HAL
open science

Physique statistique du repliement et de la dénaturation des acides nucléiques

Daniel Jost

To cite this version:

Daniel Jost. Physique statistique du repliement et de la dénaturation des acides nucléiques. Biophysique [physics.bio-ph]. Ecole normale supérieure de lyon - ENS LYON, 2010. Français. NNT : . tel-00544804

HAL Id: tel-00544804

https://theses.hal.science/tel-00544804v1

Submitted on 9 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 571
N° attribué par la bibliothèque : __ENSL571

THÈSE

en vue d'obtenir le grade de

Docteur de l'Université de Lyon - École Normale Supérieure de Lyon

spécialité : Physique

LABORATOIRE DE PHYSIQUE

École Doctorale de Physique et Astrophysique De Lyon

présentée et soutenue publiquement le 23/06/2010

par Monsieur Daniel JOST

Physique statistique du repliement et de la dénaturation des acides nucléiques

---

Directeur de thèse : Monsieur Ralf EVERAERS

Après avis de : Monsieur Ralf BLOSSEY
Monsieur Anthony MAGGS

Devant la commission d'examen formée de :
Monsieur Ralf BLOSSEY, Membre/Rapporteur
Monsieur Ralf EVERAERS, Membre
Monsieur Anthony MAGGS, Membre/Rapporteur
Monsieur Henri ORLAND, Membre
Monsieur Michel PEYRARD, Membre

Résumé

L'étude de nombreux processus biologiques et nanotechnologiques requièrent une bonne compréhension du repliement et de la dénaturation des acides nucléiques. Les travaux décrits dans cette thèse portent principalement sur le développement et l'utilisation de modèles thermodynamiques de ces mécanismes.

Nous avons tout d'abord mis en place un formalisme unifié du modèle de Poland-Scheraga qui permet de décrire la dénaturation thermique de l'ADN quelque soit la taille des molécules considérées, leur concentration et leur environnement ionique. Nous utilisons ce modèle pour décrire quelques aspects génériques de la dénaturation. En particulier, nous montrons que le comportement des observables est particulièrement sensible à l'incertitude sur les paramètres du modèle pour les longs oligomères. Nous considérons ensuite le modèle de Zimm-Bragg qui est une approximation du modèle précédent. Cela nous permet de procéder à une analyse statistique systématique des corrélations entre domaines thermodynamiquement stables et gènes dans les génomes.

Nous avons ensuite développé un modèle sur réseau du repliement de l'ARN paramétré à l'aide d'une version réduite et unifiée du modèle de Turner. L'étude du modèle sur réseau, grâce à la mise en place de plusieurs techniques avancées de Monte-Carlo, montre qu'il décrit quantitativement le repliement de structures complexes. Nous évaluons aussi l'importance des interactions stériques. En particulier, nous estimons des corrections de champ moyen utilisables dans les programmes standard traitant la structure secondaire. Enfin, nous exploitons l'aspect tridimensionnelle du modèle, pour étudier l'effet d'un confinement géométrique.

Abstract

A quantitative understanding of the folding and opening of nucleic acids is relevant for many biological and nanotechnological processes. The main goal of my PhD thesis is to understand these mechanisms by using and developping thermodynamics-based models.

In a first part, we develop a unified Poland-Scheraga model of DNA thermal denaturation. Our description covers the entire crossover from oligo- to polynucleotide melting behavior and is applicable in the full experimental range of DNA strand and salt concentrations. We use our model to discuss generic aspects of DNA melting. In particular, we emphasize that the observable features are particularly sensitive to the remaining parameterization uncertainty for long oligomers. Then, we reconsider the Zimm-Bragg model, an approximation of the previous model. This allows us to perform a statistical and systematic genome wide comparison between biologically coding domains and thermodynamically stable regions.

In a second part, we investigate the RNA folding by developping a lattice model parameterized with a unified and reduced version of the Turner model. The use of advanced Monte-Carlo techniques allows us to show that the lattice model quantitatively describe the folding of complex structures. We also evaluate the impact of steric interactions. In particular, we estimate mean-field corrections that can be used in standard programs working at the secondary structure level. Finally, the tridimensional definition of the model is exploited to study the effect of a geometrical confinement.

Aux femmes de ma vie, Chloé et Sophie,

Remerciements

Je tiens tout d'abord à remercier chaleureusement Ralf, mon directeur de thèse, pour sa gentillesse et sa disponibilité durant ces trois années. Il a su me guider tout au long de ma thèse, me posant les bonnes questions pour avancer, tout en me laissant prendre un grand nombre d'initiatives. Son exigence et son esprit critique m'ont permis de mener à bien mes travaux dans un cadre scientifique rigoureux et motivant. Travailler avec Ralf a été un privilège et un réel plaisir.

J'exprime ma gratitude envers Henri Orland, Ralf Blossey, Anthony Maggs et Michel Peyrard pour avoir accepté d'être membres de mon jury de thèse.

Je suis également très reconnaissant aux membres du laboratoire pour leur accueil et leur sympathie. En particulier, un grand merci à Thierry, Eric, Benjamin, Cendrine, Martin, Johannes et Nils pour leurs conseils avisés et les discussions fructueuses sur mes travaux ou sur la recherche en général. Je n'oublie pas non plus les secrétaires, Nadine, Laurence et Laure, ainsi que les responsables informatiques du Centre Blaise Pascal, Cerasela et Emmanuel, pour leur disponibilité et leur efficacité à résoudre ou devancer les nombreux problèmes logistiques d'un thésard.
Ma participation à l'équipe de foot du labo, les Hell's Angels, m'a permis de m'oxygéner chaque semaine et de faire mon cota de sport. Merci à toute l'équipe pour m'avoir supporter en tant que capitaine durant ces deux dernières années. Mon grand regret restera de n'avoir jamais gagné le championnat interne de l'ENS !!!

Plus personnellement, je voudrais remercier, pour leur amitié top niveau et les bons moments musicaux ou festifs passés ensemble, Las Meulas (Grand Gui et Thieu) et Adri, mais aussi les "infoensiens" (Joubs, Kiki, Gautier et Solène, JC, Xav et Cat, Tom et Sophie), les "Marseillais" (Eva et Seb, Henri et Vaness, Oliv et Claire, Cédric, God), les "TPE" (Pauline et Anatole, Philippe et Marion, Ben et Domi, Mathias, François et Katell, Julie) et les autres (Laure, Laurent, Mario, Luc, etc.).
Je suis également reconnaissant envers mes parents et mes frères et soeurs (ainsi que leur époux/épouse et enfants) pour leur soutien et leur encouragements durant ces 3 années.

Je finirai par remercier les deux personnes qui comptent le plus pour moi, ma femme Sophie, pour son amour, son soutien et sa patience (en particulier vis-à-vis de toutes les vaiselles que je n'ai pas faites !), et ma petite fille Chloé qui a illuminé les sept derniers mois de ma thèse : mon unique expérience sur l'ADN aura été un vrai succès !!

Table des matières

Introduction générale **1**

I Dénaturation de l'ADN **9**

Introduction **11**

1 Modèle de Poland-Scheraga unifié **15**
1.1 Définition du modèle 15
1.1.1 Équilibre d'association 15
1.1.2 Interactions dans le modèle de Poland-Scheraga 17
1.1.3 Résolution du modèle de Poland-Scheraga 18
1.2 Paramétrisation et erreurs 20
1.2.1 Paramètres d'association 20
1.2.2 Correction dûe au sel 22
1.2.3 Coopérativité 25
1.2.4 Propagation des erreurs 25
1.3 Comportement générique du modèle 26
1.3.1 Évolution des courbes de dénaturation 26
1.3.2 Dénaturation interne 29
1.4 Pouvoir de prédiction du modèle 31
1.4.1 Oligomères courts ($\sim$ 10 bp) 31
1.4.2 Polymères longs ($\geq$ 10 kbp) 33
1.4.3 Polymères courts (100 bp $\leq$ ... $\leq$ 10 kbp) 34
1.4.4 Oligomères longs (20 bp $\leq$ ... $\leq$ 100 bp) 36
1.5 Conclusion 37
1.5.1 Bilan 37
1.5.2 Influence de chaque paramètre 37
1.5.3 Améliorer la paramétrisation du modèle 38
1.5.4 Paramétrer σ et ω 39

2 Analyse thermodynamique des génomes **41**
2.1 Modèle de Zimm-Bragg 41
2.1.1 Une approximation du modèle PS 41
2.1.2 Méthode des matrices de transfert 43
2.1.3 Complexité et temps de calcul 44
2.2 Validation 45
2.2.1 Comparaison avec le modèle PS 45
2.2.2 Applications à la génomique 46
2.3 Application à l'analyse des génomes 49

2.3.1 Corrélations au niveau des paires de bases . . . 49
2.3.2 Corrélations au niveau des domaines . . . 52
2.3.3 Identification de gènes et exons . . . 55
2.4 Conclusion . . . 59
2.4.1 Bilan . . . 59
2.4.2 Perspective : inclure les effets de superhélicité . . . 60

II Repliement de l'ARN 63

Introduction 65

3 Modèle sur réseau 69
3.1 Définition du modèle . . . 69
3.1.1 Interactions dans le modèle de Turner et dans le modèle sur réseau . . . 69
3.1.2 Entropies des structures secondaires dans le modèle sur réseau . . . 72
3.2 Paramètres dans le modèle de Turner . . . 73
3.2.1 Paramétrisation du modèle de Turner . . . 73
3.2.2 Unification des paramètres de fourches et de nucléations de boucle . . . 73
3.2.3 Différence entre petites et grandes fourches . . . 76
3.2.4 Unification des paramètres d'empilement coaxial et de défaut d'appariement . . . 76
3.2.5 Liberté de jauge pour les termes de bords et d'initiation dans le modèle de Turner 77
3.2.6 Sensibilité des résultats . . . 78
3.3 Paramétrisation du modèle sur réseau . . . 82
3.3.1 Influence du réseau . . . 82
3.3.2 Paramétrisation naïve . . . 83
3.3.3 Paramétrisation correcte . . . 83
3.3.4 Énergie de pliage . . . 85
3.3.5 Énergie libre totale d'une conformation sur réseau . . . 87

4 Méthodes numériques 89
4.1 Modélisation *in silico* . . . 89
4.1.1 Représentation du brin d'ARN . . . 89
4.1.2 Occupation du réseau . . . 90
4.2 Notions sur les simulations de Monte-Carlo . . . 91
4.2.1 Définitions . . . 91
4.2.2 Évaluation pratique des valeurs moyennes . . . 91
4.2.3 Choix de l'algorithme . . . 92
4.3 Schémas dynamiques . . . 93
4.3.1 Règles d'acceptance . . . 93
4.3.2 Mouvements élémentaires . . . 94
4.3.3 Méthode multi-histogramme . . . 103
4.4 Schémas statiques . . . 105
4.4.1 Algorithmes de croissance de chaînes . . . 105
4.4.2 Échantillonnage d'une structure secondaire donnée . . . 115

5 Résultats 121
5.1 Validation du modèle pour des structures simples . . . 121
5.1.1 Dénaturation de petites structures en épingle . . . 121
5.1.2 Énergies libres de boucles internes et en épingle . . . 123
5.2 Prédiction de structures complexes . . . 123

5.2.1 Structures avec boucles multiples . . . 123
5.2.2 Structures avec pseudo-noeuds . . . 125
5.3 Impact du volume exclu . . . 130
5.3.1 Le modèle des brins fantômes . . . 131
5.3.2 Interactions stériques entre deux boucles . . . 133
5.3.3 Evaluation de relations de Jacobson-Stockmayer . . . 134
5.3.4 Impact sur les propriétés thermodynamiques et structurelles du repliement . . . 138
5.4 Interactions tertiaires . . . 144
5.4.1 Interactions spécifiques de contact . . . 144
5.4.2 Interactions stériques avec l'extérieur . . . 146
5.5 Conclusion . . . 149
5.5.1 Bilan . . . 149
5.5.2 Perspectives . . . 151

Conclusion générale **155**

6 Annexes **157**
6.1 Algorithme de Fixman-Freire pour le modèle de Poland-Scheraga . . . 157
6.2 Exploitation des courbes expérimentales de dénaturation obtenues par absorbance d'UV 159
6.3 Pertinence d'une modélisation . . . 160
6.4 Distributions de probabilité dans le cadre de l'analyse thermodynamique des génomes . 160
6.5 Dénaturation de l'ADN surenroulé . . . 162
6.6 Descriptions de quelques approches basées sur le modèle de Turner . . . 163
6.7 Preuve par récurrence d'une propriété topologique des structures secondaires . . . 167
6.8 Quelques notions sur les polymères . . . 168
6.9 Statistique dans les simulations de Monte-Carlo . . . 170
6.10 Lien entre probabilités et poids de Rosenbluth dans la dimérisation avec biais . . . 172
6.11 Cinétique du repliement de l'ARN . . . 173

Bibliographie **175**

Remarque : Le CD accompagnant la version papier du manuscrit contient une version électronique PDF du manuscrit, toutes les figures au format EPS, les articles déjà publiés et des fichiers annexes.

Introduction générale

L'étude des acides nucléiques mobilise une importante partie de la communauté scientifique afin de comprendre leurs rôles prépondérants dans le fonctionnement des organismes vivants et de développer de nouvelles applications nanotechnologiques. L'intérêt pour ces molécules n'a eu de cesse de s'intensifier depuis la résolution de la structure du plus connu des acides nucléiques, l'ADN, par Watson et Crick en 1953 [1]. La découverte continuelle de nouveaux mécanismes biologiques où les acides nucléiques interviennent fait que leurs champs d'études se renouvellent constamment. Surtout depuis une quinzaine d'années où l'apparition de nouvelles techniques expérimentales facilitant l'investigation et le séquençage des acides nucléiques a permis de collecter un nombre gigantesque de données. Le besoin de trouver des modèles quantitatifs qui décrivent, expliquent ou prédisent ces données fait de l'étude des acides nucléiques un terrain de jeu fantastique pour théoriciens et expérimentateurs de tous bords (biologie, physique, chimie, mathématique, informatique, ingénierie, etc.).

Composition et structure des acides nucléiques

Les acides nucléiques sont des biomolécules abondamment présentes dans les organismes vivants. Il en existe deux types : l'acide désoxyribonucléique ou ADN et l'acide ribonucléique ou ARN. Chaque brin d'acide nucléique est une chaîne polymérique dont l'unité de base est le nucléotide constitué d'un phosphate et d'un sucre (désoxyribose pour l'ADN et ribose pour l'ARN) auquel est attachée une base azotée (voir figure 1). Les nucléotides sont reliés entre eux par des liaisons phosphodiester entre un phosphate et un sucre formant ainsi le squelette du brin. Ces liaisons sont orientées chimiquement : le phosphate se lie au carbone 3 du sucre d'un nucléotide et au carbone 5 du sucre suivant, imposant ainsi une orientation intrinsèque dite $5' - 3'$ aux brins d'acides nucléiques. Il existe 5 principales bases : l'adénine (A), la guanine (G), la thymine (T), la cytosine (C) et l'uracile (U), que l'on peut classer en deux catégories : les purines (A et G) et les pyrimidines (T, C et U). L'ADN est constitué par les nucléotides A, G, T et C alors que dans l'ARN, la thymine est remplacée par l'uracile. Les purines peuvent créer des liaisons hydrogènes avec des pyrimidines complémentaires formant ainsi une paire de bases, on dit alors que les deux bases sont appariées (voir figure 1). On distingue principalement 4 paires canoniques : les 3 paires usuelles de Watson-Crick $A - T$ (2 liaisons hydrogènes), $G - C$ (3) et $A - U$ (2) qui sont stables uniquement si les deux nucléotides sont orientés de manière anti-parallèle ; et la paire dite bencale $G - U$ (2) également anti-parallèle. Les nuages électroniques des anneaux aromatiques des bases azotées adjacentes peuvent également interagir entre eux formant des liaisons de Van der Waals dites interactions d'empilement.

Les paires de bases consécutives, empilées les unes sur les autres, constituent alors une structure en double-hélice stable. Le segment de paires de bases (2 paires adjacentes) est la brique élémentaire de cette structure. L'ADN adopte typiquement une forme de type B (voir figure 2) : le plan des paires est quasiment perpendiculaire à l'axe de l'hélice, on compte 10 paires par tour d'hélice (pas de 3.4 nm) et le diamètre de l'hélice est d'environ 20 Å. Les deux espaces entre les brins de la double-hélice sont appelés sillon mineur et sillon majeur à cause de la différence de largeur entre les deux. La double-hélice d'ARN est quant à elle de type A (voir figure 2) : le plan des paires de bases forme un angle de 75 degrés avec l'axe de l'hélice qui ne passe plus par le centre des paires mais à l'intérieur du sillon majeur.

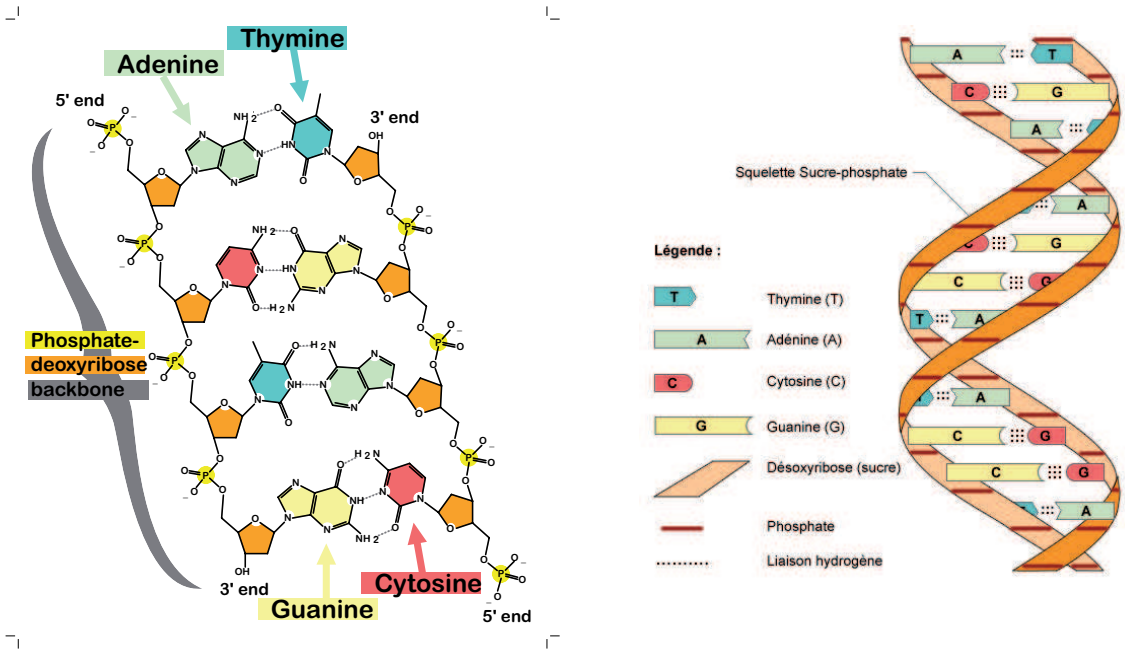

FIGURE 1 – Composition d'un nucléotide et interactions entre bases azotées [2].

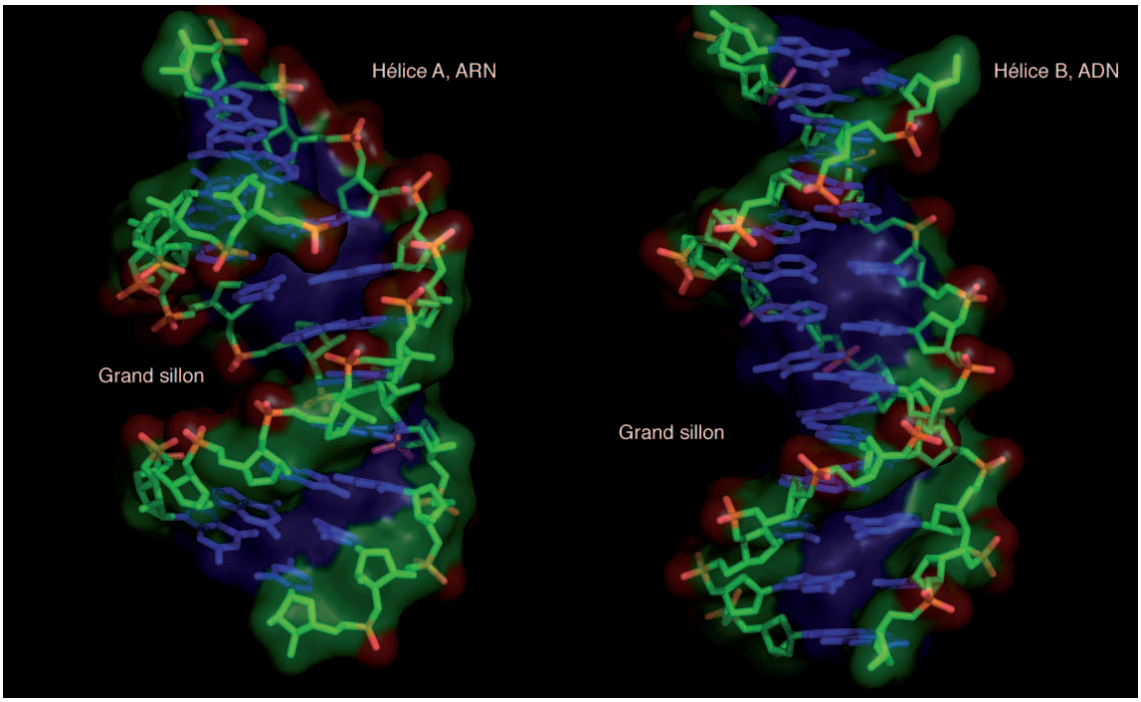

FIGURE 2 – Structure de la double-hélice de type A de l'ARN et de type B de l'ADN [2].

Cela induit une augmentation du diamètre de l'hélice (26 Å), avec 11 paires par tour d'hélice (pas de 2.8 nm).

Rôles biologiques et applications nanotechnologiques des acides nucléiques

In vivo, l'ADN se présente majoritairement sous la forme d'un complexe en double-hélice contenant deux brins d'ADN alors que l'ARN est principalement rencontré sous la forme d'un simple brin replié sur lui-même. De plus, les brins d'ARN sont en général beaucoup plus court ($\sim 10 - 10^3$ nucléotides) que ceux d'ADN ($\sim 10^6 - 10^8$ nts). Les deux jouent un rôle fondamental dans le fonctionnement des cellules.

La plupart de l'ADN cellulaire se trouve sous la forme de chromosomes composés de millions de paires de bases. Leurs séquences codent l'information génétique propre à chaque organisme. La simplicité de sa composition et la stabilité de sa structure en double-hélice fait de l'ADN un support robuste idéal pour la transmission de cette information : la transcription de la séquence est la première étape de la chaîne de production des protéines, et sa réplication est essentielle pour la conservation du patrimoine génétique lors de la division cellulaire [3].

Les rôles de l'ARN sont plus diversifiés [4]. Il intervient dans la production des protéines par l'intermédiaire de l'ARN messager issu de la transcription qui est traduit au niveau des ribosomes en protéines par l'ARN de transfert. Les molécules d'ARN ont aussi une activité enzymatique importante dans la cellule où elles facilitent la catalyse de certaines réactions. On peut citer également les micros ARN [5] qui empêchent la traduction de certains ARN messagers et régulent ainsi la production des protéines correspondantes.

Des éléments laissent penser que, d'un point de vue évolutif, l'apparition de l'ARN précéderait celle de l'ADN comme support de l'information génétique. Cette hypothèse expliquerait la plus grande variété des fonctions de l'ARN. L'ADN aurait supplanté par la suite l'ARN pour le stockage à long terme du fait sa plus grande stabilité.

Récemment, l'engouement pour les nanotechnologies ne cesse de s'intensifier et le monde des acides nucléiques n'est pas épargné, en particulier pour l'ADN. La complémentarité élémentaire requise pour lier deux bases et la solidité des paires de Watson-Crick ($-2/-3k_BT$) et de la structure en double-hélice en font un matériau robuste simple à manipuler et à contrôler.

On peut citer en particulier le travail fondateur de Seeman [9], de Rothemund [7] ou de Winfree [10] sur l'auto-assemblage de brins d'ADN. Grâce à un design précis des séquences, ils arrivent à construire des formes arbitraires (voir figures 3 A et B) comme par exemple des réseaux 2D ou 3D d'ADN qui permettraient de cristalliser plus facilement des protéines.

Une autre application remarquable est la puce à ADN (voir figure 3 C) qui consiste en un ensemble de molécules simple-brins d'ADN dont on contrôle la séquence et que l'on place sur une petite surface. La puce contient ainsi des milliers de sondes capables de détecter leur ADN complémentaire permettant d'accomplir plusieurs tests en parallèle. Cette technique est couramment utilisée en médecine pour des tests génétiques ou pour séquencer rapidement les génomes [11]. Elle est également prometteuse dans l'optique de créer des ordinateurs moléculaires à faible consommation [8] (voir figure 3 D).

