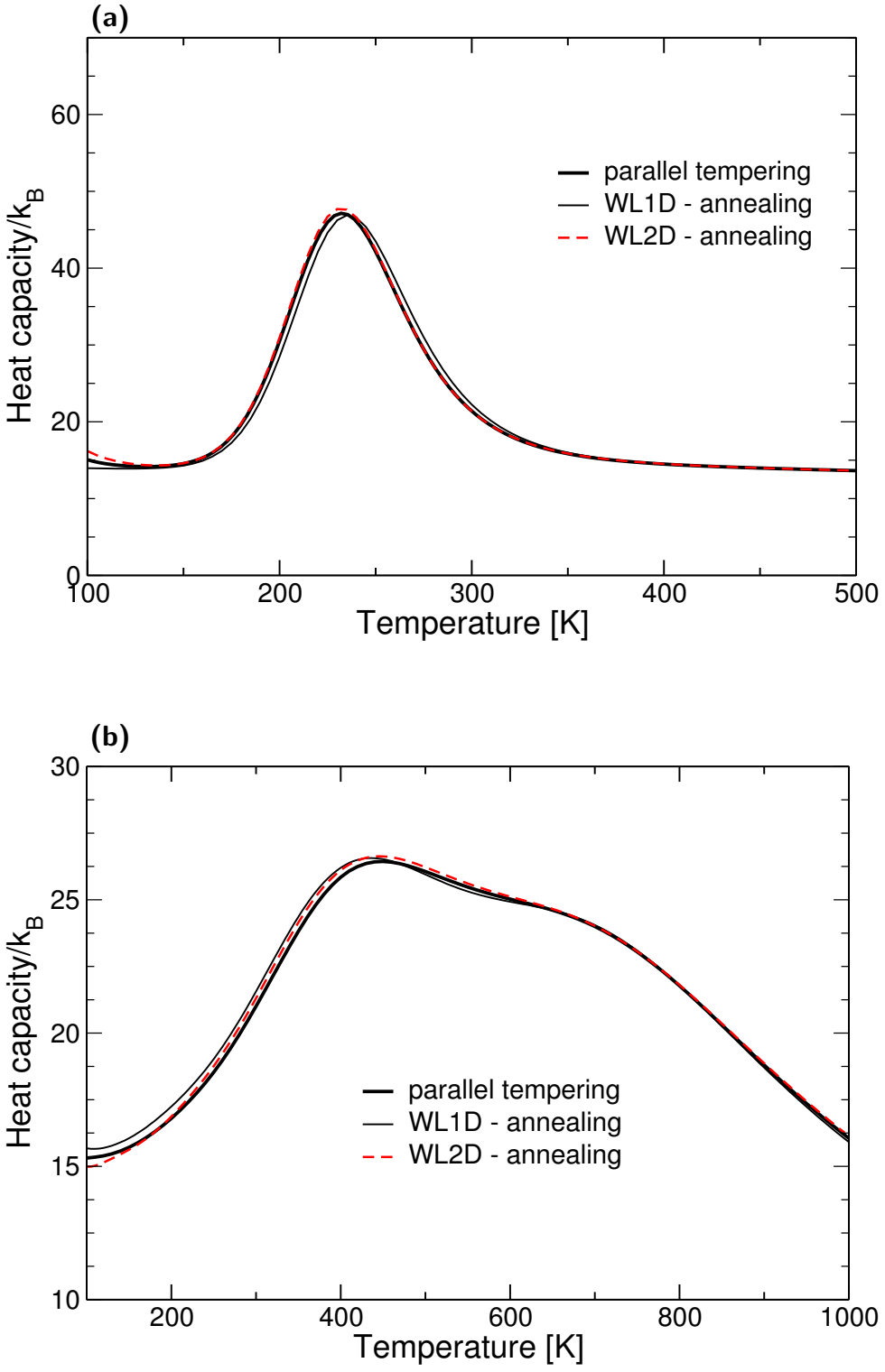


FIG. 3.12 – *Capacités calorifiques obtenues par Monte Carlo d'échange, analyse des histogrammes multiples, Wang-Landau à une dimension (E) et Wang-Landau à deux dimensions ($E, \mathcal{A}$). Pour les deux résultats Wang-Landau, l'implémentation recuit est utilisée. (a) Ala_8 avec $\mathcal{A}$ = dipôle électrique, (b) $AGWLK^+$ avec $\mathcal{A} = d$.*

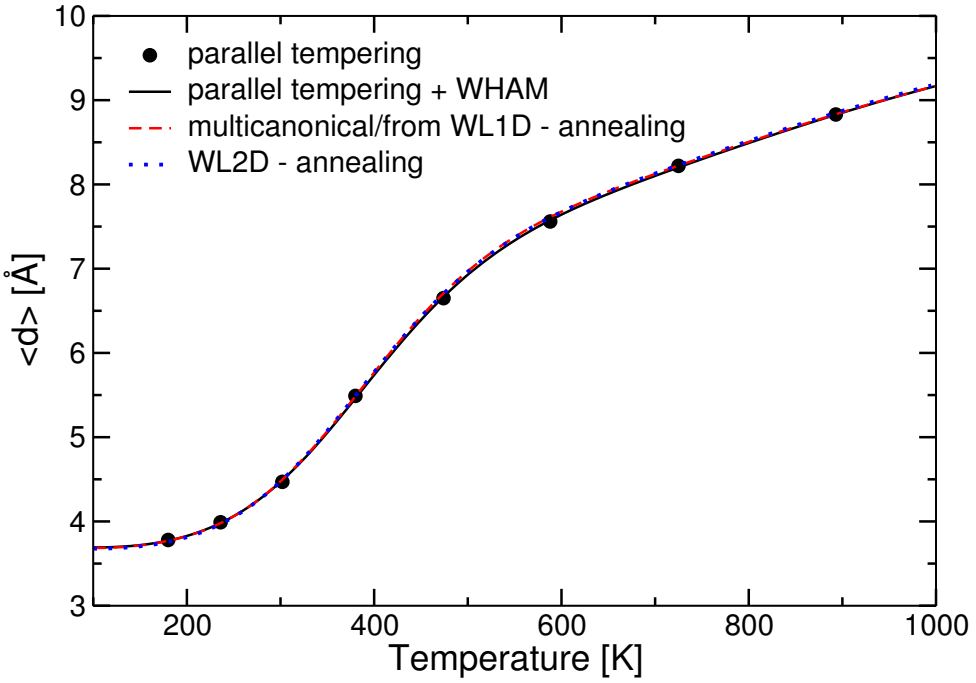


FIG. 3.13 – *Moyenne canonique $\langle d \rangle$ de la distance entre l'azote du groupe indole du tryptophane et l'azote protoné de la chaîne latérale de la lysine de $AGWLK^+$. Les résultats sont tirés d'une simulation Monte Carlo d'échange, sans et avec analyse par les histogrammes multiples, d'une simulation multicanonique dont la densité d'états provient d'une simulation Wang-Landau recuit à une dimension, et d'une simulation Wang-Landau recuit à deux dimensions.*

En particulier, la simulation avec le temps tunnel moyen le plus petit ne produit pas nécessairement le plus d'évènements.

TAB. 3.2 – *Temps tunnel moyens $\langle \tau \rangle$ et nombres d'évènements tunnels ($L \rightarrow H$ et $H \rightarrow L$) pour chaque simulation Wang-Landau.*

	Ala_8		$AGWLK^+$	
	$\langle \tau \rangle$ [pas MC]	évènements tunnel	$\langle \tau \rangle$ [pas MC]	évènements tunnel
original	56557	106	70341	168
ZB	169665	64	97879	158
JSM	228131	86	86108	234
recuit	182702	59	109359	177

La faible corrélation entre le nombre d'évènements tunnel et la qualité des courbes de capacité calorifique montre que le temps tunnel n'est pas un critère fiable pour quantifier à lui seul la bonne convergence d'une simulation Wang-Landau. Le temps tunnel détermine le temps nécessaire au système pour parcourir la surface d'énergie étudiée entre ses deux extrema. Cependant, il ne quantifie pas la qualité de la simulation puisqu'il ne prend pas en compte la manière dont l'exploration de la surface d'énergie potentielle est effectuée. En particulier, les premières étapes de la simulation Wang-Landau sont impor-

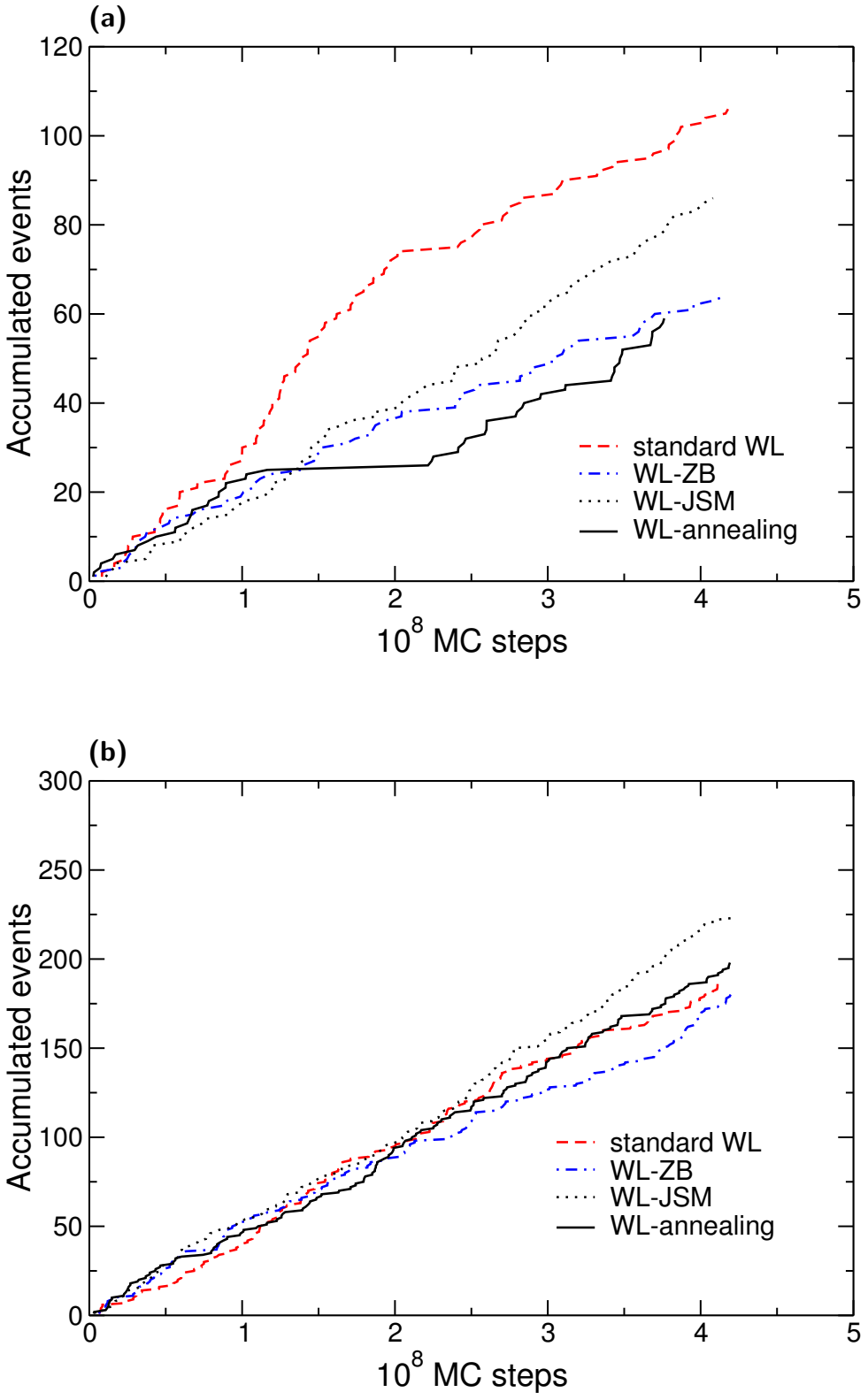


FIG. 3.14 – *Évolution du nombre d'évènements tunnel au cours des simulations de différentes implémentations Wang-Landau. (a) Ala_8. (b) $AGWLK^+$.*

tantes puisqu'elles contribuent fortement à l'allure générale de la densité d'états et donc à la convergence ultérieure de la méthode Wang-Landau. Le temps d'équilibre et le temps tunnel, bien que reliés, peuvent être significativement différents [36].

3.3 Conclusion

L'utilisation d'algorithmes dans les ensembles généralisés, parmi lesquelles la méthode Monte Carlo d'échange et la méthode Wang-Landau, permet un échantillonnage effectif de la surface d'énergie potentielle des biomolécules

Dans un premier exemple, nous avons réalisé des simulations en Monte Carlo d'échange pour comprendre l'influence d'un champ électrique intense et statique sur la conformation de biomolécules. Puis, nous avons comparé une adaptation de la méthode Wang-Landau pour les systèmes continus avec le Monte Carlo d'échange. Les algorithmes offrent des performances comparables et sont adéquates pour la simulation de peptides. La méthode Wang-Landau nécessite cependant un paramétrage parfois long.

Enfin, il faut souligner que les systèmes étudiés ici sont des systèmes pour lesquels la méthode Monte Carlo d'échange converge particulièrement bien. Des systèmes plus grands, avec plus de degrés de liberté, pourraient rendre cette stratégie moins évidente. La méthode Wang-Landau pourrait alors être une alternative pertinente, qui donnerait accès à une bonne estimation de la densité d'états pour un coût numérique inférieur à celui du Monte Carlo d'échange.

Bibliographie

[1] P. Poulain, R. Antoine, M. Broyer, and P. Dugourd. Monte Carlo simulations of flexible molecules in a static electric field: electric dipole and conformation. *Chemical Physics Letters*, 401:1–6, 2005.

[2] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[3] A. Warshel and J. Aqvist. Electrostatic Energy and Macromolecular Function. *Annual review of biophysics and biophysical chemistry*, 20:267–298, 1991.

[4] M. Costabel, D. F. Vallejo, and J. R. Grigera. Electrostatic Recognition between Enzyme and Inhibitor: Interaction between Papain and Leupeptin. *Archives of Biochemistry and Biophysics*, 394:161–166, 2001.

[5] W. G. J. Hol. The role of the alpha-helix dipole in protein function and structure. *Progress in Biophysics and Molecular Biology*, 45:149–195, 1985.

[6] A. Wada. The alpha-helix as an electric macro-dipole. *Advances in Biophysis*, 9:1–63, 1976.

[7] C. N. Schutz and A. Warshel. What are the dielectric constants of proteins and how to validate electrostatic models? *Polymer*, 44:400–417, 2001.

[8] T. Simonson, D. Perahia, and G. Bricogne. Intramolecular dielectric screening in proteins. *Journal of Molecular Biology*, 218:859–886, 1991.

[9] T. Simonson. Dielectric relaxation in proteins: Microscopic and macroscopic models. *International Journal of Quantum Chemistry*, 73:45–57, 1999.

[10] R. Antoine, I. Compagnon, D. Rayane, M. Broyer, P. Dugourd, G. Breaux, F. C. Hagemeister, D. Pippen, R. R. Hudgins, and M. F. Jarrold. Electric susceptibility of unsolvated glycine-based peptides. *Journal of American Chemical Society*, 124:6737–6741, 2002.

[11] K. D. Bonin and V. V. Kresin. *Electric-dipole polarizabilities of atoms, molecules and clusters*. World Scientific, Singapore, 1997.

[12] M. Broyer, R. Antoine, E. Benichou, I. Compagnon, P. Dugourd, and D. Rayane. Structure of nano-objects through polarizability and dipole measurements. *C.R. Physique*, 3:301–317, 2002.

[13] I. Compagnon. *Mesure de dipôle électrique en phase gazeuse : application aux agrégats et aux biomolécules*. PhD thesis, Université Claude Bernard - Lyon I, Lyon, 2003.

[14] M. Abd El Rahim. *Déflexion électrique d'un jet moléculaire : progrès expérimentaux et théoriques*. PhD thesis, Université Claude Bernard - Lyon I, Lyon, 2005.

[15] P. Debye. *Polar molecules*. Dover, New York, 1929.

[16] J. H. Van Vleck. On Dielectric Constants and Magnetic Susceptibilities in the New Quantum Mechanics. Part II Application to Dielectric Constants. *Physical Review*, 30:31–54, 1927.

[17] R. Antoine, I. Compagnon, D. Rayane, M. Broyer, Ph. Dugourd, G. Breaux, F.C. Hagemeister, D. Pippen, R. R. Hudgins, and M. F. Jarrold. Electric dipole moments and conformations of isolated peptides. *The European Physical Journal D*, 20:583–587, 2002.

[18] P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille. The Development Application of a 'Minimalist' Organic Biochemical Molecular Mechanic Force Field using a Combination of ab Initio Calculations and Experimental Data. In W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors, *Computer Simulations of Biomolecular Systems*, volume 3, pages 83–96, Dordrecht, The Netherlands, 1997. Kluwer Academic.

[19] R. Antoine, M. Broyer, P. Dugourd, G. Breaux, F. C. Hagemeister, D. Pippen, R. R. Hudgins, and M. F. Jarrold. Direct probing of zwitterion formation in unsolvated peptides. *Journal of American Chemical Society*, 125:8996–8997, 2003.

[20] W. L. Delano. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA, 2002.

[21] E. J. Sorin and V. S. Pande. Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling. *Journal of Computational Chemistry*, 26:682, 2005.

[22] T. Tabarin, R. Antoine, M. Broyer, and P. Dugourd. Specific photodissociation of peptides with multi-stage mass spectrometry. *Rapid Communications in Mass Spectrometry*, 19:2883–2892, 2005.

[23] R. Antoine, M. Broyer, J. Chamot-Rooke, C. Dedonder, C. Desfrancois, P. Dugourd, G. Gregoire, C. Jouvet, P. Poulain, T. Tabarin, and G. van der Rest. Comparison of the fragmentation pattern induced by collisions, laser excitation and electron capture. Influence of the initial excitation. *Rapid Communications in Mass Spectrometry*, 20:1648–1652, 2006.

[24] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, Jr. S. Profeta, and P. K. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of American Chemical Society*, 106:765–784, 1984.

[25] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *Journal of Chemical Physics*, 113:6042–6051, 2000.

[26] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. *Biopolymers*, 60:96–123, 2001.

[27] D. A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics*, 117:6911–6914, 2002.

[28] K. Y. Sanbonmatsu and A. E. García. Structure of Met-Enkephalin in Explicit Aqueous Solution Using Replica Exchange Molecular Dynamics. *Proteins: Structure, Function, and Genetics*, 46:225–234, 2002.

[29] D. Jayasri, V. S. S. Sastry, and K. P. N. Murthy. Wang-Landau Monte Carlo simulation of isotropic-nematic transition in liquid crystals. *Physical Review E*, 72:036702, 2005.

[30] F. Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters*, 86:2050–2053, 2001.

[31] F. Wang and D. P. Landau. Determining the density of states for classical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64:056101, 2001.

[32] U. H. E. Hansmann and Y. Okamoto. New Monte Carlo algorithms for protein folding. *Current Opinion in Structural Biology*, 9:177–183, 1999.

[33] Y. Peng and U. H. E. Hansmann. Solvation Model Dependency of Helix-Coil Transition in Polyalanine. *Biophysical Journal*, 82:3269–3276, 2002.

[34] M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos. Flat-Histogram Dynamics and Optimization in Density of States Simulations of Fluids. *Journal of Physical Chemistry B*, 108:19748–19755, 2004.

[35] P. Dayal, S. Trebst, S. Wessel, D. Würtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith. Performance Limitations of Flat-Histogram Methods. *Physical Review Letters*, 92:097201, 2004.

[36] M. D. Costa, J. Viana Lopes, and J. M. B. Lopes dos Santos. Analytical study of tunneling times in flat histogram Monte Carlo. *Europhysics Letters*, 72:802–808, 2005.

Chapitre 4

Dipôle électrique et conformations de polyalanines

Sommaire

Comprendre les facteurs qui stabilisent les éléments de structure secondaire comme les hélices α ou les feuillets β est essentiel dans l'étude du repliement des protéines. En effet, une étape préliminaire à la formation de la structure globale biologiquement active est la formation de domaines pourvus d'une structure secondaire déterminée. Un des points abordé dans ce chapitre sera la compétition entre les hélices α et les feuillets β. Cette problématique est d'ailleurs particulièrement importante pour les maladies conformationnelles où des protéines solubles peuvent adopter des structures quaternaires insolubles donnant lieu à des agrégats de fibres amyloïdes pathogènes [1]. La plupart des protéines impliquées dans ces fibres pathogènes présentent une conformation secondaire majoritairement constituée de feuillets β. Pourtant, la structure native du peptide Aβ intervenant dans la maladie d'Alzheimer [2] ainsi que celles des protéines du prion (PrP^c) [3] sont très riches en hélices α. Les phénomènes d'agrégation impliquent donc une transition conformationnelle des hélices α vers des feuillets β. Une telle transition a déjà été observée lors d'études *in vitro* de peptides dont la structure native était en hélice α et qui s'accumulaient en fibres amyloïdes de structure β si l'environnement de ces molécules était modifié [4, 5].

4.1 Étude expérimentale des polyalanines

L'étude des changements conformationnels qui s'opèrent dans les protéines est donc un problème important mais difficile à cause des nombreux paramètres qui peuvent intervenir (séquence peptidique, liaisons intermoléculaires, nature du solvant, température...). L'équipe de Philippe Dugourd [6] a utilisé une approche plus fondamentale en étudiant *in vacuo* l'influence de la taille et de la température sur le dipôle électrique d'une série de polyalanines, AceTrpAla$_n$NH$_2$ (AcWA$_n$NH$_2$) avec n = 3, 5, 10, 13 et 15 (figure 4.1). Le dispositif expérimental utilisé est le montage de déflexion électrique d'un jet moléculaire décrit dans la première partie du chapitre 3.

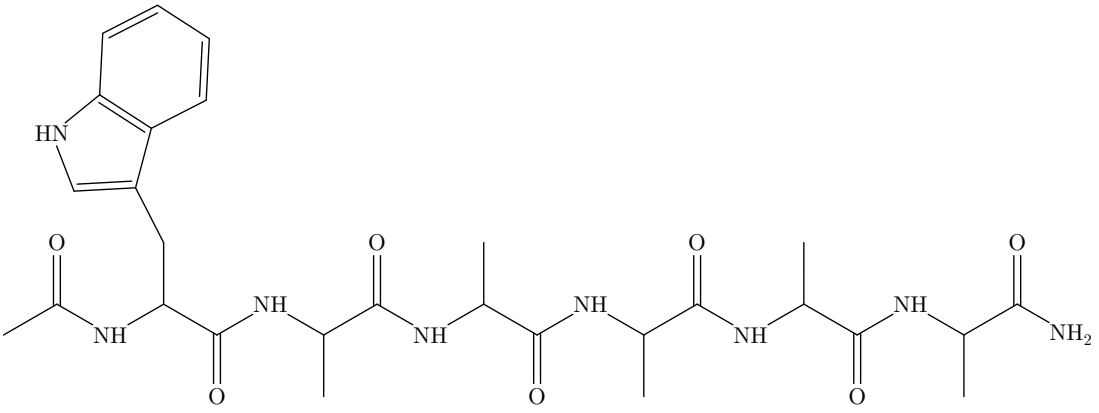


FIG. 4.1 – *Formule topologique du peptide AceTrpAla$_5$NH$_2$ (AcWA$_5$NH$_2$). La terminaison Ace représente le groupe acétyle CH$_3$CO.*

Le dipôle électrique est utilisé expérimentalement comme sonde conformationnelle. Un fort dipôle est révélateur d'une structure secondaire en hélice α alors qu'une structure en feuillet β aura un faible dipôle. Les dipôles mesurés à 300 K pour différentes tailles de peptides sont représentés figure 4.2 et comparés aux dipôles moyens attendus pour des structures de type hélice α, feuillet β et pelote statistique. Le dipôle mesuré est très proche du dipôle calculé pour une structure en feuillet β. Cette conclusion est étayée par l'évolution du dipôle électrique mesuré en fonction de la température pour le peptide AcWA$_{10}$NH$_2$ (figure 4.3). Entre 400 et 500 K, le dipôle augmente brutalement, ce qui est en faveur d'un dépliement que l'on pourrait attribuer à une transition du feuillet β vers la pelote statistique.

4.2 Approches théoriques des polyalanines

Ces observations expérimentales montrent que des polyalanines neutres isolées adoptent, à température ambiante, une structure secondaire majoritairement de type feuillet β. Ces résultats sont pourtant en désaccord avec la plupart des calculs réalisés jusqu'à présent sur ces systèmes [7, 8, 9]. Les polyalanines Ala$_n$ ont donné lieu à de nombreux travaux

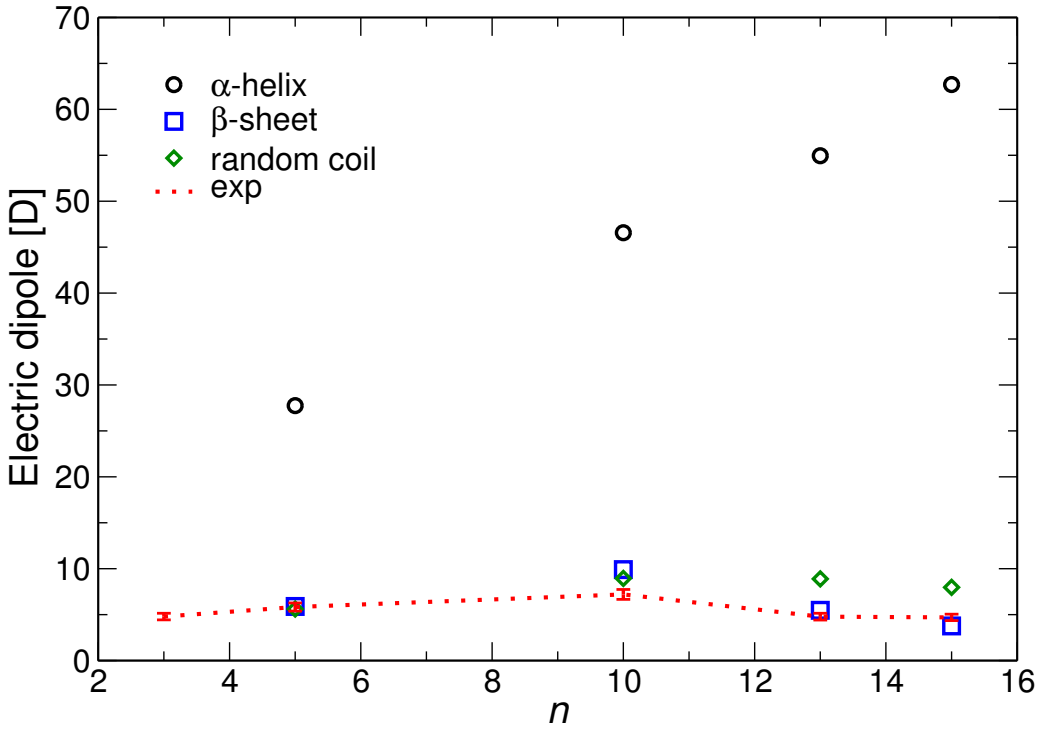


FIG. 4.2 – *Évolution expérimentale du dipôle électrique du peptide $AcWA_nNH_2$ en fonction du nombre d'alanines. Les résultats expérimentaux sont comparés aux prédictions théoriques obtenues par optimisation d'une structure secondaire spécifique.*

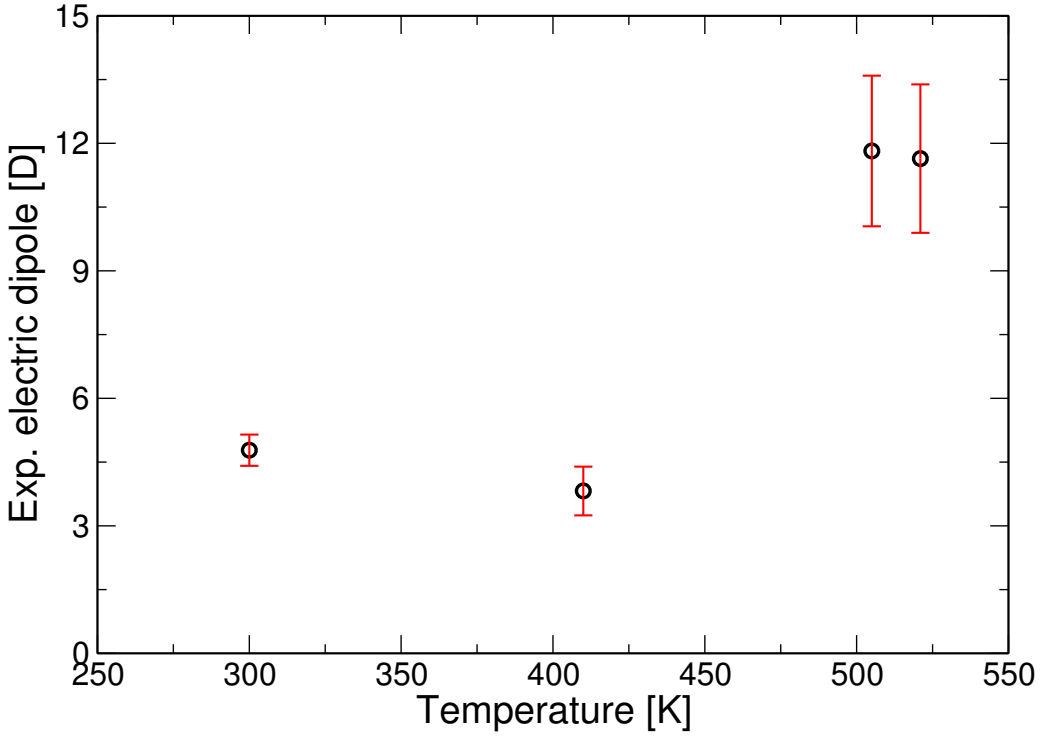


FIG. 4.3 – *Évolution expérimentale du dipôle électrique du peptide $AcWA_{13}NH_2$ en fonction de la température.*

théoriques car elles sont relativement simples à modéliser. La chaîne latérale de l'alanine se résume en effet à un groupe méthyle. Hansmann *et al.* [7] ont effectué des simulations multicanoniques sur les peptides NH_2Ala_nCOOH, avec n = 10, 15, 20 et 30, modélisés par le champ de force ECEPP/2 [10, 11, 12]. Les courbes de capacité calorifique obtenues par ces auteurs montrent un seul pic identifié comme une transition hélice α–pelote statistique. Pour Ala_{10}, par exemple, cette transition est située à 423 K [7]. Toujours en phase gazeuse, Mitsutake *et al.* [8] ont eux aussi étudié par la méthode multicanonique le peptide Ala_{10} modélisé par le champ de force ECEPP/2. Une seule transition est observée, à 420 K. Enfin plus récemment, Rathore *et al.* [9] ont calculé la capacité calorifique du même peptide modélisé par une représentation en atomes unifiés et le champ de force CHARMM 19 [13]. La méthode Wang-Landau à une dimension en énergie a été utilisée. Un pic de capacité calorifique est obtenu pour une température proche de 260 K mais les modifications structurales n'ont pas été caractérisées.