Il existe bien d'autres applications notables comme par exemple l'assemblage guidé par ADN de nanoparticules pour créer de nouveaux matériaux aux propriétés photoniques intéressantes [12, 13], ou encore la réalisation de sondes moléculaires capables d'identifier précisément des mutations génétiques [14].

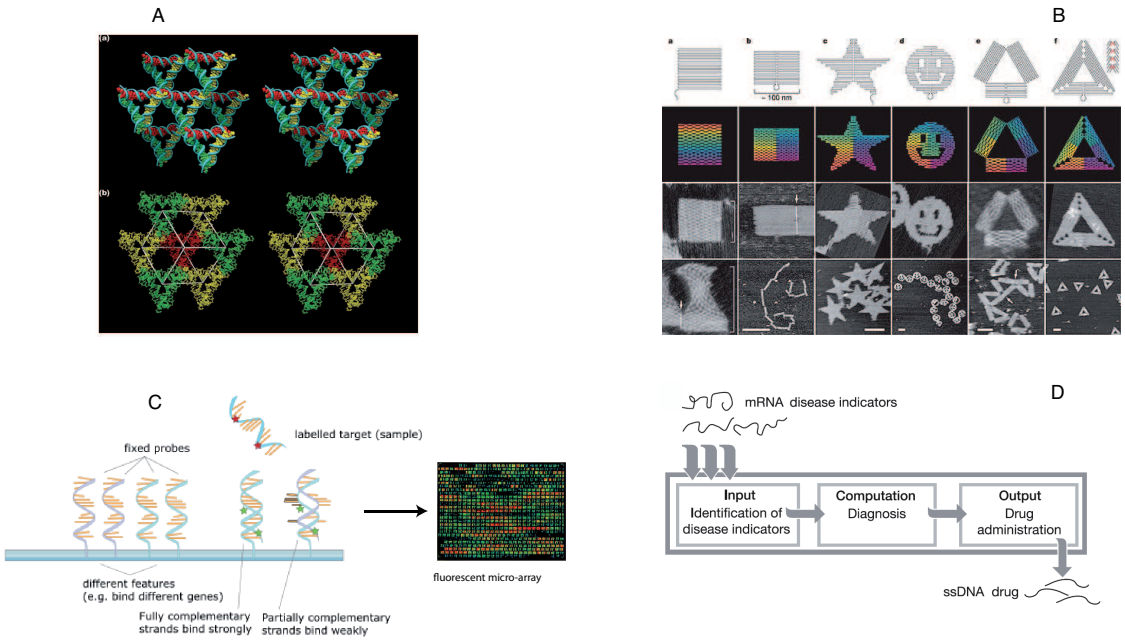

FIGURE 3 – (A) Réseaux 3D formés d'ADN [6]. (B) Différentes formes 2D d'origami d'ADN [7]. (C) Principe de la détection de fragments d'ADN par des puces à ADN [2]. (D) Principe de l'ordinateur moléculaire à ADN [8].

Étude de la dénaturation et du repliement des acides nucléiques

Une compréhension quantitative de l'appariement ou de l'ouverture des paires de bases dans les structures en double-hélice ainsi que des principes régissant la formation et le repliement des molécules d'acides nucléiques, est utile et nécessaire pour l'étude de beaucoup de processus biologiques fondamentaux comme la transcription, la réplication ou l'action enzymatique ainsi que pour le contrôle de la plupart des applications nanotechnologiques décrites ci-dessus.

L'étude des acides nucléiques se base généralement sur 4 niveaux structuraux.

1. La structure primaire spécifie la séquence des nucléotides constituant la molécule. Par convention, elle est donnée en "lisant" la séquence du côté $5'$ au côté $3'$. Pour l'exemple de la figure 4 A, la structure primaire est $5' - GCGAUU...AAUUGCACCA - 3'$.

2. La structure secondaire nous renseigne sur les paires de bases présentes dans la molécule. Elle peut être décomposée en parties double-brins ou en double-hélice constituées de paires consécutives et en parties simple-brins ou boucles (ou bulles) constituées de nucléotides non-appariés (voir figure 4 A).

3. La structure tertiaire représente la forme tridimensionnelle de la molécule (voir figure 4 A) et comment les différentes parties de la structure secondaire interagissent entre elles, souvent par l'intermédiaire d'interactions non-canoniques plus faibles.

4. La structure quaternaire caractérise les interactions et l'organisation entre biomolécules (voir figure 4 B).

L'étude de l'ouverture (ou de la dénaturation) de la double-hélice d'ADN commence en 1952 (un an avant la résolution de la structure de l'ADN par Watson et Crick) avec les travaux de Renée Thomas [16] sur l'absorption UV de l'ADN qui ont permis de caractériser le phénomène de dénaturation : soumis à des contraintes extérieures diverses (thermiques, mécaniques, ioniques), la double-hélice s'ouvre

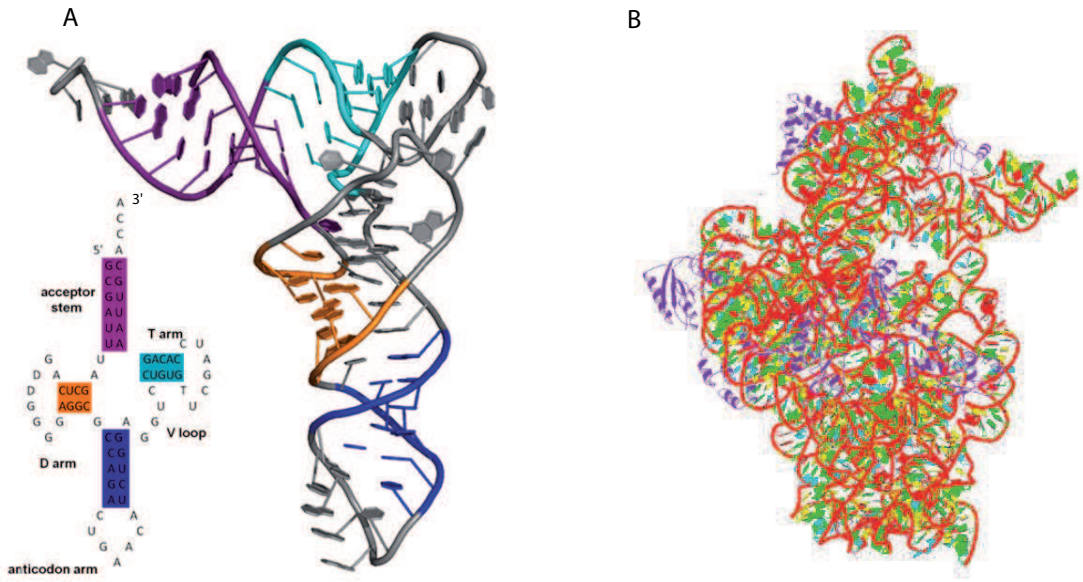

FIGURE 4 – (A) Exemple de la structure secondaire (gauche) et tertiaire (droite) native de l'ARN de transfert phenylalanine de la levure [2]. (B) Structure quaternaire de la petite sous-unité du ribosome de *Thermus thermophilus* composée de protéines (violet) et d'ARN (rouge, jaune,vert et cyan) [15].

totalement ou partiellement (formant des bulles). Depuis, des caractérisations systématiques de ce phénomène ont été menées tant au niveau expérimental qu'au niveau théorique [17].

L'étude du repliement (ou du déploiement) du simple brin d'ARN est plus récente. Dès 1965, Holley et collaborateurs séquençaient la première molécule d'ARN [18], mais il fallut attendre près de 10 ans pour déterminer par cristallographie la première structure d'un ARN de transfert [19]. La complexité et la variété des structures repliées en font un champ d'investigation beaucoup plus vaste que celui de la dénaturation de l'ADN. Là aussi, un effort considérable a été fourni pour comprendre théoriquement et expérimentalement le repliement [20]. Il est communément admis que ce processus est hiérarchique : la structure primaire se replie tout d'abord en une structure secondaire stable qui à son tour se replie pour donner la structure tertiaire.

Cette hiérarchisation permet une étude du repliement mais aussi de la dénaturation à différents niveaux.Théoriquement, les modèles physiques existants travaillent principalement aux niveaux secondaire et tertiaire, et, comme les énergies mises en jeu sont de l'ordre de celles fournies par l'agitation thermique, ils intègrent l'aspect thermodynamique dans leur analyse. Expérimentalement, il existe plusieurs méthodes utilisées pour étudier la dénaturation et le repliement des acides nucléiques. Parmi les plus courantes, on trouve

- L'absorbance UV : on mesure l'absorbance dans l'UV d'une solution d'acides nucléiques en fonction d'un paramètre ajustable (la température, la concentration en sel, etc.) ; comme l'absorbance d'une paire de bases diffère selon qu'elle est appariée ou non, on peut suivre ainsi l'évolution du nombre total de bases appariées en fonction du paramètre d'étude, et estimer des propriétés thermodynamiques comme l'énergie libre nécessaire à dénaturer ou à déplier complètement la molécule [23]. Par exemple, dans le cas classique de la dénaturation thermique de l'ADN (ouverture de la double-hélice quand on augmente la température), on observe une transition entre la double-hélice et l'état dénaturé où les deux brins sont séparés (voir figure 5 A), typiquement , elle se situe à une température de dénaturation de l'ordre de 80° C.
- La microcalorimétrie : on mesure en fonction de la température la capacité calorifique d'une

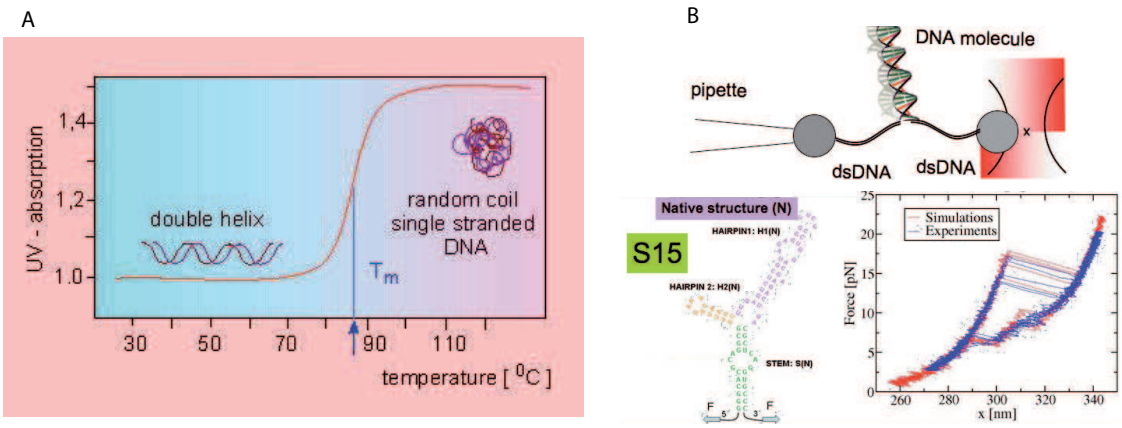

FIGURE 5 – (A) Exemple de courbe expérimentale de dénaturation en fonction de la température obtenue par absorbance UV [21]. (B) Expérience de pinces optiques dans l'étude du repliement d'une structure ARN complexe [22].

solution d'acides nucléiques en estimant la quantité de chaleur transmise à l'échantillon [24]. Cela permet également de suivre la transition de dénaturation et de mesurer des propriétés thermodynamiques.

– Les pinces optiques : on manipule des molécules uniques à l'aide de billes diélectriques piégées dans un faisceau laser. On peut alors exercer des contraintes mécaniques sur la molécule, la déformer ou la contraindre à se dénaturer et remonter ainsi à certaines propriétés thermodynamiques et cinétiques [25]. Par exemple, dans le cas du déploiement d'un ARN quand on augmente la force de tension sur la molécule, on observe dans la courbe force-extension une série de plateaux, signature de l'ouverture successive de domaines jusqu'à la dénaturation complète (voir figure 5 B).

Plan

Dans ma thèse, je me suis attelé à développer et utiliser des modèles de mécanique statistique au niveau secondaire et tertiaire afin de mieux comprendre d'un point de vue physique ces deux phénomènes, en gardant à l'esprit que nos modèles ont pour vocation à décrire ou à prédire les résultats expérimentaux (et ne doivent pas être simplement des jouets de théoriciens) tout en restant suffisamment généraux pour décrire une large gamme de séquences.
En particulier, j'expose dans ce rapport deux axes principaux sur lesquelles j'ai travaillé :

1. La dénaturation de l'ADN à l'aide de modélisations de la structure secondaire (partie I). Tout d'abord, je présente une version unifiée du modèle de Poland-Scheraga [26] qui permet de décrire la dénaturation thermique de l'ADN quelque soit la taille de la chaîne, la concentration en brins et en ions en solution (chapitre 1) ; puis, j'étudie la dénaturation d'ADN génomiques à l'aide du modèle de Zimm-Bragg [27] et je décrit l 'analyse systématique des génomes que nous avons mis en place pour tester les possibles corrélations entre propriétés thermodynamiques et génomiques (chapitre 2).
2. Le repliement de l'ARN à l'aide d'un modèle gros-grain de la structure tertiaire (partie II). Tout d'abord, je développe et paramètre un modèle sur réseau qui tient compte de la connectivité des nucléotides et du volume exclu entre eux, et rend compte quantitativement de la structure secondaire mais aussi de la structure tertiaire de manière qualitative (chapitre 3) ; puis, je détaille les méthodes numériques adaptées à l'analyse des polymères sur réseau que j'ai développé

pour étudier ce modèle (chapitre 4) ; enfin, je présente les résultats obtenus et montre qu'ils reproduisent de manière quantitative les données expérimentales et qu'ils permettent d'estimer l'importance sur le repliement d'interactions comme le volume exclu ou les contacts tertiaires spécifiques (chapitre 5)

Dans chaque partie, nous prenons un soin particulier à discuter les paramètres des modèles étudiés et à tester leur pouvoir de prédiction. Notons également que les formalismes mis en place, bien que construits initialement pour traiter l'ADN ou l'ARN, sont généraux et s'appliquent aussi bien à l'un comme à l'autre, ainsi qu'à d'autres molécules similaires.

Première partie

Dénaturation de l'ADN

Introduction

État de l'art

Depuis plus de 50 ans, de nombreux modèles ont été développés pour décrire la dénaturation et l'association de brins d'ADN. En particulier, on peut distinguer deux descriptions standard de la dénaturation thermique :

1. Le modèle plus proche voisin (NN) [28, 29, 30] qui décrit quantitativement la dénaturation thermique des oligomères courts (composés de $10 - 20$ bp) qui ont une transition à deux états entre la double-hélice et les deux simples brins séparés. L'énergie libre du complexe est décrite par la somme de termes locaux dépendant de la température et comptant pour l'appariement et l'empilement des paires de bases.

2. Le modèle de Poland-Scheraga (PS) [26, 31, 32, 17] qui décrit la dénaturation des polymères (composés d'environ 1000 bp ou plus) au niveau structure secondaire comme l'ouverture successive de domaines à l'intérieur du double-brin. L'énergie libre d'une configuration du complexe est décrite à la fois par des énergies locales d'association, semblables à celles du modèle plus proche voisin, et à la fois par des termes non-locaux comptant pour la nucléation et l'entropie de conformation des bulles (régions dénaturées) présentes dans la configuration (pour plus de détail voir section 1).

Par le passé, ces deux modèles ont été couramment employés pour décrire une grande variété d'expériences de dénaturation de l'ADN. Un formalisme similaire a également été utilisé pour étudier le repliement de l'ARN (le modèle de Turner, voir section 3.1.1). De nombreuses simplifications ou améliorations ont été apportées à ces modèles parmi lesquelles on peut citer le modèle de Zimm-Bragg (ZB) [27] (qui est en réalité plus ancien) qui simplifie le terme d'entropie de conformation du modèle PS en un terme indépendant de la taille de la bulle (voir section 2.1), ou le modèle sur réseau de Everaers, Kumar et Simm [33] qui, au contraire, estime explicitement ce terme d'entropie en modélisant une conformation du complexe par un chemin auto-évitant sur réseau (voir partie II).

Même si dans notre travail nous nous sommes inspirés majoritairement des modèles décrits ci-dessus, l'étude de la dénaturation de l'ADN ne s'est pas, bien évidemment, limitée à ces deux modèles et d'autres approches intéressantes ont également été développées. Dans le même esprit que le modèle PS, on peut citer le modèle de Peyrard-Bishop-Dauxois (PBD) [36, 37] qui réduit la molécule d'ADN en une chaîne unidimensionnelle où l'état (apparié ou dénaturé) de chaque paire de bases est donné par la position relative des deux bases (voir figure 6 A). La thermodynamique et la cinétique de la dénaturation sont alors décrites par un Hamiltonien non-linéaire décrivant l'appariement et l'empilement de ces paires de bases (il n'y pas explicitement l'équivalent des énergies de nucléation ou de conformation présentes dans le modèle PS). Il existe également des modèles gros-grains qui représentent la molécule d'ADN au niveau structure tertiaire avec une description explicite de la structure en double-hélice [34, 38, 39, 35] (voir figures 6 B et C) où l'appariement et l'empilement sont décrits par des champs de forces.

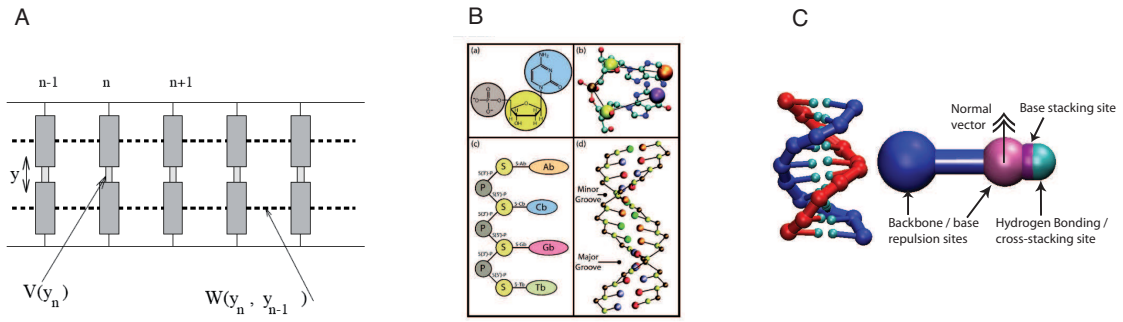

FIGURE 6 – (A) Représentation de la molécule d'ADN dans le modèle PBD : chaque paire de bases n est repérée par l'écartement relatif y_n entre les deux bases et l'Hamiltonien du système vaut $H = \sum_n (m/2)(\mathrm{d}y_n/\mathrm{d}t)^2 + W(y_n, y_{n-1}) + V(y_n)$ avec V et W des potentiels effectifs décrivant respectivement l'appariement et l'empilement. (B,C) Deux exemples de modélisation gros-grain tridimensionnelle développés par Knotts et collaborateurs [34] (B) et par Ouldridge et collaborateurs [35] (C).

Objectifs

Au fur et à mesure des avancées technologiques et des améliorations expérimentales, de nombreux efforts ont été consentis pour paramétrer les modèles NN et PS [40, 41, 42, 43, 44, 45, 46], ainsi que pour comparer les paramètres issus des différents systèmes. Il y a plus de dix ans, SantaLucia concluait dans son article référence [30] : "un jeu unifié de paramètres plus proches voisins est dorénavant disponible pour faire des prédictions précises sur la thermodynamique des oligomères et des polymères d'ADN". Cependant, la compilation des résultats de SantaLucia révèle l'existence de termes de bords ou de correction en sel qui doivent être utilisés dans certaines limites (oligomère ou polymère). Son jeu de paramètres n'est donc pas si unifié que ça. De plus, les deux descriptions standard sont limitées à l'étude de la transition de dénaturation à deux états des oligomères courts ou à celle plus progressive des longs polymères (c'est à dire, dans la limite où la dénaturation de l'ADN devient indépendante de sa concentration). La situation est plus compliquée pour des longs oligomères où une dénaturation interne partielle est observée avant la séparation des brins. Des expériences récentes réalisées par Zeng et Zocchi [47, 48] ont jeté un doute sur l'applicabilité des formalismes standard à ce type de séquences qui jouent un rôle prépondérant dans de nombreuses applications biotechnologiques (comme par exemple les puces à ADN et les balises moléculaires) et dans l'étude des phénomènes biologiques (comme l'interférence ARN). Un des objectifs de cette partie va donc être de proposer un formalisme cohérent et unifié qui permet de décrire la dénaturation de l'ADN quelque soit l'intervalle de taille, de concentration en brins d'ADN ou de concentration en sel étudié.

L'intérêt pour ce genre de modélisation est d'ailleurs d'autant plus grand que les applications possibles sont nombreuses. L'une d'elle parmi les plus originales est l'étude thermodynamique des génomes [49, 50]. Depuis, la découverte des principes gouvernant le stockage, la réplication et la translation de l'information génétique [1, 51], des efforts très importants ont été consentis pour déterminer [52, 53] et analyser [54] le génome des organismes vivants, notamment depuis l'avènement de techniques rapides de séquençage. Plusieurs techniques standard existent pour identifier les gènes à l'intérieur du génome : des approches extrinsèques qui étudient la séquence d'acides aminés des protéines produites [55, 56], des méthodes comparatives entre génomes [57] ou encore des approches ab initio qui prédisent l'emplacement des domaines fonctionnels du génome sur la base d'effet de séquence [58, 59, 60, 61, 62, 63]. De plus, l'information génétique [64, 65, 66, 67] ainsi que les propriétés physiques [68, 69, 70] de l'ADN sont toutes les deux corrélées au contenu GC. On peut donc se demander si ces dernières ne peuvent pas

être utiliser comme nouvelles techniques d'identification des gènes. La réplication et la transcription nécessitent l'ouverture partielle de la double-hélice, il est naturel, dans ce contexte, de considérer la dénaturation de l'ADN [49, 71, 72, 73, 74, 75]. Un autre objectif de cette partie va donc consister à exploiter cette idée pour déterminer si il existe une corrélation entre propriétés liées à la dénaturation thermique et annotation génomique et si l'utilisation d'un modèle thermodynamique de la dénaturation a un sens pour la prédiction de l'emplacement des gènes.

Plan

Selon les principaux objectifs que nous nous sommes fixés, la partie va être divisée en deux chapitres.

Dans un premier chapitre, nous définirons un modèle de Poland-Scheraga unifié qui décrit dans un formalisme unique la dénaturation thermique des brins d'ADN de taille quelconque pour une concentration en brin et en sel arbitraire, puis nous décrirons la méthode algorithmique utilisée pour sa résolution. Nous étudierons ensuite le comportement générique du modèle pour différentes gammes de longueur et de concentration et estimerons son pouvoir de prédiction. Nous prendrons un soin particulier à discuter de la paramétrisation du modèle et de la propagation des erreurs dans les prédictions du modèle.

Dans un deuxième chapitre, nous présenterons l'analyse thermodynamique des génomes que nous avons développée. En particulier, nous motiverons l'utilisation du modèle simplifié de Zimm-Bragg à la place du modèle PS pour le calcul des propriétés de dénaturation de longs génomes. Puis, nous introduirons et exploiterons des méthodes statistiques nous permettant de discuter de la pertinence de l'utilisation des propriétés thermodynamiques des génomes pour les annoter.

Chapitre 1

Modèle de Poland-Scheraga unifié

Dans cette partie, nous présentons une version modifiée du modèle de Poland-Scheraga qui nous permet, avec un formalisme unique, d'étudier la dénaturation de l'ADN dans toutes les gammes de tailles et de concentrations des brins d'ADN ainsi que pour une concentration quelconque en sel.
Dans une première partie, nous définissons les intéractions considérées dans le modèle unifié ainsi que les méthodes algorithmiques et numériques utilisées pour évaluer les prédictions faites par le modèle.
Dans une deuxième partie, nous présentons comment est paramétré le modèle et comment étudier l'incertitude sur les paramètres qui peut se propager aux résultats du modèle.
Puis, dans une troisième partie, nous discutons le comportement générique du modèle selon la taille de la séquence étudiée et selon la concentration en brins et en sel choisie.
Dans une quatrième partie, nous comparons quantitativement les résultats issus du modèle avec des données expérimentales afin de tester le pouvoir de prédiction selon la gamme de longueur considérée.
Enfin, nous concluons sur la nécessité d'améliorer la paramétrisation à l'aide de séquences dont les prédictions faites par le modèle seront sensibles aux erreurs de paramétrisation.