Enfin, des études prenant en compte l'influence du solvant montrent également une transition structurale hélice α–pelote statistique [8, 14]. Ces résultats peuvent être interprétés avec les modèles de mécanique statistique proposés par Zimm et Bragg [15] et Lifson et Roig [16] et qui décrivent la formation et la propagation d'une hélice α.

Pour toutes les simulations effectuées sur des polyalanines, la forme en hélice α est la structure qui présente le minimum d'énergie potentielle. L'observation de structures en feuillet β à température intermédiaire [6] pourrait être expliquée par un effet entropique qui stabiliserait ces conformations. Pour tenter de vérifier cette hypothèse et comprendre les changements de conformation liés à la taille des polyalanines Ala_n et à la température, nous avons réalisé des simulations sur ces peptides. Kohtani et Jarrold ont montré que la plus petite hélice chargée observable expérimentalement en phase gazeuse est obtenue pour 8 alanines [17]. Nous avons donc utilisé cette taille Ala_8 comme point de départ de notre étude. Enfin, pour réduire au maximum le coût numérique des simulations, nous n'avons pas pris en compte le résidu tryptophane et nous avons remplacé les extrémités protégées CH_3CO- et NH_2- des peptides expérimentaux par des extrémités neutres simples H- et HO-, comme illustré sur la figure 4.4

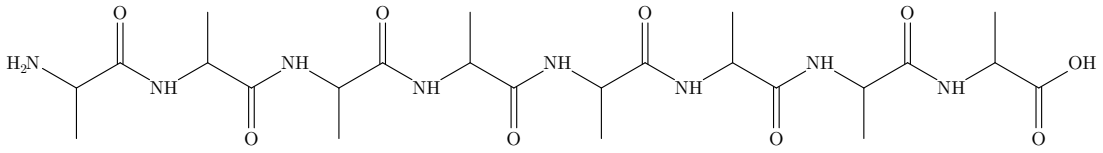


FIG. 4.4 – *Formule topologique du peptide Ala_8 étudié par simulation.*

Dans un premier temps, nous comparerons des simulations en Monte Carlo d'échange et Wang-Landau sur les peptides Ala_8 et Ala_{12} et nous verrons ensuite comment évolue le dipôle électrique moyen avec la température. Dans un deuxième temps, nous caractériserons les changements de structures observés dans le cas du peptide Ala_{12}. Nous étudierons

ensuite l'effet de la taille sur les transitions structurales de Ala_n avec $n = 8$, 12, 16 et 20. Enfin, nous tenterons de comprendre l'influence du champ électrique sur les conformations du peptide Ala_{12}.

4.3 Étude des peptides Ala_8 et Ala_{12} en Monte Carlo d'échange et Wang-Landau

Nous avons simulé les peptides Ala_8 et Ala_{12} par les méthodes Monte Carlo d'échange et Wang-Landau.

Les simulations Monte Carlo d'échange pour Ala_8 ont nécessité $7{,}41 \times 10^9$ pas Monte Carlo dont les $7{,}41 \times 10^8$ premiers pas sont utilisés pour la thermalisation et ne sont donc pas pris en compte dans les statistiques. Pour ces simulations, 23 angles de torsion sont utilisés comme coordonnées internes et 23 trajectoires entre 50 K et 1100 K explorent simultanément le paysage énergétique. Les histogrammes accumulés au cours des simulations sont retraités par la méthode des histogrammes multiples (WHAM). On calcule en particulier la capacité calorifique et la moyenne du dipôle électrique en fonction de la température.

Les simulations Wang-Landau s'effectuent à deux dimensions, énergie et dipôle électrique. Le facteur de modification varie de façon non monotone, comme dans la méthode Wang-Landau recuit détaillée dans le chapitre 2. Un total de 10^8 pas Monte Carlo est utilisé pour construire la densité d'états bidimensionnelle. Une repondération immédiate de cette dernière fournit toutes les moyennes thermiques (dont la capacité calorifique) ainsi que le dipôle électrique moyen.

Sauf indication contraire, toutes les simulations (en Monte Carlo d'échange comme avec la méthode Wang-Landau) ont été réalisées avec une géométrie de départ des polyalanines initialisée aléatoirement.

La figure 4.5 représente la variation de la capacité calorifique en fonction de la température pour les deux méthodes de simulation employées. On observe très clairement deux pics à environ 60 K et 232 K. Les deux courbes obtenues par les deux méthodes sont très proches l'une de l'autre, preuve que ces dernières ont convergé.

La figure 4.6 illustre les variations de la moyenne du dipôle électrique en fonction de la température pour les deux algorithmes de simulation. Ici encore, les deux courbes sont très proches. La moyenne du dipôle électrique décroît brutalement à 50 K, reste stable entre 100 K et 200 K puis diminue encore légèrement.

La comparaison des variations de la capacité calorifique et de la moyenne du dipôle électrique en fonction de la température montre que les deux méthodes de simulation produisent des résultats équivalents. La méthode Wang-Landau a cependant l'avantage de nécessiter un moindre effort numérique. Les courbes de capacité calorifique montrent très clairement deux transitions de phases qu'on peut corréler avec les variations de la

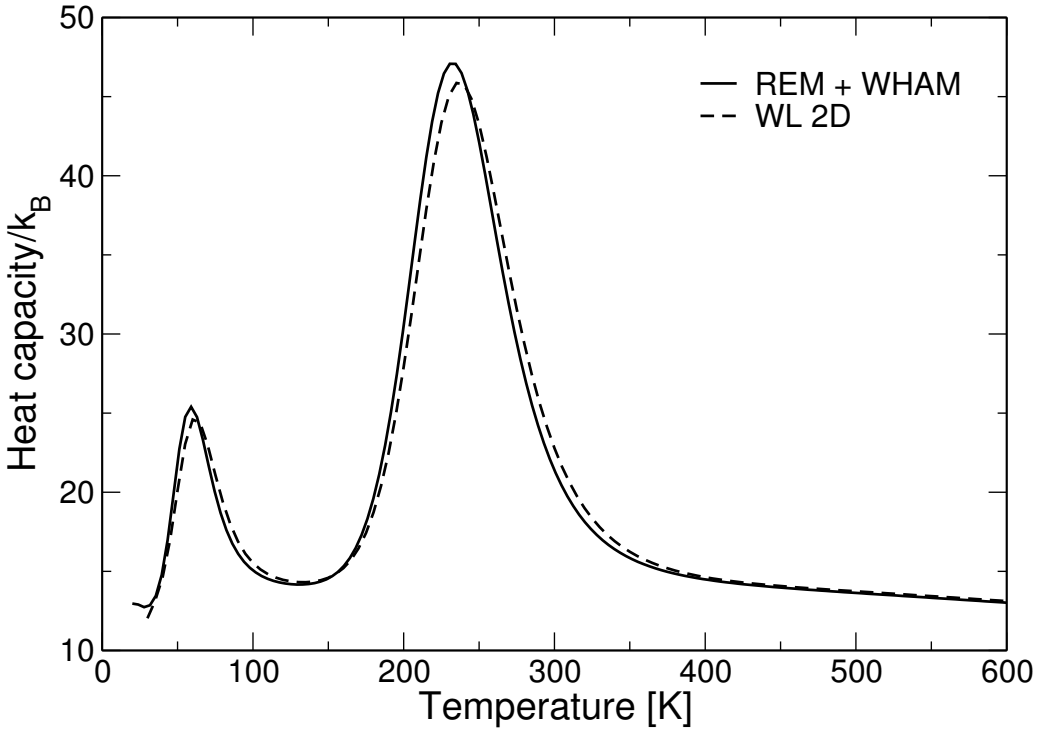


FIG. 4.5 – *Variations de la capacité calorifique de Ala_8 en fonction de la température. Résultats obtenus avec d'une part, la méthode Monte Carlo d'échange (REM) suivie d'une analyse par histogrammes multiples (WHAM) et d'autre part, avec l'algorithme Wang-Landau à deux dimensions (WL 2D).*

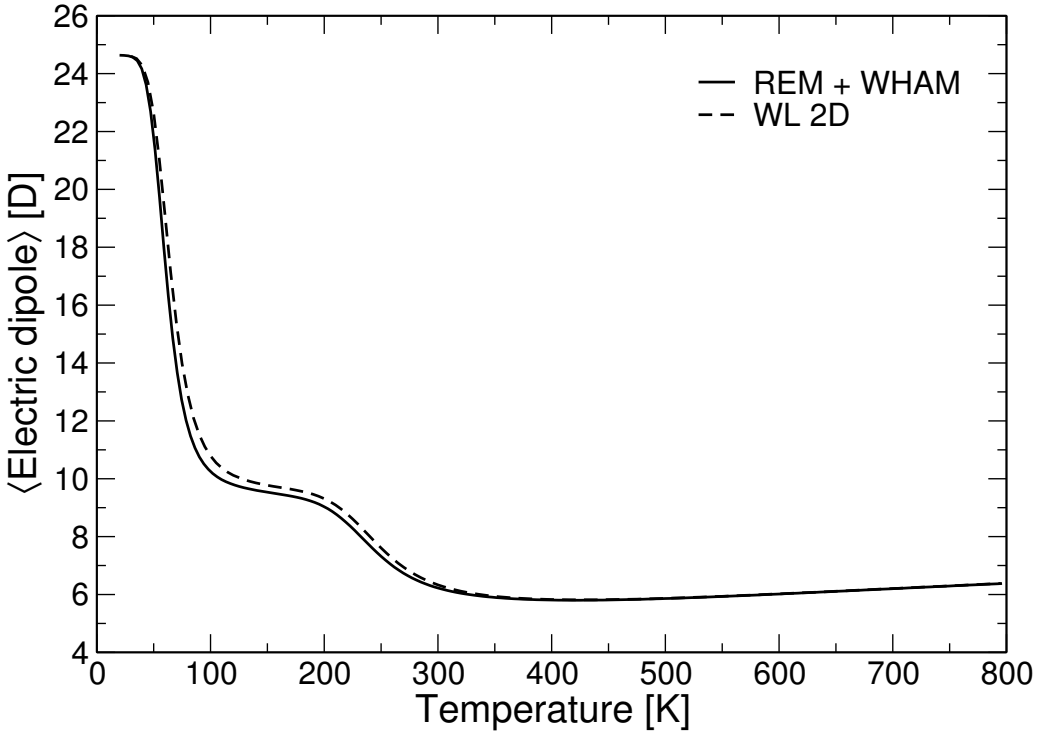


FIG. 4.6 – *Variations de la moyenne du dipôle électrique de Ala_8 en fonction de la température. Résultats obtenus avec d'une part, la méthode Monte Carlo d'échange (REM) suivie d'une analyse par histogrammes multiples (WHAM) et d'autre part, avec l'algorithme Wang-Landau à deux dimensions (WL 2D).*

moyenne du dipôle électrique. Avant la première transition, les peptides présentent un fort dipôle moyen qui correspond probablement à des structures en hélice α. Au-delà de 100 K, le dipôle moyen est très faible et les structures associées restent à caractériser.

4.3.1 Remarques sur la convergence des simulations

Les deux simulations Monte Carlo d'échange et Wang-Landau ont convergé pour le peptide Ala_8. Le coût numérique pour y parvenir est assez important et nous nous devons de discuter de la convergence des simulations qui pourra être problématique pour des systèmes plus gros.

L'effort numérique alloué pour les simulations Monte Carlo peut être particulièrement critique. Nous avons ainsi réalisé 4 simulations en Monte Carlo d'échange avec 23 températures comprises entre 50 K et 1100 K pour le peptide Ala_8. Le coût numérique est de 10^7 pas Monte Carlo pour la première simulation, 10^8 pas pour la deuxième, 10^9 pas pour la troisième et $7{,}4 \times 10^9$ pas pour la dernière. La géométrie de départ utilisée pour toutes ces simulations est une structure en hélice α. Cette initialisation permet à l'algorithme de rapidement trouver la structure la plus stable, sans pour autant biaiser les statistiques accumulées. Enfin la simulation la plus longue produit des résultats identiques, que la structure de départ soit en hélice α ou aléatoire. Pour les différents coûts numériques étudiés, plusieurs simulations indépendantes produisent des résultats équivalents et donnent l'*impression* d'avoir convergé. La figure 4.7 représente les variations de la capacité calorifique en fonction de la température pour ces 4 simulations. La courbe obtenue à partir de la simulation la plus longue présente les deux pics également observés avec l'algorithme Wang-Landau. La capacité calorifique tirée de la simulation la plus courte ne reproduit que le pic de haute température alors que celle de la simulation avec 10^8 pas Monte Carlo présente également un épaulement vers 150 K. Enfin pour la simulation à 10^9 pas Monte Carlo, on obtient un pic vers 100 K. La transition à basse température nécessite donc un effort numérique important pour être décelée. Ce comportement est en accord avec celui déjà observé pour l'agrégat Lennard-Jones LJ_{75} connu pour présenter deux entonnoirs en compétition [18]. Les simulations les plus courtes, bien que reproductibles, n'ont donc pas convergé. Cette conclusion est vérifiée par l'évolution temporelle des échanges qui montre que certaines répliques peuvent rester bloquées sur des trajectoires à basse température, d'autres répliques ne visitant jamais les trajectoires les plus froides. La figure 4.8 illustre ce phénomène pour la simulation avec 10^8 pas Monte Carlo.

Enfin, les études des polyalanines dans les références [7, 8, 9] ont été réalisées par des simulations multicanoniques. Dans le chapitre 3, nous avions comparé plusieurs variantes de la méthode Wang-Landau, en particulier la version à une dimension en énergie et à deux dimensions (énergie et paramètre d'ordre). Notre étude de la capacité calorifique du peptide Ala_8 sur la gamme de température 100 K–500 K ne présentait pas de différence significative avec la méthode Monte Carlo d'échange. Nous avons étendu cette

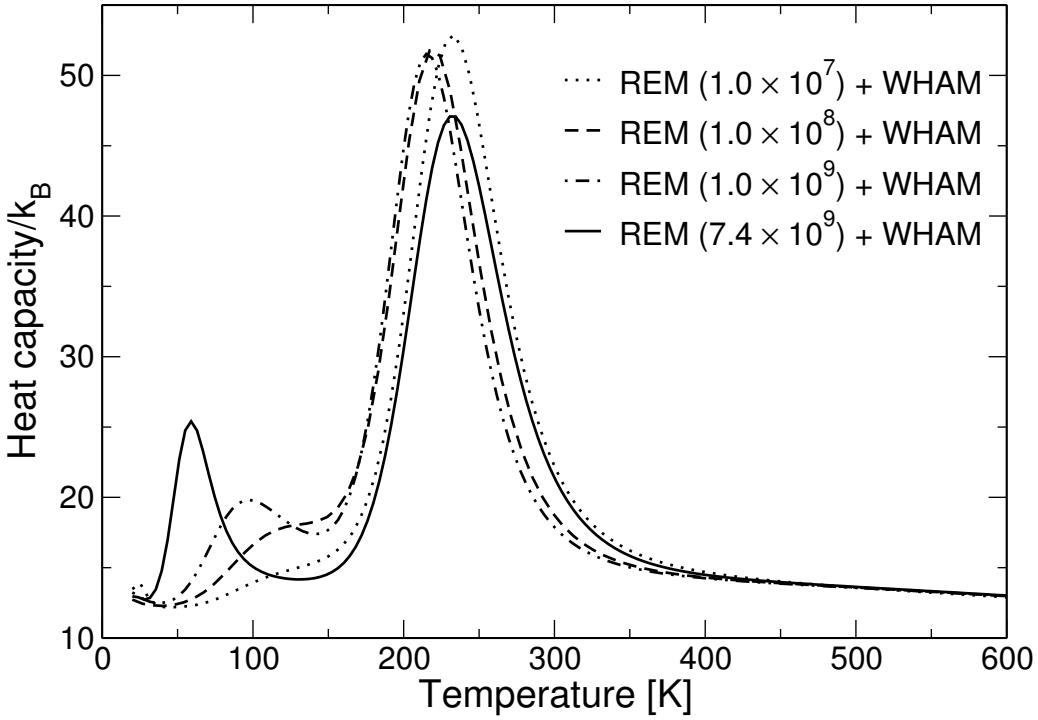


FIG. 4.7 – *Variations de la capacité calorifique du peptide Ala_8 en fonction de la température pour la méthode Monte Carlo d'échange (REM) avec 10^8 et $7{,}4\times10^9$ pas Monte Carlo. Les simulations sont suivies d'une analyse par les histogrammes multiples (WHAM).*

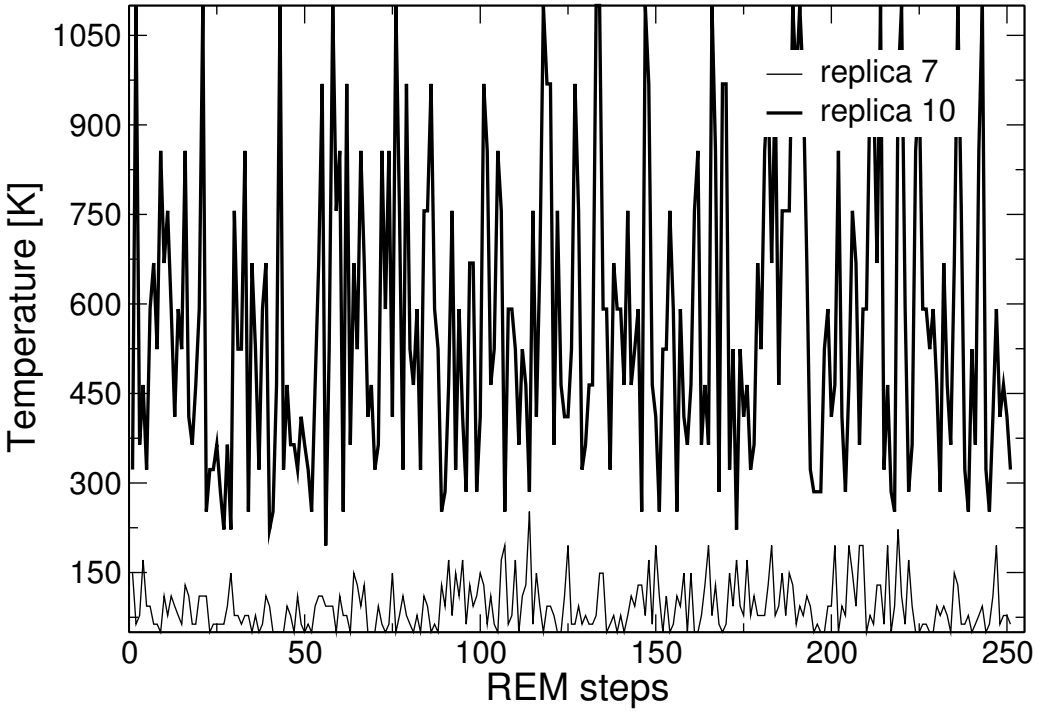


FIG. 4.8 – *Évolution des températures des trajectoires rencontrées pas les répliques 7 et 10 lors d'une simulation en Monte Carlo d'échange du peptide Ala_{12} avec 10^8 pas Monte Carlo.*

étude à un intervalle plus large (20 K–600 K) en effectuant une simulation Wang-Landau à une dimension avec $4{,}2 \times 10^8$ pas Monte Carlo et une simulation Wang-Landau à deux dimensions avec 10^8 pas Monte Carlo. On observe alors un contraste flagrant entre les résultats donnés par les deux méthodes (figure 4.9). La méthode WL 1D ne reproduit pas le premier pic de capacité calorifique, elle ne permet donc pas d'explorer complètement l'espace des configurations. On peut alors douter de l'efficacité de la méthode multicanonique à échantillonner efficacement la surface d'énergie potentielle de systèmes comme les polyalanines.

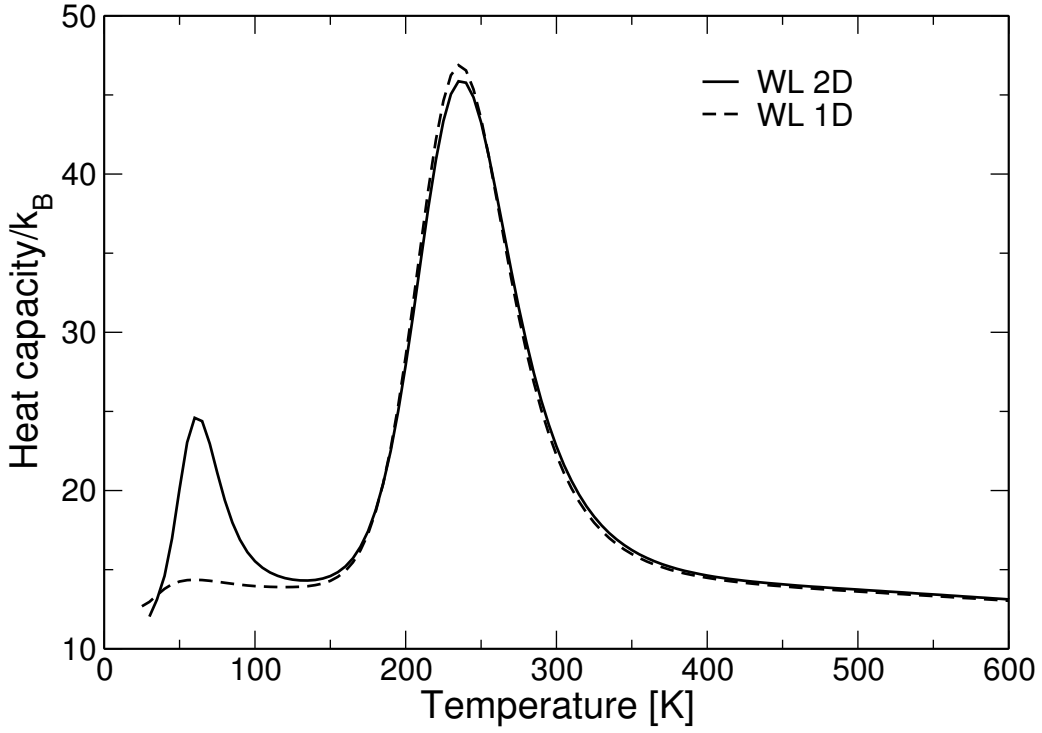


FIG. 4.9 – *Variations de la capacité calorifique du peptide Ala_8 en fonction de la température pour la méthode Wang-Landau à une (WL 1D) et deux (WL 2D) dimensions.*

4.3.2 Simulations du peptide Ala_{12}

La polyalanine Ala_{12} a également été étudiée par les deux méthodes de simulation. En Monte Carlo d'échange, les simulations ont été effectuées avec 31 températures comprises entre 80 K et 1100 K. Un total de $1{,}3 \times 10^9$ pas Monte Carlo a été nécessaire et les $1{,}3 \times 10^8$ premiers pas ont été utilisés pour la thermalisation et ne sont pas pris en compte dans les statistiques. Les simulations par la méthode Wang-Landau ont été effectuées à deux dimensions, énergie et dipôle électrique. Un total de $2{,}6 \times 10^8$ pas Monte Carlo a été utilisé pour construire la densité d'états.

La figure 4.10 présente les variations de la capacité calorifique du peptide Ala_{12} en fonction de la température pour les deux méthodes de simulation employées. On observe un désaccord entre les deux courbes à basse température. La simulation Wang-Landau

prédit deux transitions de phase. En Monte Carlo d'échange la deuxième transition est bien reproduite mais pas la première. Pour chaque méthode, plusieurs simulations indépendantes ont été réalisées et donnent des résultats sensiblement équivalents. La convergence des simulations en Monte Carlo d'échange pose problème d'autant plus que, comme pour les simulations courtes sur Ala_8, un certain nombre de répliques semblent rester bloquées dans les trajectoires de températures les plus faibles. Ainsi, et malgré un effort numérique cinq fois plus important, les simulations en Monte Carlo d'échange n'ont pas convergé. Par la suite, ceci limitera l'utilisation de l'algorithme Monte Carlo d'échange pour l'étude de plus gros peptides.

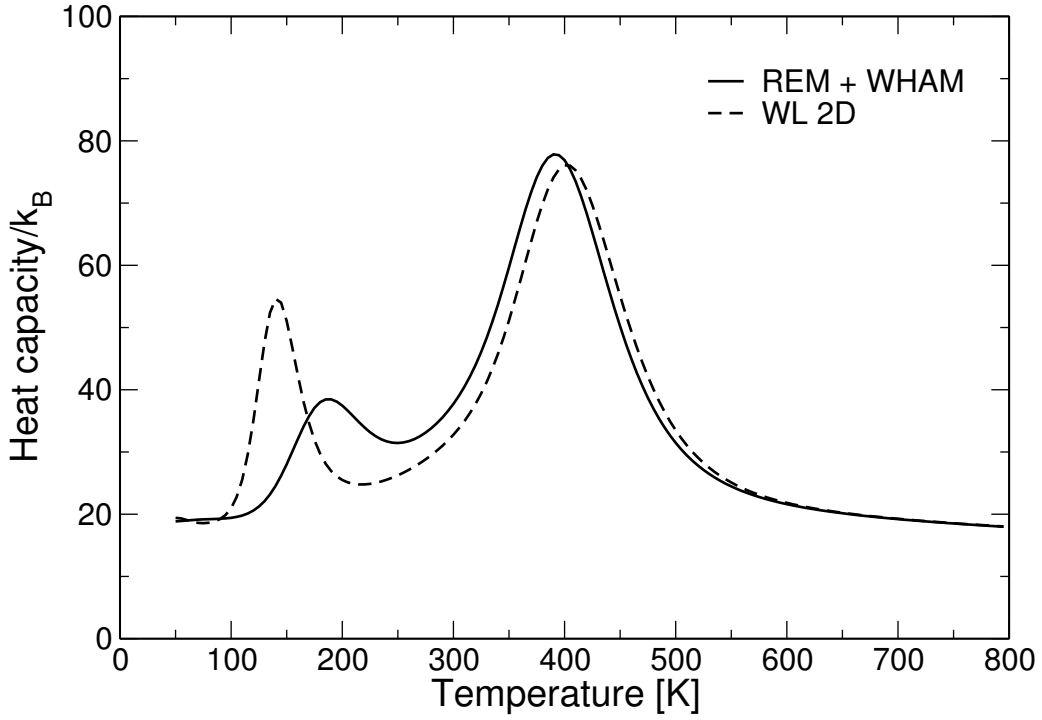


FIG. 4.10 – *Variations de la capacité calorifique de Ala_{12} en fonction de la température. Résultats obtenus avec d'une part, la méthode Monte Carlo d'échange (REM) suivie d'une analyse par histogrammes multiples (WHAM) et d'autre part, avec l'algorithme Wang-Landau à deux dimensions (WL 2D).*

Les variations de la moyenne du dipôle électrique en fonction de la température sont représentées sur la figure 4.11. Elles suivent celles de la capacité calorifique, l'écart de température entre les transitions données par les deux méthodes est reproduit.