1.1 Définition du modèle

1.1.1 Équilibre d'association

On étudie la dissociation d'un brin d'ADN considéré comme un complexe S_1S_2 en équilibre avec deux simples brins S_1 et S_2, chaque brin étant constitué de N bases (A, G, T ou C) :

$$S_1S_2 \rightleftharpoons S_1 + S_2 \tag{1.1}$$

1.1.1.1 Entropie de mélange et équilibre chimique

Pour calculer un équilibre chimique entre les simples brins et le dimère, on a besoin de tenir compte de l'entropie de translation, c'est à dire, de l'entropie de mélange avec le solvant. Pour estimer cette entropie de mélange [33], on utilise l'expression donnée par Rubinstein et Colby [76] pour le mélange de 2 espèces P et Q calculée à l'aide d'une approche sur réseau

$$-T\Delta S_{mix}/V = k_BT\left(\frac{\Phi}{V_P}\log\Phi + \frac{1-\Phi}{V_Q}\log(1-\Phi)\right) \tag{1.2}$$

où Φ est la fraction volumique en espèce P et V_P et V_Q sont respectivement les volumes moléculaires de P et Q. On a $\Phi = c_PV_P$ et $1-\Phi = c_QV_Q$ avec c_P et c_Q les concentrations en espèces P et Q ($c_P = n_P/V$ avec n_P le nombre de particules P, idem pour Q). On applique maintenant l'équation 1.2 au problème ADN/eau. P représente soit un simple brin (S_1 ou S_2) soit le double-brin S_1S_2, et Q est l'eau. Dans la limite d'une forte dilution, c'est à dire $\Phi << 1$ (ce qui est toujours le cas dans

les situations étudiées ici), on a $(1-\Phi)/V_Q \log(1-\Phi) \approx -\Phi/V_Q = -c_P V_P/V_Q$. Ainsi, l'entropie de mélange par particule P vaut

$$-\Delta S_{mix}/n_P = -T\Delta S_{mix}/V/c_P = k_B T\left(\log(c_P V_P) - \frac{V_P}{V_Q}\right) \tag{1.3}$$

$$= k_B T\left[\log\left(\frac{c_P}{c_0}\right) + \log(c_0\nu_P) - \frac{\nu_P}{\nu_Q}\right] \tag{1.4}$$

avec ν_P et ν_Q les volumes molaires de P et Q, et $c_0 = 1$ M la concentration de référence.
L'énergie libre totale par molécule est alors égale la somme de son entropie de mélange et de son énergie libre interne :

$$G_P = G_P^{int} + k_B T\left[\log\left(\frac{c_P}{c_0}\right) + \log(c_0\nu_P) - \frac{\nu_P}{\nu_{eau}}\right] \tag{1.5}$$

avec $P = S_1$, S_2 ou S_1S_2. Le potentiel chimique par molécule est alors calculé en différentiant la densité d'énergie libre par rapport à la densité

$$\mu_P = \frac{\mathrm{d}}{\mathrm{d}c_P}(c_P G_P) = G_P^{int} + k_B T\left[\log\left(\frac{c_P}{c_0}\right) + \log(c_0\nu_P) - \frac{\nu_P}{\nu_{eau}} + 1\right] \tag{1.6}$$

A l'équilibre thermodynamique, les potentiels thermodynamiques des molécules simple-brins et double-brins sont égales, c'est à dire $\mu_1 + \mu_2 = \mu_{1,2}$, et on trouve :

$$\frac{G_{1,2}^{int} - G_1^{int} - G_2^{int}}{k_B T} = \log\left(\frac{c_1 c_2}{c_{1,2}c_0}\right) + \log\left(\frac{c_0\nu_1\nu_2}{\nu_{1,2}}\right) - \frac{\nu_1 + \nu_2 - \nu_{1,2}}{\nu_{eau}} + 1 \tag{1.7}$$

Si l'on suppose que les volumes molaires s'ajoutent ($\nu_1 + \nu_2 = \nu_{1,2}$), on obtient la loi d'action de masse

$$\frac{c_1 c_2}{c_{1,2}c_0} = \exp\left[\beta(\Delta G_{int} + \Delta G_{mix}^0)\right] \equiv \exp\left(\beta\Delta G_0\right) \tag{1.8}$$

avec $\Delta G_{int} = G_{1,2}^{int} - G_1^{int} - G_2^{int}$, $\Delta G_{mix}^0 = -k_B T[\log(c_0\nu_{1,2}/4) + 1]$ et $\Delta G_0 = \Delta G_{int} + \Delta G_{mix}^0$, la différence d'énergie libre de Gibbs entre l'état fermé et l'état dénaturé à la concentration de référence. Le volume molaire $\nu_{1,2}$ du duplexe est proportionnel au nombre de segments de paires de bases dans le double-brin, soit $\nu_{1,2} = (N-1)\nu_s$ avec ν_s le volume molaire d'un segment de paires de bases qui peut être modélisé par un cylindre de hauteur 0.34 nm et de rayon 1 nm. Numériquement, on trouve $\Delta G_{mix}^0 = -k_B T \log[0.44(N-1)]$.

1.1.1.2 Définitions

Soit $c_T = c_1 + c_2 + 2c_{1,2}$, le concentration totale des brins d'ADN, on définit le degré d'association Θ_{ext} comme la probabilité qu'un brin soit sous forme double-brin :

$$\begin{cases} c_1 = (1-\Theta_{ext})c_T/2 \\ c_2 = (1-\Theta_{ext})c_T/2 \\ c_{1,2} = \Theta_{ext}\, c_T/2 \end{cases} \tag{1.9}$$

En remplaçant les concentrations dans 1.8 par leurs expressions 1.9 respectives, on trouve que

$$\Theta_{ext}(x) = 1 + x - \sqrt{x\,(2+x)} \tag{1.10}$$

avec $x = \frac{c_0}{\alpha c_T}\exp\left(\frac{\Delta G_0}{k_B T}\right)$ ($\alpha = 4$ pour des brins auto-similaires $S_1 = S_2$, ou $\alpha = 1$ sinon).

L'hybridation interne du complexe S_1S_2 et des simples brins S_1 et S_2 peut être décrite par la fraction Θ_{int} de paires fermées. Ainsi la fraction totale de paires de bases fermées est donnée par

$$\Theta = \Theta_{int,1,2}\Theta_{ext} + \frac{1}{2}\Theta_{int,1}(1-\Theta_{ext}) + \frac{1}{2}\Theta_{int,2}(1-\Theta_{ext}) \tag{1.11}$$

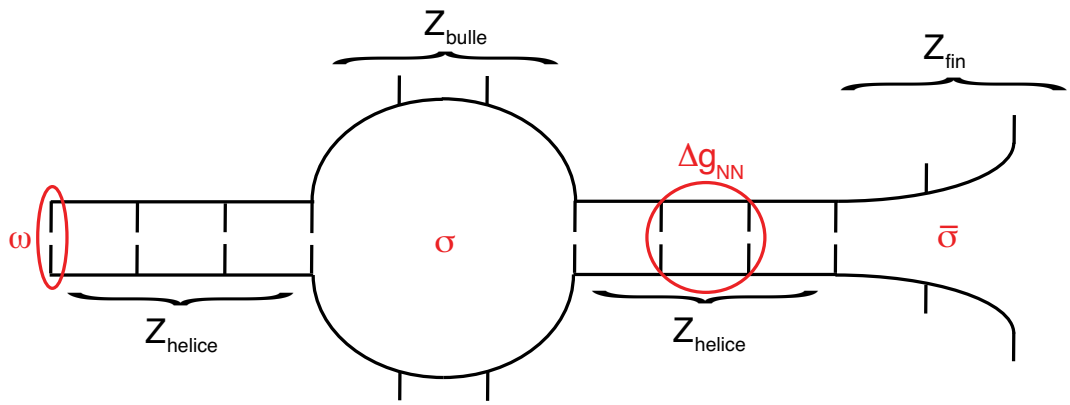

FIGURE 1.1 – Exemple d'une conformation pour la structure secondaire d'un brin d'ADN et définition des différentes contributions à la fonction de partition.

Dans la suite, on négligera l'association interne des simples brins et on supposera que $\Theta_{int,1} = \Theta_{int,2} = 0$. Notons que a priori pour un équilibre fixée, Θ_{int} ne va dépendre que de la température alors que Θ_{ext} et Θ vont dépendre à la fois de la température et de la concentration en brin.

On introduit également la notion de température de fusion (ou de dénaturation) T_m définie par $\Theta_{ext}(T_m) = 1/2$ ou $\Theta(T_m) = 1/2$ (définitions équivalentes pour des transitions à deux états [1]). Par exemple, pour un oligomère court exhibant une transition à deux états, ΔG_0 peut se décomposer sous la forme $\Delta G_0 = \Delta H_0 - T\Delta S_0$ avec ΔH_0 et ΔS_0 indépendants de la température. On a alors directement à partir de l'équation 1.10 [2]

$$T_m = \frac{\Delta H_0}{\Delta S_0 + k_B \log(\alpha c_T/(4c_0))} \tag{1.12}$$

1.1.2 Interactions dans le modèle de Poland-Scheraga

La détermination de ΔG_{int} ainsi que de Θ_{int} nécessite une description et une modélisation du complexe et des simples brins. Nous utilisons pour cela le modèle de Poland-Scheraga (PS) où la dénaturation de l'ADN est modélisée comme l'ouverture et/ou la fusion coopérative et successive de domaines dénaturés au fur et à mesure que la température augmente. Le modèle PS décrit l'ADN au niveau de sa structure secondaire comme l'agencement de parties en double-hélice et de bulles internes ou de domaines terminaux dénaturés.

L'énergie d'une partie en double-hélice est donnée par la somme, sur les segments de paires de bases la composant, des énergies libres d'association $\Delta g_{NN} = \Delta h_{NN} - T\Delta s_{NN}$ qui dépendent des 10 différents types de segment possibles (voir table 1.1). On considère également les énergies de capping $\omega = \Delta h_\omega - T\Delta s_\omega$ décrivant les terminaisons hélicales que l'on suppose dépendantes uniquement de la nature de la paire finale (A/T ou G/C), ainsi que les facteurs de coopérativité σ et $\bar{\sigma}$ pénalisant respectivement l'ouverture d'une bulle interne ou d'un domaine terminal. La figure 1.1 illustre les différentes contributions en énergie libre à la fonction de partition du système. L'état de référence étant pris comme l'état complètement fermé ($Z_{helice} = 1$), la contribution d'une bulle interne de taille n sera donnée par $Z_{bulle} = \sigma n^{-c} \exp(\beta \sum \Delta g_{NN})$ et celle d'un domaine terminal dénaturé de taille $\bar{n}$ par $Z_{fin} = \bar{\sigma}\bar{n}^{c'} \exp(\beta \sum \Delta g_{NN} + \beta\omega)$. Ici, n^{-c} et $\bar{n}^{c'}$ rendent compte respectivement du nombre de polygones auto-évitants de taille $2n$ et du nombre de chemins auto-évitants de taille $2\bar{n}$. Les valeurs de c et c' décrivent les interactions stériques dans ces domaines dénaturés et sont dérivées de la théorie

1. En effet, pour une transition à deux état, l'unique état en duplexe est l'état complètement fermé soit $\Theta_{int} = 1$.
2. $\Theta_{ext} = 1/2$ donne $x(T_m) = 1/4$, soit $(c_0/(\alpha c_T)) \exp[\Delta H_0/(k_B T_m) - \Delta S_0/k_B] = 1/4$, ce qui donne en inversant la relation $T_m = \Delta H_0/(\Delta S_0 + k_B \log(\alpha c_T/(4c_0)))$.

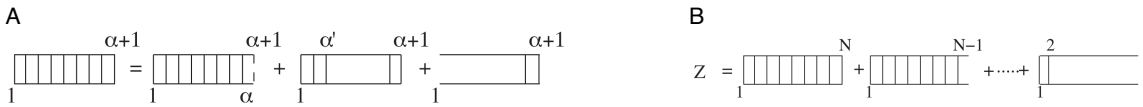

Figure 1.2 – (A) Représentation graphique de la relation de récurrence pour $Z_f(\alpha+1)$ (équation 1.13). (B) Représentation graphique de la fonction de partition thermodynamique Z (équation 1.17).

des polymères. L'exposant c vaut environ 1.76 [26, 77] si l'on considère que les bulles n'interagissent pas avec les autres sous-structures ou 2.15 [78, 79, 80] si on considère des interactions stériques entre bulles. L'exposant c' vaut quant à lui environ 1/6 [81, 82]. La fonction de partition totale décrivant une conformation est alors calculée en effectuant le produit des différentes fonctions de partition partielles. Pour l'exemple de la figure 1.1, on aura $Z = Z_{helice}Z_{bulle}Z_{helice}Z_{fin}$.

Notons que l'on se limite à la formation de contacts natifs pour des brins d'ADN parfaitement complémentaires (pas de défauts d'appariement ou "mismatch") et on ne tient pas compte d'éventuelles formations de bulles en épingle ou de renflements, mais le cas général peut également être abordé avec le même formalisme [83, 84]. Il est de plus également possible de traiter la formation de structure secondaire dans les simples brins [85, 86, 87] (voir le modèle de Turner et le modèle sur réseau à la section 3.1.1).

1.1.3 Résolution du modèle de Poland-Scheraga

1.1.3.1 Relations de récurrence

Originellement, le modèle PS s'intéresse à des relations de récurrence sur des probabilités thermodynamiques conditionelles. Récemment [83, 84, 88], des relations de récurrence sur les fonctions de partition ont été développées pour étudier la dénaturation de l'ADN. On suit ici la procédure définie par Garel et Orland [84]. Ces relations sont un peu plus compliquées que les relations originales car nous considérons que la longueur minimale d'une section hélicoidale est de 2 bp au lieu de 1 bp dans [84].
On modélise un simple brin comme une chaîne de N bases A, T, C ou G, numérotées de 1 à N : $5'-1-2-3-...-N-3'$. On suppose que les deux simples brins sont parfaitement complémentaires : la base i du brin S_1 ne peut s'apparier qu'avec la base $N-i+1$ du brin S_2. On considère l'état double brin totalement fermé comme état de référence.

Tout d'abord, on regarde la fonction de partition en sens direct $Z_f(\alpha+1)$ comptant le nombre d'états de la partie du brin compris entre les bases 1 et $\alpha+1$, les bases α et $\alpha+1$ étant fermées. Il y a trois moyens d'avoir ces deux bases fermées : soit la paire ($\alpha-1$,α) est appariée, soit il y a une boucle commencant à la base α' et se finissant en α, soit le complexe est complètement ouvert de la base 1 à la base α (voir figure 1.2 A)

$$Z_f(\alpha+1) = Z_f(\alpha) + \sigma \sum_{\alpha'=2}^{\alpha-2} (\alpha-\alpha')^{-c} e^{\beta g_{\alpha',\alpha-1}} Z_f(\alpha') + \bar{\sigma}(\alpha-1)^{c'} e^{\beta g_{1,\alpha-1}+\beta\omega_1} \tag{1.13}$$

avec $g(\alpha',\alpha) = \sum_{i=\alpha'}^{\alpha} \Delta g_{NN}(i,i+1)$ où $\Delta g_{NN}(i,i+1)$ est l'énergie libre d'association (ou d'appariement) de la paire $(i,i+1)$, et c et c' tiennent compte des interactions stériques des boucles ou des bouts libres.
De manière analogue à Z_f, on introduit la fonction de partition en sens inverse $Z_b(\alpha)$ estimant le

nombre d'états de la partie comprise entre les bases α et N, la base α étant fermée. Ainsi,

$$Z_b(\alpha) = Z_b(\alpha+1) + \sigma \sum_{\alpha'=\alpha+2}^{N-1} (\alpha'-\alpha)^{-c} e^{\beta g_{\alpha,\alpha'-1}} Z_b(\alpha'+1) + \bar{\sigma}(N-\alpha)^{c'} e^{\beta g_{\alpha,N-1}+\beta\omega_N} \tag{1.14}$$

Enfin, on considère $Z_{sf}(\alpha)$ la seconde fonction de partition en sens direct, débutant à la base 1 et se finissant à la base α, la base α étant fermée et la base $\alpha-1$ étant ouverte. D'où

$$Z_{sf}(\alpha) = \sigma \sum_{\alpha'=2}^{\alpha-2} (\alpha-\alpha')^{-c} e^{\beta g_{\alpha',\alpha-1}} Z_f(\alpha') + \bar{\sigma}(\alpha-1)^{c'} e^{\beta g_{1,\alpha-1}+\beta\omega_1} \tag{1.15}$$

Avec ces trois fonctions de partition, on peut exprimer la probabilité $p(\alpha)$ pour que la base α soit fermée

$$p(\alpha) = \frac{Z_{sf}(\alpha)Z_b(\alpha+1) + Z_f(\alpha)Z_b(\alpha)}{Z} \tag{1.16}$$

avec Z la fonction de partition totale (voir figure 1.2 B).

$$Z = Z_f(N) + \bar{\sigma} \sum_{\alpha=2}^{N-1} (N-\alpha)^{c'} e^{\beta g_{\alpha,N-1}+\beta\omega_N} Z_f(\alpha) \tag{1.17}$$

où l'on s'est restreint aux configurations avec au moins un segment de paires de bases fermées (puisque l'on s'intéresse à l'association interne et non à la dissociation).
Ainsi on peut exprimer Θ_{int} comme la moyenne des $p(\alpha)$

$$\Theta_{int} = \frac{1}{N} \sum_{\alpha=1}^{N} p(\alpha) \tag{1.18}$$

et la différence d'énergie libre interne

$$\Delta G_{int} = \sum_{i=1}^{N-1} \Delta g_{NN}(i,i+1) + \omega_1 + \omega_N - k_B T \log Z \tag{1.19}$$

1.1.3.2 Algorithme de Fixman-Freire

Le challenge est maintenant de résoudre les équations 1.13, 1.14 et 1.15. Une résolution directe de ces relations nécessiterait $\mathcal{O}(N^2)$ opérations. L'algorithme de Fixman-Freire [89] permet d'accélérer de manière significative cette résolution. Nous décrivons brièvement ici son principe. Une description de l'algorithme et une discussion sur le temps de calcul plus détaillées sont données respectivement en annexe 6.1 et à la section 2.1.3.

L'algorithme consiste à approcher les termes non-locaux x^{-c} par un développement multi-exponentiel :

$$x^{-c} \approx \sum_{k=1}^{I} a_k e^{-b_k x} \tag{1.20}$$

Le nombre I d'éléments dans la somme dépend de la précision voulue et est proportionnel à $\log n_{max}$ avec n_{max} une taille maximale de bulle jusqu'où l'approximation précédente reste valide. Typiquement pour $n_{max} = 5$ kbp, $I = 14$, et pour $n_{max} = 10^8$ bp, on a $I = 21$. Les valeurs de $\{a_k, b_k\}$ sont déterminées par la résolution d'équations non-linéaires [89].
En insérant l'approximation 1.20 dans les relations de récurrence 1.13, 1.14 et 1.15, et en définissant $2I$ nouvelles variables, on obtient de nouvelles relations de récurrence qui se résolvent en $\mathcal{O}(N \times I)$ opérations. Ainsi, si pour chaque taille de séquence on désire la meilleure approximation on aura finalement un algorithme en $\mathcal{O}(N \log N)$.

1.2 Paramétrisation et erreurs

Malgré le nombre important de paramètres ajustables (10 Δh_{NN}, 10 Δs_{NN}, 2 Δh_ω, 2 Δs_ω, σ et $\bar{\sigma}$), le modèle PS représente clairement une simplification drastique du problème ; l'hypothèse principale étant de considérer que toutes les contributions à la différence d'énergie libre s'écrivent comme la somme de termes (locaux comme Δg_{NN} ou non locaux comme $-k_B T(\log\sigma - c\log n)$) indépendants entre eux. De plus, pour l'ADN, la dépendance en la séquence des termes de bords n'est souvent traitée que partiellement. Par exemple, comme les bouts $5'$ et $3'$ sont chimiquement différents, il n'y a aucune raison à avoir des énergies ω identiques pour $\begin{smallmatrix}5'-A\\3'-T\end{smallmatrix}$ et $\begin{smallmatrix}5'-T\\3'-A\end{smallmatrix}$, ainsi que pour $\begin{smallmatrix}5'-G\\3'-C\end{smallmatrix}$ et $\begin{smallmatrix}5'-C\\3'-G\end{smallmatrix}$. De même, on s'attendrait (comme pour l'ARN) à l'existence d'énergies de fourches (au niveau des jonctions entre les parties en double-hélice et les bulles) dépendantes de la séquence et de la concentration en sel ; ou encore à la dépendance en la composition des domaines dénaturés de σ et $\bar{\sigma}$. Cependant les données expérimentales ne permettent pas a priori de telles considérations. En général, les hypothèses faites peuvent seulement se justifier a posteriori en évaluant le succès du modèle à décrire et à prédire les résultats expérimentaux. Ainsi, un soin particulier doit être pris pour estimer les paramètres du modèle et leurs barres d'erreurs correspondantes pour pouvoir comparer les prédictions faites avec les expériences.

C'est pourquoi, dans cette section, nous décrivons tout d'abord la paramétrisation des différentes contributions du modèle PS, puis, nous discutons de l'influence du sel sur ces mêmes paramètres, enfin nous expliquons comment nous étudions la propagation des erreurs faites sur les paramètres aux résultats prédits par le modèle.

1.2.1 Paramètres d'association

Les oligomères courts ($N \sim 10$ bp) exhibant une transition de dénaturation à deux états peuvent être utilisés pour paramétrer les énergies d'association et de capping (car les états partiellement dénaturés pour lesquels la coopérativité interviendrait sont très largement minoritaires). Dans le modèle de plus proche voisin, pour un oligomère court, $\Delta G_0 = \sum \Delta g_{NN} + 2\Delta g_{ini}$, où Δg_{ini} est l'énergie libre d'initiation [30]. Dans le modèle PS unifié, on a $\Delta G_0 = \sum \Delta g_{NN} + 2\omega + \Delta G^0_{mix}$. En égalisant les deux expressions, on trouve

$$\omega = \Delta h_{ini} - T(\Delta s_{ini} - \Delta S^0_{mix}/2) \tag{1.21}$$

où $\Delta S^0_{mix} = 1.5 k_B$ tient compte de l'entropie de mélange pour $N = 10$ bp, la taille typique des oligomères utilisés dans les expériences de paramétrisation du modèle plus proche voisin. Reste donc à déterminer les 24 paramètres indépendants du modèle plus proche voisin (10 Δh_{NN}, 10 Δs_{NN}, 2 Δh_{ini}, 2 Δs_{ini}).

Prenons tout d'abord le cas général où les énergies d'initiation pour $\begin{smallmatrix}5'-A\\3'-T\end{smallmatrix}$ et $\begin{smallmatrix}5'-T\\3'-A\end{smallmatrix}$ ne sont a priori pas égales (idem pour $\begin{smallmatrix}5'-G\\3'-C\end{smallmatrix}$ et $\begin{smallmatrix}5'-C\\3'-G\end{smallmatrix}$), soit 28 paramètres. Dans un oligomère court donné, le fait que le nombre de fois où une certaine paire de bases (A/T ou G/C) précède une autre soit égale au nombre de fois où elle succède à une autre, impose les contraintes sur les nombres d'occurrences $n(j)$ dans l'oligomère du même type de segment de paires de bases j [90] avec $j \in \{\begin{smallmatrix}5'-AA-3'\\3'-TT-5'\end{smallmatrix}, \begin{smallmatrix}5'-AT-3'\\3'-AT-'5\end{smallmatrix}, \begin{smallmatrix}5'-TA-3'\\3'-TA-5'\end{smallmatrix}, \begin{smallmatrix}5'-CA-3'\\3'-GT-5'\end{smallmatrix}, \begin{smallmatrix}5'-GT-3'\\3'-CA-5'\end{smallmatrix}, \begin{smallmatrix}5'-CT-3'\\3'-GA-5'\end{smallmatrix}, \begin{smallmatrix}5'-GA-3'\\3'-CT-5'\end{smallmatrix}, \begin{smallmatrix}5'-CG-3'\\3'-GC-5'\end{smallmatrix}, \begin{smallmatrix}5'-GC-3'\\3'-CG-5'\end{smallmatrix}, \begin{smallmatrix}5'-GG-3'\\3'-CC-5'\end{smallmatrix}, \begin{smallmatrix}5'-A\\3'-T\end{smallmatrix}, \begin{smallmatrix}5'-T\\3'-A\end{smallmatrix}, \begin{smallmatrix}5'-G\\3'-C\end{smallmatrix}, \begin{smallmatrix}5'-C\\3'-G\end{smallmatrix}\}$. Soit,

$$\begin{aligned} & n(\tfrac{5'-AA-3'}{3'-TT-5'}) + n(\tfrac{5'-AT-3'}{3'-TA-5'}) + n(\tfrac{5'-GT-3'}{3'-CA-5'}) + n(\tfrac{5'-CT-3'}{3'-GA-5'}) + n(\tfrac{5'-A}{3'-T}) \\ = \; & n(\tfrac{5'-AA-3'}{3'-TT-5'}) + n(\tfrac{5'-TA-3'}{3'-AT-'5}) + n(\tfrac{5'-CA-3'}{3'-GT-5'}) + n(\tfrac{5'-GA-3'}{3'-CT-5'}) + n(\tfrac{5'-T}{3'-A}) \end{aligned} \tag{1.22}$$

$$\begin{aligned} & n(\tfrac{5'-GG-3'}{3'-CC-5'}) + n(\tfrac{5'-GC-3'}{3'-CG-5'}) + n(\tfrac{5'-GA-3'}{3'-CT-5'}) + n(\tfrac{5'-GT-3'}{3'-CA-5'}) + n(\tfrac{5'-G}{3'-C}) \\ = \; & n(\tfrac{5'-GG-3'}{3'-CC-5'}) + n(\tfrac{5'-CG-3'}{3'-GC-5'}) + n(\tfrac{5'-CT-3'}{3'-GA-5'}) + n(\tfrac{5'-CA-3'}{3'-GT-5'}) + n(\tfrac{5'-C}{3'-G}) \end{aligned} \tag{1.23}$$

Table 1.1 – Paramètres standard du modèle Poland-Scheraga unifié et leurs barres d'erreur pour une concentration en sel de 1 M.

Sequence	Δh (kcal/mol)	Δs (cal/mol/K)
$\begin{smallmatrix}5'-AA-3'\\3'-TT-5'\end{smallmatrix}$	-7.93 ± 0.31	-22.4 ± 1.0
$\begin{smallmatrix}5'-AT-3'\\3'-TA-5'\end{smallmatrix}$	-7.15 ± 0.78	-20.2 ± 2.6
$\begin{smallmatrix}5'-TA-3'\\3'-AT-'5\end{smallmatrix}$	-7.23 ± 0.82	-21.6 ± 2.7
$\begin{smallmatrix}5'-CA-3'\\3'-GT-5'\end{smallmatrix}$	-8.44 ± 0.77	-22.9 ± 2.5
$\begin{smallmatrix}5'-GT-3'\\3'-CA-5'\end{smallmatrix}$	-8.47 ± 0.66	-22.9 ± 2.2
$\begin{smallmatrix}5'-CT-3'\\3'-GA-5'\end{smallmatrix}$	-7.73 ± 0.66	-20.9 ± 2.2
$\begin{smallmatrix}5'-GA-3'\\3'-CT-5'\end{smallmatrix}$	-8.29 ± 0.61	-22.6 ± 2.0
$\begin{smallmatrix}5'-CG-3'\\3'-GC-5'\end{smallmatrix}$	-10.54 ± 0.82	-27.1 ± 2.7
$\begin{smallmatrix}5'-GC-3'\\3'-CG-5'\end{smallmatrix}$	-9.81 ± 0.73	-24.6 ± 2.4
$\begin{smallmatrix}5'-GG-3'\\3'-CC-5'\end{smallmatrix}$	-8.02 ± 0.68	-19.6 ± 2.3
capping $\begin{smallmatrix}G\\C\end{smallmatrix}$	0.08 ± 0.99	-3.7 ± 3.4
capping $\begin{smallmatrix}A\\T\end{smallmatrix}$	2.22 ± 1.02	2.8 ± 3.3
$\log \sigma$	-9.0 ± 2.7	
$\log \bar{\sigma}$	-5.7 ± 1.4	
γ_S (cal/mol/K)	-8.9 ± 2.7	
ΔS^0_{mix} (cal/mol/K)	3.0 ± 0.9	
a_1 (K^{-1})	$(4.3 \pm 0.3) \times 10^{-5}$	
a_2 (K^{-1})	$(-3.9 \pm 0.2) \times 10^{-5}$	
a_3 (K^{-1})	$(9.4 \pm 0.3) \times 10^{-6}$	
b_1 (K^{-1})	$(2.1 \pm 0.2) \times 10^{-5}$	
b_2 (K^{-1})	$(-1.4 \pm 0.1) \times 10^{-5}$	
b_3 (K^{-1})	$(4.0 \pm 0.7) \times 10^{-6}$	
b_4 (K^{-1})	$(1.0 \pm 0.1) \times 10^{-6}$	
b_5 (K^{-1})	$(-5.4 \pm 0.1) \times 10^{-7}$	
b_6 (K^{-1})	$(1.1 \pm 0.1) \times 10^{-8}$	
b_7 (K^{-1})	$(6.3 \pm 0.4) \times 10^{-5}$	
b_8 (K^{-1})	$(1.8 \pm 0.1) \times 10^{-5}$	
b_9 (K^{-1})	$(-5.7 \pm 0.3) \times 10^{-8}$	
b_{10} (K^{-1})	$(-1.1 \pm 0.1) \times 10^{-7}$	

Supposons maintenant qu'expérimentalement nous avons déterminé ΔH_0 et ΔS_0 pour N_s oligomères de composition et de longueurs différentes ($\Delta G_0 = \Delta H_0 - T\Delta S_0$). Les données peuvent alors être écrites sous la forme

$$\mathbf{H} = \mathbf{P} \cdot \mathbf{H_{NN}} \tag{1.24}$$

$$\mathbf{S} = \mathbf{P} \cdot \mathbf{S_{NN}} \tag{1.25}$$

où $\mathbf{H_{NN}}$ est le vecteur des 14 paramètres enthalpiques (idem pour $\mathbf{S_{NN}}$), $\mathbf{P}$ est la matrice de paires, c'est à dire, $P_{ij} = n_i(j)$ est le nombre de segments de paires de type j dans la séquence i et $\mathbf{H}$ le vecteur des ΔH_0 (idem pour $\mathbf{S}$). Du fait des contraintes 1.22 et 1.23, la matrice $\mathbf{P}$ sera de rang 12 (et non de rang 14). Ce qui signifie que 2 paramètres enthalpiques (et 2 entropiques) ne pourront pas être déterminés indépendamment des autres. C'est pour cela que dans le modèle plus proche voisin (et donc

dans le modèle PS), on suppose que $\Delta h_{ini}\binom{5'-A}{3'-T} = \Delta h_{ini}\binom{5'-T}{3'-A}$ et que $\Delta h_{ini}\binom{5'-G}{3'-C} = \Delta h_{ini}\binom{5'-C}{3'-G}$ (idem pour Δs_{ini}) pour avoir 12 paramètres que l'on pourra déterminer de manière unique [91, 43, 90, 92].

Les solutions des équations 1.24 et 1.25 sont obtenues avec la méthode de décomposition en valeur singulière (SVD) [93] qui inverse $\mathbf{P}$ et minimise

$$\chi_H^2 = \sum_i \left(\frac{H_i - \sum_j P_{ij} H_{NN,j}}{\sigma_{H,i}} \right)^2 \tag{1.26}$$

$$\chi_S^2 = \sum_i \left(\frac{S_i - \sum_j P_{ij} S_{NN,j}}{\sigma_{S,i}} \right)^2 \tag{1.27}$$

avec $\sigma_{H,i}$ et $\sigma_{S,i}$ les déviations standard de ΔH_0 (5%) et ΔS_0 (6%).
Soit $\mathbf{U}$ (matrice unitaire de taille $N_s \times N_s$), $\mathbf{V}$ (matrice unitaire de taille 12×12) et $\mathbf{W}$ (matrice diagonale positive de taille $N_s \times 12$) les trois matrices obtenues par la décomposition SVD de $\mathbf{A}$ ($A_{ij} = P_{ij}/\sigma_i$)

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^\dagger \tag{1.28}$$

alors la minimisation des équations 1.26 et 1.27 a pour solution

$$\mathbf{H_{NN}} = \sum_{i=1}^{N_s} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b_H}}{W_i} \right) \mathbf{V}_{(i)} \tag{1.29}$$

$$\mathbf{S_{NN}} = \sum_{i=1}^{N_s} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b_S}}{W_i} \right) \mathbf{V}_{(i)} \tag{1.30}$$

où $\mathbf{U}_{(i)}$ représente la colonne i de $\mathbf{U}$ (idem pour $\mathbf{V}$) et $b_{H,i} = H_i/\sigma_{H,i}$ (idem pour $\mathbf{b_S}$).

C'est avec cette méthode qu'Allawi et SantaLucia ont pu déterminer les paramètres plus proche voisin avec $N_s = 108$ oligomères pour une concentration en sel $[Na^+] = 1$ M [43]. Afin d'avoir accès à la matrice de covariance entre les différents paramètres (voir section 1.2.4), nous avons refait cette minimisation. Les paramètres obtenus sont donnés dans la table 1.1 et restent très proches de ceux trouvés par Allawi et SantaLucia.

1.2.2 Correction dûe au sel

Des observations expérimentales [94] montrent que changer la concentration en sel de 1 M à 0.1 M décale la température de dénaturation d'environ -10 K. L'ADN étant une molécule très chargée (négativement), une telle dépendance n'est pas surprenante et est formellement le résultat d'une intégration sur les degrés de liberté microscopiques de l'ADN, des molécules de solvant et des ions salins. Ces variations sont décrites phénoménologiquement par des corrections différentes pour les oligomères courts et les polymères [30, 94]. Une approche systématique en mécanique statistique devrait pouvoir prédire les températures de dénaturation observées à partir de paramètres dépendant du sel.

1.2.2.1 Effet des ions monovalents

La correction la plus utilisée dans les programmes bioinformatiques (DINAmelt [88], MELTING [95]) est celle donnée par SantaLucia pour les paramètres plus proche voisin des oligomères [30]

$$\Delta s_{NN}([Na^+]) = \Delta s_{NN}(1\text{M}) + 0.368 \log[Na^+] \quad (\text{cal/mol/K}) \tag{1.31}$$

et Δh_{NN} est supposée invariante par changement de concentration en sel. On remarque que cette correction est indépendante de la séquence. Or Owczarzy et collaborateurs [94] ont montré que les

températures de dénaturation des oligomères exhibant une transition à deux états suivaient la loi phénoménologique dans la gamme 50 mM $\leq [Na^+] \leq$ 1 M :

$$\frac{\mathrm{d}}{\mathrm{d}(\log[Na^+])}\left(\frac{1}{T_m}\right) = (a_1 f(GC) + a_2) + 2\, a_3 \log[Na^+] \tag{1.32}$$

où a_1, a_2 et a_3 sont des constantes d'intensité comparables (voir table 1.1) et $f(GC)$ est la fraction en GC dans la séquence. L'équation 1.32 ne peut pas être directement utilisée dans notre modèle. Pour en déduire la dépendance en sel des paramètres locaux Δh_{NN}, Δs_{NN}, Δh_ω et Δs_ω, on procède en deux étapes :
1) Premièrement, on suppose que l'enthalpie est indépendante de la concentration en sel [96, 97] mais que l'entropie elle en dépend. En effet, le gain en entropie de mélange des contre-ions libérés lors de la dénaturation de partie en double-hélice doit dépendre de la concentration en sel [98, 99]. Ainsi, pour des séquences ayant une transition à deux états, la température de dénaturation suit également l'équation 1.12, on peut donc réécrire l'équation 1.32

$$\frac{\mathrm{d}}{\mathrm{d}(\log[Na^+])}(\Delta S_0) = \Delta H_0 \times \big((a_1 f(GC) + a_2) + 2\, a_3 \log[Na^+]\big) \tag{1.33}$$

2) Deuxièmement, au lieu d'appliquer l'équation 1.33 à la séquence toute entière, on l'utilise pour décrire les termes entropiques locaux d'association et de capping

$$\Delta s([Na^+]) = \Delta s(1\mathrm{M}) + \Delta h((a_1 f_l(GC) + a_2)\log[Na^+] + a_3 \log^2[Na^+]) \tag{1.34}$$

où $f_l(GC)$ est la fraction locale en GC pour chaque segment de paires. L'équation 1.34 révèle l'effet stabilisateur sur l'énergie libre de la hausse de $[Na^+]$, dûe à l'élévation de l'écrantage des contre-ions [94, 99]. Le nombre de voisins pris en compte pour le calcul de $f_l(GC)$ ne semble pas influer sur les résultats pour des nombres inférieurs à 5 qui correspond typiquement à la longueur de Debye dans la solution ($\lambda_D = \sqrt{\epsilon\, k_B\, T/\rho_e} \approx 10\dot{A}$) . Nous fixons ce nombre à 0, c'est à dire que $f_l(GC)$ ne dépend que de la composition du segment de paires de base considéré : pour $\begin{smallmatrix}5'-AT-3'\\3'-TA-5'\end{smallmatrix}$ $f_l(GC) = 0$, pour $\begin{smallmatrix}5'-GT-3'\\3'-CA-5'\end{smallmatrix}$ $f_l(GC) = 0.5$, ou encore pour $\begin{smallmatrix}5'-GC-3'\\3'-CG-5'\end{smallmatrix}$ $f_l(GC) = 1$. Des valeurs typiques de la correction $\delta s[Na^+] = \Delta s([Na^+]) - \Delta s(1\mathrm{M})$ pour les différentes contributions locales sont données dans la table 1.2.

Table 1.2 – Correction en sel issue de l'équation 1.34

	$-\delta s([Na^+])$ (cal/mol/K)			
Sequence	0.01M	0.05M	0.1M	0.5M
$\begin{smallmatrix}5'-AA-3'\\3'-TT-5'\end{smallmatrix}$	3.0 ± 0.2	1.6 ± 0.1	1.1 ± 0.05	0.3 ± 0.01
$\begin{smallmatrix}5'-AT-3'\\3'-TA-5'\end{smallmatrix}$	2.7 ± 0.3	1.5 ± 0.2	1.0 ± 0.1	0.2 ± 0.03
$\begin{smallmatrix}5'-TA-3'\\3'-AT-'5\end{smallmatrix}$	2.8 ± 0.3	1.5 ± 0.2	1.0 ± 0.1	0.2 ± 0.03
$\begin{smallmatrix}5'-CA-3'\\3'-GT-5'\end{smallmatrix}$	2.4 ± 0.2	1.2 ± 0.1	0.8 ± 0.1	0.1 ± 0.02
$\begin{smallmatrix}5'-GT-3'\\3'-CA-5'\end{smallmatrix}$	2.4 ± 0.2	1.2 ± 0.1	0.8 ± 0.05	0.1 ± 0.02
$\begin{smallmatrix}5'-CT-3'\\3'-GA-5'\end{smallmatrix}$	2.2 ± 0.2	1.1 ± 0.1	0.7 ± 0.05	0.1 ± 0.02
$\begin{smallmatrix}5'-GA-3'\\3'-CT-5'\end{smallmatrix}$	2.3 ± 0.2	1.1 ± 0.1	0.8 ± 0.05	0.1 ± 0.02
$\begin{smallmatrix}5'-CG-3'\\3'-GC-5'\end{smallmatrix}$	1.9 ± 0.2	0.8 ± 0.1	0.4 ± 0.05	0.0 ± 0.02
$\begin{smallmatrix}5'-GC-3'\\3'-CG-5'\end{smallmatrix}$	1.8 ± 0.2	0.7 ± 0.1	0.4 ± 0.05	0.0 ± 0.02
$\begin{smallmatrix}5'-GG-3'\\3'-CC-5'\end{smallmatrix}$	1.5 ± 0.2	0.6 ± 0.1	0.3 ± 0.05	0.0 ± 0.02
capping $\begin{smallmatrix}G\\C\end{smallmatrix}$	0.0 ± 0.2	0.0 ± 0.1	0.0 ± 0.05	0.0 ± 0.01
capping $\begin{smallmatrix}A\\T\end{smallmatrix}$	0.8 ± 0.4	0.5 ± 0.2	0.3 ± 0.1	0.0 ± 0.03

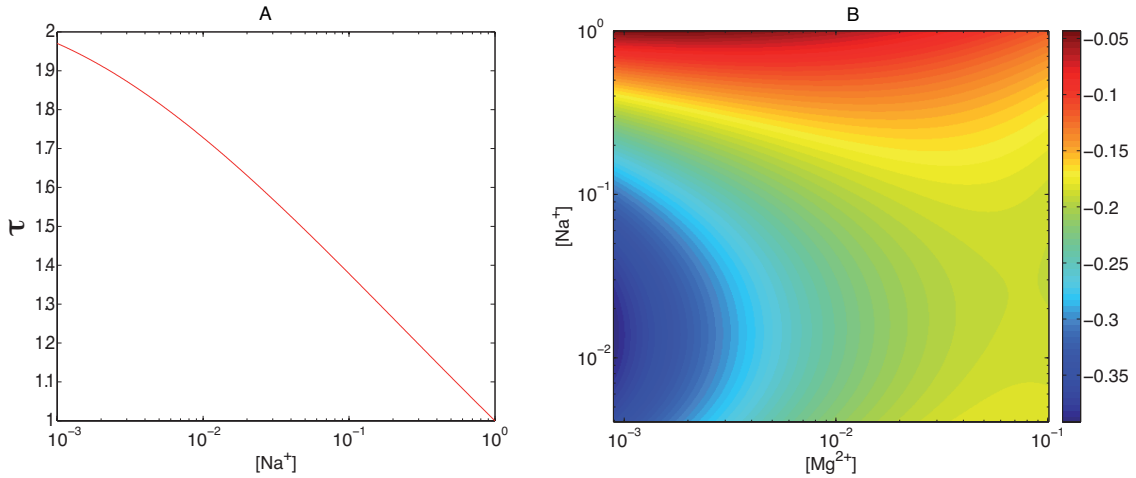

Figure 1.3 – (A) τ en fonction de la concentration en sel $[Na^+]$ (voir équation 1.37). (B) δs pour le segment CA/GT en fonction de la concentration en $[Na^+]$ et en $[Mg^{2+}]$ donnée par l'équation 1.38.

Une des particularités majeures de la correction en sel introduite dans l'équation 1.34 par rapport à celle de l'équation 1.31 est qu'elle tient compte de la nature même du segment de paires de bases. Ainsi les bases de type AT seront plus affectées (pertes plus importantes d'entropie) que celles de type GC par une baisse de la concentration en sel (voir table 1.2). Les régions riches en AT seront donc plus instables à des concentrations physiologiques ($[Na^+] \approx 154$ mM). En effet, la probabilité d'ouverture $P_{open}(XY)$ du segment XY est proportionnelle à $\exp\left[\beta \Delta g_{NN}(XY)\right]$. Soit, par exemple,

$$\frac{P_{open}(AT)}{P_{open}(GC)}([Na^+]) = \frac{P_{open}(AT)}{P_{open}(GC)}(1\,\mathrm{M}) \times \tau([Na^+]) \tag{1.35}$$

avec

$$\begin{aligned} \tau([Na^+]) &= \exp\Big\{\big[(\Delta h_{NN}(GC)(a_1+a_2) - \Delta h_{NN}(AT)a_2)\log[Na^+] \\ &\quad + (\Delta h_{NN}(GC) - \Delta h_{NN}(AT))\, a_3 \log^2[Na^+]\big]/k_B\Big\} \qquad (1.36) \\ &\approx \exp\left(-1.64\times 10^{-1}\log[Na^+] - 9.53\times 10^{-3}\log^2[Na^+]\right) \qquad (1.37) \end{aligned}$$

Sur la figure 1.3 A, on remarque que par rapport à 1 M, la probabilité d'ouverture de AT devient de plus en plus grande par rapport à celle de GC lorsque l'on diminue $[Na^+]$.

1.2.2.2 Effet combiné des ions mono- et divalents

La correction 1.34 rend compte uniquement des effets des ions monovalents (via $[Na^+]$). Récemment, Owczarzy et collaborateurs [100] ont également étudié l'effet d'un changement combiné de concentration en $[Na^+]$ et $[Mg^{2+}]$ dans la gamme $\{1\text{ mM} \leq [Na^+] \leq 1\text{ M}, 0.5\text{ mM} \leq [Mg^{2+}] \leq 125\text{ mM}\}$ et en ont tiré une relation phénoménologique équivalente à l'équation 1.32. De la même manière que précédemment, on peut en déduire une correction sur les paramètres locaux valable dans la gamme de concentration considérée

$$\begin{aligned} \Delta s([Na^+],[Mg^{2+}]) &= \Delta s(1\text{M } [Na^+]) + \Delta h\Big[b_1 + b_2\sqrt{[Na^+]}\log[Na^+] + b_3\log[Mg^{2+}] \\ &\quad + (b_4 + b_5\log[Na^+] + b_6\log^3[Na^+])\log^2[Mg^{2+}] \qquad (1.38) \\ &\quad + f_l(GC)\left\{b_7 + (b_8 + b_9\log[Na^+] + b_{10}\log^2[Na^+])\log[Mg^{2+}]\right\}\Big] \end{aligned}$$

Les paramètres b_i sont donnés dans la table 1.1. La figure 1.3 B représente la correction $\delta s = \Delta s([Na^+],[Mg^{2+}]) - \Delta s(1\text{M } [Na^+])$ pour le segment $\begin{smallmatrix}5'-CA-3'\\3'-GT-5'\end{smallmatrix}$. On remarque que la correction n'est pas symétrique par rapport à $[Na^+]$ et $[Mg^{2+}]$: ions monovalents et divalents contribuent différemment à la stabilisation du complexe. De plus, la compétition entre les ions $[Na^+]$ et $[Mg^{2+}]$ pour se lier au duplexe ou aux simple-brins ainsi que les effets de bord ou de saturation peuvent expliquer [100] les comportements surprenants de δs : apparition d'un minimum quand, à $[Mg^{2+}]$ fixée, on augmente $[Na^+]$ ou d'un maximum quand, à $[Na^+]$ fixée, on augmente $[Mg^{2+}]$.

1.2.3 Coopérativité

Il n'y a a priori aucune raison qui justifie que dans le modèle PS nous supposions que les facteurs de coopérativité n'aient pas de contribution enthalpique ou ne dépendent pas de la séquence. Cependant, le manque de précision sur leurs valeurs ne permet pas rendre compte de telles dépendances. Le facteur σ est dans la gamme $10^{-4} - 10^{-5}$ [46, 80] ce qui correspond à une énergie de nucléation $-k_BT\log\sigma \sim 10k_BT$. Everaers et collaborateurs [33] interprètent $-k_BT\log\sigma$ comme étant égale à deux fois un terme de fourche $-\gamma_S T$ décrivant l'énergie interfaciale à la jonction entre une partie double-hélice et une partie dénaturée. Ainsi, comme un domaine terminal dénaturé n'a qu'une jonction fourche, cela donne que $\bar{\sigma} \sim \sqrt{\sigma}$ [17]. Dans la suite, on supposera $\bar{\sigma} = 0.3\sqrt{\sigma}$. Cependant Garel et Orland [84] ont montré que tant que $\bar{\sigma}$ n'était pas nulle, sa valeur n'avait qu'une faible incidence sur les résultats du modèle. Les valeurs numériques des différents paramètres sont données dans la table 1.1. Le manque de données concernant σ nous force à supposer une barre d'erreur arbitrairement grande de 30% pour $\log\sigma$.

1.2.4 Propagation des erreurs

Tous les paramètres du modèle ne sont pas déterminés avec une précision absolue, mais sont connus plus ou moins une certaine barre d'erreur. Traditionnellement, les effets de ces incertitudes sur les prédictions faites par le modèle ne sont pas étudiés et on se limite à utiliser uniquement les paramètres standard. Mais cela est-il justifié ? Pour répondre à cette question, on s'intéresse donc à la propagation des erreurs faites sur les paramètres sur les résultats prédits par le modèle. Ce qui nous permettra en plus d'estimer des barres d'erreurs pour les prédictions. Pour une séquence donnée, la méthode consiste à calculer les différentes observables considérées (Θ, T_m, etc.) pour différents jeux de paramètres (typiquement une centaine) en cohérence avec les barres d'erreur individuelles de chacun d'eux, puis d'évaluer les déviations standard de ces prédictions.

Pour générer ces différents jeux de paramètres, on pourrait tout simplement pour chaque paramètre tirer aléatoirement une valeur distribuée de manière gaussienne autour de sa valeur standard suivant son écart-type. Cependant, certains paramètres sont fortement corrélés (par exemple Δh_{NN} et Δs_{NN} pour un même segment de paires sont corrélés à près de 99%) et il faut donc tenir compte de ces corrélations lors du tirage d'un nouveau jeu de paramètres. Pour cela, supposons que l'on connaisse $\mathbf{C}$ la matrice de covariance des n différents paramètres $\{x_i\}$

$$C_{i,j} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{1.39}$$

Le problème est de trouver n variables $\{X_i\}$ indépendantes entre elles et fonctions des $\{x_i\}$ afin d'effectuer des tirages indépendants sur les $\{X_i\}$ puis de revenir aux $\{x_i\}$. Cherchons les $\{X_i\}$ sous la forme

$$X_i = \sum_j a_{ij} x_j \tag{1.40}$$

On veut que les nouvelles variables ne soient pas corrélées, on impose donc

$$\langle (X_l - \langle X_l \rangle)(X_k - \langle X_k \rangle) \rangle = 0 \quad \forall l \neq k \tag{1.41}$$

On a donc $n(n-1)/2$ équations et a priori n^2 inconnues, les $\{a_{ij}\}$. On peut donc fixer $n(n+1)/2$ coefficients : on va supposer que la matrice $\mathbf{a}$ est triangulaire supérieure avec des 1 sur sa diagonale. Les équations (1.39-1.41) reviennent alors à résoudre

$$\mathbf{a_l} \cdot \mathbf{C} \cdot (\mathbf{a_k})^\dagger = 0 \quad \forall l \neq k \tag{1.42}$$

avec $\mathbf{a_l}$ la ligne l de la matrice $\mathbf{a}$. On résout ces équations "ligne par ligne", en débutant par $n-1$ puis $n-2$, et ainsi de suite, pour calculer les $\{a_{ij}\}$. Cela nous permet alors d'évaluer la valeur moyenne et l'écart-type des X_i

$$\langle X_i \rangle = \sum_j a_{i,j} \langle x_j \rangle \tag{1.43}$$

$$\sigma^2_{X_i} = \mathbf{a_i} \cdot \mathbf{C} \cdot (\mathbf{a_i})^\dagger \tag{1.44}$$

La génération d'un jeu de paramètres suit alors les 2 étapes suivantes : 1) Pour chaque X_i, tirer aléatoirement une valeur suivant la distribution gaussienne définie par sa valeur moyenne $\langle X_i \rangle$ et son écart-type σ_{X_i} (voir équations 1.43 et1.44). 2) Revenir aux $\{x_i\}$ en inversant la relation 1.40 : $\mathbf{x} = \mathbf{a}^{-1} \cdot \mathbf{X}$.