4.4 Caractérisation des transitions structurales

Dans cette section, nous nous proposons de caractériser les transitions structurales associées aux deux pics de la capacité calorifique observée précédemment. Nous étudierons particulièrement le peptide Ala_{12} mais les résultats obtenus sont généralisables aux autres polyalanines étudiées. La simulation Wang-Landau à deux dimensions, en énergie et en

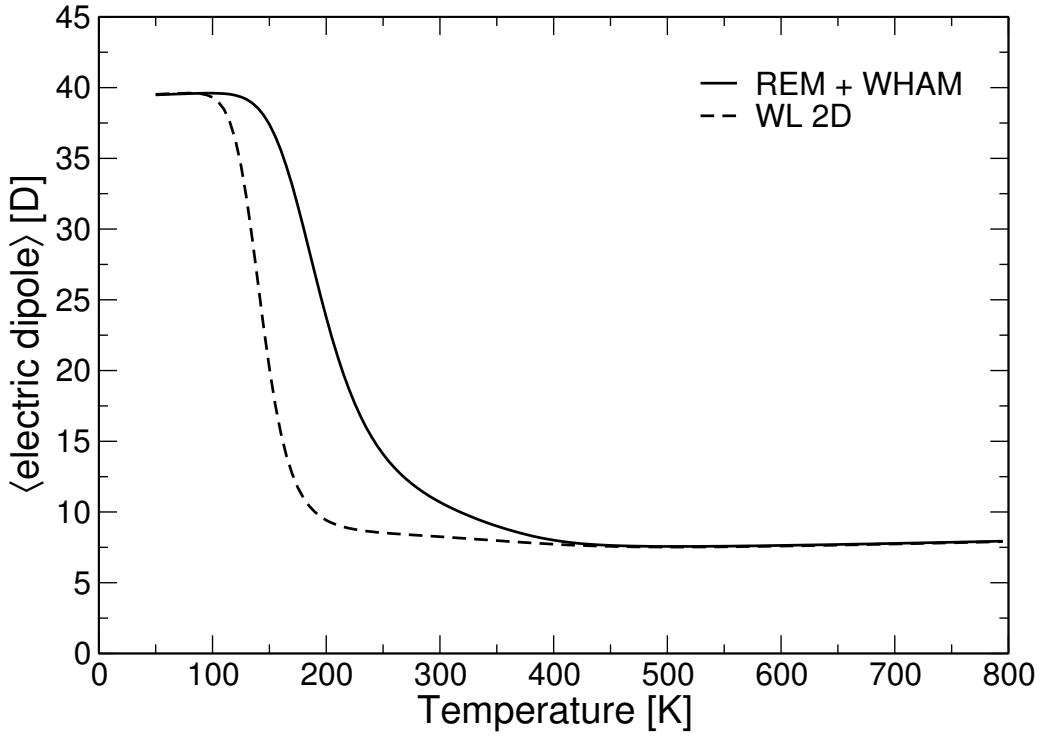


FIG. 4.11 – *Variations de la moyenne du dipôle électrique de Ala_{12} en fonction de la température. Résultats obtenus avec d'une part, la méthode Monte Carlo d'échange (REM) suivie d'une analyse par histogrammes multiples (WHAM) et d'autre part, avec l'algorithme Wang-Landau à deux dimensions (WL 2D).*

dipôle électrique, fournit immédiatement les moyennes thermiques dont la moyenne du dipôle. Pour obtenir les moyennes canoniques d'autres observables, il faut effectuer une simulation multicanonique supplémentaire. Nous avons ainsi échantillonné 4×10^7 configurations du peptide Ala_{12}, pour lesquelles nous avons enregistré le nombre de résidus en conformation hélice α, le nombre de résidus en conformation feuillet β et la distance bout-à-bout. La repondération décrite dans le chapitre 2 nous permet de calculer les moyennes canoniques de ces observables.

Les figures 4.12 et 4.13 représentent les variations canoniques du nombre moyen de résidus en hélice α, $\langle N_\alpha \rangle$, et en feuillet β, $\langle N_\beta \rangle$, ainsi que les variations de la distance bout-à-bout d_{bb} avec la température. Comme énoncé dans le chapitre 1, un résidu est en hélice α si les angles (Φ, Ψ) sont dans l'intervalle (-70° ± 30°, -37° ± 30°). De même, on définit un résidu en feuillet β si les angles (Φ, Ψ) du squelette peptidique sont dans l'intervalle (-140° ± 40°, 140° ± 40°) [19]. La distance bout-à-bout mesure l'étirement de la chaîne peptidique et elle est définie ici comme la distance entre l'azote du résidu N-ter et l'hydrogène du groupe carboxylique en C-ter.

À faible température, la moyenne du dipôle électrique est importante (figure 4.11). Le nombre moyen de résidus en α vaut 10, ce qui est le maximum pour Ala_{12} dans la mesure où les résidus aux extrémités N-ter et C-ter n'adoptent pas de structure secondaire particulière. Par ailleurs, $\langle N_\beta \rangle$ est minimal et d_{bb} adopte une valeur moyenne de l'ordre

de 17 Å. Les structures associées sont en hélice α, comme représenté sur la figure 4.14(a). À température intermédiaire, $\langle N_\beta \rangle$ est maximale avec une valeur de 8 résidus et $\langle N_\alpha \rangle$ est minimale, tout comme la moyenne du dipôle électrique et la distance bout-à-bout. Ceci caractérise une structure en feuillet β pour laquelle les extrémités des deux brins β sont proches [figure 4.14(b)]. Enfin à température élevée, $\langle N_\alpha \rangle$ est très faible et $\langle N_\beta \rangle$ adopte une valeur non négligeable car les structures désordonnées [figure 4.14(c)] sont structuralement proches des feuillets β. Le dipôle électrique est également petit comme attendu pour une structure aléatoire. La valeur maximale de d_{bb} traduit une structure désordonnée plutôt étirée.

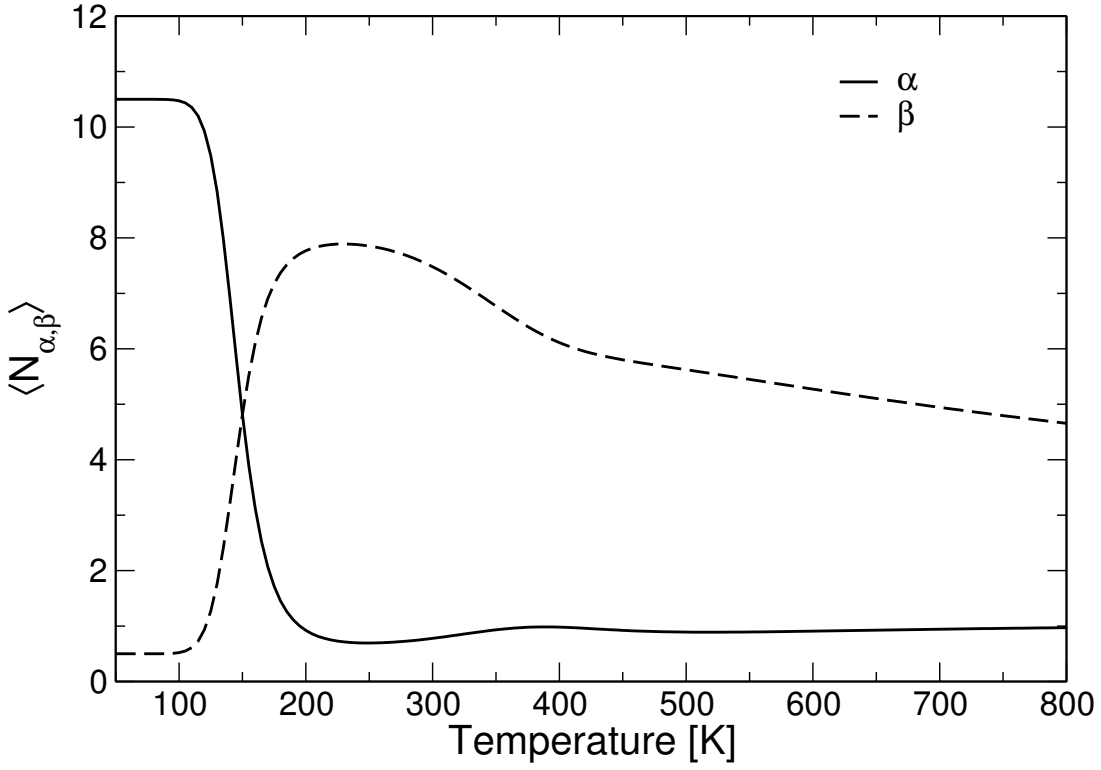


FIG. 4.12 – *Variations du nombre moyen de résidus en structure hélice α et en feuillet β pour Ala_{12} en fonction de la température.*

On retrouve également ces conclusions dans les cartes d'énergie libre à température fixée à deux dimensions, dipôle électrique et distance bout-à-bout, qui illustrent la corrélation entre ces deux observables. Ces cartes sont représentées figure 4.15 pour les températures de 50, 300 et 600 K. À faible température, les structures obtenues ont un dipôle élevé pour une valeur de d_{bb} intermédiaire, ce qui correspond effectivement à des structures en hélice α. À température ambiante, on obtient majoritairement des structures avec un faible dipôle et une courte distance bout-à-bout ce qui est révélateur d'une géométrie en feuillet β. Enfin, à température élevée, un dipôle faible ainsi qu'une d_{bb} importante traduisent des structures désordonnées étirées.

Les figures 4.15 (a, b et c) nous permettent donc de caractériser les deux transitions

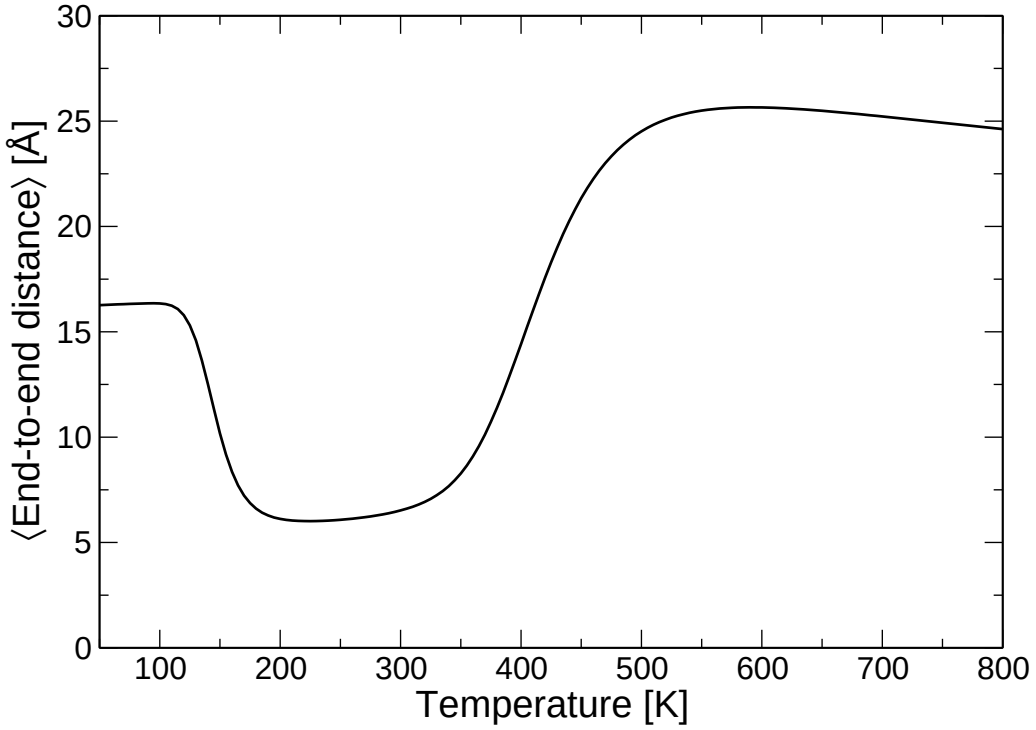


FIG. 4.13 – *Évolution de la distance bout-à-bout moyenne de Ala_{12} en fonction de la température.*

structurales observées dans la figure 4.10, faisant passer des structures en hélice α vers des structures en feuillet β, puis vers des structures étirées.

4.5 Effets de taille

Nous nous intéressons maintenant aux comportements des transitions identifiées précédemment pour des systèmes de taille plus importante. Pour cela, nous avons simulé le comportement des polyalanines Ala_8, Ala_{12}, Ala_{16} et Ala_{20} par la méthode Wang-Landau à deux dimensions. L'évolution de la capacité calorifique en fonction de la température est représentée figure 4.16 pour ces peptides. D'une manière générale, les deux pics pour chaque courbe de capacité calorifique se décalent vers les températures plus élevées à mesure que la taille du peptide augmente. Cette évolution correspond à une stabilisation des différentes structures secondaires avec la taille qui s'explique par une augmentation des liaisons hydrogènes intramoléculaires formées. Par ailleurs, le pic de la première transition devient de plus en plus fin et haut, ce qui traduit une transition du premier ordre arrondie par des effets de taille finie. Le pic de la seconde transition a tendance à s'élargir, il est plus difficile de conclure pour ce dernier cas.

Borrmann *et al.* ont proposé une méthode pour déterminer l'ordre d'une transition de phase pour des systèmes finis [20]. Cette méthode repose sur la classification de Lee et Yang [21, 22] reliant la distribution des zéros de la fonction de partition dans le plan complexe à l'ordre de la transition. Alves et Hansmann ont appliqué cette idée afin d'identifier

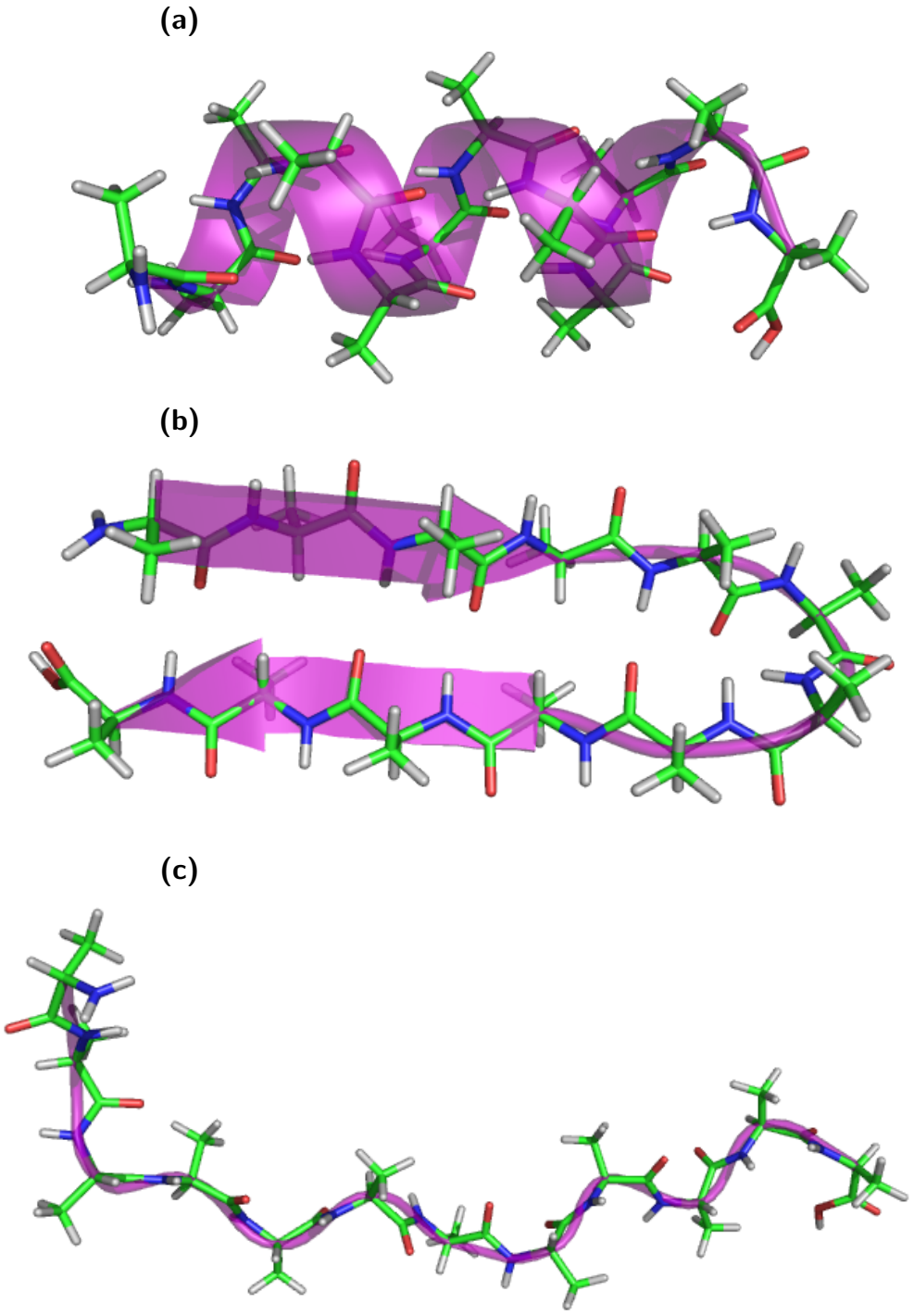


FIG. 4.14 – *Structures du peptide Ala_{12}. (a) Hélice α, E = 29,1 kcal/mol, dipôle = 39,2 D et distance d_{bb} = 16,4 Å. (b) Feuillet β, E = 34,4 kcal/mol, dipôle = 6,7 D et d_{bb} = 4,7 Å. (c) Structure désordonnée étirée, E = 70,7 kcal/mol, dipôle = 8,5 D et d_{bb} = 28,2 Å.*

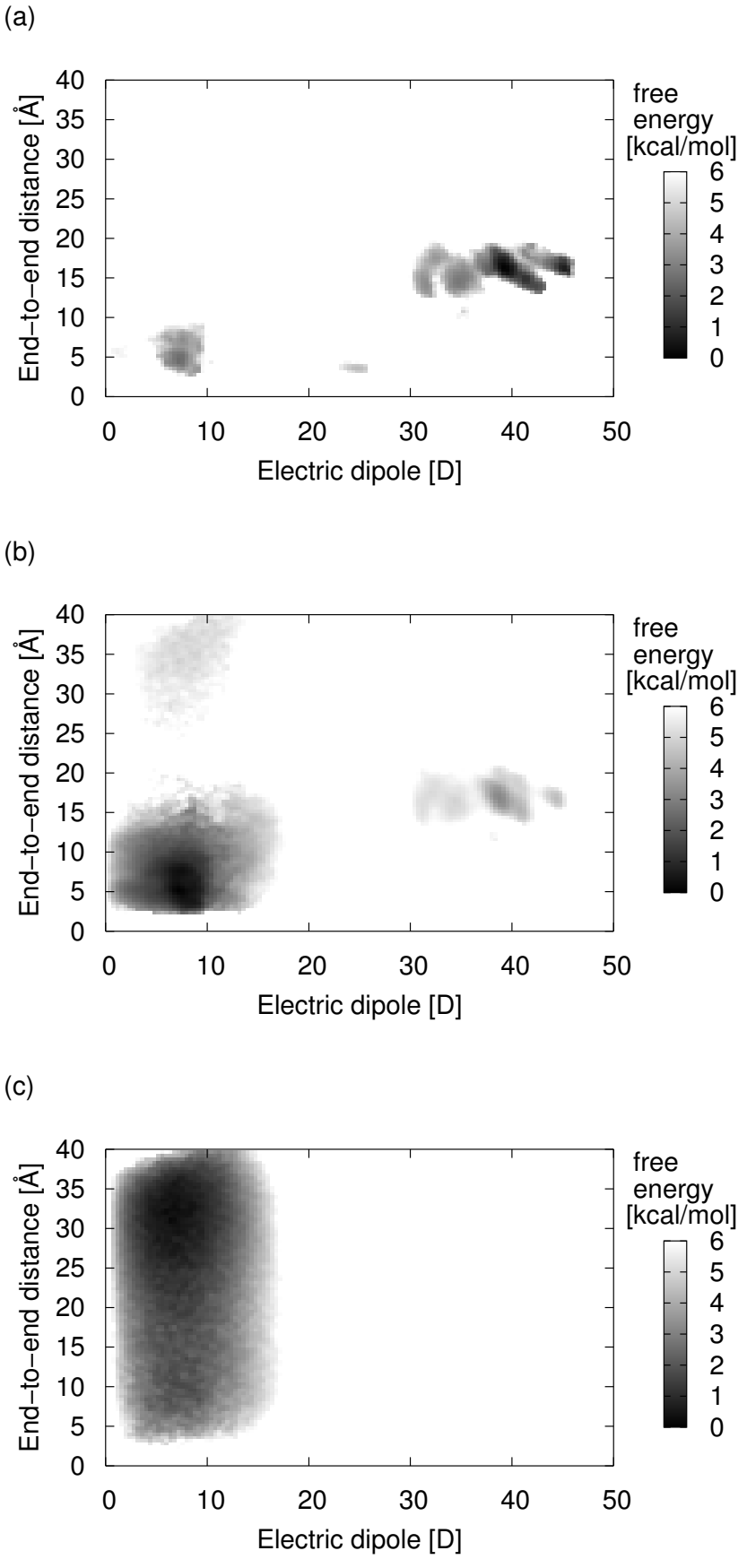


FIG. 4.15 – *Cartes d'énergie libre de Ala_{12} en fonction du dipôle électrique et de la distance bout-à-bout pour les températures de (a) 50 K, (b) 300 K et (c) 600 K.*

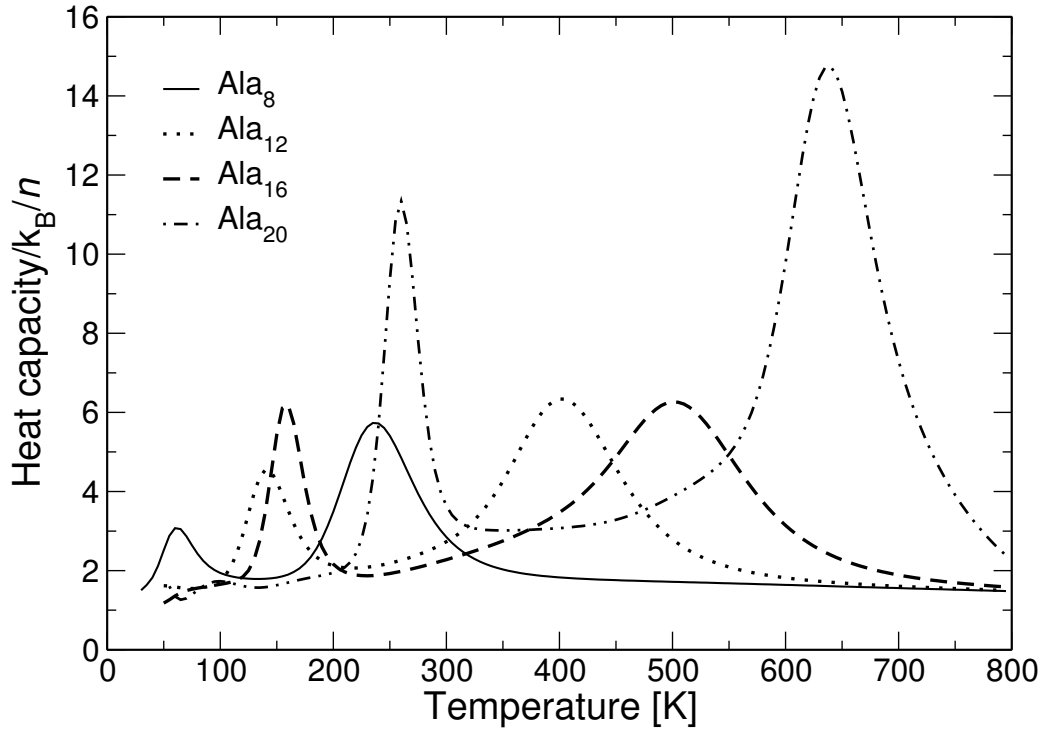


FIG. 4.16 – *Évolution de la capacité calorifique en fonction de la température pour les peptides Ala_8, Ala_{12}, Ala_{16} et Ala_{20}.*

l'unique transition de phase qu'ils observaient pour des polyalanines [23]. La complexité des systèmes étudiés n'a cependant pas permis de déterminer l'ordre de la transition de phase. Dans la mesure où nous observons non pas une mais deux transitions de phase qui pourraient s'influencer mutuellement, nous ne tenterons pas de caractériser leur ordre.

Il est par contre possible d'étudier les effets de taille finie sur les transitions calculées. Comme proposé dans la référence [24], on trace l'évolution de la température de transition (tableau 4.1) en fonction de l'inverse du nombre d'alanines (figure 4.17). Les premiers points obtenus pour Ala_{20} semblent aberrants car la température des deux transitions augmente rapidement par rapport à ce qui avait été obtenu pour des tailles plus petites. Par contre, rien ne laisse penser que les deux transitions puissent se rejoindre, autrement dit, les trois types de structures sont conservés.

TAB. 4.1 – *Températures respectives des deux transitions de phase pour chaque taille de peptide Ala_n.*

n	T_1 [K]	T_2 [K]
8	60	232
12	140	402
16	158	501
20	261	638

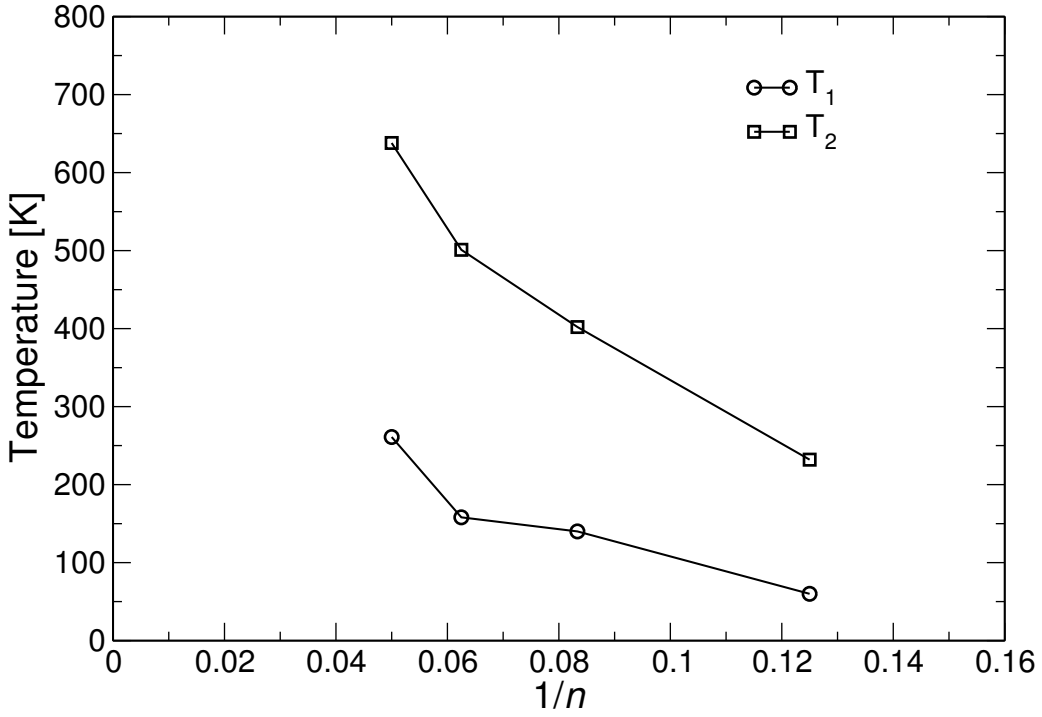


FIG. 4.17 – *Évolutions des températures de la 1re (T1) et de la 2nde (T2) transition de phase en fonction de l'inverse du nombre d'alanines.*

Finalement, les variations du dipôle électrique moyen en fonction de la température et du nombre d'alanines sont données sur la figure 4.18. À faible température, dans le domaine de stabilité des hélices α, on retrouve le dipôle moyen de chaque résidu, soit environ 3,5 D. Lors de la première transition, le dipôle chute brutalement vers une valeur très faible (entre 0,50 D et 0,75 D par résidu), ce qui correspond aux structures en feuillet β. Le dipôle remonte ensuite légèrement, particulièrement pour Ala_{16} et Ala_{20}. On se situe alors dans le domaine de prédominance des structures désordonnées étirées.

La figure 4.18 montre également que l'évolution du dipôle moyen par résidu pour Ala_8 ne suit pas exactement celle observée pour les peptides plus grands. Cette différence est due à la petite taille de l'octaalanine qui rend plus difficile la stabilisation des structures en hélice α et en feuillet β. Le premier et le dernier résidus ne participent généralement pas à une structure secondaire particulière, ce qui signifie que l'hélice α est ainsi réduite à un tour et demi (à raison de 3,6 résidus par tour) et que seulement 4 résidus stabilisent le feuillet β sachant que les 2 résidus médians sont impliqués dans le demi-tour.

4.6 Influence du champ électrique

Dans le chapitre 3, nous avions étudié l'influence d'un champ électrique statique et intense sur le dipeptide WG. Nous avions montré que de forts champs électriques favorisent les conformations de dipôle élevé. Nous poursuivons cette étude sur les polyalanines pour lesquelles on espère observer un effet important sur les structures en hélice α possédant

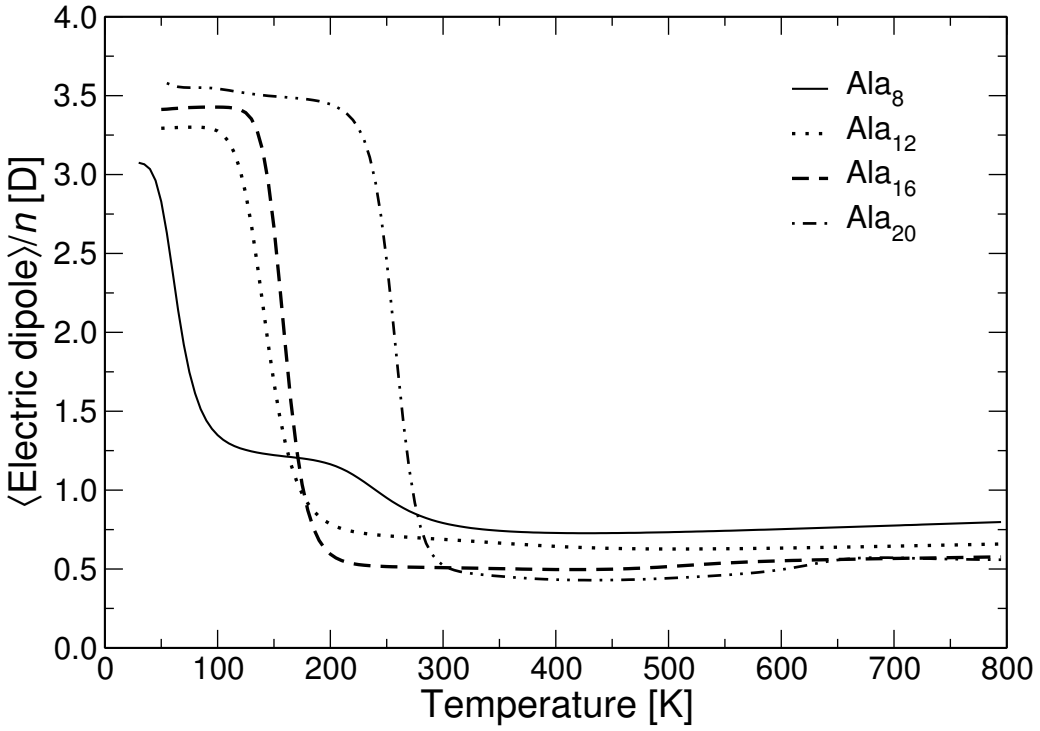


FIG. 4.18 – *Évolution du dipôle électrique moyen en fonction de la température pour les peptides Ala_8, Ala_{12}, Ala_{16} et Ala_{20}.*

un macrodipôle. Nous avons pour cela effectué des simulations Wang-Landau sur Ala_{12} pour trois valeurs du champ électrique : 10^8, 5×10^8 et 10^9 V/m. Toutes les simulations ont eu le même coût numérique, à savoir $2{,}6 \times 10^8$ pas Monte Carlo.