Reste dorénavant à estimer la matrice de covariance $\mathbf{C}$. Pour avoir accès à cette matrice, on utilise la méthode Bootstrap [93, 101]. Elle consiste à choisir aléatoirement 68 séquences différentes (63% du total) sur les 108, de tirer ensuite uniformément 40 séquences sur les 108 (il peut donc y avoir des doublons), puis d'appliquer la méthode de détermination des paramètres vue à la section 1.2.1 sur ces 108 données [43]. Les nouveaux paramètres sont alors distribués de manière gaussienne autour de leur valeur standard (celle donnée dans la table 1.1). Ces distributions permettent ainsi de calculer la matrice covariante avec l'équation 1.39, la moyenne étant effectuée sur les différents jeux de paramètres obtenus par la méthode Bootstrap (plus de 100000 essais ont été réalisés). En particulier, on peut déterminer la barre d'erreur de chaque paramètre ($\sqrt{C_{i,i}}$). Les résultats obtenus (voir table 1.1) dévient légèrement de ceux fournis par [43]. Ceci peut s'expliquer par le faible nombre d'essais de Bootstrap (30) réalisés par Allawi et SantaLucia, ce qui n'est pas suffisant pour permettre une évaluation fiable des erreurs.
La détermination des paramètres des corrections en sel a été réalisée indépendamment de celle des paramètres plus proche voisin, donc on suppose qu'il n'y a pas de corrélation entre ces deux types de paramètres. Par contre, des corrélations entre les paramètres des corrections en sel existent et sont considérées dans l'analyse de propagation des erreurs.
Pour les facteurs de coopérativité, le manque de donnée impose de négliger les possibles corrélations entre σ et les autres paramètres du modèle.

1.3 Comportement générique du modèle

Dans cette section, nous discutons de quelques aspects génériques des prédictions du modèle de Poland-Scheraga unifié. Dans un premier temps, nous étudions le comportement des courbes de dénaturation pour des séquences de différentes tailles et pour plusieurs concentrations en brin et en sel. Puis, dans un deuxième temps, nous nous intéressons à l'évolution de la structure interne des double-brins lors de la transition de dénaturation.

1.3.1 Évolution des courbes de dénaturation

La figure 1.4 représente l'évolution de l'observable $-\mathrm{d}\Theta/\mathrm{d}T$ pour des séquences aléatoires de différentes tailles en fonction de la concentration en brin et de la concentration en sel.

On remarque tout d'abord que la hauteur et la largeur des pics dépendent de la longueur de chaînes. Comme l'intégrale sur la courbe de dénaturation doit être constante (égale à 1) : plus l'intervalle de

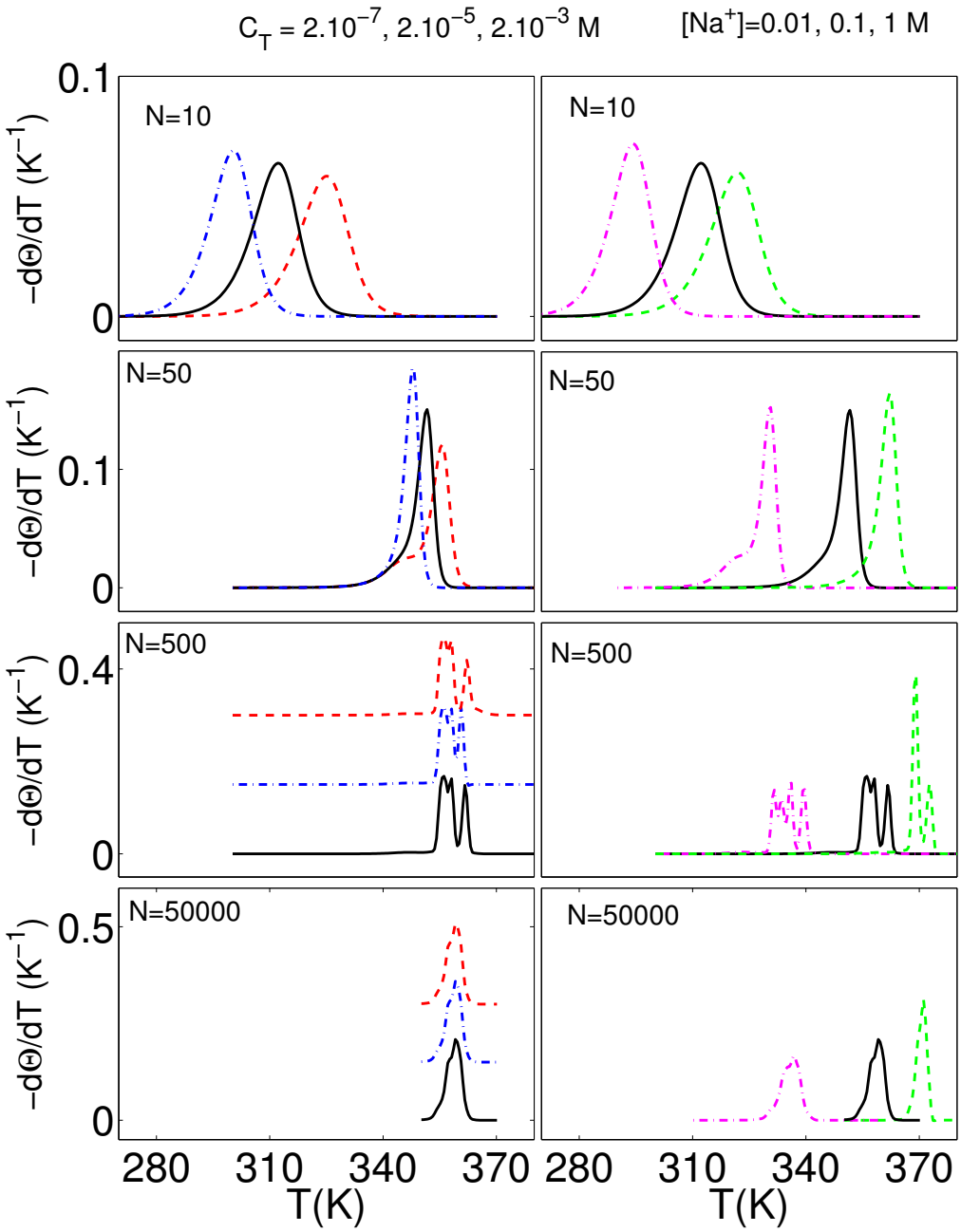

FIGURE 1.4 – $-\mathrm{d}\Theta/\mathrm{d}T$ en fonction de la température T pour différentes séquences aléatoires ($f(GC) \approx 0.5$: $N = 10$ (première ligne), $N = 50$ (deuxième ligne), $N = 500$ (troisième ligne) et $N = 50000$ (quatrième ligne). (A gauche) $[Na^+] = 0.1$ M et $c_T = 2 \times 10^{-3}$ M (lignes tiretées rouges), $c_T = 2 \times 10^{-5}$ M (lignes pleines noires) et $c_T = 2 \times 10^{-7}$ M (lignes pointillées bleues). Pour $N = 500$ et $N = 50000$, certaines courbes ont été décalées verticalement pour pouvoir être clairement distinguer. (A droite) $c_T = 2 \times 10^{-5}$ M et $[Na^+] = 1$ M (lignes tiretées vertes), $[Na^+] = 0.1$ M (lignes pleines noires) et $[Na^+] = 0.01$ M (lignes pointillées mauves).

température sur lequel a lieu la transition est petit, plus la hauteur moyenne des pics de transitions est grande. Alors que pour des oligomères courts ($N \sim 10$), cet intervalle est de l'ordre de $\Delta T \sim 40$ K et est centré autour de températures physiologiques (310 K), sa taille diminue pour des longs polymères ($\Delta T \sim 10K$ pour $N = 50000$). Ainsi, la transition de dénaturation devient plus abrupte quand la longueur des séquences augmente, indiquant une dénaturation de plus en plus coopérative. On remarque également que le nombre de pics varient avec N. Les courbes de dénaturation des oligomères ($N = 10$ et 50) exhibent un ou deux pics, signature d'une transition à deux ou trois états. Pour des petits ($N = 500$) polymères, la courbe contient plusieurs pics dus à l'ouverture successive de domaines. Pour des polymères très longs ($N = 50000$), $-\mathrm{d}\Theta/\mathrm{d}T$ ne contient plus qu'un seul pic à cause de la superposition d'un grand nombre d'ouvertures simultanées de domaines.

Les courbes de dénaturation pour plusieurs concentrations en brin montrent une influence forte de la concentration en brins c_T sur les oligomères mais un effet négligeable pour des chaînes plus longues. Comme $\Theta = \Theta_{int} \times \Theta_{ext}$, la transition globale est le résulat d'une compétition entre la dénaturation interne et la dissociation. La concentration c_T agit uniquement sur les effets dissociatifs (voir équation 1.10) via $x = \exp[\beta(\Delta G_0 - k_B T \log[c_T/c_0])]$, où $-k_B \log[c_T/c_0]$ peut être interprété comme l'entropie de translation des simples brins (voir équation 1.2). De plus fortes concentrations vont alors diminuer le gain en entropie de mélange des simples brins et favoriser les duplexes. L'importance relative de $-k_B T \log[c_T/c_0]$ par rapport à ΔG_0 est plus forte pour des chaînes courtes à faibles concentrations pour lesquelles $x >> 1$ et donc $\Theta_{ext} \sim 0$ donnant lieu à des transitions à deux états. Pour des oligomères plus longs à des plus fortes concentrations, la dénaturation interne concurrence la dissociation, et le caractère à deux états de la transition peut être perdu. Pour des plus longues chaînes, $|\Delta G_0| >> |k_B T \log[c_T/c_0]|$ soit $x << 1$, seule la dénaturation interne compte et les effets de dissociation ne jouent aucun rôle majeur dans la transition.

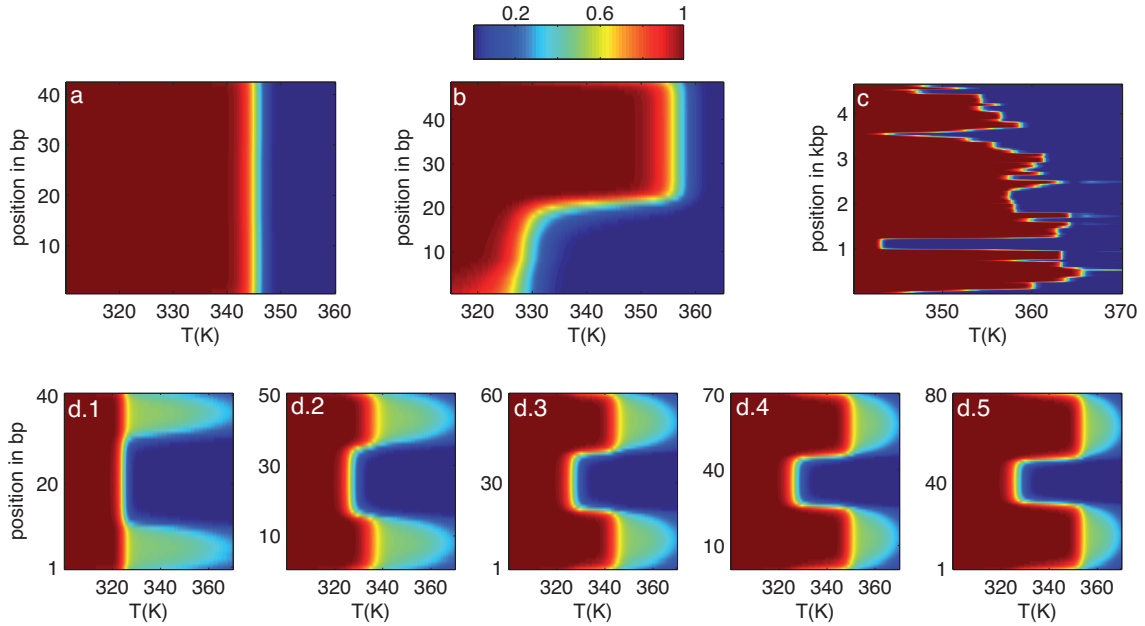

Figure 1.5 – (a,b,c) Probabilité individuelle totale $p(\alpha) \times \Theta_{ext}$ en fonction de la température T et de la position α de la paire de bases calculée avec les paramètres standard pour la séquence $L42B18$ (a, $N = 42$), $L48AS$ (b, $N = 48$) [47] et PN/MCS-13 (c, $N = 4660$) [46]. (d) Probabilité individuelle interne $p(\alpha)$ en fonction de la température T et de la position α de la paire de bases pour la séquence $G_n A_{20} G_n / C_n T_{20} C_n$ ($[Na^+] = 0.1$ M) pour $n = 10$ (1), 15 (2), 20 (3), 25 (4) et 30 (5).

Concernant l'influence de la concentration en sel, on observe qu'une variation de $[Na^+]$ conduit à un décalages identique des courbes de dénaturation pour toutes les gammes de tailles étudiées et à seulement quelques petites différences dans la forme des courbes. Les faibles concentrations ont tendance à stabiliser les bulles et les simples brins à cause d'un écrantage moins efficace par les contre-ions. De plus, les effets du sel dépendent de la séquence (équation 1.34). Ainsi, comme précisé dans la section 1.2.2, diminuer la concentration en sel favorise l'ouverture de domaines riches en AT.

1.3.2 Dénaturation interne

Un des atouts du modèle PS est la possibilité d'étudier le comportement individuel de chaque paire de bases (via le calcul de $p(\alpha)$). Ceci est d'autant plus avantageux que de nouvelles techniques expérimentales permettent d'avoir accès à la probabilité locale d'ouverture de la double-hélice [102, 103].

1.3.2.1 Ouverture successive de domaines

Sur les figures 1.5 a, b et c, la probabilité individuelle qu'une paire de base soit fermée $p(\alpha) \times \Theta_{ext}$ est tracée en fonction de la position de la paire et de la température. La figure (a) montre typiquement une transition à deux états où toutes les paires s'ouvrent simultanément. La figure (b) quant à elle illustre l'ouverture progressive d'un domaine terminal jusqu'à stabilisation d'un état avec les 20 premières bases ouvertes puis l'ouverture simultanée des bases restantes. Enfin, la figure (c) illustre une transition multiple d'un polymère court où l'on observe l'ouverture successive de domaines dans l'ADN. Cette méthode de visualisation permet une étude locale de la dénaturation et en particulier met en relief l'existence de domaines à l'intérieur de l'ADN ayant des propriétés de stabilité thermodynamique propres et dont les paires de bases qui le composent ont un comportement coopératif lors de sa dénaturation. On utilisera pratiquement cette propriété dans la section 2. La stabilité de ces domaines est à la fois dépendante de leur composition (un domaine riche en AT s'ouvrira avant un domaine riche en GC), de la concentration en sel et en brins (voir section précédente) mais aussi des domaines adjacents (par exemple une région interne riche en AT formera une bulle stable que si les bords qui l'entourent sont suffisamment longs, voir ci-dessous).

1.3.2.2 Cas d'un bulle interne riche en AT

Afin d'illustrer l'ouverture interne de domaine et d'estimer l'influence de la taille des bords sur la stabilité de la région interne, nous étudions ici le cas de séquences composées d'une région centrale riche en AT entourée de deux bords riches en GC via les exemples $\begin{smallmatrix} 5'-G_nA_lG_n-3' \\ 3'-C_nT_lC_n-5' \end{smallmatrix}$ avec n la taille d'un bord et l la taille du domaine central. On ne s'intéresse pas dans cette partie aux effets de dissociation ni à l'influence de c_T ou du sel, on regarde donc uniquement les observables internes comme $p(\alpha)$ calculée à $[Na^+] = 0.1$ M.

La figure 1.5 (d) illustre l'importance d'avoir des bords suffisamment longs pour observer, sur une gamme de température significative, un état avec une bulle interne stable (figure 1.6 A.2). Si les bords sont trop petits, l'ouverture de la bulle se fera en simultané avec un des bords et l'on observera une transition entre le double-brin totalement fermé (figure 1.6 A.1) et le double-brin où seul un bord est clos (figure 1.6 A.3). A partir de l'évolution de $p(\alpha)$, on peut déterminer pour chaque paire de bases sa température de dénaturation $T_m(\alpha)$ définie par $p(\alpha)(T_m(\alpha)) = 1/2$ et calculer la valeur moyenne de cette température T_m^{bulle} pour les bases de la région interne et T_m^{bord} pour celle des bords. La figure 1.6 B représente l'évolution de T_m^{bulle} et T_m^{bord} en fonction de n pour plusieurs l. On remarque que T_m^{bulle} converge rapidement vers une valeur limite T_{bulle}^{∞}, fonction de l (voir figure 1.7 A), alors que T_m^{bord} converge également vers une valeur limite $T_{bord}^{\infty} \sim 400$ K indépendante de l mais de manière beaucoup plus lente. Pour interpréter ces valeurs limites, on modélise la dénaturation interne par l'intermédiaire des trois états principaux : le double-brin totalement clos (figure 1.6 A.1), la bulle interne stable (figure

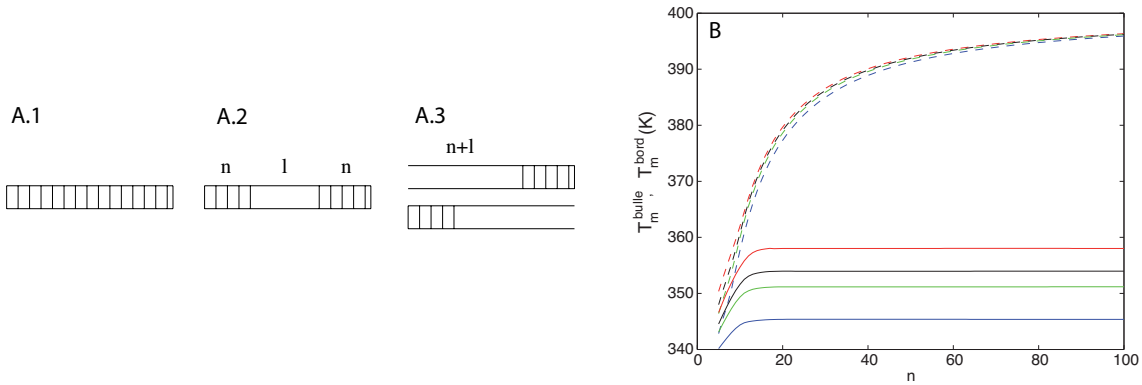

Figure 1.6 – (A) États principaux pour la dénaturation interne de $G_nA_lG_n/C_nT_lC_n$. (B) Température de dénaturation de la bulle interne T_m^{bulle} (lignes pleines) et des bords T_m^{bord} (lignes tiretées) en fonction de la taille n des bords pour plusieurs tailles l de bulles : 20 (rouge), 25 (noir), 30 (vert) et 50 (bleu).

1.6 A.2), et la bulle interne et un des bords ouvert (figure 1.6 A.3), définis par leur fonction de partition respective

$$Z_1 = 1 \tag{1.45}$$

$$Z_2 = \sigma l^{-c} \exp\left\{\beta\left[l\Delta g_{NN}\binom{5'-AA-3'}{3'-TT-5'} + \Delta g_{NN}\binom{5'-AG-3'}{3'-TC-5'} + \Delta g_{NN}\binom{5'-GA-3'}{3'-CT-5'}\right]\right\} \tag{1.46}$$

$$Z_3 = \bar{\sigma}(l+n)^{c'} \exp\left\{\beta\left[l\Delta g_{NN}\binom{5'-AA-3'}{3'-TT-5'} + n\Delta g_{NN}\binom{5'-GG-3'}{3'-CC-5'} + \Delta g_{NN}\binom{5'-AG-3'}{3'-TC-5'} + \Delta g_{NN}\binom{5'-GA-3'}{3'-CT-5'} + \omega\binom{G}{C}\right]\right\} \tag{1.47}$$

La probabilité pour que la bulle soit fermée vaut alors $\mathcal{P}_{bulle} = Z_1/(Z_1+Z_2+2Z_3)$ et par définition on a $\mathcal{P}_{bulle}(T_m^{bulle}) = 1/2$. Pour $n \to \infty$, comme $T_{bulle}^{\infty} < T_m\binom{5'-GG-3'}{3'-CC-5'} \equiv \Delta h_{NN}\binom{5'-GG-3'}{3'-CC-5'}/\Delta s_{NN}\binom{5'-GG-3'}{3'-CC-5'}$ = 400.6 K (voir figure 1.6 B), $\Delta g_{NN}\binom{5'-GG-3'}{3'-CC-5'} < 0$ et donc $Z_3 \to 0$. Ainsi, on a $Z_2(T_{bulle}^{\infty}) = 1$ ce qui donne

$$T_{bulle}^{\infty} = \frac{l\Delta h_{NN}\binom{5'-AA-3'}{3'-TT-5'} + \Delta h_{NN}\binom{5'-AG-3'}{3'-TC-5'} + \Delta h_{NN}\binom{5'-GA-3'}{3'-CT-5'}}{l\Delta s_{NN}\binom{5'-AA-3'}{3'-TT-5'} + \Delta s_{NN}\binom{5'-AG-3'}{3'-TC-5'} + \Delta s_{NN}\binom{5'-GA-3'}{3'-CT-5'} + k_B(c\log l - \log\sigma)} \tag{1.48}$$

Sur la figure 1.7 A, on remarque que l'équation précédente rend bien compte des résultats issus du modèle PS unifié pour les grandes valeurs de l. Quand l diminue, les effets de bords des bulles deviennent de plus en plus importants et l'approximation simple utilisée n'est plus valable. On remarque également que T_{bulle}^{∞} tend vers la température de fusion par paire de base $T_m\binom{5'-AA-3'}{3'-TT-5'} = 335.9$ K quand $l \to \infty$. De même que pour T_{bulle}^{∞}, si l'on s'intéresse à la probabilité qu'un bord (disons le premier par exemple) soit fermé $\mathcal{P}_{bord} = Z_3/(Z_1+Z_2+2Z_3)$, quand $n \to \infty$, $\mathcal{P}_{bord}(T_{bord}^{\infty}) = 1/2$ implique $Z_3 \to \infty$, soit $\Delta g_{NN}\binom{5'-GG-3'}{3'-CC-5'} > 0$, d'où $T_{bord}^{\infty} = T_m\binom{5'-GG-3'}{3'-CC-5'}$ qui est bien indépendant de l.

Le plateau observé pour T_m^{bulle} (figure 1.6B) correspond à la stabilisation de la bulle, la taille des bords ne jouant plus sur l'ouverture de la région A_l. La valeur de n pour laquelle, T_{bulle}^{∞} est atteinte à 99.9% près, est donc la taille minimum des bords $n_{min}(l)$ pour avoir un état intermédiaire avec bulle stable (voir figure 1.7 B). Si $n \gtrsim n_{min}$ la stabilisation de la bulle interne sera observée sur un intervalle significatif de température, par contre si $n < n_{min}$, la bulle aura tendance à s'ouvrir en même temps qu'un des bords.

Le fait de rajouter les effets associatifs (c'est à dire, de s'intéresser à $p(\alpha) \times \Theta_{ext}$ au lieu de $p(\alpha)$) ne changent pas l'allure des courbes, ni l'influence des divers paramètres. La principale modification est

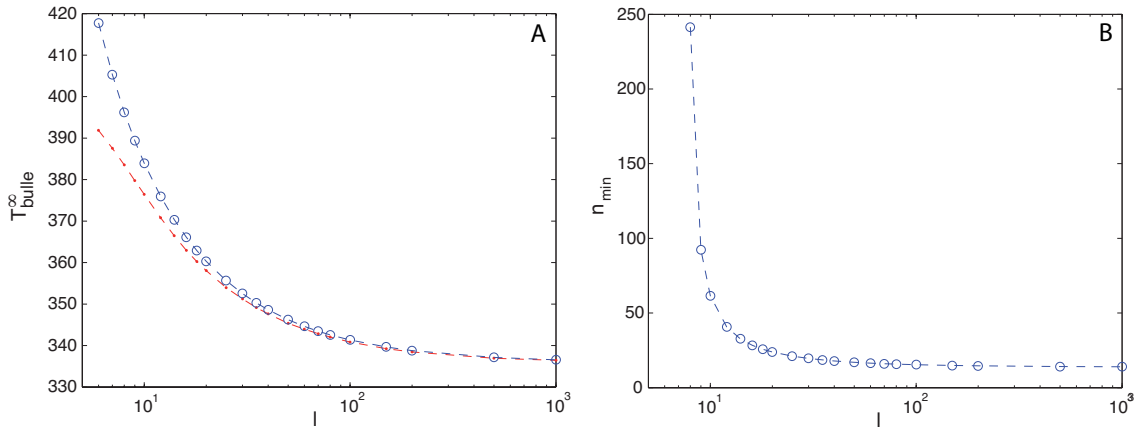

FIGURE 1.7 – (A) Température d'ouverture de la bulle interne T^{∞}_{bulle} en fonction de la taille l de la bulle interne calculée avec le modèle PS unifié (ronds bleus) ou d'après l'équation 1.48 (points rouges). (B) Taille minimale n_{min} des bords permettant d'observer la stabilité de la bulle de taille l sur un intervalle significatif de température.

d'augmenter n_{min} par rapport à précédemment : les effets associatifs ont tendance à séparer le double brin pour des températures inférieures aux températures internes de fusion des différents domaines.

1.4 Pouvoir de prédiction du modèle

Dans cette partie, nous comparons des données expérimentales aux prédictions faites par le modèle de Poland-Scheraga unifié pour différentes tailles d'ADN. On s'attend à avoir un bon accord pour les oligomères courts, la gamme de taille de chaînes utilisées pour la paramétrisation du modèle. Les comparaisons pour des séquences plus grandes constituent alors des tests cruciaux pour la validation du modèle.

1.4.1 Oligomères courts ($\sim$ 10 bp)

Pour les oligomères courts [43, 42, 94] que nous avons utilisé lors de la paramétrisation du modèle PS unifié, on trouve une erreur moyenne sur les températures de dénaturation prédites T_m^{th} par rapport à celles expérimentales T_m^{exp}

$$\langle \Delta T_m \rangle = \frac{1}{N_s} \sum_i |T_m^{exp} - T_m^{th}| = 1.7 \text{ K} \tag{1.49}$$

En utilisant la correction en sel de SantaLucia [30] (équation 1.31), on a $\langle \Delta T_m \rangle = 2.4$ K, ce qui justifie l'utilisation de notre correction en sel. De plus, l'étude de la propagation des erreurs donne une déviation standard moyenne des prédictions pour chaque T_m d'environ 2 K (avec une erreur expérimentale de 0.3 K). Une comparaison plus détaillée entre T_m^{th} et T_m^{exp} (figure 1.8 A) montre qu'il n'y a pas de concentration en sel ou en brin privilégiée sur toute la gamme accessible expérimentalement $[Na^+] \in [0.01, 1]$ M et $c_T \in [10^{-6}, 10^{-3}]$ M : tous les points sont uniformément concentrés autour de la bissectrice avec l'erreur faite sur chaque prédiction qui est de l'ordre de la déviation typique par rapport aux données expérimentales.

En plus, de la valeur absolue de la température de dénaturation, le modèle reproduit aussi fidèlement la courbe complète de dénaturation (voir figure 1.8 B). Le comportement à deux états de la transition peut être évalué en calculant le maximum Σ_{max} de la fonction $\Sigma = \Theta_{ext} - \Theta$ [47]. Si la séquence présente

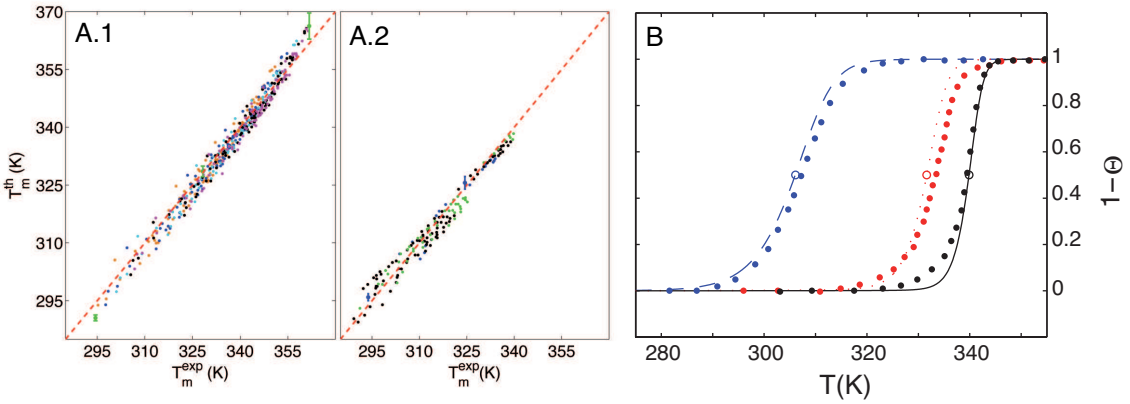

Figure 1.8 – (A) Température de dénaturation T_m^{th} prédite par le modèle en fonction de celle expérimentale T_m^{exp} pour (A.1) 92 séquences étudiées dans [94] avec $c_T = 2 \times 10^{-6}$ M et pour 5 concentrations en sel différentes : $[Na^+] = 69$ mM (orange), 119 mM (bleu), 220 mM (cyan), 621 mM (noir) et 1.02 M (violet) (les points verts sont représentatifs des déviations standard estimées par la méthode de propagation des erreurs, voir section 1.2.4) ; et pour (A.2) 20 séquences étudiées dans [42] avec $[Na^+] = 1$ M et pour plusieurs gammes de concentrations en brins différentes : $c_T \in [10^{-6}, 10^{-5}]$ M (bleu), $[10^{-5}, 10^{-4}]$ M (noir) et $[10^{-4}, 10^{-3}]$ M (vert). (B) Courbes de dénaturation calculées (lignes) et expérimentales [94] (points) pour des oligomères courts ($c_T = 2 \times 10^{-6}$ M et $[Na^+] = 69$ mM) : $ATCGTCTGGA$ (bleu), $TACTTCCAGTGCTCAGCGTA$ (rouge) et $TCGGAGAAATCACTGAGCTGCCTGAGAAGA$ (noir).

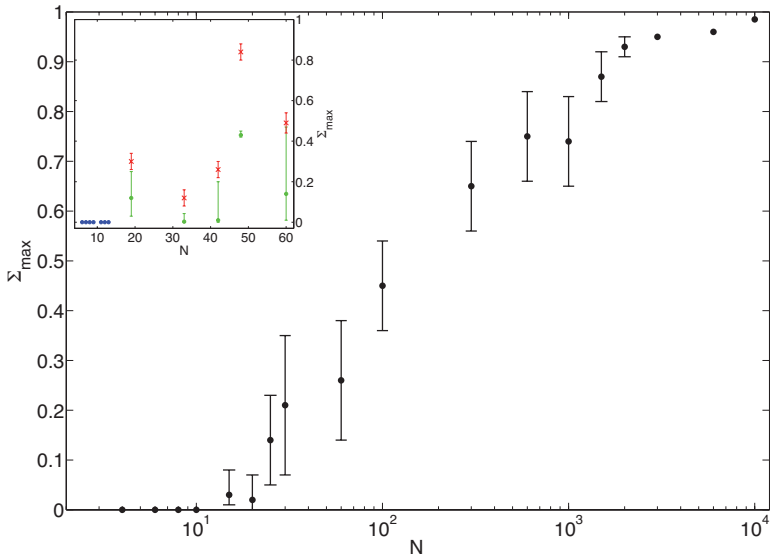

Figure 1.9 – Écart maximal Σ_{max} entre Θ_{ext} et Θ pour des séquences aléatoires de plusieurs tailles avec $c_T = 2 \times 10^{-4}$ M et $[Na^+] = 0.1$ M (points noirs). (Encart) Σ_{max} pour les oligomères étudiés par Zeng et collaborateurs [47] avec $c_T = 2 \times 10^{-6}$ M et $[Na^+] = 0.05$ M (croix rouges : résultats expérimentaux, points verts : prédictions) et pour des oligomères courts étudiés par Allawi et SantaLucia [43] avec $c_T = 2 \times 10^{-6}$ M et $[Na^+] = 1$ M (points bleus). Pour chaque point, nous avons tracé les déviations standard dues à la propagation des erreurs lorsqu'elles étaient plus grandes que la taille du point.

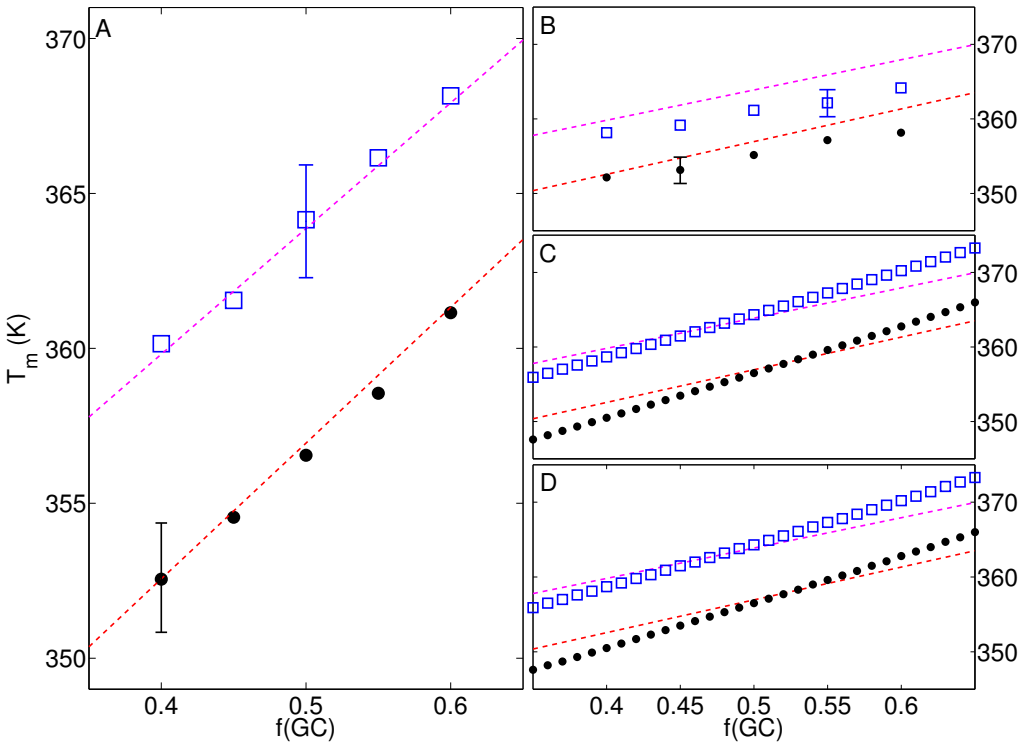

FIGURE 1.10 – (A, B) Température de dénaturation T_m calculée avec le modèle PS unifié pour plusieurs hétéropolymères aléatoires ($N = 50000$) de contenu GC $f(GC)$ différents pour un concentration en sel de 74.5 mM (points noirs) et de 220 mM (carrés bleus). La correction en sel utilisée est celle définie par l'équation 1.34 (A) ou l'équation 1.31 (B). Les barres d'erreur tracées sont représentatives des déviations standard observées pour T_m. (C,D) Température de dénaturation T_m théorique calculée pour des séquences aléatoires si la transition était à deux états (C) ou si l'entropie de nucléation était nulle ($\log \sigma = 0$) (D). Nous avons utilisé ici la correction en sel de l'équation 1.34. Les lignes pointillées représentent les relations empiriques de Frank-Kamenetskii définies par l'équation 1.50.

une transition à deux états, $\Theta_{ext} = \Theta$, soit $\Sigma_{max} \approx 0$. Au contraire, pour de longues chaînes, $\Theta_{ext} \sim 1$, soit $\Sigma_{max} \sim 1$. La figure 1.9 confirme bien que nous reproduisons le comportement à deux états des transition des oligomères courts. Notons que la comparaison est faite avec le modèle complet incluant les fluctuations, qui ont été négligées lors de la paramétrisation. Négliger les facteurs de coopérativité (c'est à dire, mettre $\log \sigma = \log \bar{\sigma} = 0$) conduirait à des résultats drastiquement différents [33].

1.4.2 Polymères longs (≥ 10 kbp)

Dans la limite opposée des très longues chaînes d'ADN, les courbes de dénaturation redeviennent relativement lisses (voir figure 1.4) et peuvent ainsi être également caractérisée par une température de dénaturation, qui dépend essentiellement du contenu en bases GC ($f(GC)$) de la séquence. Il y a plus de 30 ans, Frank-Kamenetskii et collaborateurs [69, 104] proposaient la relation empirique

$$T_m = T_m^{AT} + f(GC)\left(T_m^{GC} - T_m^{AT}\right) \tag{1.50}$$

avec

$$\begin{cases} T_m^{AT} &= (355.55 + 7.9 \log[Na^+])\ \text{K} \\ T_m^{GC} &= (391.55 + 4.89 \log[Na^+])\ \text{K} \end{cases}$$

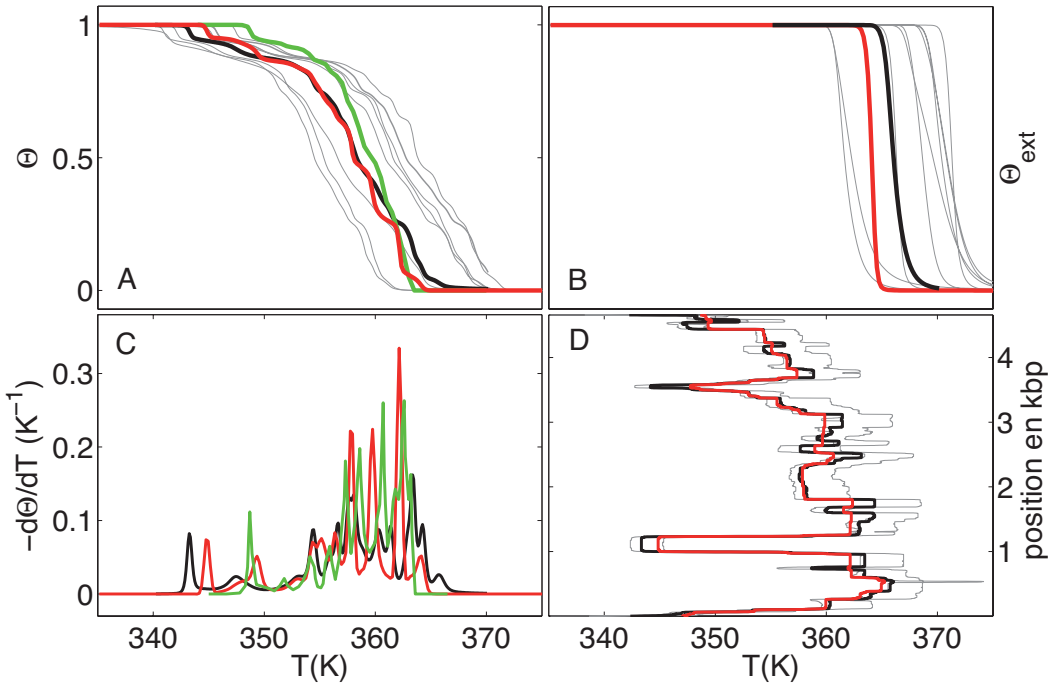

Figure 1.11 – Courbes de dénaturation Θ (A), Θ_{ext} (B), $-\mathrm{d}\Theta/\mathrm{d}T$ (C) et évolution de la température de dénaturation locale (D) pour PN/MCS-13 [46] pour une concentration en sel de $[Na^+] = 74.5$ mM. Les lignes vertes représentent les données expérimentales, les lignes noires les résultats calculés par le modèle avec les paramètres standard (table 1.1) et les autres lignes colorées (rouges et grises) les résultats calculés avec d'autres jeux de paramètres aléatoires.

Nous avons généré plusieurs séquences ADN aléatoires de longueur $N = 50000$ avec un contenu GC compris dans l'intervalle $[0.4, 0.6]$. La figure 1.10 A montre un accord excellent entre nos résultats calculés avec la correction en sel introduite dans la section 1.2.2 et l'équation 1.50, alors que ceux issus de la correction en sel de SantaLucia (équation 1.31) dévient de la courbe empirique (figure 1.10 B). Cette accord représente une validation essentielle de notre modèle unifié (comprenant la correction en sel) puisqu'il démontre qu'à partir d'une paramétrisation faite à l'aide de données sur des oligomères courts et d'un modèle de mécanique statistique, on retrouve de manière systématique les comportements expérimentaux de très longs ADN. En particulier, la figure 1.10 A fournit des preuves évidentes de la validité de notre correction en sel pour les paramètres plus proche voisin.

Sur les figures 1.10 C et D, on compare les relations de Frank-Kamenetskii avec l'évolution de T_m calculée pour notre correction en sel, si l'on suppose que la transition est à deux états ou que la pénalité de nucléation de bulle est nulle ($\log \sigma = 0$). Les déviations observées démontrent que ces relations empiriques ne peuvent pas se retrouver trivialement mais nécessite la résolution complète du modèle PS unifié en prenant compte en les fluctuations thermiques et la coopérativité.

1.4.3 Polymères courts (100 bp $\leq$ ... $\leq$ 10 kbp)

Les courbes de dénaturation des polymères courts montrent une structure assez riche qui peut être discuter indépendamment de la concentration en brins d'ADN (voir figure 1.4). Pour tester le pouvoir de prédiction du modèle pour cette gamme de longueur, nous avons choisi d'étudier la séquence PN/MCS-

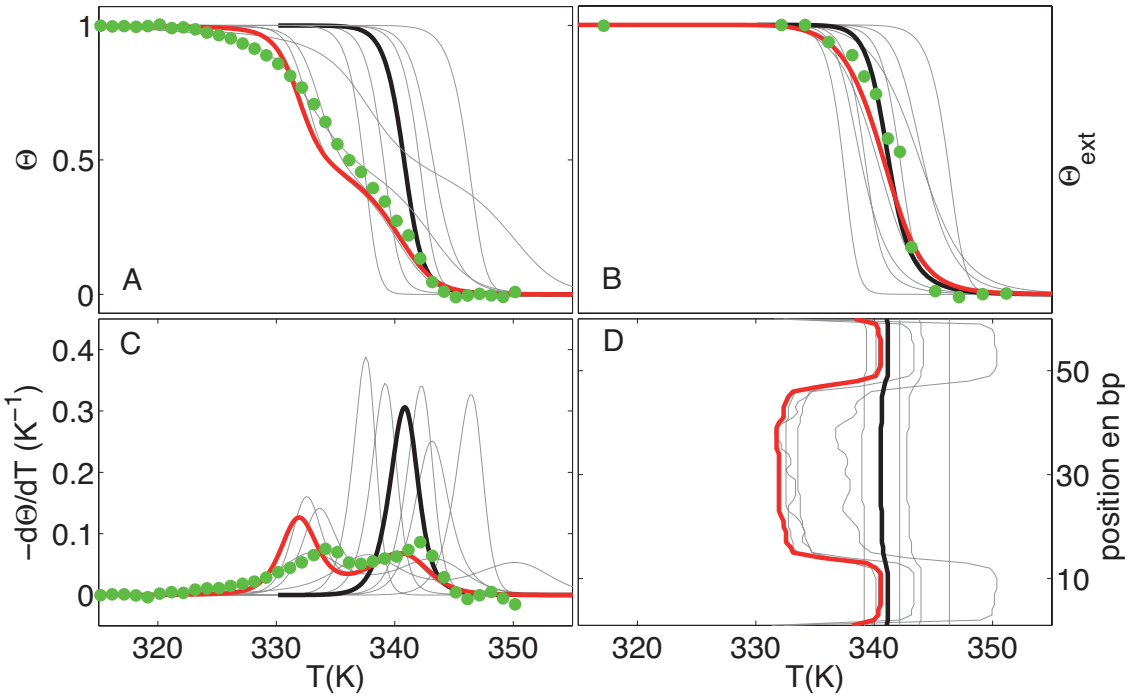

FIGURE 1.12 – Courbes de dénaturation Θ (A), Θ_{ext} (B), $-\mathrm{d}\Theta/\mathrm{d}T$ (C) et évolution de la température de dénaturation locale (D) pour la séquence L60B36 étudiée par Zeng et collaborateurs [47] pour une concentration en sel de $[Na^+] = 50$ mM et une concentration en brin de $c_T = 2 \times 10^{-6}$ M. Les points verts représentent les données expérimentales, les lignes noires les résultats calculés par le modèle avec les paramètres standard (table 1.1) et les autres lignes colorées (rouges et grises) les résultats calculés avec d'autres jeux de paramètres aléatoires. La ligne rouge souligne un jeu de paramètres qui reproduit quantitativement les données expérimentales. Les données expérimentales ont été normalisées par la procédure décrite dans l'annexe 6.2.

13 [3]. La figure 1.11 montre que le modèle de Poland-Scheraga unifié reproduit assez bien la température de mélange de PN/MCS-13, mais échoue à prédire la structure fine des courbes. En particulier, le modèle ne décrit que qualitativement la courbe de dénaturation différentielle expérimentale (figure 1.11 C). D'autres programmes bioinformatiques tels que DINAmelt [88] ou MELTSIM [105] donnent des résultats similaires pour cette séquence.

Comment dès lors interpréter ces déviations ? En plus des courbes calculées avec les paramètres standard, la figure 1.11 contient d'autres courbes calculées avec des jeux de paramètres tirés aléatoirement dans les barres d'erreurs corrélées issues de la paramétrisation (voir section 1.2.4). Ces autres courbes marchent aussi bien (ou aussi mal) que les courbes standard. En particulier, les déviations par rapport aux données expérimentales sont contenues dans l'incertitude des prédictions théoriques et ne révèlent pas une défaillance profonde du modèle. Par exemple, les courbes rouges de la figure 1.11 illustrent les résultats issus d'un jeu de paramètres décrivant assez bien les courbes expérimentales.

Cela soulève la question de savoir si le modèle a un quelconque pouvoir de prédiction pour décrire la dénaturation des polymères (en dehors de la seule relation 1.50). En regardant les différentes courbes obtenues pour plusieurs jeux de paramètres, la réponse semble être négative. Cependant, si l'on regarde ce qu'il se passe localement, les prédictions faites par le modèle semblent assez robustes. A partir de la

3. PN/MCS-13 est une séquence de 4660 bp composée par pBR322 et par une séquence répétitive de 245 bp [46]. pBR322 est associée au numéro d'accession primaire J01749.