L'évolution de la capacité calorifique en fonction de la température est représentée figure 4.19 pour différentes valeurs du champ électrique. À 10^8 V/m, les deux pics de capacité calorifique sont conservés avec cependant un décalage du premier pic vers les températures plus élevées. Pour des champs de 5×10^8 V/m et 10^9 V/m, on n'obtient qu'un seul pic de capacité calorifique. Il semblerait que le premier pic observé à champ faible soit suffisamment décalé pour recouvrir le second à 5×10^8 V/m. Pour un champ plus important, un pic unique se déplace vers les températures croissantes.

Pour tenter d'identifier les modifications structurales opérées sous l'influence du champ électrique, nous nous sommes intéressés aux variations du dipôle électrique moyen en fonction de la température (figure 4.20). La chute du dipôle moyen observée à 140 K pour un champ électrique nul se décale vers des températures plus élevées à mesure que le champ électrique s'intensifie. On étend ainsi le domaine de stabilité de l'hélice α et pour un champ de 5×10^8 V/m, il est possible d'observer le peptide en hélice α à température ambiante. Inversement, le domaine de stabilité des structures en feuillet β est d'autant plus réduit que le champ électrique croît. Pour une valeur de 5×10^8 V/m, les deux pics de capacité calorifique sont superposés. Les structures en feuillet β n'apparaissent plus au profit d'une unique transition hélice α–structures étirées.

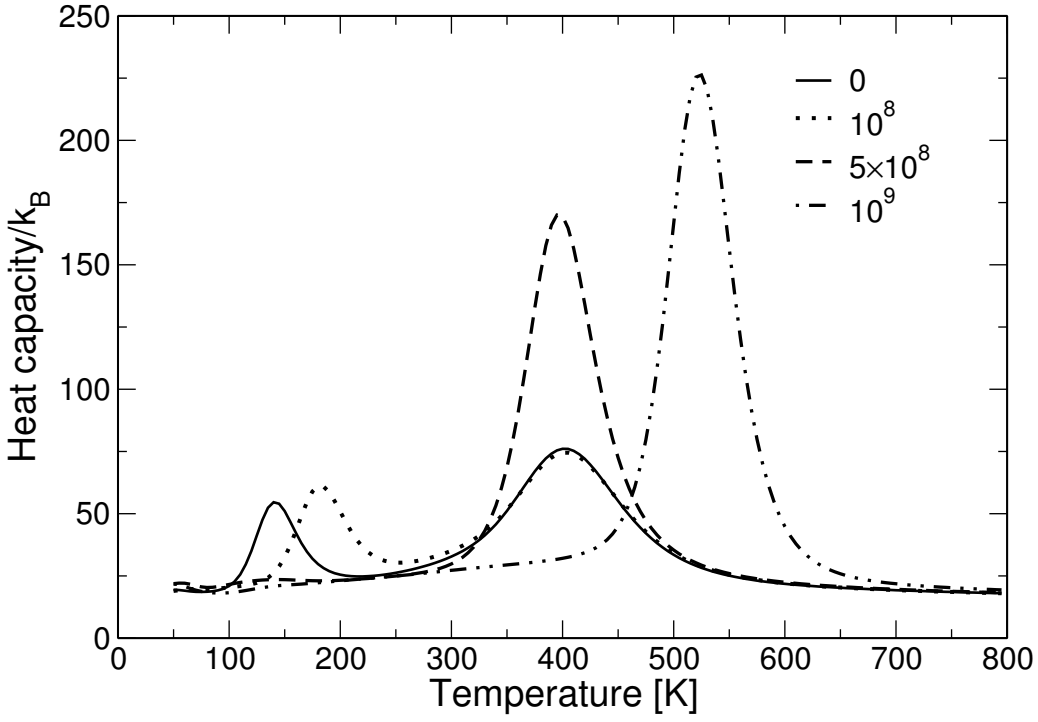


FIG. 4.19 – *Variations de la capacité calorifique du peptide Ala_{12} en fonction de la température pour des champs électriques de* 0, 10^8, 5×10^8 *et* 10^9 *V/m.*

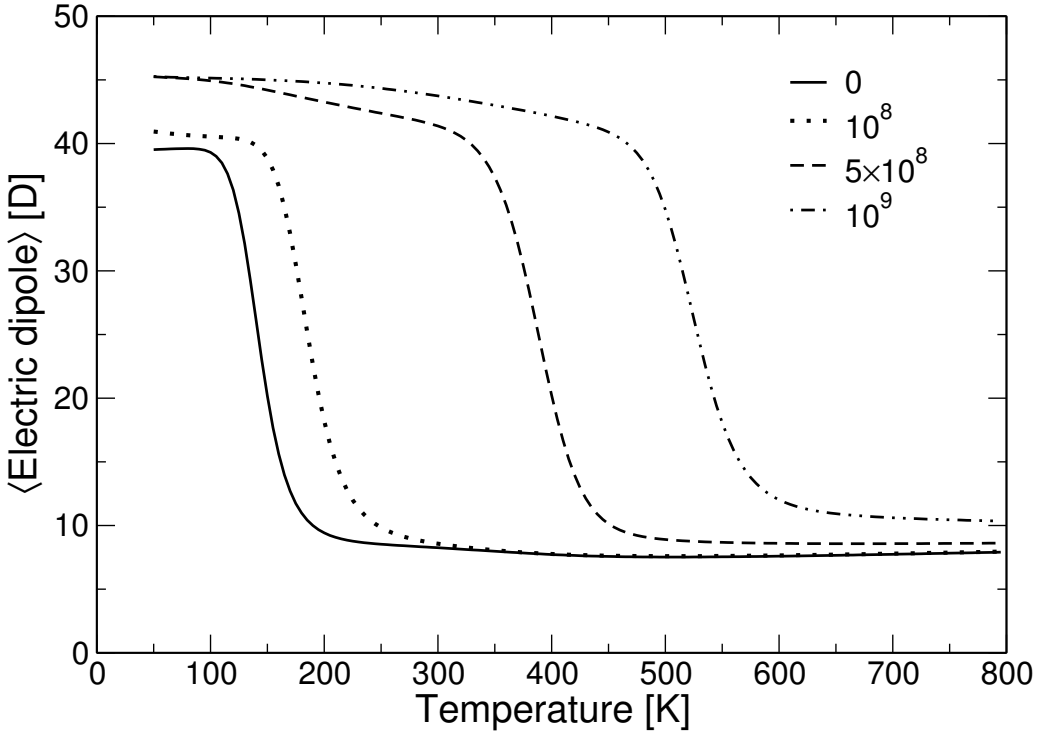


FIG. 4.20 – *Variations de la moyenne du dipôle électrique du peptide Ala_{12} en fonction de la température pour des champs électriques de* 0, 10^8, 5×10^8 *et* 10^9 *V/m.*

Enfin, il est intéressant de remarquer que le dipôle moyen à 100 K augmente avec la valeur du champ. Or, Ala_{12} compte 11 liaisons peptidiques et devrait donc présenter au maximum un dipôle de $11 \times 3{,}5 = 38{,}5$ D. Des valeurs supérieures proches de 45 D obtenues, par exemple, pour un champ de 10^9 V/m s'expliquent par l'alignement supplémentaire du dipôle du groupe carboxylique avec le macrodipôle peptidique. Cet alignement optimal est illustré par la structure de la figure 4.21.

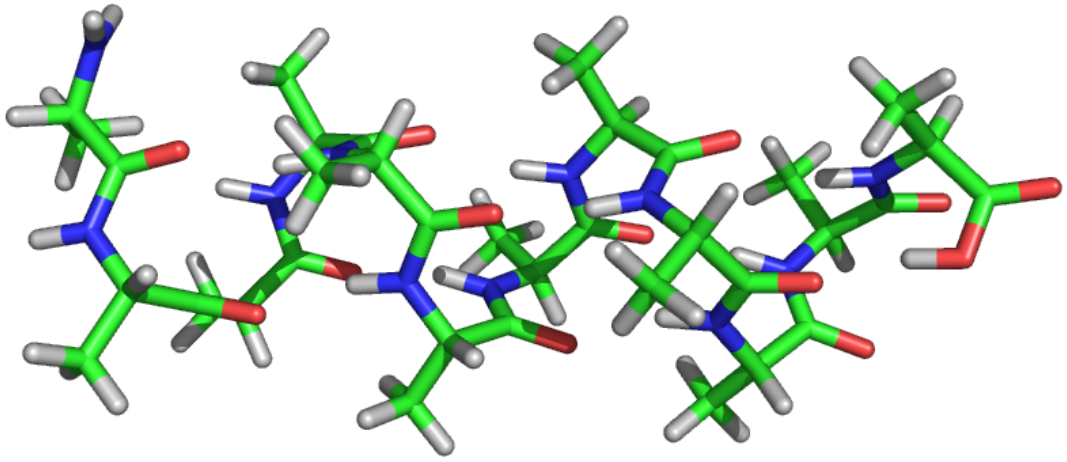


FIG. 4.21 – *Structure du peptide Ala_{12} en hélice α présentant un dipôle maximal de 45 D.*

4.7 Discussion et conclusion

L'étude des polyalanines par des simulations Wang-Landau à deux dimensions a permis de mettre en évidence deux pics de capacité calorifique qui correspondent à une double transition structurale. À basse température, les structures en hélice α sont stabilisées car elles présentent le minimum d'énergie potentielle. À température intermédiaire, les géométries de type feuillet β sont favorisées entropiquement. Enfin, à température élevée, des structures désordonnées, plutôt étirées, sont majoritairement observées.

Ces résultats sont en accord avec les mesures expérimentales de dipôle électrique effectuées par Dugourd *et al.* [6]. D'un point de vue théorique, la stabilité des feuillets β a été également étudié par Nguyen *et al.* [25], Ding *et al.* [26] et Levy *et al.* [27]. Ces derniers ont en particulier montré que la surface d'énergie potentielle du peptide Ala_{12} dans le vide présente un bassin profond et étroit associé aux hélices α et un bassin moins profond mais plus large attribué aux feuillets β. Cette topologie traduit la plus grande stabilité des hélices α par rapport aux feuillets β due au nombre plus important de liaisons hydrogènes. Par contre, une structure en feuillet est plus flexible qu'une structure hélicoïdale et l'entropie

associée sera donc plus importante. Par des calculs de dynamique moléculaire en solvant implicite, Ding *et al.* ont confirmé la stabilité des feuillets β dans une phase intermédiaire entre les hélices α et les structures désordonnées [26]. Enfin, dans une étude très complète, Nguyen *et al.* ont caractérisé l'influence des interactions hydrophobes, autrement dit du solvant, sur les conformations du peptide $AcKA_{14}KNH_2$. Lorsque les interactions hydrophobes sont faibles (comme pour une molécule isolée), le peptide présente une transition à deux états hélice α–pelote statistique. Pour des interactions modérées, l'apparition de l'état correspondant aux feuillets β se traduit par une stabilisation des hélices α à basse température, des feuillets β à température intermédiaire et des structures désordonnées à température élevée. Pour des interactions hydrophobes importantes, le peptide peut occuper deux états, en feuillet à basse température et une structure en pelote statistique à température plus élevée. Ces résultats pourraient qualitativement expliquer les nôtres puisqu'une constante diélectrique de 2 augmente les interactions hydrophobes. Par contre, ils n'expliquent pas l'écart avec les résultats théoriques obtenus par d'autres groupes [7, 8, 9] qui utilisent le même constante diélectrique. Pour essayer de comprendre ces différences, trois points peuvent être évoqués.

1. Le champ de force ECEPP/2, relativement ancien (1984), pourrait favoriser les hélices α au détriment des feuillets β.
2. La définition des degrés de liberté mobiles et surtout des modes gelés pourrait induire des différences significatives. En particulier, nous avons montré qu'une simulation partant d'une structure optimisée en hélice α et pour laquelle on ne fait bouger que les angles de torsion va produire une unique transition hélice α–pelote statistique. Réciproquement, une simulation avec une structure initiale optimisée en feuillet β va plutôt donner une seule transition feuillet β–pelote statistique.
3. Enfin, dans une dernière hypothèse, nous ne pouvons évidemment pas conclure sans évoquer un problème de convergence dû à un temps de simulation insuffisant, comme discuté précédemment.

Bibliographie

[1] J. W. Kelly. The alternative conformations of amyloidogenic proteins and their multi-steps assembly pathways. *Current Opinion in Structural Biology*, 8:101–106, 1998.

[2] D. J. Selkoe. Amyloid beta-protein and the genetics of Alzheimer's desease. *Journal of Biological Chemistry*, 271:18295–18298, 1996.

[3] S. B. Prusiner. Prion diseases and the BSE crisis. *Science*, 278:245–251, 1997.

[4] Y. Takahashi, A. Ueno, and H. Mihara. Design of a peptide undergoing α–β structural transition and amyloid fibrillogenesis by the introduction of a hydrophobic defect. *Chemistry - A European Journal*, 4:2475–2484, 1998.

[5] R. Cerpa, F. E. Cohen, and I. D. Kuntz. Conformational switching in designed peptides: the helix/sheet transition. *Folding and Design*, 1:91–101, 1996.

[6] P. Dugourd, R. Antoine, G. Breaux, M. Broyer, and M. F. Jarrold. Entropic Stabilization of Isolated β-Sheets. *Journal of American Chemical Society*, 127:4675–4679, 2005.

[7] U. H. E. Hansmann and Y. Okamoto. Finite-size scaling of helix-coil transitions in polyalanine studied by multicanonical simulations. *Journal of Chemical Physics*, 110:1267, 1999.

[8] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. *Biopolymers*, 60:96–123, 2001.

[9] N. Rathore, T. A. Knotts IV, and J. J. de Pablo. Density of states simulations of proteins. *Journal of Chemical Physics*, 118:4285–4290, 2003.

[10] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy Parameters in Polypeptides VII. Geometric Parameters, Partial Charges, Non-bonded Interactions, Hydrogen Bond Interactions and Intrensic Torsional Potentials for Naturally Occurring Amino Acids. *Journal of Physical Chemistry*, 79:2361–2381, 1975.

[11] G. Némethy, M. S. Pottle, and H. A. Scheraga. Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *Journal of Physical Chemistry*, 87:1883–1887, 1983.

[12] M. J. Sippl, G. Némethy, and H. A. Scheraga. Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H...O=C hydrogen bonds from packing configurations. *Journal of Physical Chemistry*, 88:6231–6233, 1984.

[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[14] Y. Peng and U. H. E. Hansmann. Solvation Model Dependency of Helix-Coil Transition in Polyalanine. *Biophysical Journal*, 82:3269–3276, 2002.

[15] H. B. Zimm and J. K. Bragg. Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *Journal of Chemical Physics*, 31:526–535, 1959.

[16] S. Lifson and A. Roig. On the Theory of Helix–Coil Transition in Polypeptides. *Journal of Chemical Physics*, 34:1963–1974, 1961.

[17] M. Kohtani and M. F. Jarrold. Water Molecule Adsorption and Short Alanine Peptides: How Short Is the Shortest Gas-Phase Alanine-Based Helix? *Journal of American Chemical Society*, 126:8454–8458, 2004.

[18] V. A. Mandelshtam, P. A. Frantsuzov, and F. Calvo. Structural Transitions and Melting in LJ_{74-78} Lennard-Jones Clusters from Adaptive Exchange Monte Carlo Simulations. *Journal of Physical Chemistry A*, 110:5326–5332, 2006.

[19] Y. Peng and U. H. E. Hansmann. Helix versus sheet formation in a small peptide. *Physical Review E*, 68:041911, 2003.

[20] P. Borrmann, O. Mülken, and J. Harting. Classification of Phase Transitions in Small Systems. *Physical Review Letters*, 84:3511–3514, 2000.

[21] C. N. Yang and T. D. Lee. Statistical Theory of Equations of State and Phase Transitions. I. Theory of Condensation. *Physical Review*, 87:404–409, 1952.

[22] T. D. Lee and C. N. Yang. Statistical Theory of Equations of State and Phase Transitions. II. Lattice Gas and Ising Model. *Physical Review*, 87:410–419, 1952.

[23] N. A. Alves and U. H. E. Hansmann. Partition Function Zeros and Finite Size Scaling of Helix-Coil Transitions in a Polypeptide. *Physical Review Letters*, 84:1836–1839, 2000.

[24] F. Calvo and J. P. K. Doye. Entropic tempering: A method for overcoming quasiergodicity in simulation. *Physical Review E*, 63:010902, 2000.

[25] H. D. Nguyen, A. J. Marchut, and C. K. Hall. Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Science*, 13:2909–2924, 2004.

[26] F. Ding, J. M. Borreguero, S. V. Buldyrey, H. E. Stanley, and N. V. Dokholyan. Mechanism for the α-Helix to β-Hairpin Transition. *Proteins: Structure, Function, and Genetics*, 53:220–228, 2003.

[27] Y. Levy, J. Jortner, and O. M. Becker. Solvent effects on the energy landscapes and folding kinetics of polyalanine. *Proceedings of the National Academy of Sciences of the United States of America*, 98:2188–2193, 2001.

Chapitre 5

Structure et fragmentation de polypeptides

Sommaire

L'étude expérimentale de la structure de biomolécules est très difficile comme l'atteste le différentiel entre le nombre de séquences et de structures connues. Les techniques RMN et de diffraction de rayons X sont cependant devenues routinières pour déterminer la structure d'une biomolécule en phase condensée. Par contre, l'étude de protéine isolée reste un défi. La spectroscopie infrarouge permet de mesurer les fréquences de vibration d'une molécule qu'on espère relier à sa structure [1, 2, 3], mais d'autres sondes conformationnelles doivent être développées. La mesure du dipôle électrique permet déjà de discriminer les structures secondaires de certains peptides. Une méthode alternative, développée au LASIM, repose sur des expériences de photofragmentation. Celles-ci permettent de mesurer les spectres de photofragmentation de molécules en fonction de la longueur d'onde d'excitation. La comparaison des spectres obtenus avec les spectres calculés peut permettre d'identifier la conformation de la molécule étudiée [4]. Il s'agit de la stratégie utilisée habituellement en spectroscopie pour déterminer la géométrie d'une

molécule mais appliquée ici à des systèmes relativement complexes. La distribution des fragments obtenus à partir de ces expériences de photofragmentation pourrait également être utilisée pour déterminer des structures de biomolécules. Nous espérons pour cela comprendre comment la fragmentation peut produire une information utilisable pour l'analyse de la molécule observée et réciproquement, de savoir s'il est possible de prédire cette fragmentation. Ces objectifs sont bien sûr ambitieux.

Dans ce chapitre, nous allons commencer à répondre à ces questions en étudiant la fragmentation par plusieurs méthodes d'excitation sur des systèmes modèles dont les structures sont connues. Nous avons ainsi comparé les dissociations induites par collision, par laser et par capture d'électron pour le pentapeptide AlaGlyTrpLeuLys ainsi que pour la famille de peptides déclinée à partir de TrpValValValVal. Les fragments observés par spectrométrie de masse dépendent du mécanisme d'excitation et de l'état de charge du peptide. Sans nous focaliser sur la dynamique explicite de l'excitation, qui met en jeu de multiples états électroniques, nous avons comparé les résultats expérimentaux avec les structures obtenues par des simulations en Monte Carlo d'échange. Cette comparaison suggère que le mécanisme de désexcitation, qui suppose une dynamique rapide, consécutif à une excitation par laser ou par capture d'électron est relié à la géométrie initiale de la molécule. Ce premier travail a été réalisé en collaboration avec J. Chamot-Rooke et G. van der Rest du laboratoire des mécanismes réactionnels de Palaiseau, avec C. Dedonder et C. Jouvet du laboratoire de photophysique moléculaire d'Orsay, avec C. Desfrançois et G. Grégoire du laboratoire de physique des lasers de Villetaneuse et enfin avec D. Onidas du laboratoire des collisions atomiques et moléculaires d'Orsay.

Une deuxième étude, plus qualitative mais effectuée sur un grand nombre de peptides, concerne la fragmentation de peptides extraits de la digestion enzymatique de protéines. Ce projet est l'objet du stage de master de L. Joly, en collaboration avec J. Lemoine de l'unité des sciences analytiques de Lyon.

5.1 Fragmentation de biomolécules

L'étude de la fragmentation des polypeptides pourrait permettre de remonter à leur structure. Elle est également d'un intérêt fondamental pour mieux comprendre les mécanismes fondamentaux de redistribution d'énergie dans un système à grand nombre de degrés de liberté excités électroniquement et initialement à l'état fondamental. Elle peut aussi déboucher sur d'importantes applications en sciences analytiques. Un large spectre d'expériences dans le domaine de la protéomique est, en effet, basé sur l'identification d'un peptide par l'observation de son schéma de fragmentation. Une meilleure compréhension des différents mécanismes impliqués entre l'excitation initiale et la fragmentation observée dans une expérience de spectrométrie de masse MS/MS pourrait permettre de prédire et

d'agir sur les fragments observés mais aussi de développer de nouvelles stratégies pour l'analyse peptidique.

La fragmentation d'une biomolécule peut se produire au niveau du squelette peptidique ou au niveau des chaînes latérales. La fragmentation du squelette peut engendrer deux types d'ions [5, 6, 7] détaillés dans la figure 5.1 :

- les ions de type a, b et c pour lesquels la charge positive est portée par la partie N-terminale ;
- les ions de type x, y et z pour lesquels la charge positive est portée par la partie C-terminale.

À basse énergie, les spectres de fragmentation présentent principalement des ions de type b et y qui correspondent à la rupture de la liaison peptidique.

La fragmentation des chaînes latérales produit préférentiellement des pertes de molécules neutres comme l'ammoniac NH_3 ou l'eau.

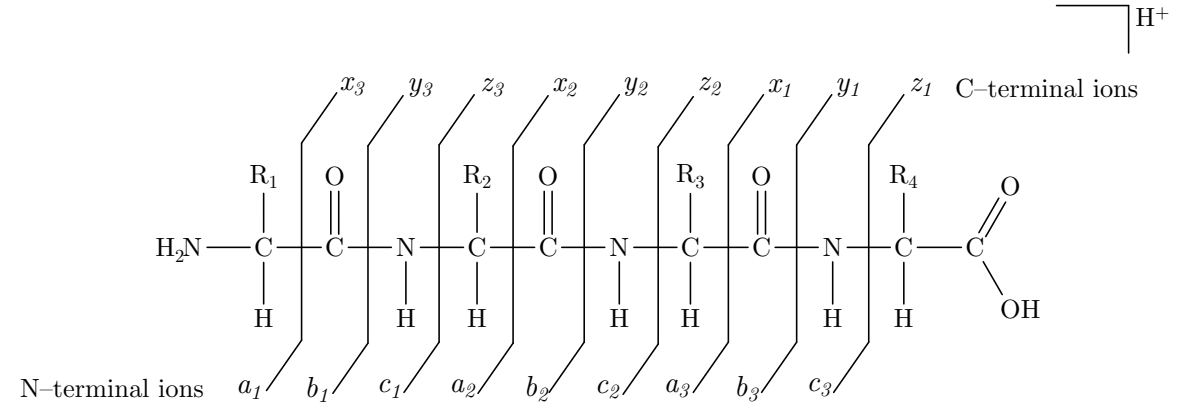


FIG. 5.1 – *Nomenclature des ions issus de la fragmentation du squelette peptidique [5, 6, 7].*

5.1.1 Différentes méthodes expérimentales

Différentes méthodes existent pour exciter et fragmenter des biomolécules. Dans la dissociation à basse énergie induite par collision (*collision-induced dissociation* ou CID) et dans la dissociation infra-rouge multi-photonique (*infrared multiphoton dissociation* ou IRMPD), le peptide est chauffé par un processus multi-étapes. Dans le cas précis de la CID, le peptide entre en collision avec des atomes d'hélium qui cèdent une partie de leur énergie cinétique au peptide, comme représenté sur la figure 5.2. L'excitation résultante est similaire à un chauffage du peptide pour lequel la redistribution de l'énergie vibrationnelle (*intramolecular vibrational-energy relaxation* ou IVR) est importante. Les réarrangements structuraux sont en compétition avec les chemins de fragmentation qui reposent sur le modèle du proton mobile [8]. Dans cette dernière hypothèse, la molécule a suffisamment d'énergie pour que le proton localisé sur le site le plus basique « saute » sur la chaîne peptidique. Le transfert d'un proton affaiblit alors localement la liaison peptidique, ce qui

conduit principalement à des fragments de type a, b et y. Ce type d'excitation est *a priori* peu sensible à la géométrie initiale du peptide car elle s'opère globalement sur toute la structure.

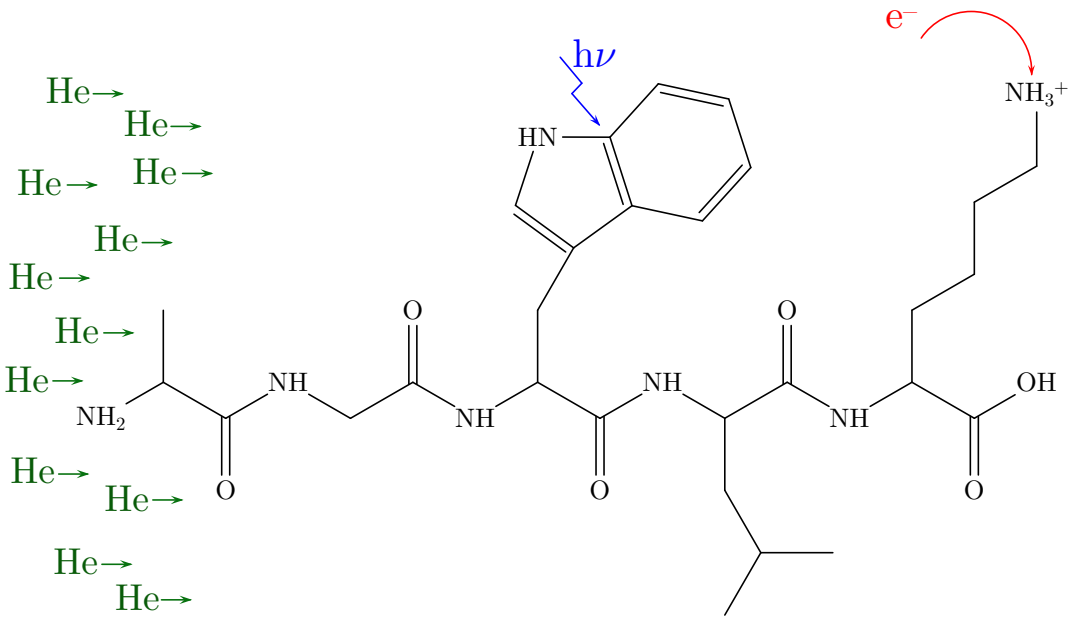


FIG. 5.2 – *Excitation par collision, par laser et par capture d'électron du peptide AGWLK.*

À l'opposé, on pense que la dissociation par capture d'électron (*electron-capture dissociation* ou ECD) et la dissociation induite par laser UV (*laser-induced dissociation* ou LID) engendrent une fragmentation non-ergodique [9, 10, 11], autrement dit, la rupture de liaison est plus rapide que la redistribution de l'énergie sur tous les degrés de liberté. En ECD, un électron thermique neutralise un proton (figure 5.2). Cette excitation en une seule étape conduit à la formation d'une structure hypervalente avec un excès d'énergie d'à peu près 5 eV dans le peptide [12]. Un atome d'hydrogène est ainsi libéré et transféré au plus proche groupe carbonyle du squelette peptidique et engendre une dissociation avec des fragments de type c / z [13, 14], comme illustré sur la figure 5.3.