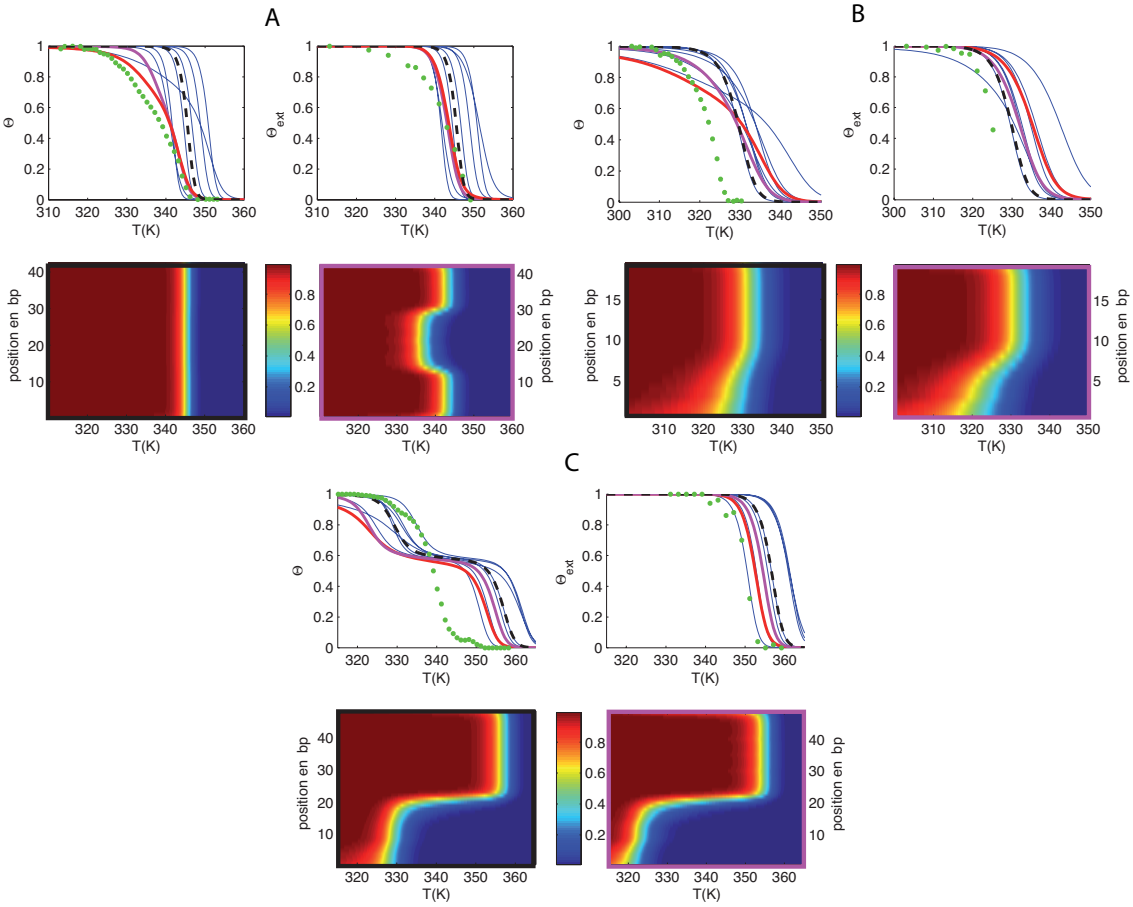

Figure 1.13 – Courbes de dénaturation pour les séquences L42B18 (A), L19AS2 (B) et L48AS (C) dans une solution à 50 mM en sel et contenant $c_T = 2 \times 10^{-6}$ M d'ADN. Les points verts représentent les données expérimentales. Les lignes donnent les résultats calculés pour les paramètres standard (pointillés noirs) ou pour d'autres jeux de paramètres. Les lignes rouges et violettes soulignent des paramètres qui permettent de bien reproduire les données pour L60B36. On a tracé également les probabilités $p(\alpha)\Theta_{ext}$ pour les paramètres standard (encadré noir) ou pour les paramètres de la courbe violette (encadré violet).

probabilité qu'une certaine paire de base soit fermé $p(\alpha) \times \Theta_{ext}$ (voir figure 1.5 c), on peut calculer pour chaque paire, sa température locale de dénaturation définie par $(p(\alpha)\Theta_{ext})(T_m(\alpha)) = 1/2$. La figure 1.11 D montre l'évolution de $T_m(\alpha)$ pour différents jeux de paramètres. On remarque que l'allure de la courbe reste plus ou moins identique : la définition de domaines formés par les paires de bases consécutives s'ouvrant simultanément est robuste [106], seule la valeur absolue de la température de dénaturation de ces domaines fluctue. Les erreurs faites sur les températures locales sont d'ailleurs du même ordre que celles réalisées pour les oligomères courts (~ 2 K).

1.4.4 Oligomères longs (20 bp $\leq$ ... $\leq$ 100 bp)

Alors que les oligomères courts ont pour la plupart une transition à deux états, cela n'est plus le cas pour des oligomères plus longs. Comparée au cas des polymères (section précédente), la difficulté vient de la dépendance des résultats en la concentration en brins (voir figure 1.4). Récemment, la communauté scientifique s'est concentrée sur cette gamme de longueur à cause de résultats expérimentaux particulièrement intéressants fournis par l'équipe de Giovanni Zocchi [47, 48]. Zocchi et collaborateurs

ont développé un protocole expérimental permettant de mesurer le degré d'association Θ_{ext} indépendamment de l'observable classique Θ. Dans la figure 1.12, on compare les résultats expérimentaux et ceux calculés avec le modèle pour la séquence L60B36, un oligomère de 60 bp ayant un domaine central riche en AT [47]. Les paramètres standard prédisent de manière incorrecte une transition à deux états. Cependant, d'autres combinaisons de paramètres reproduisent correctement l'ouverture d'une bulle interne avant la séparation des brins. La figure 1.13 A montre également que les prédictions théoriques (dans la limite de confiance) rendent compte des courbes de dénaturation d'une autre séquence L42B18 avec bulle interne pour le même jeu de paramètres que L60B36.

Cependant, les deux séquences L48AS et L19AS2 [47] (figures 1.13 B et C) contenant des domaines terminaux riches en AT ne sont bien décrites par aucun jeux de paramètres. Ces résultats pourraient être dus à une défaillance de la modélisation concernant un effet oublié par le modèle PS et qui jouerait un rôle important pour ces séquences. Peyrard et collaborateurs [103] interprètent ce phénomène par l'existence d'une intéraction non-locale par laquelle une région riche en $\frac{A}{T}$ pourrait influer sur l'ouverture de paires de bases situées juqu'à 10 bp de cette région. La cause physico-chimique de cet effet demeure cependant assez floue.

Dans tous les cas, on remarque une très forte sensibilité des prédictions par rapport à l'incertitude sur la paramétrisation pour cette gamme de taille d'ADN. En particulier, la figure 1.9 compare les valeurs théoriques et expérimentales de $\Sigma_{max} = \max(\Theta_{ext} - \Theta)$ pour les oligomères longs étudiés par l'équipe de Zocchi. L'existence de grandes barres d'erreur asymétriques souligne cette forte sensibilité du comportement des courbes de dénaturation. Par contre, comme pour les polymères, la prédiction de la structure interne semble assez robuste.

1.5 Conclusion

1.5.1 Bilan

Dans les sections précédentes, nous avons présenté un modèle unifié de Poland-Scheraga pour la dénaturation thermique de l'ADN. Contrairement aux autres formulations existantes, notre description couvre toutes les gammes de tailles d'ADN, des oligomères courts aux polymères longs, et est applicable pour toutes les concentrations en sel ou brins usuellement étudiées expérimentalement. En particulier, nous avons utilisé ce modèle pour discuter différents aspects génériques de la dénaturation thermique. Nous avons mis en relief la cohérence de notre formalisme en montrant le lien systématique qu'il existait entre la paramétrisation faite avec des oligomères courts et les résultats du modèle pour des très longues séquences. Dans la marge d'erreur imposée par la paramétrisation, le modèle reproduit les données expérimentales pour des ADN de tailles arbitraires incluant le cas des longs oligomères. Cependant, pour ce type de chaînes (et aussi pour les polymères courts), les prédictions faites pour les observables expérimentales sont particulièrement sensibles aux incertitudes rémanentes de la paramétrisation. Ce dernier point souligne que le pouvoir de prédiction de modèle pour ce genre de séquences est assez limité même si les prédictions locales semblent être plus robustes.
Ce travail a abouti à la publication d'un article dans *Biophysical Journal* [107].

1.5.2 Influence de chaque paramètre

La figure 1.14 montre comment l'incertitude sur la connaissance de paramètres (ou de classes de paramètres) particuliers affecte les prédictions des courbes de dénaturation selon les différents régimes de tailles étudiés. Dans la plupart des cas, la contribution dominante vient des 20 paramètres plus proche voisin d'association. Le paramètre d'initiation n'influe que sur les oligomères où les termes de bords ne sont pas négligeables. Le capping quant à lui ne joue un rôle que pour les oligomères longs exhibant une transition à plusieurs états où au moins un des bords va s'ouvrir stablement. Les facteurs de coopérativité ont une contribution importante pour les polymères mais aussi pour les oligomères

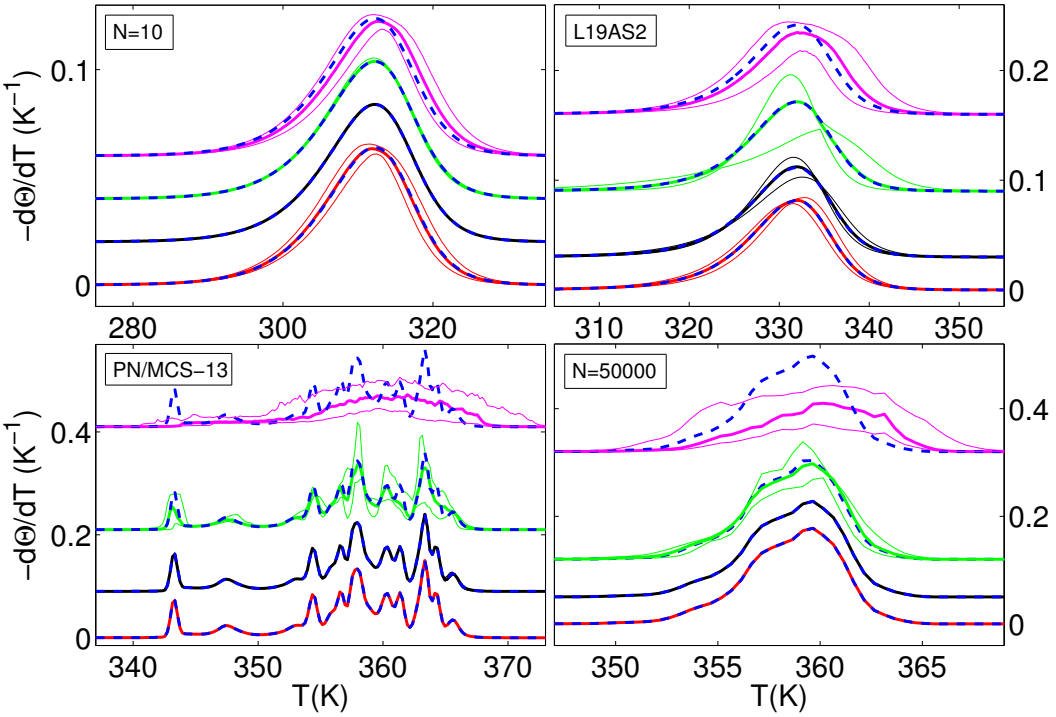

Figure 1.14 – Courbes de dénaturation $-\mathrm{d}\Theta/\mathrm{d}T$ pour des oligomères courts (aléatoire, $N = 10$) ou longs (L19AS2) et pour des polymères courts (PN/MCS-13) ou longs (aléatoire, $N = 50000$). Pour chaque séquence, on fait varier différents paramètres : initiation (rouge), capping (noir), coopérativité (vert) et paramètres plus proche voisin (violet). Les lignes épaisses représentent les courbes moyennées sur les différentes réalisations de paramètres, les lignes fines estiment la déviation standard sur chaque classe de paramètres et les lignes pointillées bleues sont les courbes calculées avec les paramètres standard.

longs où la compétition entre dénaturation interne et dissociation est forte.

1.5.3 Améliorer la paramétrisation du modèle

Au vue des différentes contributions de chaque type de paramètres, dont certaines ne sont pas négligeables, dans l'incertitude des courbes de dénaturation, on se pose la question de savoir comment on peut améliorer la paramétrisation du modèle pour diminuer les erreurs ou les incertitudes des prédictions.

Pour les oligomères courts, seuls les paramètres plus proche voisin jouent un rôle significatif, c'est pourquoi, bien sûr, on les utilise pour estimer ces paramètres. Insistons sur le fait que les relations de Frank-Kamenetskii (équation 1.50) ne peuvent pas être utilisées directement pour déterminer (comme étant une combinaison linéaire de) les paramètres microscopiques du modèle PS [90]. En effet, elles représentent une moyenne sur un grand nombre d'ouvertures coopératives de domaines et ne peuvent pas être trivialement décrites par une moyenne sur les paramètres plus proche voisin (voir figures 1.10 C et D). Cependant, les figures 1.10 A et 1.14 suggèrent qu'une comparaison avec les températures de dénaturation calculées avec le modèle complet pourrait être un excellent moyen (bien qu'assez lourd en calcul) pour affiner l'évaluation des paramètres plus proche voisin.

Pour les facteurs de coopérativité σ et $\bar{\sigma}$, la forte sensibilité des longs oligomères (ou de polymères courts) par rapport à ces paramètres, doit être utilisée pour minimiser les erreurs lors du processus de paramétrisation (voir figures 1.9 et 1.14). En effet, plus une observable va être sensible à la valeur

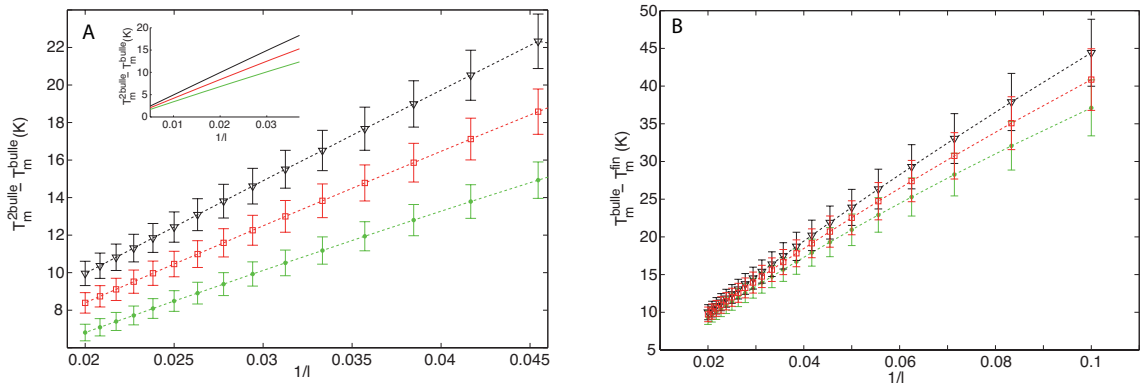

FIGURE 1.15 – (A) Évolution de la différence de température de dénaturation $T_m^{2bulle} - T_m^{bulle}$ entre $\mathcal{S}_{2bulle}$ et $\mathcal{S}_{bulle}$ en fonction de $1/l$ pour trois valeurs de γ_S : -10.9 cal/mol/K (triangles noirs), -8.9 cal/mol/K (carrés rouges) et -6.9 cal/mol/K (points verts). Les barres d'erreurs représentent l'incertitude dues aux autres paramètres. (B) Évolution de la différence de température de dénaturation $T_m^{bulle} - T_m^{fin}$ entre $\mathcal{S}_{bulle}$ et $\mathcal{S}_{fin}$ en fonction de $1/l$ pour trois valeurs de ΔS_{mix}^0 : 5 cal/mol/K (triangles noirs), 3 cal/mol/K (carrés rouges) et 1 cal/mol/K (points verts). Les barres d'erreurs représentent l'incertitude due aux autres paramètres.

d'un paramètre, plus elle sera utile pour le paramétrer [108]. La meilleure stratégie pour déterminer leurs valeurs serait donc de concevoir des expériences de dénaturation comparatives qui permettraient d'isoler leurs effets, en suivant les mêmes idées que Blake et Delcourt [46] (voir section suivante).

Il faut être conscient tout de même qu'il existera toujours des barres d'erreur inhérentes au modèle qui reflètent les hypothèses simplificatrices que l'on a faites. Pour réduire ses erreurs intrinsèques, une solution serait d'accroître le nombre de paramètres afin de décrire des effets chimiques ou physiques qui ne sont pas pris en compte ici (comme l'élasticité des brins, la dépendance de la coopérativité en la composition de la bulle, etc.). Néanmoins, cela rendrait le modèle difficilement paramétrable et plus lourd en calcul.

Ainsi, une amélioration de la paramétrisation du modèle pourrait passer par une modélisation simultanée d'une base de donnée étendue comprenant des séquences de toutes tailles. Cette option est d'autant plus envisageable que le modèle de Poland-Scheraga unifié permet de décrire tous les cas avec le même formalisme de mécanique statistique.

1.5.4 Paramétrer σ et ω

Dans cette partie, nous développons quelques idées utiles pour envisager une amélioration de la paramétrisation de σ et ω.

Pour paramétrer σ (via $\gamma_S = k_B(\log\sigma)/2$), on propose d'étudier deux types de séquences : $\mathcal{S}_{bulle} = {}^{G_n A_l G_n}_{C_n T_l C_n}$ (une bulle centrale) et $\mathcal{S}_{2bulle} = {}^{G_{2n/3} A_{l/2} G_{2n/3} A_{l/2} G_{2n/3}}_{C_{2n/3} T_{l/2} C_{2n/3} T_{l/2} C_{2n/3}}$ (deux bulles internes) avec n suffisamment grand ($n = 100$) pour pouvoir négliger les effets de bords (voir section 1.3.2) et les effets de dissociation. Suivant l'idée de Blake et Delcourt [46], pour chaque séquence, on calcule la température de fusion des bulles (respectivement T_m^{2bulle} et T_m^{bulle} pour $\mathcal{S}_{2bulle}$ et $\mathcal{S}_{bulle}$). Comme les deux types de séquences ont la même composition en paires de bases, on s'attend à ce que la différence $T_m^{2bulle} - T_m^{bulle}$ ne soit pas trop sensible aux paramètres plus proche voisin mais qu'elle le soit vis à vis de la différence d'énergie entre les deux séquences qui vaut environ $-2\gamma_S T$ (voir équation 1.48). La figure 1.15 A montre l'évolution de $T_m^{2bulle} - T_m^{bulle}$ en fonction de $1/l$ pour différentes valeurs de γ_S, les barres d'erreurs étant calculées en faisant fluctuer tous les paramètres sauf γ_S. On remarque que l'observable

est d'autant plus sensible à γ_S que $1/l$ augmente. Ainsi, une bonne série d'expériences sur $\mathcal{S}_{2bulle}$ et $\mathcal{S}_{bulle}$ devra être réalisée avec des bulles petites (il faut néanmoins que l soit suffisamment grande pour observer la transition vers un état où les bulles internes soient stables). L'erreur due à l'incertitude sur les paramètres plus proche voisin va tout de même limiter la précision sur la paramétrisation de γ_S à environ 10%. Cependant, comme l'erreur expérimentale est typiquement de 0.3 K, une évaluation plus précise (que les 30% supposés) de γ_S peut être envisagée avec des séquences dans l'intervalle $l \in [20, 40]$.

De la même manière, pour paramétrer ω (via ΔS^0_{mix}), on étudie $\mathcal{S}_{bulle}$ et $\mathcal{S}_{fin} = {}^{A_{l/2}G_{2n}A_{l/2}}_{T_{l/2}C_{2n}T_{l/2}}$ (deux domaines terminaux). On s'attend ici à ce que $T_m^{bulle} - T_m^{fin}$ dépendent essentiellement de la différence d'énergie entre les deux structures, soit environ 2ω. La figure 1.15 B illustre la sensibilité de $T_m^{bulle} - T_m^{fin}$ par rapport à ΔS^0_{mix}. Malheureusement, les barres d'erreurs ne permettent pas une évaluation précise de l'entropie de mélange. En effet, l'incertitude sur les paramètres plus proche voisin d'initiation (Δh_{ini} et Δs_{ini}) est très importante et cela se reflète automatiquement dans les erreurs théoriques. Sans une amélioration préalable de ces paramètres, il paraît impossible d'envisager de paramétriser ΔS^0_{mix} avec ce type d'expériences.

Chapitre 2

Analyse thermodynamique des génomes

Dans cette partie, nous nous intéressons à l'étude de la dénaturation de l'ADN pour des génomes entiers. Plus précisément, nous développons des méthodes algorithmiques et statistiques efficaces qui nous permettent de discuter de la pertinence de l'utilisation des propriétés thermodynamiques de l'ADN pour annoter les génomes.
Dans une première partie, nous définissons le modèle de Zimm-Bragg (ZB) qui est une approximation du modèle de Poland-Scheraga et nous décrivons la méthode des matrices de transfert utilisée pour calculer les prédictions du modèle.
Dans une deuxième partie, nous validons cette approximation en la comparant aux résultats du modèle PS et nous discutons de l'efficacité d'un tel modèle pour l'étude des génomes.
Puis, dans une troisième partie, nous appliquons le modèle ZB à plusieurs génomes entiers et nous étudions la possible existence de corrélations entre les propriétés codantes et thermodynamiques des séquences que ce soit au niveau des paires de bases individuelles ou au niveau des domaines. Nous discutons également de la fiabilité des prédictions de l'analyse thermodynamique pour l'identification de gènes ou d'exons.
Enfin, nous concluons sur l'applicabilité d'une telle méthode à l'annotation des génomes et sur la perspective de prendre en compte les effets de superhélicité pour une étude plus réaliste de la dénaturation dans les génomes.

2.1 Modèle de Zimm-Bragg

2.1.1 Une approximation du modèle PS

Le modèle de Zimm-Bragg [27, 109] a la même structure mathématique que le modèle d'Ising 1D hétérogène et, même s'il lui est antérieur, peut être vu comme une approximation du modèle PS. Le facteur de coopérativité σn^{-c}, dépendant de la taille n de la bulle, est alors remplacé par un terme constant

$$\sigma \bar{n}^{-c} \equiv \exp(-2\beta\gamma) \tag{2.1}$$

où $\bar{n}$ est une taille typique fixée et γ peut être interprétée comme une énergie d'interface (ou de fourche) entre une partie double-brin et une bulle. A première vue, cette approximation peut paraître brutale car elle néglige les effets génériques dus à la nature polymérique de l'ADN. En particulier, le modèle ZB ne prédit pas la transition de phase du premier ordre observé pour des homopolymères [78] (En effet, il n'y a pas de transition de phase pour le modèle d'Ising 1D). Cependant, l'hétérogénéité des séquences masquent généralement ces effets [110]. En effet, pour des séquences naturelles, la valeur de l'exposant c (issu de la théorie des polymères) ne semble pas influer sur les prédictions faites par le modèle PS pourvu que σ soit modifiée de manière appropriée [80].

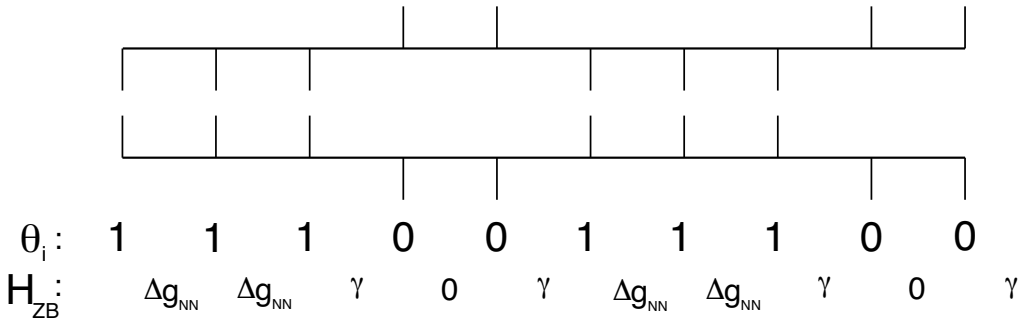

Figure 2.1 – Exemple d'une configuration d'Ising 1D décrivant une structure secondaire du double-brin d'ADN et description des différentes contributions énergétiques dans le modèle ZB.

Avec cette approximation, l'hamiltonien d'une séquence de taille N bp est décrit par (voir figure 2.1)

$$\begin{aligned} \mathcal{H}_{ZB} &= \sum_i \{\Delta g_{NN}(i,i+1)\,\theta_i\,\theta_{i+1} + \gamma\,[\theta_i\,(1-\theta_{i+1}) + \theta_{i+1}\,(1-\theta_i)]\} \\ &= \sum_i \{2\gamma\theta_i + (\Delta g_{NN}(i,i+1) - 2\gamma)\,\theta_i\,\theta_{i+1}\} \end{aligned} \tag{2.2}$$

où $\theta_i = 0$ (1) si la paire de bases i est dénaturée (ou appariée). $\Delta g_{NN}(i,i+1)$ représente l'énergie libre d'association du segment de paires de bases $(i,i+1)$ (idem que pour le modèle PS, voir table 1.1) et dépend de la concentration en sel (voir section 1.2.2). γ est l'énergie de fourche définie précédemment et dépend du choix de $\bar{n}$ (voir table 2.1). Notons que les termes de capping sont ici négligés (ce qui est justifié vu que l'on s'intéresse à de très grandes séquences de l'ordre de Mbp) et que l'on utilise des conditions périodiques "$N+1 \equiv 1$". La fonction de partition du système est alors définie par

$$Z = \sum_{\theta_i} \exp\left(-\beta\mathcal{H}_{ZB}(\{\theta_j\})\right) \tag{2.3}$$

Les propriétés de dénaturation étudiées par la suite sont la probabilité individuelle qu'une paire de base α soit fermée, $p(\alpha) = \langle\theta_\alpha\rangle$ (qui dépend de la température), et la température de dénaturation locale définie par $p(\alpha)(T_m(\alpha)) = 1/2$.

Table 2.1 – Estimations de γ et $\sigma\bar{n}^{-c}$ pour différentes valeurs de $\bar{n}$ calculées avec $c = 2.15$ et $\sigma = 1.23 \times 10^{-4}$.