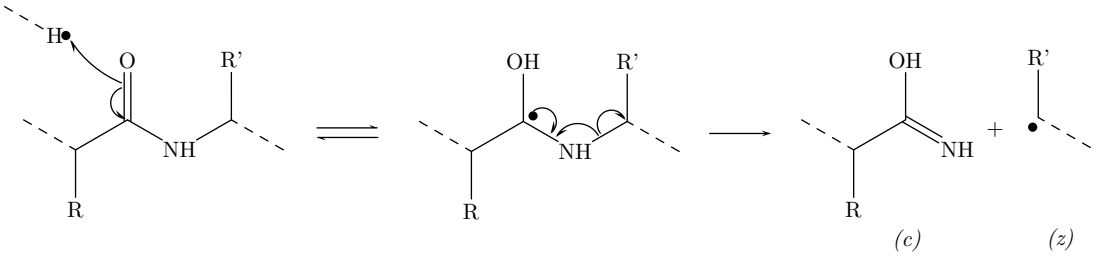


FIG. 5.3 – *Mécanisme de réarrangement observé en ECD. Le site protoné capte un électron, ce qui engendre alors un réarrangement sur le squelette peptidique et une rupture au niveau du carbone C_α [15].*

Zubarev *et al.* ont montré pour la protéine ubiquitine que les liaisons intramoléculaires faibles peuvent être conservées suite à une dissociation par capture d'électron [16]. De plus,

le repliement et le dépliement de peptides multi-chargés ont également été caractérisés par ECD [17, 18], ce qui prouve que des processus de fragmentation non-ergodiques peuvent permettre de déterminer les structures secondaires et tertiaires de peptides [19, 20, 21].

En LID, le photon produit une excitation électronique du peptide (figure 5.2). Après cette étape initiale d'excitation, la dissociation directe dans l'état excité rentre en compétition avec un retour à l'état fondamental par une conversion interne, et avec une désexcitation radiative. À 266 nm (4,66 eV), les électrons des résidus aromatiques peuvent être excités et un modèle impliquant un transfert de charge vers l'état dissociatif a été proposé [22, 23]. Les mesures effectuées sur une expérience femtoseconde de type pompe/sonde montrent que les durées de vie des états excités du tryptophane et des peptides contenant du tryptophane sont courtes [19]. Elles sont de l'ordre de quelques centaines de femtosecondes à quelques dizaines de picosecondes, mettant en évidence un fort couplage entre l'état localement excité $\pi\pi^*$ et l'état prédissociatif. De plus, des canaux spécifiques de fragmentation sont détectés, comme la formation de cations radicaux après une perte d'atome d'hydrogène, et une fragmentation sur la liaison C_α-C_β [10, 24, 25]. Pour une énergie plus élevée (157 nm, 7,9 eV), des fragments additionnels, similaires à ceux observés en ECD, sont détectés [26]. On peut également observer des fragments de la chaîne peptidique similaires à ceux obtenus en CID, ce qui suggère une compétition entre une fragmentation rapide non ergodique et une fragmentation sur des temps plus longs avec une redistribution de l'énergie.

En résumé, les mécanismes d'excitation semblent exercer une forte influence sur les fragments produits. Bien qu'elle ne soit pas spécifiquement prise en compte dans les modèles, la structure secondaire des peptides peut également avoir une influence sur le processus de relaxation consécutif à une excitation électronique. Pour mieux comprendre les propriétés structurales et la fragmentation des peptides, nous avons entrepris une étude complète de peptides modèles par différentes méthodes expérimentales, CID, ECD et LID.

Parallèlement à ces expériences réalisées par mes collègues, j'ai effectué des simulations Monte Carlo d'échange avec le champ de force AMBER pour explorer la surface d'énergie potentielle de l'état fondamental du pentapeptide simplement et doublement chargé. Les configurations obtenues en phase gazeuse montrent des différences significatives suivant l'état de charge et fournissent une piste pour expliquer qualitativement les résultats expérimentaux.

5.1.2 Piège à ions quadripolaire

Les expériences de dissociation induite par collision et par laser ont, en partie, été réalisées au laboratoire avec un spectromètre LCQ^{Duo} (ThermoFinningan) modifié pour également permettre l'injection d'un laser nanoseconde OPO accordable [27]. Le schéma de fonctionnement est représenté figure 5.4.

Les peptides sont ionisés par un électrospray et conduits par des octapôles dans un

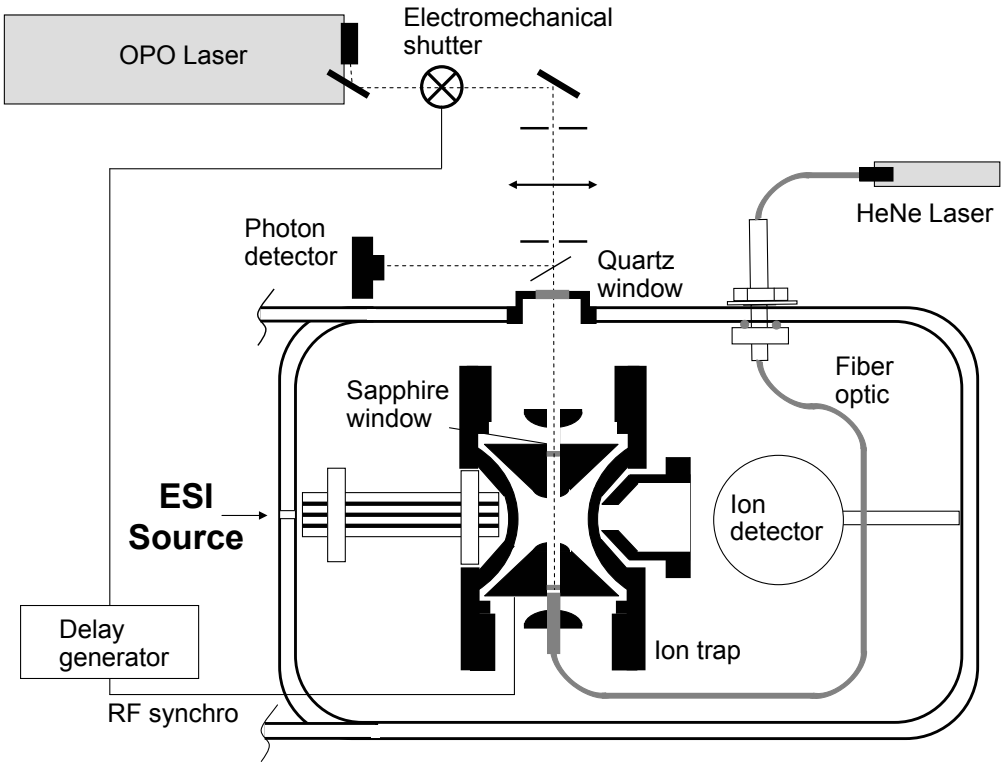


FIG. 5.4 – *Schéma de fonctionnement de la trappe à ions LCQ^{Duo} tiré de [27].*

piège quadripolaire. Pour les expériences de CID, de l'hélium présent dans le piège entre en collision avec les ions. Pour la LID, un trou dans l'électrode du piège permet d'exciter les ions par un laser OPO.

Les mesures d'ECD et de CID ont été obtenues par un instrument de résonance cyclotron à ions équipé d'une transformée de Fourier (laboratoire des mécanismes réactionnels de Palaiseau). D'autres expériences de LID ont utilisé un jet d'ions qui coupe un faisceau laser femtoseconde à 266 nm (laboratoire de photophysique moléculaire d'Orsay).

5.2 Pentapeptide AlaGlyTrpLeuLys

Mes collègues ont étudié le pentapeptide AlaGlyTrpLeuLys (AGWLK) par différentes méthodes expérimentales, CID, ECD et LID. Ce peptide a été choisi car il possède un résidu tryptophane nécessaire pour une excitation UV efficace, et un acide aminé C-terminal basique, la lysine, nécessaire pour former une espèce doublement chargée utilisable en ECD.

5.2.1 Résultats expérimentaux et interprétation

Les expériences de LID et CID ont été réalisées par deux membres du groupe avec la trappe à ions.

La figure 5.5 représente les spectres de masse CID et LID/MS^2 du peptide AGWLK

simplement protoné (m/z 574). Le spectre CID [figure 5.5(a)] présente des fragments du type a, b et y. Le spectre LID à 260 nm [figure 5.5(b)] présente des pics correspondant aux ions b_3 et b_4 ainsi qu'à la perte d'une molécule d'eau (perte de 18 Da) et à la cassure de la chaîne latérale du tryptophane (perte de 130 Da). Ce radical cation, qui n'est pas observé en CID, est un fragment habituel en LID. Le spectre de LID à 220 nm [figure 5.5(b)] présente deux canaux de fragmentation principaux à m/z 573 (perte de 1 Da) et m/z 444 (perte de 130 Da). Le premier fragment correspond à la perte d'un atome d'hydrogène provenant de la formation du radical $M^{\bullet +}$. Le pic à m/z 444 correspond à la brisure de la chaîne latérale du tryptophane, déjà observée à 260 nm. Les autres fragments obtenus à 260 nm sont aussi détectés mais avec une intensité moindre. Des fragments similaires ont été observés avec les impulsions d'un laser femtoseconde dans l'expérience à croisement de jets.

Le spectre CID du peptide doublement protoné (m/z 288), représenté sur la figure 5.6(a), offre, comme pour le simplement chargé, les ruptures habituelles du squelette peptidique (ions de type a, b et y). Par contre, le spectre LID à 220 nm [figure 5.6(b)] ne présente pratiquement pas de fragmentation. Un résultat similaire est obtenu pour le spectre de LID à 260 nm. Pour ces deux longueurs d'onde, l'efficacité de fragmentation est 50 fois plus petite pour le peptide doublement protoné que pour le simplement protoné. Finalement, le spectre ECD/MS^2 [figure 5.6(c)] est marqué par la présence d'ions a, b, y, de l'ion du peptide simplement protoné et de l'ion w_2. Mis à part l'ion w_2, ce dernier spectre de masse est très proche du spectre CID du peptide simplement chargé, tant par les pics présents que par leur intensité relative.

Ces premiers résultats expérimentaux montrent que les chemins de fragmentation dépendent fortement de la méthode employée pour apporter l'énergie dans le peptide. Les excitations par CID, LID et ECD conduisent à des canaux de fragmentation distincts. Ils dépendent également de l'état électronique excité par le photon comme illustré, par exemple, sur les figures 5.5(b) et 5.5(c). La variété des canaux de fragmentation qui peuvent être observés et la sensibilité du processus d'excitation montrent qu'il est difficile de construire un modèle général pour la fragmentation de peptides.

Les résultats les plus surprenants de ces expériences sont la différence de taux de LID entre les peptides simplement et doublement protonés, ainsi que l'absence d'ions de type c ou z dans le spectre ECD. L'écart de rendement de photofragmentation est d'un ou deux ordres de grandeur et pourrait être expliqué par une différence de sections efficaces d'absorption. Or, l'absorption est principalement due à la transition $\pi\pi$–$\pi\pi^*$ du chromophore indole et l'énergie associée paraît insensible à l'environnement. En effet, l'acide aminé Trp, l'ion $[TrpH]^+$, les peptides contenant du tryptophane en phase gazeuse et le Trp en solution aqueuse absorbent tous dans la même région d'énergie [27, 28, 29, 30, 31, 32]. De plus, le peptide doublement protoné ne fragmente pas, que ce soit à 220 nm ou à 260 nm. Si une faible absorption $\pi\pi$–$\pi\pi^*$ était responsable du bas taux de fragmentation

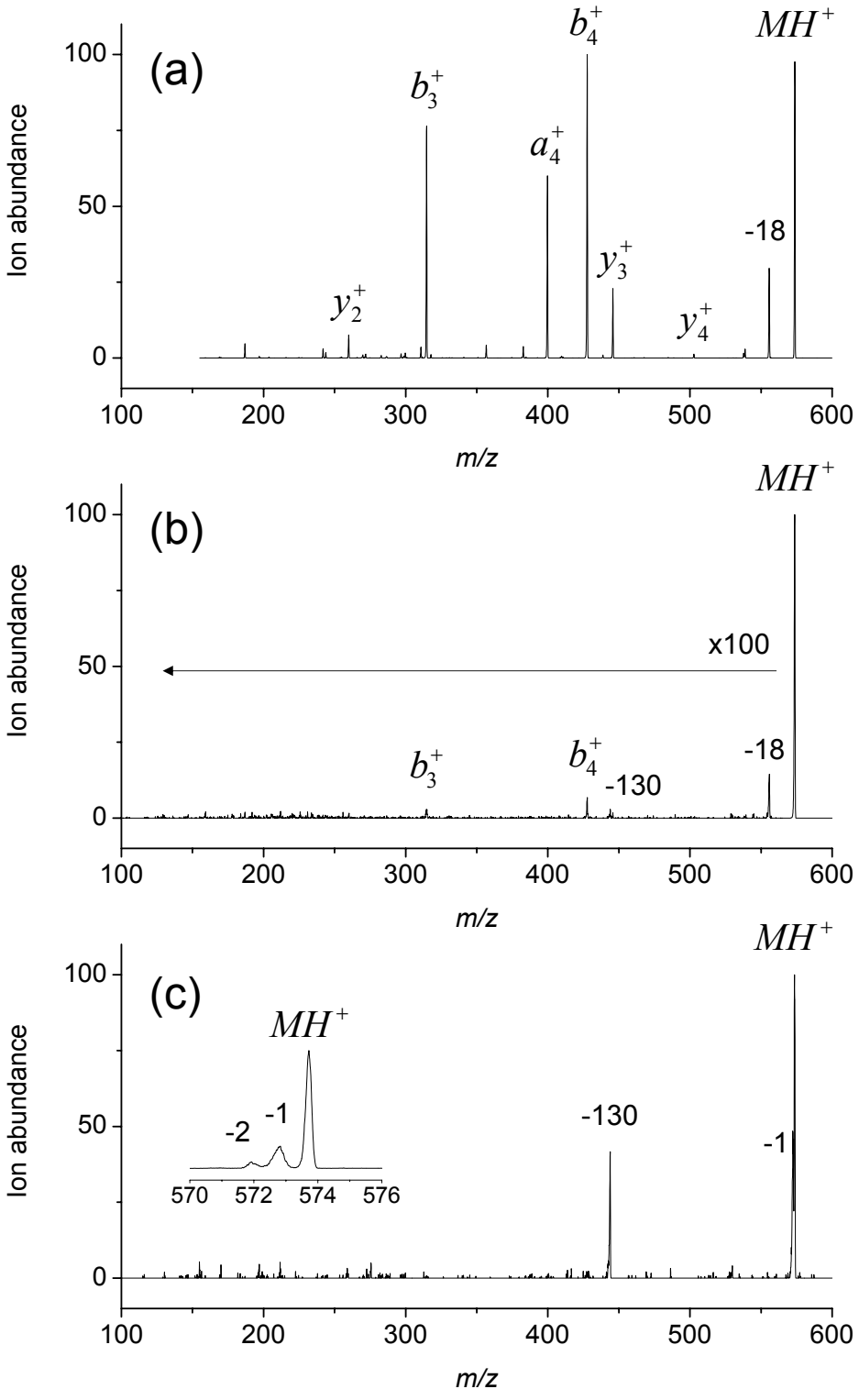


FIG. 5.5 – *Spectres de masse du peptide [AGWLK + H]$^+$ obtenus par (a) CID, (b) LID à λ= 260 nm et (c) LID à λ= 220 nm. L'insert de la figure (c), entre m/z 570 et m/z 576, met en évidence les pertes d'hydrogène.*

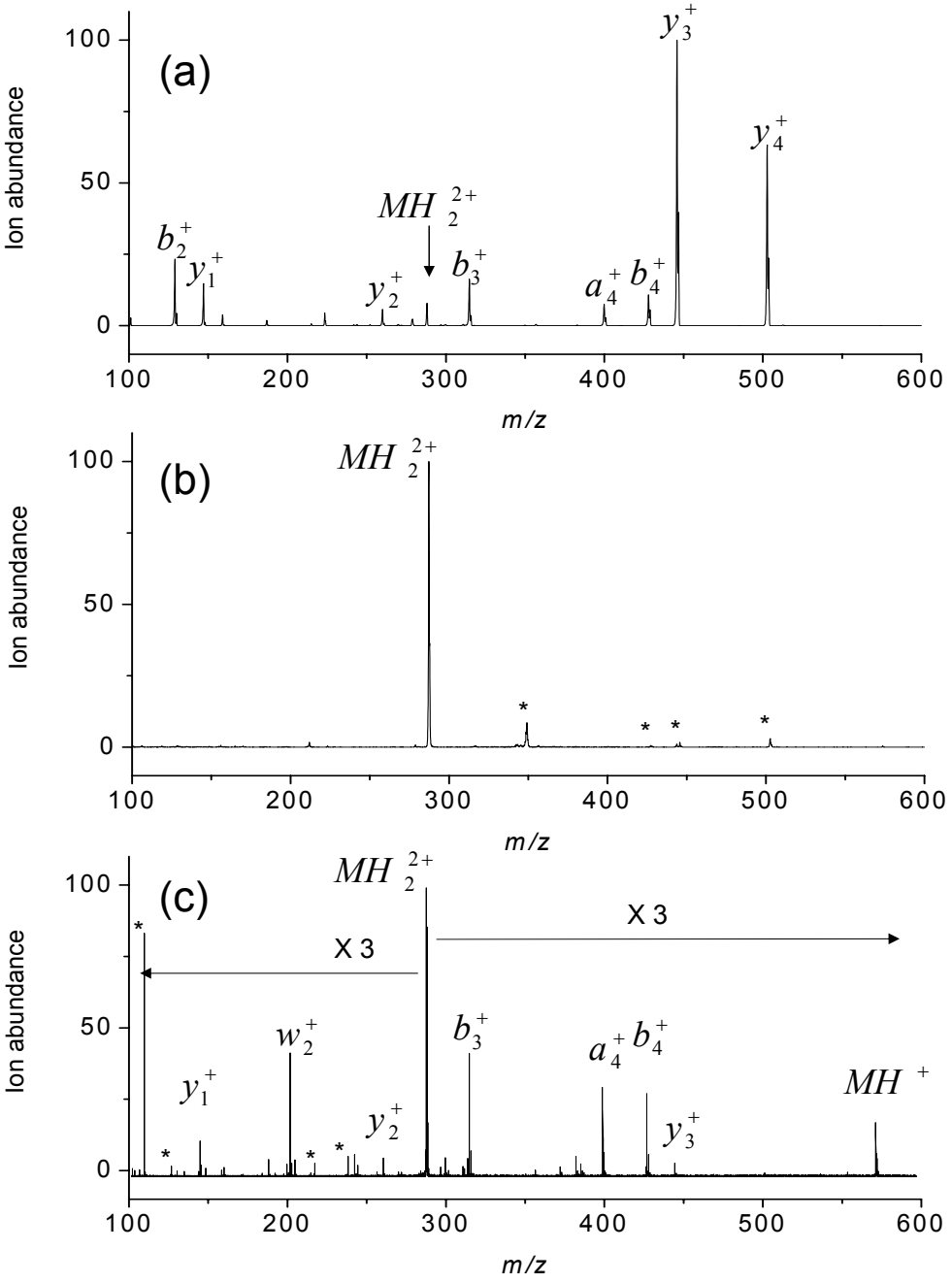


FIG. 5.6 – *Spectres de masse du peptide [AGWLK + 2H]$^{2+}$ obtenus par (a) CID, (b) LID à λ= 260 nm et (c) ECD MS^2. Dans la figure (b), les pics étoilés sont également observés sans laser. Figure (c), les pics étoilés sont des pics de bruit.*

du peptide doublement protoné, alors la transition devrait être déplacée de plus d'un électron-volt, ce qui semble improbable.

Nous proposons alors une interprétation qui repose sur un mécanisme général développé dans un précédent travail sur la photodissociation d'acides aminés et de dipeptides contenants du tryptophane [22, 23, 33]. Dans ce modèle schématisé sur la figure 5.7, le mécanisme de fragmentation est contrôlé par le couplage entre l'état électronique excité $\pi\pi^*$ localisé sur le cycle de l'indole et l'état dissociatif $\pi\sigma^*$ où l'électron se trouve sur le groupe amine protoné. Ce couplage peut être compris comme un transfert d'électron de l'indole vers le groupe NH_3^+. La présence de l'électron sur ce groupe forme alors un radical hypervalent similaire au radical ammonium NH_4. Ce radical est instable et produit une dissociation rapide le long de la coordonnée NH [34]. On obtient ensuite soit une perte d'atome d'hydrogène, soit une conversion interne par le couplage entre l'état $\pi\sigma^*$ et l'état fondamental. Pour que ce couplage ait lieu, le groupe NH_3^+ doit être à proximité du cycle de l'indole. Sans couplage, aucune fragmentation ne se produit. Dans le cadre de ce modèle, et même si le temps d'apparition des fragments peut être long, la première étape vers la fragmentation spécifique de l'indole est courte et dépend de la géométrie initiale du peptide.

5.2.2 Simulations en Monte Carlo d'échange du peptide AGWLK

Avec le modèle précédent et en considérant uniquement les conformations adoptées par les deux peptides, nous tentons d'expliquer qualitativement les variations importantes du taux de fragmentation entre le peptide simplement protoné et celui doublement protoné. Intuitivement, la répulsion électronique entre les deux charges sur le doublement chargé devrait privilégier, par rapport au simplement chargé, des structures moins compactes.

Nous avons donc réalisé des simulations Monte Carlo d'échange avec le champ de force AMBER *ff96* et une constante électrostatique $\epsilon = 2$ pour explorer la surface d'énergie potentielle des peptides simplement et doublement protonés.

Les deux sites les plus basiques et les plus sensibles à la protonation sont l'atome d'azote de la chaîne latérale de la lysine et le N-terminal de l'alanine. À priori, le proton peut être localisé indifféremment sur un des deux sites pour le simplement protoné. Nous avons effectué des calculs en Monte Carlo d'échange suivis de calculs semi-empiriques AM1 qui montrent que l'espèce $AGWLK^+$ est plus stable que A^+GWLK. La molécule doublement protonée l'est sur les deux sites basiques : A^+GWLK^+. La figure 5.8 représente les formules topologiques du pentapeptide simplement et doublement protoné.

Les simulations en Monte Carlo d'échange sont constituées de 9 répliques aux températures de 180, 206, 302, 380, 474, 588, 725, 893 et 1100 K. La distribution des températures est à 85 % géométrique. Chaque simulation compte un total de $6{,}75 \times 10^8$ pas Monte Carlo, dont les $6{,}75 \times 10^7$ premiers pas sont utilisés pour la thermalisation et non pris en compte dans les statistiques. Nos simulations sont initialisées par une conformation aléatoire. Les

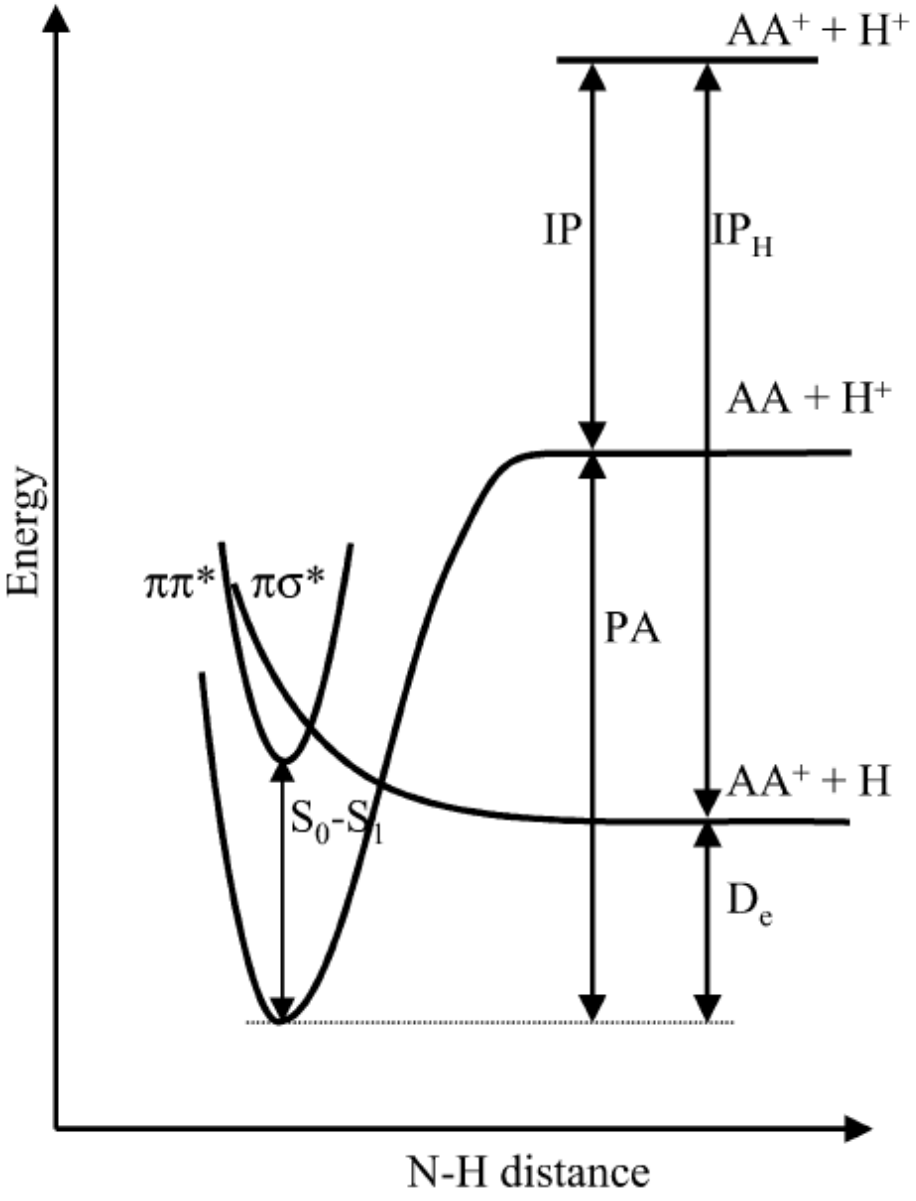


FIG. 5.7 – *Schéma du mécanisme de transfert d'électron entre l'indole du tryptophane et la charge (figure tirée de l'article de Kang* et al. *[22]). S_0–S_1 est la transition depuis l'état fondamental vers l'état excité $\pi\pi^*$, supposée identique pour les molécules neutres et protonées. IP désigne le potentiel d'ionisation du peptide, IP_H est le potentiel d'ionisation de l'hydrogène (13,6 eV) et PA l'affinité du peptide pour le proton. Enfin, D_e est l'énergie de dissociation de l'atome d'hydrogène quittant le peptide.*

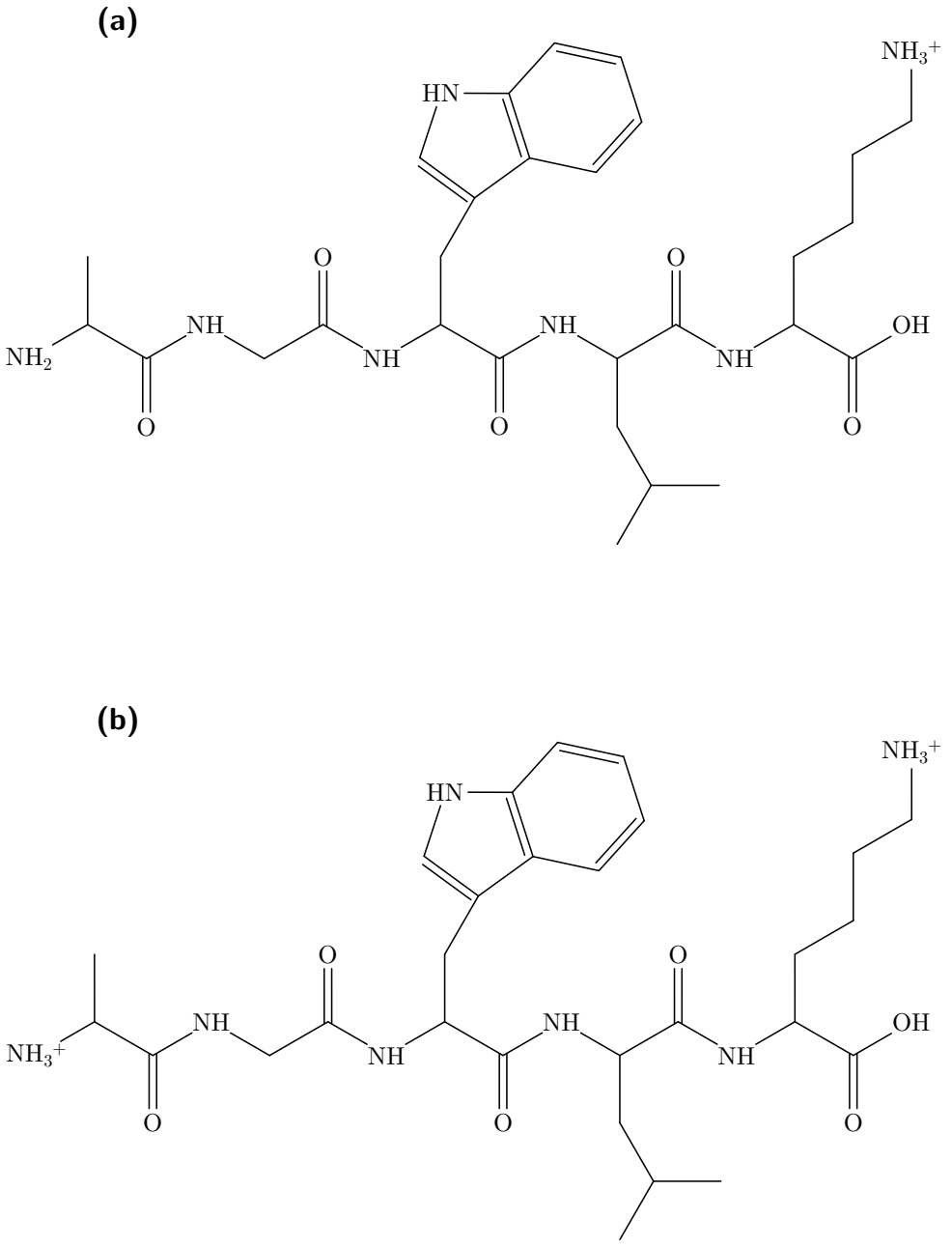


FIG. 5.8 – *Formule topologique du pentapeptide (a) simplement et (b) doublement protoné.*

échanges de répliques sont tentés tous les 100 cycles Monte Carlo. Un cycle Monte Carlo consiste à modifier chaque angle dièdre du squelette peptidique et des chaînes latérales (soit un total de 25 angles de torsion pour les molécules considérées ici). Les longueurs des liaisons et les angles de flexion sont gardés constants. Au cours de ces simulations, nous avons suivi l'évolution de la distance entre l'azote du groupe indole du tryptophane et les atomes d'azote des sites de protonation, sur la lysine [$d(\mathrm{N_{indole}}$–$\mathrm{N_{Lys}})$] et l'alanine N-terminale [$d(\mathrm{N_{ter}}$–$\mathrm{N_{indole}})$], comme représenté sur la figure 5.9.

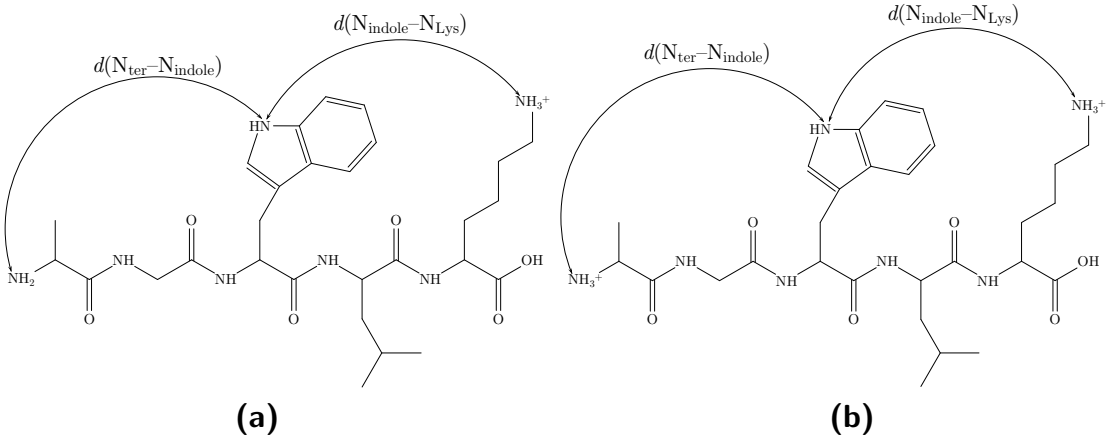


FIG. 5.9 – *Distance $d(N_{\mathrm{indole}}$–$N_{\mathrm{Lys}})$ et $d(N_{\mathrm{ter}}$–$N_{\mathrm{indole}})$ pour le pentapeptide (a) simplement et (b) doublement protoné.*

Les résultats pour la température de 302 K sont présentés sur la figure 5.10 par des cartes à trois dimensions, $d(\mathrm{N_{indole}}$–$\mathrm{N_{Lys}})$, $d(\mathrm{N_{ter}}$–$\mathrm{N_{indole}})$ et intensité relative. Les différentes zones peuplées sur ces deux cartes correspondent à différentes familles de structures. Le couplage entre les états $\pi\pi^*$ et $\pi\sigma^*$ ne peut se produire que pour des structures dont une charge est proche du cycle de l'indole. Pour le peptide simplement protoné, la structure la plus probable [notée (1) sur la figure 5.10(a)] est une structure repliée où l'azote protoné est en proche interaction avec le cycle de l'indole [figure 5.11(a)], ce qui peut conduire à un transfert de charge efficace. Pour le doublement protoné, il n'y a clairement pas de maximum correspondant à une structure où la charge est en proche interaction avec le cycle indole. Le maximum noté (2) sur la carte de la figure 5.10(b) correspond à des structures où les deux charges sont partiellement solvatées par les groupes carbonyles et n'ont pas d'interaction directe avec le cycle indole [figure 5.11(b)]. En moyenne, la distance indole–proton est également plus longue, ce qui explique l'absence de transfert de charge.

Ces résultats sont basés sur des calculs par champ de force et sont donc à prendre avec précautions. On peut cependant penser que ces structures rendent en partie compte de la différence d'efficacité dans le couplage entre les états excités $\pi\pi^*$ et $\pi\sigma^*$ et ainsi expliquer la différence de fragmentation entre les peptides simplement et doublement protonés.

Pour les structures représentatives du doublement protoné, les groupes carbonyles solvatent partiellement les charges. Ce comportement peut aussi expliquer les résultats

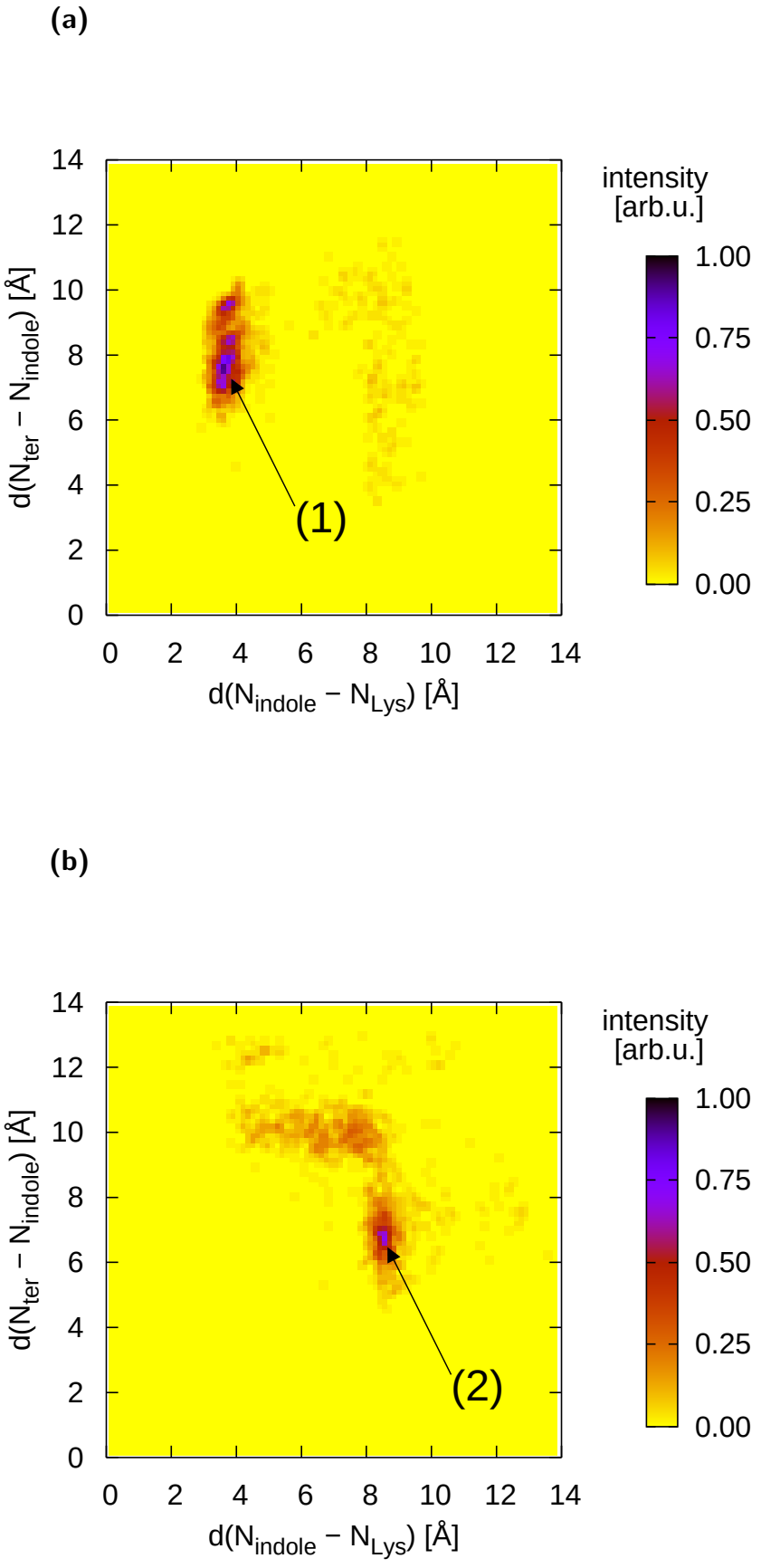


FIG. 5.10 – *Carte à trois dimensions, $d(N_{\text{indole}}\text{-}N_{\text{Lys}})$, $d(N_{\text{ter}}\text{-}N_{\text{indole}})$ et intensité, de la répartition des différentes structures obtenues pour le peptide (a) $AGWLK^+$ et (b) A^+GWLK^+ à 302 K.*

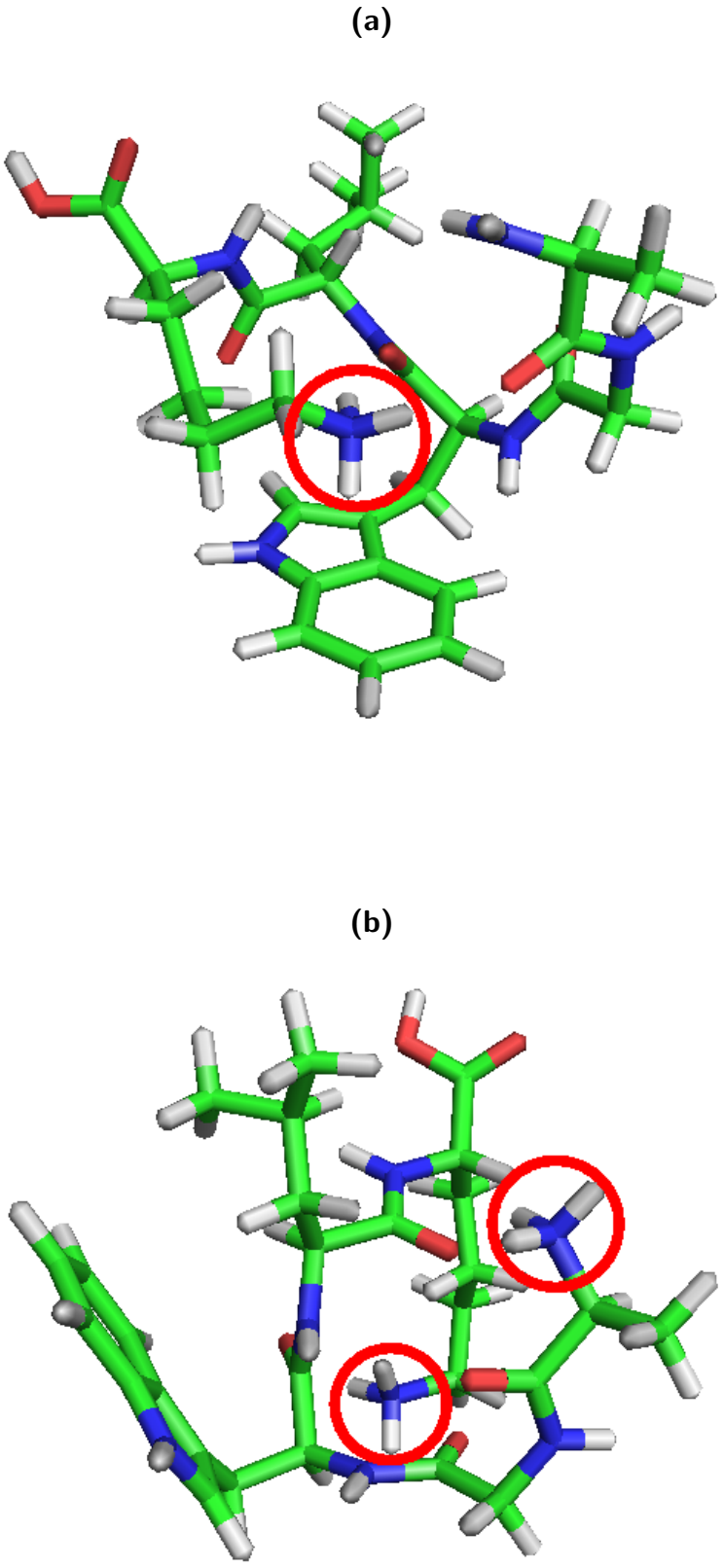


FIG. 5.11 – *Structures de (a) $AGWLK^+$ et (b) A^+GWLK^+ représentatives respectivement des maxima (1) et (2) des cartes (a) et (b) de la figure 5.10. Les cercles rouges indiquent les sites protonés.*

observés en ECD où aucun ion c ou z n'était détecté. En effet, l'obtention d'ions de type a, b, y et w ainsi que d'ions c et z est qualifiée d'expérience d'ECD « chaude » [35], pour laquelle des électrons avec une énergie de 10 eV ou plus sont utilisés. Dans un travail récent sur l'apparition d'ions de type b dans les spectres ECD, Cooper [36] a montré que ces fragmentations ne proviennent pas de la fragmentation secondaire d'ions c et il a suggéré que le mécanisme d'apparition de ces fragments pourrait être un dépôt d'énergie suivi d'une perte d'atome d'hydrogène. Nos résultats confortent cette hypothèse puisqu'aucun ion c ou z n'est observé pour le doublement protoné et que l'augmentation de l'énergie des électrons de 1 à 16 eV ne modifie pas significativement le taux de fragmentation. Dans le mécanisme de capture d'électron, la neutralisation d'un proton induit le transfert d'un atome d'hydrogène sur le squelette peptidique, en compétition avec l'évaporation d'un radical hydrogène. Ici, le pic à m/z 574 de la figure 5.6(c) et un modèle de fragmentation similaire à celui observé en CID pour le simplement protoné peuvent être interprétés comme la signature de cette perte.

Les simulations Monte Carlo d'échange sur les deux peptides montrent que le nombre d'interactions $NH^+ - CO$ est plus petit pour le doublement chargé que pour le simplement chargé. En prenant une distance de coupure de 2,5 Å entre l'hydrogène et l'oxygène du groupe carbonyle, on constate qu'en moyenne, 1,1 atomes d'hydrogène par groupe NH_3^+ est engagé dans une liaison hydrogène $NH^+ - CO$ avec un des 5 oxygènes carbonyles accessibles. L'évaporation d'atomes d'hydrogène est donc favorisée dans ce type de structures. En ECD, une telle fragmentation préférentielle a déjà été observée pour des systèmes plus grands [21].

Ces simulations nous ont permis de vérifier que la distance indole–charge était plus petite pour le simplement que pour le doublement protoné. Ce résultat est en faveur de l'hypothèse émise précédemment, à savoir, que la perte du chromophore en LID était due à un couplage entre l'état excité $\pi\pi^*$ et l'état dissociatif $\pi\sigma^*$, ce qui nécessite une faible distance indole–charge pour un transfert d'électron efficace. On peut ainsi expliquer, pour le cas du pentapeptide AGWLK, la différence de fragmentation LID entre les peptides simplement et doublement chargés.

5.3 Famille de polyvalines TrpValValValVal

Afin d'essayer de mieux comprendre l'influence de la distance charge–indole dans le mécanisme de fragmentation des peptides, nous avons étudié une famille de peptides par des simulations en Monte Carlo d'échange et des expériences de CID et LID.

Les peptides choisis pour cette étude sont des pentapeptides constitués d'un tryptophane et de 4 acides aminés valines. Le tryptophane joue le rôle de chromophore pour l'excitation initiale. Les valines possèdent une chaîne latérale encombrée qui va occasionner une gène stérique et limiter le nombre de conformations de la molécule. La famille de

peptides est obtenue en déclinant la position du tryptophane, de l'acide aminé N-terminal au C-terminal, comme représenté sur la figure 5.12. Les 5 peptides sont simplement protonés. Pour les besoins des simulations, le proton est localisé sur l'atome d'azote de l'acide aminé N-terminal, qui est le site le plus basique.

5.3.1 Simulations en Monte Carlo d'échange

Comme pour le peptide AGWLK, les simulations sont constituées de 9 répliques aux températures de 180, 206, 302, 380, 474, 588, 725, 893 et 1100 K. Chaque simulation compte un total de $1{,}16 \times 10^8$ pas Monte Carlo, dont les $1{,}16 \times 10^7$ premiers pas sont utilisés pour la thermalisation et non pris en compte dans les statistiques. Nos simulations sont initialisées par une conformation aléatoire. Les échanges de répliques sont tentés tous les 100 cycles Monte Carlo. Un cycle Monte Carlo consiste à modifier chaque angle dièdre du squelette peptidique et des chaînes latérales (soit un total de 29 angles de torsion). Au cours de ces simulations, nous avons suivi l'évolution de la distance [$d(N_{ter}–N_{indole})$] entre l'azote du groupe indole du tryptophane et l'atome d'azote protoné sur la valine N-terminale.

La figure 5.13 présente la distance $d(N_{ter}–N_{indole})$ pour les 5 peptides à 302 K. Les distributions de probabilité de distance sont toutes piquées, sauf celle de V^+VWVV qui est complètement plate. La distance la plus courte est obtenue pour V^+VVVW. Les trois peptides V^+VVWV, V^+WVVV et W^+VVVV offrent une distribution de distance intermédiaire.

Quelques structures représentatives des pics observés sur la figure 5.13 sont rassemblées dans la figure 5.14.

5.3.2 Résultats expérimentaux

Des expériences de CID et LID ont été réalisées sur les polyvalines avec, à Lyon, la trappe à ions décrite précédemment (figure 5.4) et l'expérience sur jets à Orsay. Les spectres LID et CID à 260 nm sont représentés respectivement figures 5.15 et 5.16. Les spectres de masse obtenus par CID offrent pour tous les peptides une perte d'eau (-18 Da), ainsi que les fragments de type b et y résultant de la rupture du squelette peptidique. Les taux de dissociation rassemblés dans le tableau 5.1 sont relativement uniformes, entre 67 et 92 %. Les spectres LID présentent également tous des fragments de type b et y, correspondant à la brisure de la chaîne peptidique. Le spectre de masse de $[WVVVV + H]^+$ est le seul à offrir la perte du chromophore (-130 Da) situé sur l'indole du tryptophane. Les spectres des peptides $[WVVVV + H]^+$ et $[VWVVV + H]^+$ présentent une évaporation d'ammoniac (-17 Da) alors que pour les spectres des trois autres peptides, nous observons uniquement des pertes d'eau (-18 Da). Les taux de dissociation (tableau 5.1) sont plus faibles et plus dispersés, de 1 à 6 %.

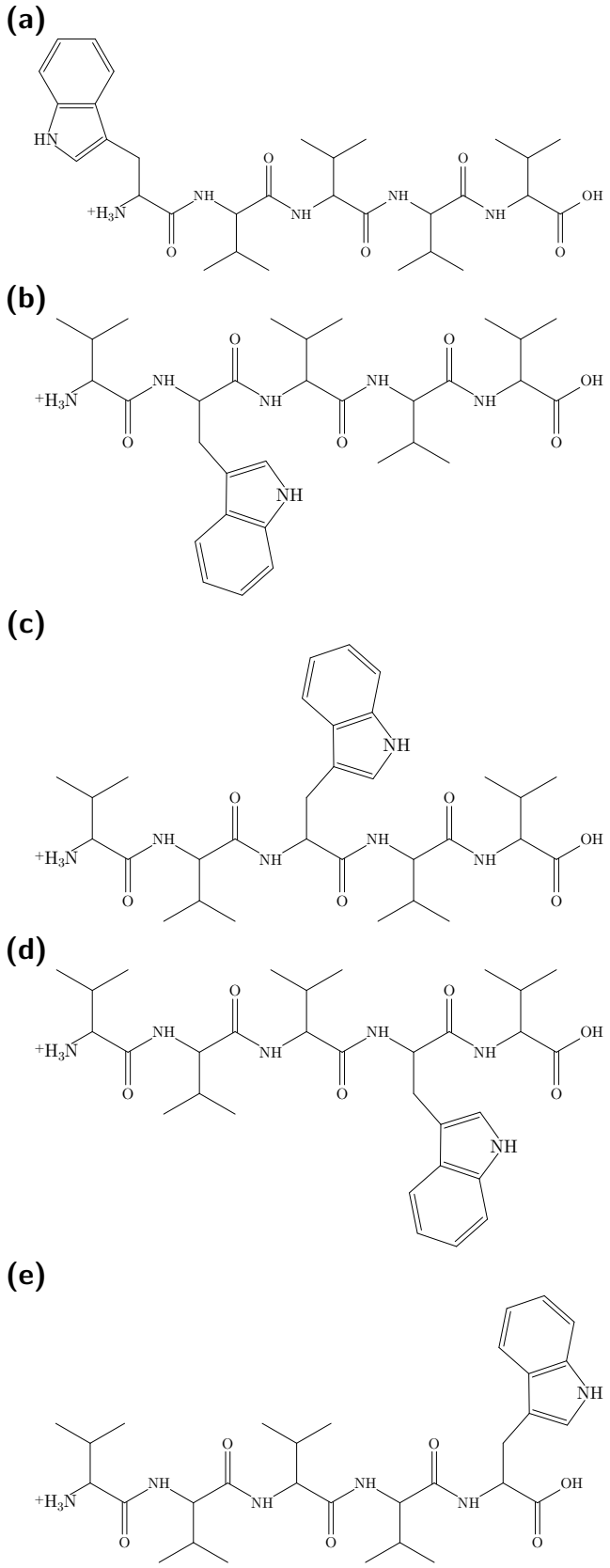


FIG. 5.12 – *Formules topologiques des polyvalines étudiées, (a) $Trp^{+}ValValValVal$ ou $W^{+}VVVV$, (b) $Val^{+}TrpValValVal$ ou $V^{+}WVVV$, (c) $Val^{+}ValTrpValVal$ ou $V^{+}VWVV$, (d) $Val^{+}ValValTrpVal$ ou $V^{+}VVWV$ et (e) $Val^{+}ValValValTrp$ ou $V^{+}VVVW$.*

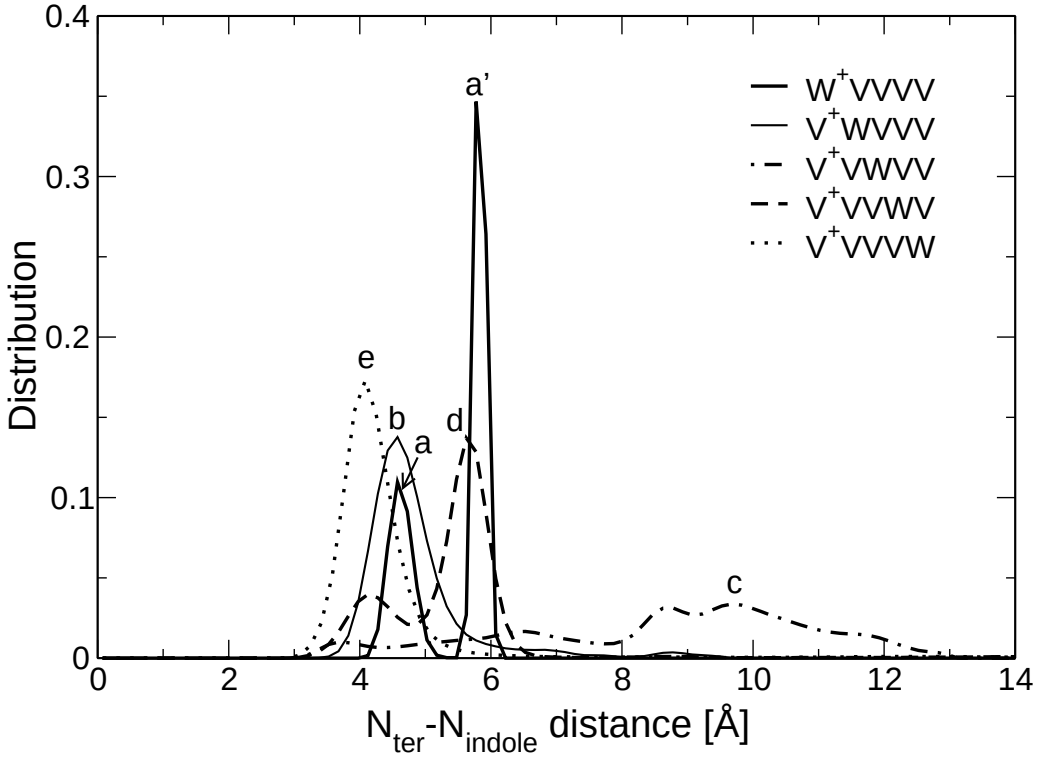


FIG. 5.13 – *Distribution de la distance d(N_{ter}–N_{indole}) pour la famille de polyvalines à 302 K.*

Les spectres LID des trois derniers peptides sont très proches des spectres CID correspondants. De plus, les taux de fragmentation LID sont plus importants pour [WVVVV + H]$^+$ et [VWVVV + H]$^+$ alors qu'ils sont tous homogènes en CID.

TAB. 5.1 – *Taux de dissociation des polyvalines pour les fragmentations CID et LID.*

	Taux de dissociation	
Peptide	CID [%]	LID [%]
[WVVVV + H]$^+$	67	6
[VWVVV + H]$^+$	92	6
[VVWVV + H]$^+$	89	3
[VVVWV + H]$^+$	76	1
[VVVVW + H]$^+$	91	2

5.3.3 Analyse des résultats

Les résultats expérimentaux présentés précédemment montrent que la fragmentation après une excitation CID est différente de celle obtenue après une excitation par laser. Les deux premiers peptides présentent clairement un comportement de LID où la fragmentation rapide fait intervenir le mode d'excitation. Les trois autres peptides fragmentent plutôt dans un mode d'IVR proche de la CID. On a donc deux mécanismes très différents

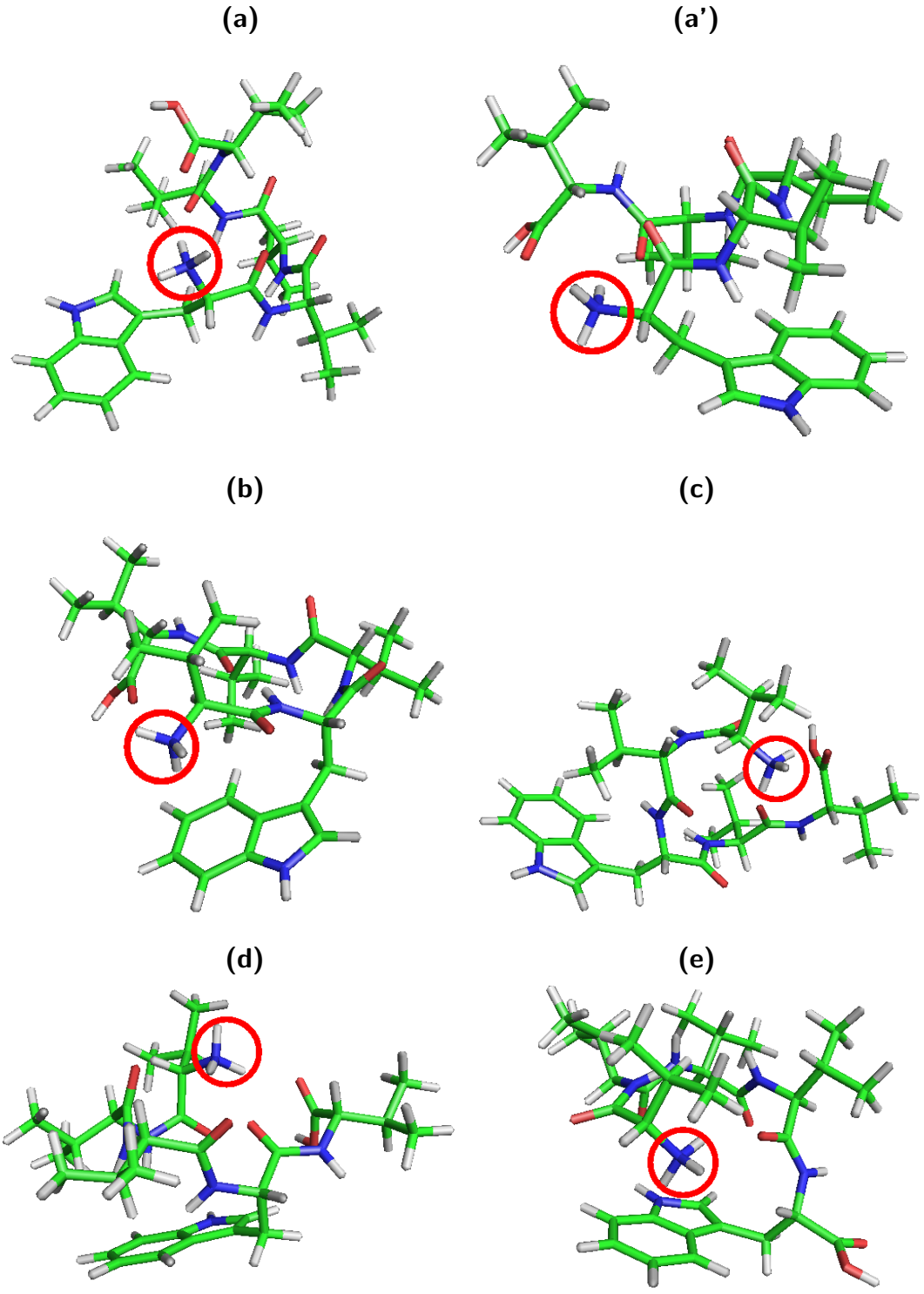


FIG. 5.14 – *Structures représentatives correspondant aux maxima notés a, a', b, c, d et e sur la figure 5.13. La distance d(N_{ter}–N_{indole}) vaut respectivement 4,48 Å, 5,77 Å, 4,80 Å, 9,00 Å, 5,60 Å et 4,08 Å. Les cercles rouges indiquent les sites de protonation.*

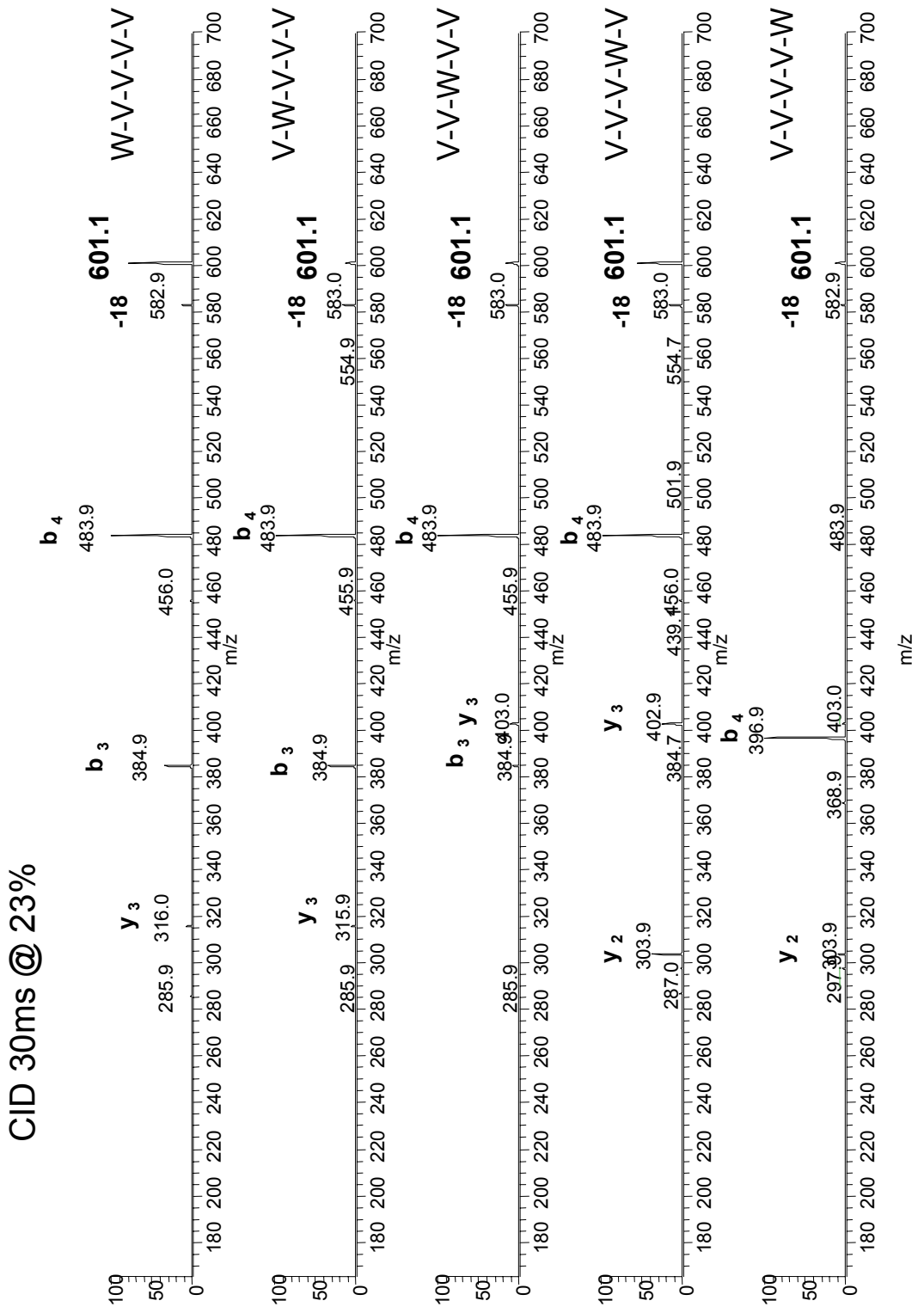


FIG. 5.15 – *Spectres CID des peptides [WVVVV + H]$^+$, [VWVVV + H]$^+$, [VVWVV + H]$^+$, [VVVWV + H]$^+$ et [VVVVW + H]$^+$.*

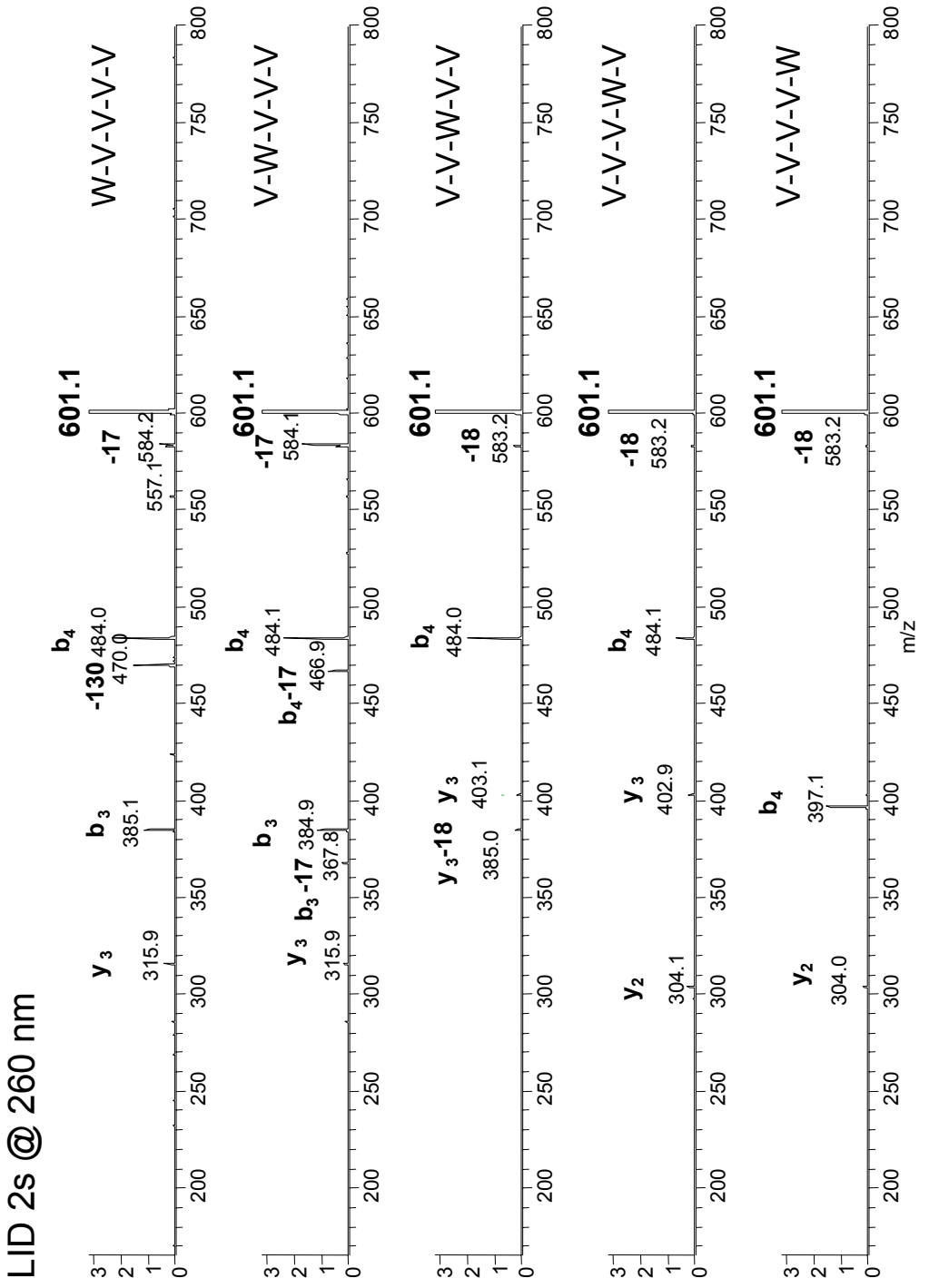


FIG. 5.16 – *Spectres LID à 260 nm des peptides [WVVVV + H]$^+$, [VWVVV + H]$^+$, [VVWVV + H]$^+$, [VVVWV + H]$^+$ et [VVVVW + H]$^+$.*

suivant la position du tryptophane dans le peptide. Les simulations montrent qu'il est difficile d'obtenir une corrélation nette entre, d'une part la distance entre l'azote de l'indole et l'azote protoné, et d'autre part, les comportements de fragmentation, comme observé pour le pentapeptide AGWLK. Plusieurs hypothèses pourraient expliquer cette difficulté à interpréter les résultats à partir de ce modèle de transfert d'électron [22, 23, 33].

La première hypothèse est le modèle lui même, qui est très réducteur car il ne considère la surface d'énergie que sur une seule coordonnée. De plus, il ne prend pas en compte, par exemple, la compétition avec d'autres mécanismes de réarrangement (la coupure C_α-C_β se produit-elle dans l'état excité ou dans l'état fondamental ?).

Une deuxième hypothèse pourrait être que les géométries calculées ne soient pas les bonnes, compte-tenu du modèle utilisé. Le champ de force peut, en effet, ne pas être adapté pour décrire les peptides étudiés.

Ensuite, la localisation de la charge, pourrait être erronée car le proton serait mobile. En particulier, la basicité de l'azote N_{ter} serait fortement atténuée dans une série de résidus valine par l'importance des chaînes carbonées. Le proton pourrait alors facilement atteindre la chaîne peptidique.

Enfin, on ne peut exclure qu'une partie de la redistribution d'énergie se fasse par relaxation radiative (fluorescence).

5.4 Peptides extraits d'une digestion enzymatique

En parallèle aux études effectuées par L. Joly pour son stage de master, nous avons abordé le problème du lien entre dissociation et structure par une approche totalement différente, de type statistique, en considérant de nombreux peptides. J'ai essayé d'établir une relation entre la fragmentation et la structure primaire, voire secondaire. Cette dernière n'est d'ailleurs plus déterminée par calcul mais tirée d'une structure cristallisée en solution. Les expériences se déroulant en phase gazeuse, les structures cristallines utilisées ne sont bien évidemment qu'indicatives.

Ainsi, nous avons particulièrement étudié le précurseur de la sérotransferrine humaine (entrée Swiss-Prot P02787 [37]) dont la séquence est donnée figure 5.17. Cette protéine a été digérée par l'enzyme trypsine, qui a la capacité de couper la chaîne peptidique à gauche d'une arginine (R) ou d'une lysine (K). Une chromatographie sur colonne a ensuite permis d'isoler et de purifier les peptides résultant de la digestion.

Ces peptides sont ensuite injectés dans le piège à ions puis excités par un laser kHz (YLF) à 262 nm . Le tableau 5.2 rassemble quelques peptides identifiés par spectrométrie de masse. Pour les peptides 1 à 5, une fragmentation est observée avec un pic à -107 Da qui correspond à une perte du chromophore de la tyrosine. Les peptides 6 à 10 ne fragmentent pas. Pour tous ces peptides, aucun pic à -130 Da n'est observé, ce qui signifie qu'il n'y a pas de perte du chromophore tryptophane. Les acides aminés les plus basiques sont

MRLAVGALLVCAVLGLCLAVPDKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVK
KASYLDCIRAIAANEADAVTLDAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVV
KKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPC
ADGTDFPQLCQLCPGCGCSTLNQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRD
QYELLCLDNTRKPVDEYKDCHLAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKE
FQLFSSPHGKDLLFKDSAHGFLKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKP
VKWCALSHHERLKCDEWSVNSVGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGK
CGLVPVLAENYNKSDNCEDTPEAGYFAVAVVKKSASDLTWDNLKGKKSCHTAVGRTAGWN
IPMGLLYNKINHCRFDEFFSEGCAPGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAF
RCLVEKGDVAFVKHQTVPQNTGGKNPDPWAKNLNEKDYELLCLDGTRKPVEEYANCHLAR
APNHAVVTRKDKEACVHKILRQQQHLFGSNVTDCSGNFCLFRSETKDLLFRDDTVCLAKL
HDRNTYEKYLGEEYVKAVGNLRKCSTSSLLEACTFRRP

FIG. 5.17 – *Séquence de la sérotransferrine humaine (entrée Swiss-Prot P02787). Cette protéine a une masse de 77050 Da et compte 698 acides aminés.*

l'arginine (R) et la lysine (K). Ils sont susceptibles d'être facilement protonés et donc de porter la charge.

TAB. 5.2 – *Peptides identifiés par spectrométrie de masse. Les résidus entre parenthèses correspondent aux résidus voisins N-ter et C-ter avant la digestion enzymatique. Les résidus en gras sont les tyrosines.*

N°	m/z	Position dans la séquence	Coupures manquées	Séquence	% fragmentation
1	1000,4991	669–676	0	(K)**Y**LGEE**Y**VK(A)	3,4
2	1283,5697	531–541	0	(K)EG**YY**G**Y**TGAFR(C)	2,8
3	1478,7354	332–343	0	(K)M**Y**LG**Y**E**Y**VTAIR(N)	0,8
4	1577,8150	476–489	0	(R)TAGWNIPMGLL**Y**NK(I)	1,0
5	1923,9349	328–343	1	(R)MDAKM**Y**LG**Y**E**Y**VTAIR(N)	2,2
6	827,4052	565–571	0	(K)NPDPWAK(N)	0,0
7	1249,6065	454–464	0	(K)SASDLTWDNLK(G)	0,0
8	1377,7014	453–464	1	(K)KSASDLTWDNLK(G)	0,0
9	1610,8542	669–682	1	(K)**Y**LGEE**Y**VKAVGNLR(K)	0,0
10	1629,8164	108–121	0	(K)EDPQTF**YY**AVAVVK(K)	0,5

En gardant, l'hypothèse précédente d'un transfert d'électron entre le chromophore excité par le laser et la charge située sur le site le plus basique, une faible distance entre le chromophore et la charge est requise pour un transfert efficace. Les peptides numérotés 1 à 5 sur le tableau 5.2 fragmentent et ils possèdent au moins une tyrosine. On ne sait pas à priori quel chromophore et donc quelle tyrosine va être excitée et donner lieu à une fragmentation. Cependant on remarque facilement que le nombre de résidus entre la charge (portée par K ou R) et la plus proche tyrosine est plus petit pour les peptides qui

fragmentent que pour ceux qui ne fragmentent pas. Ces résultats sont rassemblés dans le tableau 5.3. Cette notion de distance séquentielle n'est bien sur que qualitative puisque pour une même distance, le taux de fragmentation peut varier de manière significative.

Tab. 5.3 – *Nombre de résidus entre les lysines (K) ou les arginines (R), et la tyrosine (Y) (en gras).*

N°	Séquence	Distance	% fragmentation
1	**Y**LGEE**Y**VK	1	3,4
2	EG**YY**G**Y**TGAFR	4	2,8
3	M**Y**LG**Y**E**Y**VTAIR	4	0,8
4	TAGWNIPMGLL**Y**NK	1	1,0
5	MDAKM**Y**LG**Y**E**Y**VTAIR	4	2,2
6	NPDPWAK	–	0,0
7	SASDLTWDNLK	–	0,0
8	KSASDLTWDNLK	–	0,0
9	**Y**LGEE**Y**VKAVGNLR	7	0,0
10	EDPQTF**YY**AVAVVK	5	0,5

Une comparaison des structures tridimensionnelles des peptides sera alors pertinente. Pour cela, nous avons aligné la séquence du précurseur de la sérotransférine humaine avec les 36121 séquences des structures présentes dans la base de donnée PDB. Le meilleur score obtenu par l'algorithme *basic local alignment search tool* (BLAST) est pour la structure 1JNF [38] avec 78 % d'identité (figure 5.18). Cette structure possède 676 acides aminés, ce qui est comparable aux 698 résidus de la séquence initiale. Par contre, cette protéine n'a pas la même taxinomie puisqu'elle provient du lapin ! On supposera que malgré les mutations de certains résidus, la structure tertiaire de la sérotransférine du lapin sera très proche de celle de l'humain. Parmi les résultats proposés par BLAST, nous avons également cherché une séquence ayant une identité maximale (figure 5.19). La structure 1A8E [39] répond à un tel critère car elle correspond effectivement à la sérotransférine humaine. Par contre, elle ne possède que 329 résidus.

Notre approche a donc été la suivante. Dans un premier temps, nous avons essayé de déterminer la structure tertiaire des peptides présents dans le tableau 5.2 à partir de la structure humaine puis, à partir de la structure du lapin si la séquence du peptide recherché n'était pas présente chez l'humain (tableau 5.4).

La figure 5.20 rassemble les structures des peptides 1 à 10. Les structures des peptides 2 et 5 montrent clairement la proximité d'une arginine ou d'une lysine avec une tyrosine, ce qui peut expliquer la fragmentation observée (dont les taux respectifs sont de 2,8 et 2,2 %). Les peptides 1 et 4 ne présentent pas, dans la structure PDB, d'interaction immédiate entre les résidus basiques et un des chromophores. La fragmentation s'explique alors par la proximité séquentielle (voir tableau 5.3) et le fait que la tyrosine et la lysine

```
>lcl|1JNF:1
          Length = 676

 Score = 1135 bits (2936), Expect = 0.0
 Identities = 531/676 (78%), Positives = 594/676 (87%), Gaps = 3/676 (0%)

Query: 22  DKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTL 81
           +KTVRWCAV++HEA+KC +FRD MK V+P DGP + CVKKASYLDCI+AIAA+EADAVTL
Sbjct: 3   EKTVRWCAVNDHEASKCANFRDSMKKVLPEDGPRIICVKKASYLDCIKAIAAHEADAVTL 62

Query: 82  DAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGL 141
           DAGLV++A L PNNLKPVVAEFYGSKE+P+TFYYAVA+VKK S FQ+N+L+GKKSCHTGL
Sbjct: 63  DAGLVHEAGLTPNNLKPVVAEFYGSKENPKTFYYAVALVKKGSNFQLNELQGKKSCHTGL 122

Query: 142 GRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQLCQLCPGCGCSTL 201
           GRSAGWNIPIGLL CDLPEPRKPLEKAVA+FFSGSC PCADG DFPQLCQLCPGCGCS++
Sbjct: 123 GRSAGWNIPIGLLLCDLPEPRKPLEKAVASFFSGSCVPCADGADFPQLCQLCPGCGCSSV 182

Query: 202 NQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCH 261
             YFGYSGAFKCLKDG GDVAFVK  TIFENL +K +RDQYELLCLDNTRKPVDEY+ CH
Sbjct: 183 QPYFGYSGAFKCLKDGLGDVAFVKQETIFENLPSKDERDQYELLCLDNTRKPVDEYEQCH 242

Query: 262 LAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGF 321
           LA+VPSH VVARS+ GKEDLIWELLNQAQEHFGKDKS +FQLFSSPHGK+LLFKDSA+GF
Sbjct: 243 LARVPSHAVVARSVDGKEDLIWELLNQAQEHFGKDKSGDFQLFSSPHGKNLLFKDSAYGF 302

Query: 322 LKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKPVKWCALSHHERLKCDEWSVNS 381
            KVPPRMDA +YLGYEYVTA+RNLREG CP+   DECK VKWCAL HHERLKCDEWSV S
Sbjct: 303 FKVPPRMDANLYLGYEYVTAVRNLREGICPDPLQDECKAVKWCALGHHERLKCDEWSVTS 362

Query: 382 VGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGKCGLVPVLAENYNKSDNCEDTP 441
            G IEC SAET EDCIAKIMNGEADAMSLDGG+VYIAG+CGLVPVLAENY  +D C+  P
Sbjct: 363 GGLIECESAETPEDCIAKIMNGEADAMSLDGGYVYIAGQCGLVPVLAENYESTD-CKKAP 421

Query: 442 EAGYFAVAVVKKSASDLTWDNLKGKKSCHTAVGRTAGWNIPMGLLYNKINHCRFDEFFSE 501
           E GY +VAVVKKS  D+ W+NL+GKKSCHTAV RTAGWNIPMGLLYN+INHCRFDEFF +
Sbjct: 422 EEGYLSVAVVKKSNPDINWNNLEGKKSCHTAVDRTAGWNIPMGLLYNRINHCRFDEFFRQ 481

Query: 502 GCAPGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAFRCLVEKGDVAFVKHQTVPQNT 561
           GCAPGS+K+SSLC+LC+G   ++C PNN+EGYYGYTGAFRCLVEKGDVAFVK QTV QNT
Sbjct: 482 GCAPGSQKNSSLCELCVGP--SVCAPNNREGYYGYTGAFRCLVEKGDVAFVKSQTVLQNT 539

Query: 562 GGKNPDPWAKNLNEKDYELLCLDGTRKPVEEYANCHLARAPNHAVVTRKDKEACVHKILR 621
           GG+N +PWAK+L E+D+ELLCLDGTRKPV E  NCHLA+APNHAVV+RKDK ACV + L
Sbjct: 540 GGRNSEPWAKDLKEEDFELLCLDGTRKPVSEAHNCHLAKAPNHAVVSRKDKAACVKQKLL 599

Query: 622 QQQHLFGSNVTDCSGNFCLFRSETKDLLFRDDTVCLAKLHDRNTYEKYLGEEYVKAVGNL 681
             Q  FG+ V DCS  FC+F S+TKDLLFRDDT CL  L  +NTYEKYLG +Y+KAV NL
Sbjct: 600 DLQVEFGNTVADCSSKFCMFHSKTKDLLFRDDTKCLVDLRGKNTYEKYLGADYIKAVSNL 659

Query: 682 RKCSTSSLLEACTFRR 697
           RKCSTS LLEACTF +
Sbjct: 660 RKCSTSRLLEACTFHK 675
```

FIG. 5.18 – *Alignement de la séquence de la sérotransférine humaine (P02787) avec la séquence de la sérotransférine du lapin (1JNF) avec l'algorithme BLAST.*

```
>lcl|1A8E:1
          Length = 329

 Score =  691 bits (1782), Expect = 0.0
 Identities = 329/329 (100%), Positives = 329/329 (100%)

Query: 22  DKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTL 81
           DKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTL
Sbjct: 1   DKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTL 60

Query: 82  DAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGL 141
           DAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGL
Sbjct: 61  DAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGL 120

Query: 142 GRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQLCQLCPGCGCSTL 201
           GRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQLCQLCPGCGCSTL
Sbjct: 121 GRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQLCQLCPGCGCSTL 180

Query: 202 NQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCH 261
           NQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCH
Sbjct: 181 NQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCH 240

Query: 262 LAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGF 321
           LAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGF
Sbjct: 241 LAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGF 300

Query: 322 LKVPPRMDAKMYLGYEYVTAIRNLREGTC 350
           LKVPPRMDAKMYLGYEYVTAIRNLREGTC
Sbjct: 301 LKVPPRMDAKMYLGYEYVTAIRNLREGTC 329
```

FIG. 5.19 – *Alignement de la séquence de la sérotransférine humaine (P02787) avec la séquence de la structure humaine correspondante (1A8E).*

TAB. 5.4 – *Correspondance entre la séquence de la sérotransférine humaine et les séquences des structures humaine (1A8E) et du lapin (1JNF).*

N°	Séquence P02787		Structure humaine		Structure du lapin	
		Position		Position	Séquence	Position
1	YLGEEYVK	669–676	non	–	YLGADYIK	647–654
2	EGYYGYTGAFR	531–541	non	–	EGYYGYTGAFR	509–519
3	MYLGYEYVTAIR	332–343	oui	311–322	–	–
4	TAGWNIPMGLLYNK	476–489	non	–	TAGWNIPMGLLYNR	456–469
5	MDAKMYLGYEYVTAIR	328–343	oui	307–322	–	–
6	NPDPWAK	565–571	non	–	NSEPWAK	543–549
7	SASDLTWDNLK	454–464	non	–	SNPDINWNNLE	434–444
8	KSASDLTWDNLK	453–464	non	–	KSNPDINWNNLE	433–444
9	YLGEEYVKAVGNLR	669–682	non	–	YLGADYIKAVSNLR	647–660
10	EDPQTFYYAVAVVK	108–121	oui	87–100	–	–

soient parmi les résidus qui présentent le plus de rotamères, les chaînes latérales étant très flexibles. En phase gazeuse, on peut d'autant plus facilement imaginer que le chromophore se replie sur la chaîne peptidique. Les taux de fragmentation observés sont aussi plus hétérogènes, respectivement de 3,4 et 1,0 %. Enfin, la structure du peptide 3 montre un rapprochement possible entre un des chromophores et le résidu arginine. Cependant, ces deux résidus sont bloqués dans une hélice α et l'interaction ne pourra pas être optimale, ce qui pourrait expliquer un taux de dissociation assez faible (0,8 %).

Les seuls chromophores des peptides 6 à 8 sont des tryptophanes qui semble-t-il ne donnent pas lieu à une fragmentation dans le cas présent. D'après les structures correspondantes, une interaction entre l'éventuel chromophore et un résidu basique semble improbable. La structure du peptide 9 possède plusieurs tyrosines et résidus basiques, mais la plupart sont bloqués dans une hélice α non flexible. Enfin, la structure du dernier peptide est très allongée car sa structure secondaire est en brin β. Cette structure n'est stabilisée que par des interactions avec d'autres brins β, qui ne sont pas présents ici. Un repliement de la chaîne peptidique peut alors être envisagé, rapprochant ainsi une des tyrosines avec la lysine et donnant lieu à une faible fragmentation (0,5 %).

Cette approche est bien sûr très qualitative car basée sur des structures obtenues en phase condensée. Cependant, il semblerait qu'il existe une certaine corrélation entre la structure primaire et secondaire et la fragmentation des peptides étudiés. Ce travail pourrait être poursuivi d'un point de vue théorique en relaxant en phase gaz les structures utilisées. Nous pourrions également calculer les conformations de ces peptides en partant uniquement de la séquence, comme nous l'avons fait pour le pentapeptide AGWLK et la famille de polyvalines. D'un point de vue expérimental enfin, il serait intéressant d'étudier un plus grand nombre de peptides, ce qui semble envisageable par la technique de digestion protéique développée au laboratoire.

5.5 Conclusion

Comme énoncé en introduction, l'utilisation d'expériences de photofragmentation pour rendre compte de la structure d'une biomolécule est difficile. Cependant, les résultats obtenus sont encourageants et fournissent déjà une première approche. La dissociation dépend indéniablement du mode d'excitation, qu'il soit court comme en LID ou en ECD, ou plus long avec une relaxation vibrationnelle de l'énergie comme en CID. Notre équipe dispose d'ailleurs du seul dispositif pouvant travailler sur les deux modes d'excitation CID et LID [27]. Le mécanisme de couplage entre l'état excité $\pi\pi^*$ et l'état dissociatif $\pi\sigma^*$ tente d'expliquer la fragmentation observée en LID. Ce couplage qui se traduit par un transfert d'électron entre l'indole et la charge ne peut se produire que si l'indole est suffisamment proche de la charge et donc si le peptide est dans une conformation bien particulière. Cette hypothèse est bien vérifiée pour le pentapeptide AGWLK, mais reste insuffisante

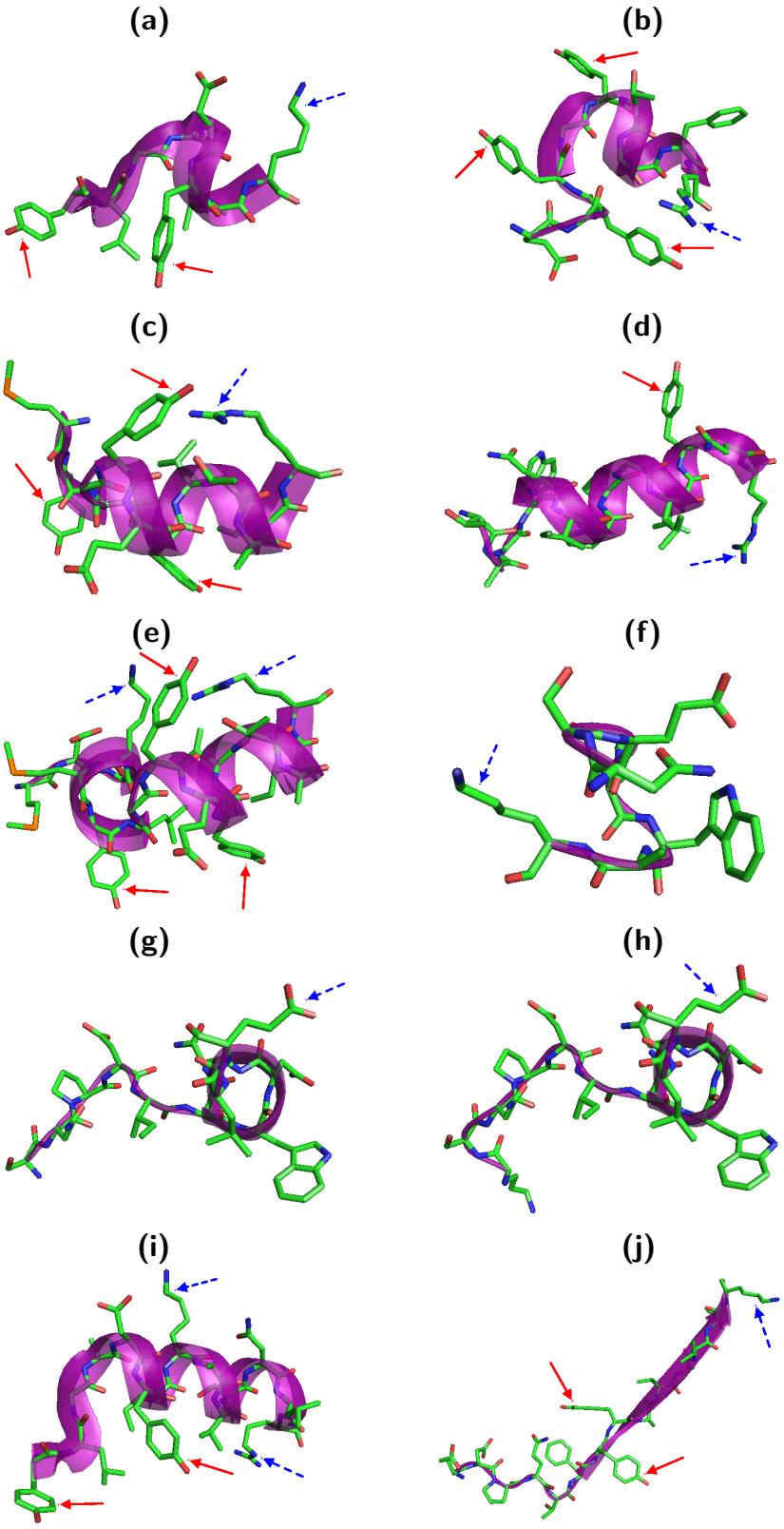


FIG. 5.20 – *Structures des peptides 1 à 10, notées dans l'ordre, de (a) à (j). Les tyrosines sont marquées d'une flèche continue rouge et les résidus arginine et lysine (ou leurs mutants pour la structure du lapin) d'une flèche discontinue bleue.*

pour la famille de polyvalines. Le modèle assez simple utilisé ainsi que la localisation de la charge peuvent être en partie responsables de ce désaccord. Pour être rigoureux, un calcul de dynamique dans les états excités devrait être effectué.

Enfin, une approche statistique sur un grand nombre de peptides en utilisant les structures cristallines peut également aider à comprendre la relation qui existe entre structure et photofragmentation. Ce travail mérite donc d'être poursuivi avant de pouvoir conclure.

Bibliographie

[1] P. Crozet, A. J. Ross, and M. Vervloet. Gas-phase molecular spectroscopy. *Annual Reports on the Progress of Chemistry, Section C*, 98:33–86, 2002.

[2] E. G. Robertson and J. P. Simons. Getting into shape: Conformational and supramolecular landscapes in small biomolecules and their hydrated clusters. *Physical Chemistry Chemical Physics*, 3:1–18, 2001.

[3] W. Chin, F. Piuzzi, I. Dimicoli, and M. Mons. Probing the competition between secondary structures and local preferences in gas phase isolated peptide backbones. *Physical Chemistry Chemical Physics*, 8:1033–1048, 2006.

[4] I. Compagnon, T. Tabarin, R. Antoine, M. Broyer, P. Dugourd, R. Mitrić, J. Peterson, and V. Bonačić-Koutecký. Silver Cluster-Tryptophan in the Gas Phase: A Model for Absorption Enhancement in Nanoparticle-Biomolecule Hybrid Systems. *submitted*, 2006.

[5] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry*, 11:601, 1984.

[6] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical Chemistry*, 59:2621–2625, 1987.

[7] M. Kinter and N. E. Sherman. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. Wiley Inter Sciences, New York, 2000.

[8] V. H. Wysocki, G. Tsaprailis, L. L. Smith, and L. A. Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35:1399–1406, 2000.

[9] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of American Chemical Society*, 120:3265–3266, 1998.

[10] T. Tabarin, R. Antoine, M. Broyer, and P. Dugourd. Specific photodissociation of peptides with multi-stage mass spectrometry. *Rapid Communications in Mass Spectrometry*, 19:2883–2892, 2005.

[11] G. Grégoire, H. Kang, C. Dedonder-Lardeux, C. Jouvet, C. Desfrançois, D. Onidas, V. Lepere, and J. A. Fayeton. Statistical vs. non-statistical deactivation pathways in the UV photo-fragmentation of protonated tryptophani-leucine dipeptide. *Physical Chemistry Chemical Physics*, 8:122–128, 2006.

[12] R. A. Zubarev, K. F. Haselmann, B. A. Budnik, F. Kjeldsen, and F. Jensen. Towards and understanding of the mechanism of electron-capture dissociation: a historical

perspective and modern ideas. *European Journal of Mass Spectrometry*, 8:337–349, 2002.

[13] F. Turecek. N-C_α Bond Dissociation Energies and Kinetics in Amide and Peptide Radicals. Is the Dissociation a Non-ergodic Process? *Journal of American Chemical Society*, 125:5954–5963, 2003.

[14] E. A. Syrstad, D. D. Stephens, and F. Turecek. Hydrogen Atom Adducts to the Amide Bond. Generation and Energetics of Amide Radicals in the Gas Phase. *Journal of Physical Chemistry A*, 107:115–126, 2003.

[15] R. A. Zubarev, N. A. Kruger, E. K. Fridriksson, M. A. Lewis, D. M. Horn, B. K .Carpenter, and F. W. McLafferty. Electron Capture Dissociation of Gaseous Multiply-Charged Proteins Is Favored at Disulfide Bonds and Other Sites of High Hydrogen Atom Affinity. *Journal of American Chemical Society*, 121:2857–2862, 1999.

[16] R. A. Zubarev, E. K. Friddriksson, D. M. Horn, N. L. Kelleher, N. A. Kruger, B. K. Carpenter, and F. W. McLafferty. Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations. *Analytical Chemistry*, 72:563–573, 2000.

[17] D. M. Horn, K. Breuker, A. J. Frank, and F. W. McLafferty. Kinetic Intermediates in the Folding of Gaseous Protein Ions Characterized by Electron Capture Dissociation Mass Spectrometry. *Journal of American Chemical Society*, 123:9792–9799, 2001.

[18] K. Breuker, H. Oh, D. M. Horn, B. A. Cerda, and F. W. McLafferty. Detailed Unfolding and Folding of Gaseous Ubiquitin Ions Characterized by Electron Capture Dissociation. *Journal of American Chemical Society*, 124:6407–6420, 2002.

[19] H. Oh, K. Breuker, S. K. Sze, Y. Ge, B. K. Carpenter, and F. W. McLafferty. Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 99:15863–15868, 2002.

[20] C. M. Adams, F. Kjeldsen, R. A. Zubarev, B. A. Budnik, and K. F. Haselmann. Electron capture dissociation distinguishes a single D-amino acid in a protein and probes the tertiary structure. *Journal of the American Society for Mass Spectrometry*, 15:1087–1098, 2004.

[21] N. C. Polfer, K. F. Haselmann, P. R. R. Langridge-Smith, and P. E. Barran. Structural investigation of naturally occurring peptides by electron capture dissociation and AMBER force field modelling. *Molecular Physics*, 15:1481–1489, 2005.

[22] H. Kang, C. Jouvet, C. Dedonder-Lardeux, S. Martrenchard, G. Grégoire, C. Desfrançois, J.-P. Schermann, M. Barat, and J. A. Fayeton. Ultrafast deactivation mechanisms of protonated aromatic amino acids following UV excitation. *Physical Chemistry Chemical Physics*, 7:394–398, 2005.

[23] A. L. Sobolewski, W. Domcke, C. Dedonder-Lardeux, and C. Jouvet. Excited-state hydrogen detachment and hydrogen transfer driven by repulsive $^1\pi\sigma^*$ states: A new paradigm for nonradiative decay in aromatic biomolecules. *Physical Chemistry Chemical Physics*, 4:1093–1100, 2002.

[24] E. R. Williams, J. J. P. Furlong, and F. W. McLafferty. Efficiency of Collisionally-activated dissociation and 193-nm photodissociation of peptide ions in fourier transform mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 1:288–294, 1990.

[25] D. C. Barbassi and D. H. Russell. Sequence and side-chain specific photofragment (193 nm) ions from protonated substance P by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 10:1038–1040, 1999.

[26] T.-Y. Kim, M. S. Thomson, and J. P. Reilly. Peptide photodissociation at 157 nm in a linear ion trap mass spectrometer . *Rapid Communications in Mass Spectrometry*, 19:1657–1665, 2005.

[27] F. O. Talbot, T. Tabarin, R. Antoine, M. Broyer, and P. Dugourd. Photodissociation spectroscopy of trapped protonated tryptophan. *Journal of Chemical Physics*, 122:074310, 2005.

[28] S. Arnold and M. Sulkes. Spectroscopy of solvent complexes with indoles: induction of 1La-1Lb state coupling. *Journal of Physical Chemistry*, 96:4768–4778, 1992.

[29] C. Dedonder-Lardeux, C. Jouvet, S. Perun, and A. L. Sobolewski. External electric field effect on the lowest excited states of indole: ab initio and molecular dynamics study. *Physical Chemistry Chemical Physics*, 5:5118–5126, 2003.

[30] J. R. Lakowicz. *Principles of fluorescence spectroscopy.* Kluwer Academic/Plenum, New-York, 2^{nd} edition, 1999.

[31] D. Nolting, C. Marian, and R. Weinkauf. Protonation effect on the electronic spectrum of tryptophan in the gas phase. *Physical Chemistry Chemical Physics*, 6:2633–2640, 2004.

[32] T. R. Rizzo, Y. D. Park, L. A. Peteanu, and D. H. Levy. The electronic spectrum of the amino acid tryptophan in the gas phase. *Journal of Chemical Physics*, 84:2534, 1986.

[33] H. Kang, C. Jouvet, C. Dedonder-Lardeux, S. Martrenchard, C. Charrière, G. Grégoire, C. Desfrançois, J.-P. Schermann, M. Barat, and J. A. Fayeton. Photoinduced processes in protonated tryptamine. *Journal of Chemical Physics*, 122:84307, 2005.

[34] C. Yao and F. Turecek. Hypervalent ammonium radicals. Competitive NC and NH bond dissociations in methyl ammonium and ethyl ammonium. *Physical Chemistry Chemical Physics*, 7:912–920, 2005.

[35] K. F. Haselmann, B. A. Budnik, F. Kjeldsen, M. L. Nielsen, J. V. Olsen, and R. A. Zubarev. Electronic excitation gives informative fragmentation of polypeptide cations and anions. *European Journal of Mass Spectrometry*, 8:117–121, 2002.

[36] H. J. Cooper. Investigation of the Presence of b Ions in Electron Capture Dissociation Mass Spectra. *Journal of the American Society for Mass Spectrometry*, 16:1932–1940, 2005.

[37] F. Yang, J. B. Lum, J. R. McGill, C. M. Moore, S. L. Naylor, P. H. van Bragt, W. D. Baldwin, and B. H. Bowman. Human transferrin: cDNA characterization and chromosomal localization. *Proceedings of the National Academy of Sciences of the United States of America*, 81:2752–2756, 1984.

[38] D. R. Hall, J. M. Hadden, G. A. Leonard, S. Bayley, M. Neu, M. Winn, and P. F. Lindley. The crystal and molecular structures of diferric porcine and rabbit serum transferrins at resolutions of 2.15 and 2.60 A respectively. *Acta Crystallographica, Section D*, 58:70–80, 2002.

[39] R. T. MacGillivray, S. A. Moore, J. Chen, B. F. Anderson, H. Baker, Y. Luo, M. Bewley, C. A. Smith, M. E. Murphy, Y. Wang, A. B. Mason, R. C. Woodworth, G. D. Brayer, and E. N. Baker. Two high-resolution crystal structures of the recombinant N-lobe of human transferrin reveal a structural change implicated in iron release. *Biochemistry*, 37:7919–7928, 1998.

Conclusion et perspectives

On est au milieu indécis d'une sieste éveillée, avec un magazine à parcourir, ou mieux : une vieille bande dessinée qu'on a pas lue depuis longtemps. Le temps s'étire vaguement. Il est deux ou trois heures de l'après-midi, un jour d'août accablant de canicule. On n'a pas même le léger remords de gâcher un infime quelque chose : de toute façon, il fait beaucoup trop chaud pour se promener.

Philippe Delerm, *La sieste assassinée.*

Ce travail de thèse est une étude théorique des propriétés thermodynamiques de polypeptides en phase gazeuse avec comme objectif une meilleure compréhension des mécanismes fondamentaux impliqués dans le repliement des protéines. Face à la complexité de ce problème, tant liée aux systèmes (les protéines) qu'au phénomène lui même (le repliement de ces molécules), nous avons adopté la stratégie suivante. Nous avons, d'une part, étudié de petits polypeptides modèles (d'une dizaine d'acides d'aminés) modélisés par le champ de force AMBER 96 et nous avons, d'autre part, utilisé des algorithmes sophistiqués capables d'explorer largement une surface d'énergie potentielle tourmentée. Deux méthodes, dites dans les ensembles généralisés, ont été particulièrement employées. Le Monte Carlo d'échange repose sur la simulation Monte Carlo de M répliques à M températures avec un échange de répliques tenté périodiquement. La méthode Wang-Landau vise à estimer avec précision la densité d'états microcanonique en réalisant une marche aléatoire sur le paysage énergétique et en pénalisant les états au fur et à mesure qu'ils sont visités. Cette dernière méthode a été développée et adaptée à la simulation de systèmes à degrés de liberté continus comme les biomolécules.

Ce projet a été réalisé en étroite interaction avec les avancées expérimentales du groupe. Dans un premier temps, nous avons essayé de comprendre les facteurs qui stabilisent les éléments de structure secondaire comme les hélices α et les feuillets β, ceux-ci intervenant très tôt dans le mécanisme de repliement d'une protéine dans sa structure biologiquement active. La compétition entre ces deux structures est d'autant plus critique qu'elle intervient dans les maladies conformationnelles, comme la maladie d'Alzheimer et plus généralement les maladies à prion. Nous avons ainsi étudié une famille de polyalanines, de Ala_8 à Ala_{20}, et mis en évidence qu'à basse température, les structures en hélice α sont stabilisées car

elles présentent le minimum d'énergie potentielle. À température intermédiaire, les géométries de type feuillet β sont favorisées entropiquement car elles sont plus flexibles. Enfin, à température élevée, des structures désordonnées, plutôt étirées, sont majoritairement observées. Ces résultats sont en parfait accord avec les mesures expérimentales de dipôle électrique effectuées dans l'équipe. Nous avons également repris un travail précédent sur l'influence d'un champ électrique statique et intense sur le dipeptide WG pour étudier le comportement des polyalanines dans un tel champ. Nous montrons qu'il est alors possible d'étendre le domaine de stabilité d'une hélice α au détriment de celui du feuillet β. Ces études en fonction de la température, de la taille des polyalanines et du champ électrique ont pour objectif de produire à terme un diagramme de stabilité des différentes structures secondaires en compétition. Cette étude sur les polyalanines a impliqué des simulations de systèmes de grande taille (jusqu'à 200 atomes) avec un coût numérique important. Nous avons rencontré des problèmes de convergence, notamment pour le Monte Carlo d'échange qui nécessite une statistique élevée et le suivi des échanges entre répliques. En ce qui concerne la méthode Wang-Landau, le temps tunnel n'est pas un critère de convergence évident. Nous avons par contre montré que l'utilisation d'un deuxième paramètre d'ordre, en plus de l'énergie, est particulièrement efficace pour décrire les propriétés considérées à basse température. Pour les systèmes étudiés, le dipôle électrique est un paramètre d'ordre naturel. Une étude avec d'autres coordonnées de réaction comme le rayon de giration ou la distance bout-à-bout serait pertinente pour analyser l'influence du paramètre d'ordre sur les résultats obtenus. Par ailleurs, la stabilisation entropique des feuillets β est en contradiction avec d'autres études théoriques et le choix du champ de force peut être discuté. Nous projetons alors de poursuivre cette étude à des champs de force différents (notamment ECEPP/2). Il semble également que le choix des structures initiales dans une simulation en coordonnées internes soit particulièrement critique. Pour autant, une simulation en coordonnées cartésiennes uniquement a de fortes chances de ne pas converger. Une alternative pourrait alors être d'effectuer des simulations à la fois en coordonnées internes et en coordonnées cartésiennes. L'introduction des angles de flexion comme coordonnées internes pourraient également être une approche prometteuse. Ensuite, nous envisageons de prendre en compte la polarisabilité des atomes car elle ne peut pas être négligée par exemple, dans des structures en hélice α qui présentent un macrodipôle ou lorsque la molécule est sous l'influence d'un fort champ électrique. Nous pensons pour cela recalculer les charges atomiques au cours de la simulation, l'introduction d'un champ de force polarisable nécessite cependant un paramétrage fin. Un dernier travail consiste aussi à essayer de calculer les propriétés thermodynamiques à champ électrique non nul à partir de celles obtenues pour des simulations Monte Carlo d'échange à champ nul. D'un point de vue expérimental enfin, une étude systématique des polyalanines en fonction de la taille et de la température est sur le point d'être réalisée.

Dans un deuxième temps nous nous sommes intéressés à la photofragmentation de biomolécules en phase gazeuse pour essayer d'établir un lien entre la distribution des canaux de fragmentation et la structure secondaire du peptide étudié. Cette relation n'est cependant pas évidente puisque la fragmentation dépend de la dynamique des états excités que nous n'avons pas calculé. Par contre, si cette dynamique est suffisamment rapide, comme c'est le cas après une excitation laser, la structure initiale joue un rôle important dans la fragmentation. Nous avons ainsi réalisé des calculs statistiques à l'état fondamental sur un pentapeptide simplement et doublement protoné. Les résultats théoriques montrent qu'il existe un couplage entre l'état excité $\pi\pi^*$ et l'état dissociatif $\pi\sigma^*$ qui explique la fragmentation observée pour la dissociation induite par laser. Ce couplage se traduit par un transfert d'électron entre l'indole et la charge. Il ne peut se produire que si l'indole est suffisamment proche de la charge et donc que si le peptide est dans une conformation bien particulière. Ce modèle reste cependant insuffisant pour expliquer les résultats obtenus pour la famille de polyvalines, la localisation de la charge pouvant être en partie responsable de ce désaccord. En effet, il est tout à fait envisageable que le proton, initialement sur le N-ter saute d'un groupe carbonyle à un autre, le long du squelette peptidique. Finalement, une approche statistique sur un grand nombre de peptides en utilisant les structures cristallines peut aussi qualitativement aider à comprendre la relation qui existe entre structure et photofragmentation. Cette démarche mérite d'être poursuivie avant de pouvoir initier une conclusion, mais les résultats déjà obtenus sont très encourageants.

Pour conclure, notre objectif a été d'une part, d'étudier la compétition entre les différentes structures secondaires d'un peptide ainsi que le rôle joué par l'entropie, et d'autre part, de comprendre l'influence de la charge et de la structure d'un peptide sur sa photofragmentation. Ce travail qui dépasse le cadre de cette thèse nécessite une très forte interaction théorie–expérience.

Annexe A

Bestiaire des systèmes étudiés

Les principales caractéristiques des différents systèmes étudiés au cours de ce travail de thèse sont résumées dans le tableau ci-dessous :

Molécule	Notation	Nombre d'atomes	Masse [uma]	Charge	Formule topologique (figure)
TrpGly	WG	34	261,28	0	A.1
$[AlaGlyTrpLeuLys + H]^+$	$AGWLK^+$	85	574,70	+1	A.2
$[AlaGlyTrpLeuLys + 2H]^{2+}$	A^+GWLK^+	86	575,71	+2	A.3
$[TrpValValValVal + H]^+$	W^+VVVV	92	601,77	+1	A.4
$[ValTrpValValVal + H]^+$	V^+WVVV	92	601,77	+1	A.5
$[ValValTrpValVal + H]^+$	V^+VWVV	92	601,77	+1	A.6
$[ValValValTrpVal + H]^+$	V^+VVWV	92	601,77	+1	A.7
$[ValValValValTrp + H]^+$	V^+VVVW	92	601,77	+1	A.8
Ala_8	Ala_8	83	586,65	0	A.9
Ala_{12}	Ala_{12}	123	870,96	0	A.10
Ala_{16}	Ala_{16}	163	1155,28	0	A.11
Ala_{20}	Ala_{20}	203	1439,59	0	A.12

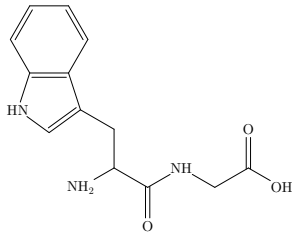


FIG. A.1 – *Formule topologique du peptide WG.*

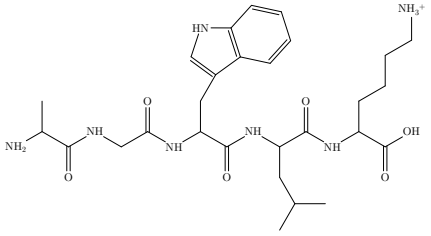


FIG. A.2 – *Formule topologique du peptide $AGWLK^+$.*

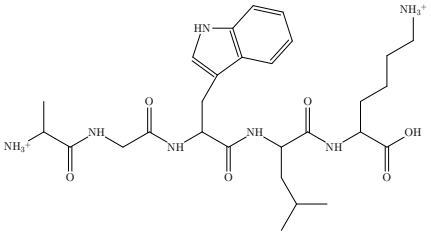


FIG. A.3 – *Formule topologique du peptide A^+GWLK^+.*

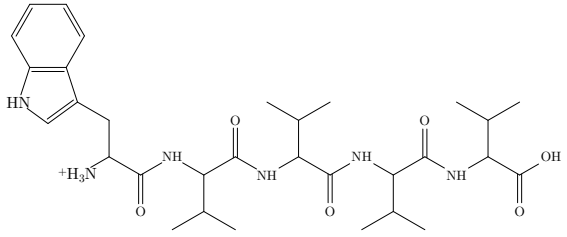


FIG. A.4 – *Formule topologique du peptide W^+VVVV.*

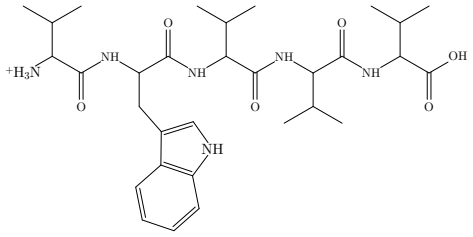


FIG. A.5 – *Formule topologique du peptide V^+WVVV.*

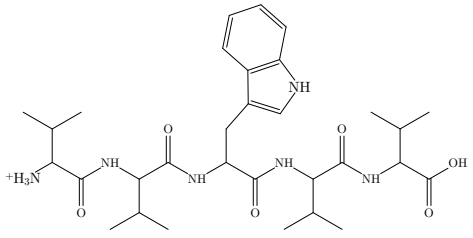


FIG. A.6 – *Formule topologique du peptide V^+VWVV.*

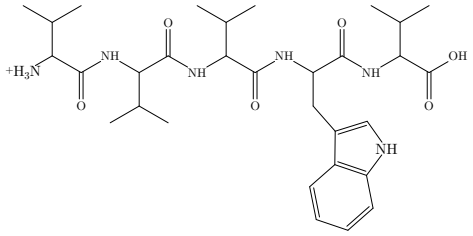


FIG. A.7 – *Formule topologique du peptide $V^{+}VVWV$.*

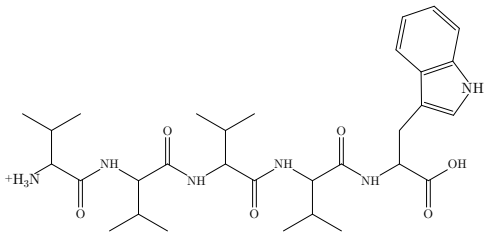


FIG. A.8 – *Formule topologique du peptide $V^{+}VVVW$.*

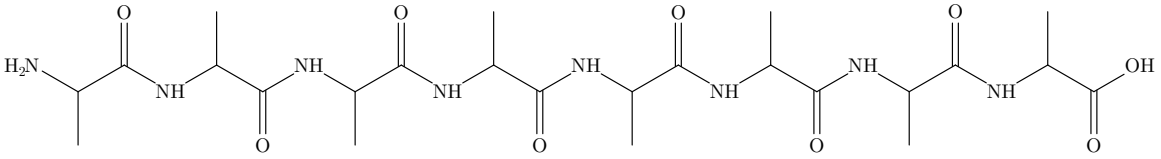


FIG. A.9 – *Formule topologique du peptide Ala_8.*

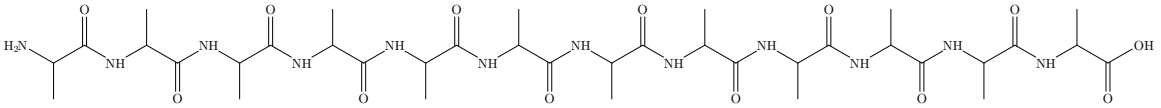


FIG. A.10 – *Formule topologique du peptide Ala_{12}.*

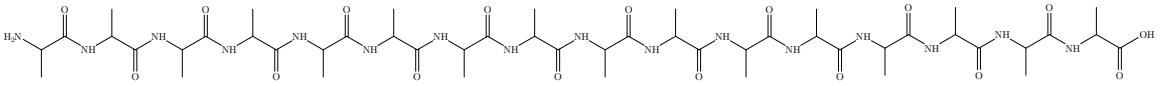


FIG. A.11 – *Formule topologique du peptide Ala_{16}.*

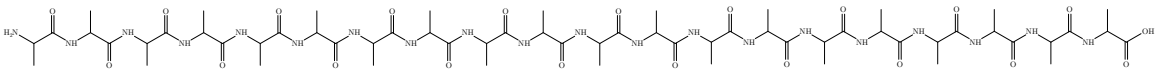


FIG. A.12 – *Formule topologique du peptide Ala_{20}.*

Annexe B

Moyens de calcul utilisés

Les simulations réalisées pour ce travail de thèse ont nécessité d'importants moyens de calcul, tant parallèles pour le Monte Carlo d'échange, qu'en série pour l'algorithme Wang-Landau. Les ressources en temps de calcul ont été obtenues localement ainsi que dans les deux centres nationaux :

- Laboratoire de Spectrométrie Ionique et Moléculaire (LASIM), Lyon, `http://lasim.univ-lyon1.fr/` ;
- Pôle Scientifique de Modélisation Numérique (PSMN), Lyon, `http://www.psmn.ens-lyon.fr/` ;
- Institut du Développement et des Ressources en Informatique Scientifique (IDRIS), Orsay, `http://www.idris.fr/` ;
- Centre Informatique National de l'Enseignement Supérieur (CINES), Montpellier, `http://www.cines.fr/`.

Le tableau B.1 recense les ressources et les machines utilisées.

Tab. B.1 – *Ressources en temps de calcul et machines utilisées.*

Institutions	Machines	Temps de calcul
LASIM	PC Dell 1 processeur Intel Pentium 2,6 GHz	« illimité »
LASIM	PC Dell 2 processeurs Intel Xeon 2,0 GHz	« illimité »
LASIM	cluster de PC Intel 16 processeurs – 8 × 2 processeurs Xeon 3,0 GHz	« illimité »
PSMN	cluster AMD 64 processeurs – 32 × 2 processeurs Opteron 2,6 Ghz	« illimité »
PSMN	cluster AMD 32 processeurs – 8 × 4 processeurs Opteron 2,6 Ghz	
IDRIS	cluster IBM Regatta Power4 1024 processeurs – 8 × 32 processeurs Power4 1,3 GHz – 12 × 32 processeurs Power4 1,3 GHz – 6 × 16 × 4 processeurs Power4 1,3 GHz	20 000 h
CINES	cluster IBM P1600 Power4 288 processeurs – 2 × 32 processeurs Power4 1,3 GHz – 7 × 32 processeurs Power4+ 1,7 GHz	10 000 h
CINES	cluster de PC AMD 32 processeurs – 16 × 2 processeurs Opteron 1,8 GHz	5 000 h

RÉSUMÉ en français :

Ce travail de thèse est une étude théorique des propriétés thermodynamiques de polypeptides en phase gazeuse avec comme objectif une meilleure compréhension des mécanismes fondamentaux impliqués dans le repliement des protéines. Une approche statistique basée sur des algorithmes Monte Carlo dans les ensembles généralisés, comme le Monte Carlo d'échange ou la méthode Wang-Landau, a été utilisée pour échantillonner le paysage énergétique complexe de ces systèmes. Les peptides étudiés comprenant de 2 à 20 acides aminés ont été modélisés par le champ de force AMBER 96. Les simulations ont été réalisées en étroite interaction avec les avancées expérimentales du groupe. Nous avons ainsi tenté de comprendre l'influence de la structure secondaire sur les mécanismes de photofragmentation, le rôle de l'entropie dans la stabilisation des feuillets beta à température ambiante et l'effet d'un champ électrique intense sur la conformation de peptides.

TITRE en anglais :

Structure and dynamics of isolated proteins : statistical approaches

RÉSUMÉ en anglais :

We present a theoretical study of the thermodynamical properties of polypeptides in the gas phase. The aim of this work is a better understanding of the fundamental mechanisms involved in protein folding. A statistical approach based on Monte Carlo algorithms applied in generalised ensembles, such as Replica Exchange Method or Wang-Landau method, has been used to sample the rugged energy landscape of those molecules. The peptides were composed of 2 to 20 amino acids and have been modelised by the AMBER 96 forcefield. Simulations have been realised in strong relation with experimental progress of the group. We thus tried to understand the influence of the secondary structure on photofragmentation mechanism, the role of entropy in the stabilisation of beta sheets and the effect of intense electric field on peptide conformation.

DISCIPLINE :

Physique

MOTS-CLÉS :

protéine, peptide, dipôle électrique, hélice alpha, feuillet beta, entropie, photofragmentation, Monte Carlo, Monte Carlo d'échange, Wang-Landau

INTITULÉ ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de spectrométrie ionique et moléculaire,
Unité Mixte de Recherche (UMR 5579) CNRS/UCB Lyon I,
Domaine scientifique de la Doua - Université Claude Bernard Lyon I,
Bâtiment Alfred Kastler,
43, bd du 11 Novembre 1918,
69622 Villeurbanne,
France