$\bar{n}$	$\gamma/(k_BT)$	$\sigma\bar{n}^{-c}$
10	7.0	8.9×10^{-7}
170	10.0	2.0×10^{-9}
500	11.2	2.0×10^{-10}
1000	11.9	4.5×10^{-11}
5000	13.6	1.4×10^{-12}
10000	14.4	3.2×10^{-13}

2.1.2 Méthode des matrices de transfert

De par sa formulation sous la forme d'un modèle d'Ising 1D, la résolution du modèle et le calcul des observables $p(\alpha)$ et $T_m(\alpha)$ peuvent être facilement réalisés par la méthode des matrices de transfert [27, 111]. La méthode consiste à remarquer que

$$Z = \text{Trace}\left(\prod_{i=1}^{N} T_i\right) \tag{2.4}$$

$$p(\alpha) = \frac{1}{Z}\text{Trace}\left[\left(\prod_{i=1}^{\alpha-1} T_i\right) T'_\alpha \left(\prod_{i=\alpha+1}^{N} T_i\right)\right] \tag{2.5}$$

avec

$$T_i = \begin{pmatrix} e^{-\beta\Delta g_{NN}(i,i+1)} & e^{-2\beta\gamma} \\ 1 & 1 \end{pmatrix}$$

$$T'_i = \begin{pmatrix} e^{-\beta\Delta g_{NN}(i,i+1)} & e^{-2\beta\gamma} \\ 0 & 0 \end{pmatrix}$$

Pratiquement, on calcule $\prod_{i=1}^{N} T_i$ itérativement soit en partant de $M_1^{pre} \equiv T_1$, puis successivement $M_{i+1}^{pre} = M_i^{pre} \times T_{i+1}$; soit en partant de $M_N^{post} \equiv T_N$ puis $M_{i-1}^{post} = T_{i-1} \times M_i^{post}$. Ainsi, en prenant la trace de M_N^{pre} ou de M_1^{post} on a accès à Z (voir équation 2.4), puis la trace de $M_{\alpha-1}^{pre} \times T'_\alpha \times M_{\alpha+1}^{post}$, on trouve $p(\alpha)$ (voir équation 2.5).

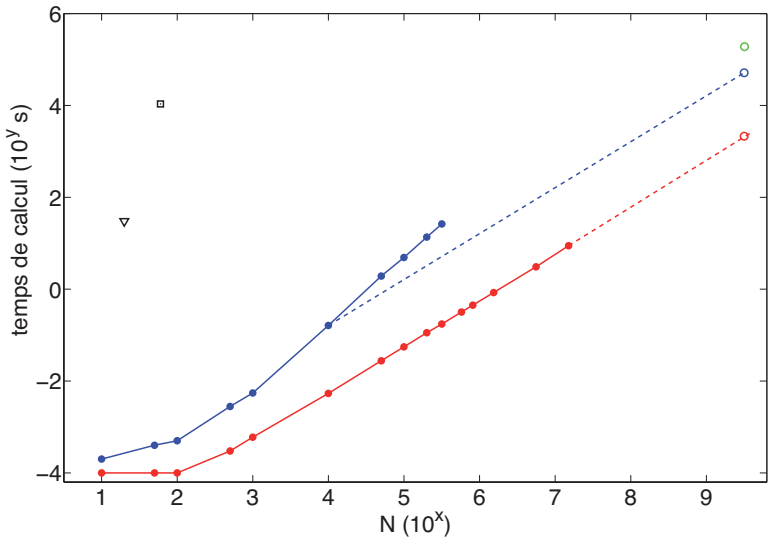

FIGURE 2.2 – Temps de calcul (en seconde) utile pour estimer tous les $p(\alpha)$ à température fixée en fonction de la taille de la séquence avec un processeur 2.4 GHz Intel Core 2 Duo, à l'aide due modèle PS-FF pour $N_s = n_{max} = N$ (ligne bleue) ou du modèle ZB (ligne rouge). Les cercles représentent le temps de calcul nécessaire pour étudier le génome humain avec le modèle PS-FF (vert : $N_s \sim 10^8$ bp, $n_{max} = 10^8$ bp [75] ; bleu : $N_s = 10$ kbp, $n_{max} = 10$ kbp, temps extrapolé) ou avec le modèle ZB (rouge, temps extrapolé). Le carré noir correspond au temps typique de résolution du modèle de Peyrard-Bishop-Dauxois [73] et le triangle noir à celui du modèle sur réseau [33].

2.1.3 Complexité et temps de calcul

La méthode des matrices de transfert permet ainsi d'évaluer Z et tous les $p(\alpha)$ pour une séquence de taille N en $\mathcal{O}(N)$ opérations. Pour faciliter la résolution numérique, le modèle ZB est résolu pour des tranches de séquence de 100 kbp en quinconce afin d'éviter les effets de bords.

Pour la même séquence, la solution exacte du modèle PS requiert $\mathcal{O}(N^2)$ opérations, alors qu'en utilisant l'algorithme de Fixman-Freire, on peut se restreindre à $\mathcal{O}(N \times I)$ opérations où $I \propto \log n_{max}$, la taille maximale de bulle approximée par l'équation 1.20 (voir section 1.1.3). Pour des raisons numériques, la résolution du modèle PS peut être accélérée en découpant la séquence en blocks d'une certaine taille N_s se chevauchant et en utilisant l'algorithme de Fixman-Freire sur chaque block. Dans le cadre d'une application génomique, la taille caractéristique des domaines étant de l'ordre du kbp ou moins (voir figure 2.10), il faut donc choisir $N_s >$ kbp et $n_{max} \sim$ kbp. Par exemple, Yeramian et Jones [106] utilisent $N_s = 10$ kbp et $n_{max} = 5$ kbp ; par contre Liu et collaborateurs [75] prennent $N_s \sim 10^8$ bp et $n_{max} = 10^8$ bp.

La figure 2.2 compare le temps de calcul nécessaire au modèle de PS avec l'algorithme de Fixman-Freire (PS-FF), au modèle ZB et au modèle Peyrard-Bishop-Dauxois pour calculer les propriétés thermodynamiques d'une séquence. On remarque que le modèle ZB nous permet d'accélérer le calcul d'un facteur 30 par rapport au modèle PS-FF avec $N_s = 10$ kbp et $n_{max} = 10$ kbp et d'un facteur 90 pour $N_s = 100$ kbp et $n_{max} = 10$ kbp. Par exemple, le calcul de la carte de dénaturation du génome humain ($\sim 10^{10}$ bp) réalisé par Liu et collaborateurs [75] ($N_s \sim 10^8$ bp, $n_{max} = 10^8$ bp) a pris 22 jours sur un HP SuperDome (64 processeurs Itanium 2, 1.5 GHz, 6 MB cache), alors que cela aurait pris environ 6 jours à la même machine pour résoudre le modèle PS-FF avec $N_s = 10$ kbp et $n_{max} = 10$ kbp, et seulement 5h pour le modèle ZB. A titre de comparaison, la résolution du modèle de Peyrard-Bishop-Dauxois par une méthode d'intégration numérique directe [73, 112] ou une simulation du modèle sur réseau [33] prendrait des heures pour une séquence de seulement 60 bp. Dans le même temps de calcul, des séquences de tailles $\sim 10^8$ bp pourraient être étudiées avec le modèle PS-FF et de tailles $\sim 10^{10}$ bp avec le modèle ZB.

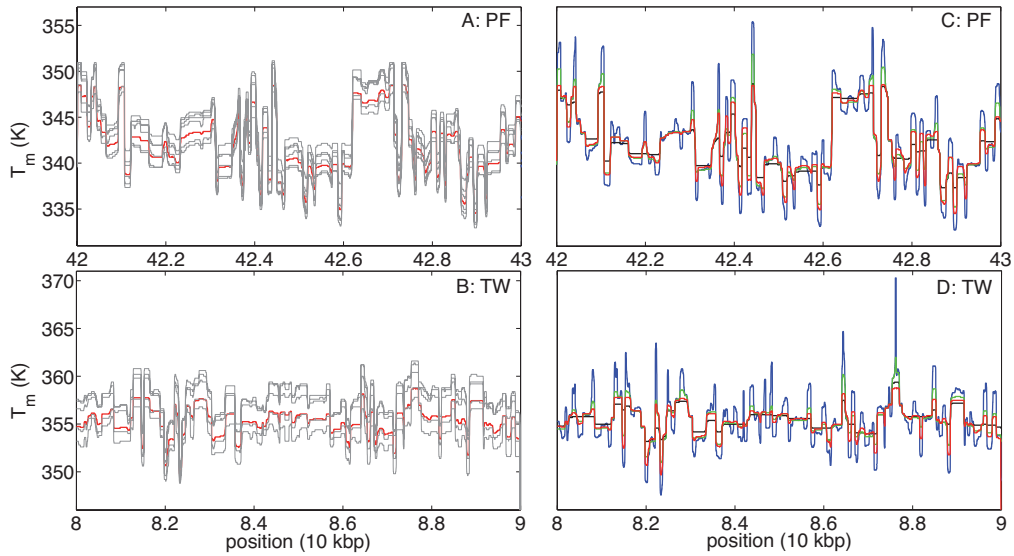

Figure 2.3 – Evolution de $T_m(\alpha)$ pour une échantillon du génome de *Plasmodium falciparum* (A,C) et de *Tropheryma whipplei* (B,D) calculée par le modèle PS avec les paramètres standard (lignes rouges) ou avec différents jeux de paramètres aléatoirement générés dans la limite de confiance (voir section 1.2.4) (lignes grises) (A,B) et par le modèle ZB pour différentes valeurs de γ (C,D) : $7k_BT$ (bleu), $10k_BT$ (vert) et $13k_BT$ (noir).

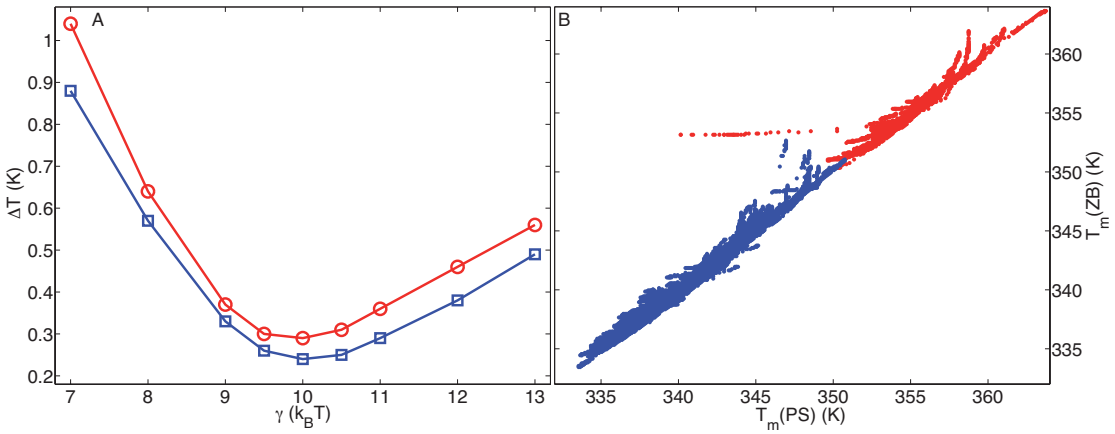

FIGURE 2.4 – (A) Erreur systématique moyenne ΔT en fonction de la valeur de γ calculée pour le génome de PF (carrés bleus) et de TW (ronds rouges). (B) Diagramme de corrélation entre $T_m^{ZB}(\alpha)$ et $T_m^{PS}(\alpha)$ pour PF (bleu) et TW (rouge) calculé pour $\gamma = 10k_BT$.

2.2 Validation

Dans un premier temps, on compare le modèle ZB au modèle PS pour différentes valeurs de γ afin de trouver la valeur optimale. Puis, dans un deuxième temps, nous regardons les résultats donnés par le modèle ZB pour deux exemples où le modèle PS a été utilisé pour étudier des propriétés génomiques de l'ADN.

Dans cette partie et dans la suivante (sauf mention particulière), la concentration en sel est fixée à $[Na^+] = 0.1$ M. Vu que l'on s'intéresse à des séquences longues, les effets de concentration (voir section 1.3.1) sont négligeables et ne seront pas considérés.

2.2.1 Comparaison avec le modèle PS

Dans les figures 2.3 C et D, on a tracé les profils de dénaturation $T_m(\alpha)$ calculés avec le modèle PS et avec le modèle ZB pour trois différentes valeurs de γ (correspondant à $\bar{n} \sim 10, 150, 2500$) pour deux séquences génomiques : une extraite du chromosome 11 du parasite *Plasmodium falciparum* (PF) [1] et l'autre extraite du génome de la bactérie *Tropheryma whipplei* TW08/27 (TW) [2]. L'évolution de $T_m(\alpha)$ le long de la séquence permet facilement de distinguer des domaines de dénaturation. La définition de ces domaines semble assez robuste par rapport aux incertitudes sur les paramètres (voir figures 2.3 A et B). On observe qu'une grande énergie d'interface γ va avoir tendance à accroître la taille de ces domaines, illustrant la coopérativité dans le modèle ZB. De plus, les résultats semblent indiquer que le modèle ZB reproduit assez fidèlement les profils calculés avec le modèle PS. Pour des faibles valeurs de $\bar{n}$ (ou de γ), le modèle ZB a tendance à prédire des petites bulles qui ne sont pas présentes dans les résultats de PS ; pour des fortes valeurs de $\bar{n}$, au contraire, il tend à grouper ensemble plusieurs petits domaines prédits par le modèle PS. Le profil calculé avec $\gamma = 10k_BT$ semble être en excellent accord avec les prédictions faites par le modèle PS.

Afin d'être plus précis sur l'erreur systématique introduite par l'approximation de ZB, on calcule

1. Le génome de PF est accessible sur le site du National Center for Biotechnology Information (NCBI) (http ://www.ncbi.nlm.nih.gov) au numéro d'accession NC004315.
2. Son numéro d'accession sur le site du NCBI est BX251411.

en fonction de γ

$$\Delta T = \frac{1}{N}\sum_{\alpha=1}^{N} |T_m^{ZB}(\alpha) - T_m^{PS}(\alpha)| \tag{2.6}$$

La figure 2.4 A montre qu'il existe, pour chaque organisme étudié (PF et TW), une valeur de l'énergie d'interface γ pour laquelle ΔT est minimale. La valeur optimale de γ vaut environ $10k_BT$ et correspond à une taille typique de $\bar{n} \approx 170$ bp. L'erreur systématique minimale vaut alors environ 0.3 K. Par comparaison, l'incertitude due aux paramètres est de l'ordre de 2 K (voir section 1.4.1 et figures 2.3 A et B). La figure 2.4 B illustre la très forte corrélation entre T_m^{ZB} et T_m^{PS} pour $\gamma = 10k_BT$ (facteur de corrélation de 0.97 pour TW et 0.99 pour PF). D'ailleurs, dans la suite, on utilisera cette valeur pour calculer les résultats prédits par le modèle ZB.

2.2.2 Applications à la génomique

Dans cette section, on vérifie qu'à l'aide du modèle ZB, on retrouve des résultats précédemment obtenus avec le modèle PS sur de possibles corrélations entre propriétés thermodynamiques et génomiques.

2.2.2.1 Identification de gènes dans l'ADN génomique

La transition entre double-hélice et bulle prédite par les modèles décrivant la dénaturation étant assez abrupte, les paires de bases adjacentes s'ouvrant simultanément forment des domaines le long de la séquence que l'on peut comparer avec les propriétés génomiques de l'ADN. Comme une paire GC est plus stable qu'une paire AT, les températures de dénaturation dépendent du contenu local en GC [69] (voir équation 1.50). De plus, généralement, les régions codantes des génomes sont riches en GC comparées au contenu GC moyen [64, 65] (voir table 2.2). Ainsi, on s'attend à observer des corrélations entre la position des domaines dénaturés et des parties codantes du génome [113, 106, 114, 115, 116, 49, 50].

Pour plusieurs températures, Yeramian calcule $p(\alpha)$ à l'aide du modèle PS. Pour chaque température, l'évolution de $p(\alpha)$ le long de la séquence fait apparaître des régions qui sont comparées aux parties codantes (CDS) [106]. On effectue le même genre de comparaison avec le modèle ZB pour les génomes complets de PF et TW. La figure 2.2.2.1, en plus de confirmer le bon accord entre les modèles PS et ZB, illustre le fait qu'une bonne corrélation entre les domaines fermés et codants à 68 ou 70° C existent pour PF mais aucune preuve de cette corrélation n'est observée pour TW.

2.2.2.2 Identification de bords d'exon dans l'ADN complémentaire

Carlon et Blossey [71, 117] ont étudié les propriétés de dénaturation de l'ADN complémentaire (ADNc). L'ADNc est la transcription inverse de l'ARNm d'un gène, il correspond donc à la séquence d'un gène privé des parties correspondant aux introns. Les figures 2.6 A, B et C montrent la courbe de dénaturation $-\mathrm{d}\Theta/\mathrm{d}T$ pour trois ADNc humains. Là encore, les résultats corroborent la validité de l'approximation de ZB et montrent qu'il existe, pour certaines frontières entre exons, une bonne correspondance avec les propriétés de dénaturation. Lorsque la séquence devient grande, la courbe de dénaturation devenant plus complexe, il est plus pratique d'évaluer, pour une position donnée le long de la séquence, l'intervalle de température pour lequel cette même position délimite deux domaines de dénaturation et de comparer cela aux frontières entre exons. La figure 2.6 D confirme l'hypothèse d'une correspondance entre frontières de domaines thermodynamiques et d'exons.

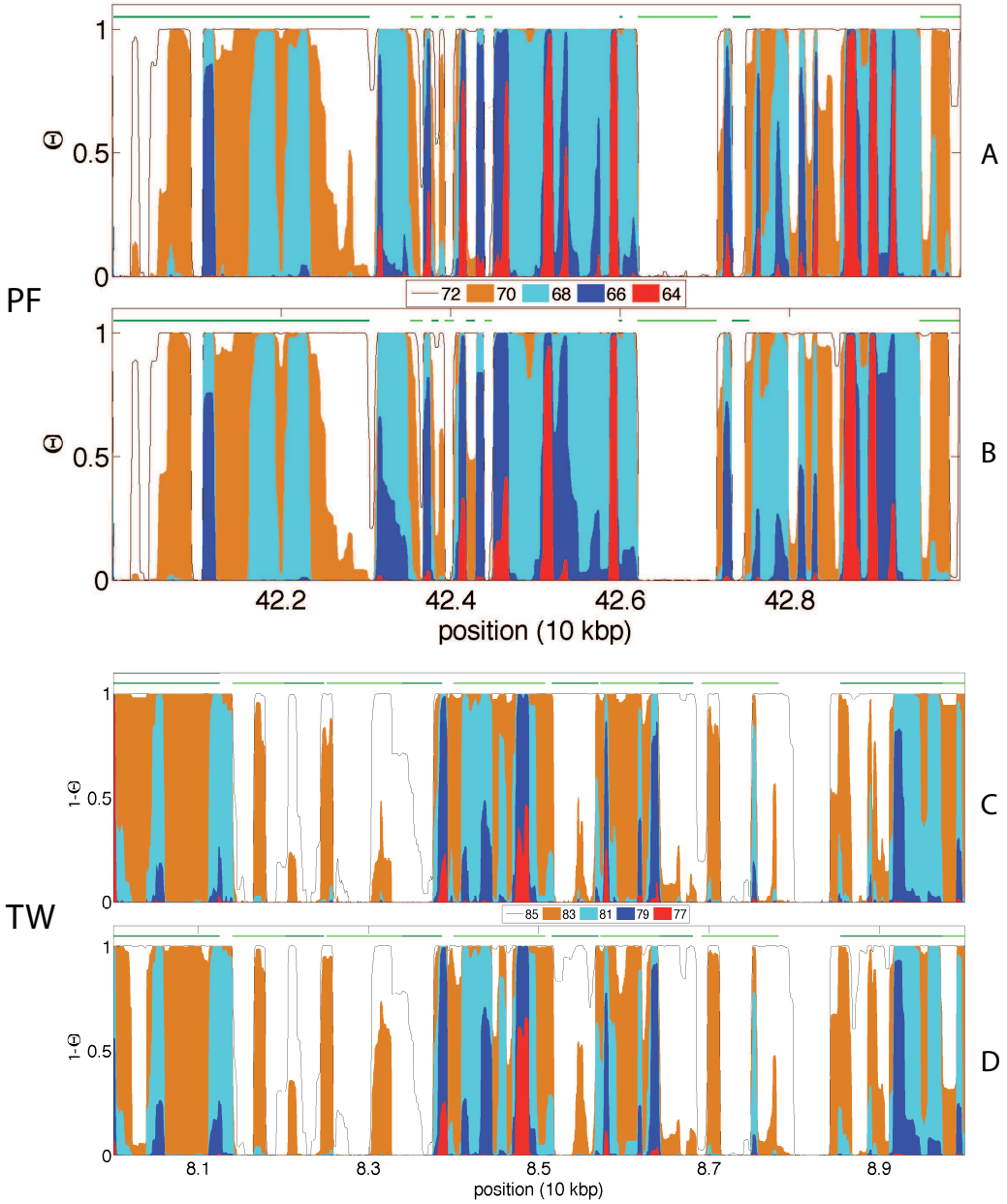

FIGURE 2.5 – Probabilité d'ouverture de la double-hélice $1 - \Theta$ le long d'un extrait de séquence de PF (A,B) et de TW (C,D) à plusieurs températures (voir la légende des couleurs) calculée avec le modèle PS (A,C) et ZB (B,D). Les domaines codants sont représentés en vert au-dessus des courbes.

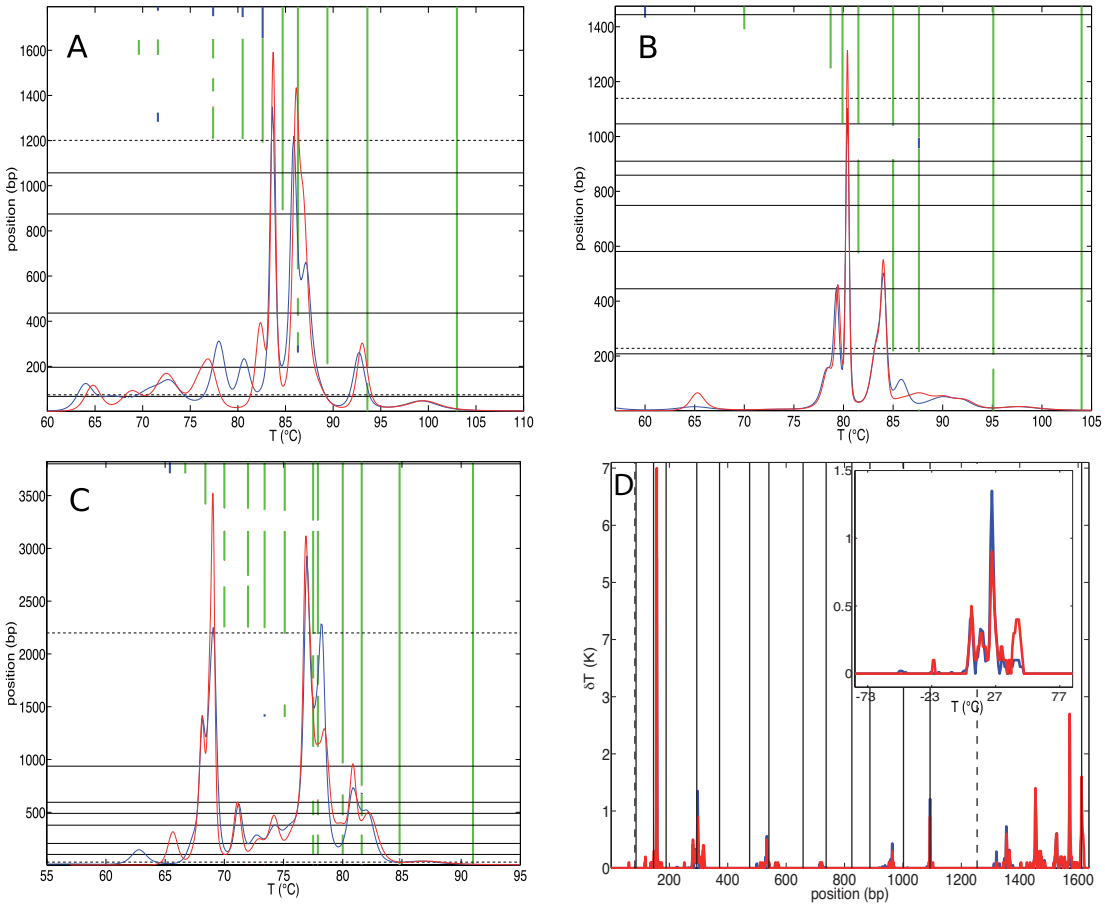

Figure 2.6 – (A,B,C) $-\mathrm{d}\Theta/\mathrm{d}T$ et domaines de dénaturation pour les ADNc humains de l'actine β (A, numéro d'accesion NCBI : NM001101), de la CDK4 (B, NM000075) et du gène EHHADH (NM001966) calculés pour une concentration en sel de 0.05 M avec le modèle PS (lignes bleus) et le modèle ZB (lignes rouges). Les barres verticales représentent les paires de bases pour lesquelles $T_m(\alpha) < T$ dans le modèle PS uniquement (bleu), dans le modèle ZB uniquement (rouge) et dans les deux (vert). Les barres horizontales montrent les frontières entre exons (lignes pleines) et les frontières entre la partie codante et les régions non-traduites (lignes pointillés). (D) Intervalle de température δT pendant lequel une position particulière dans la séquence délimite deux domaines de dénaturation calculé pour l'ADNc d'un facteur favorisant la fixation de l'interleukine (ILF2, NM004515) pour une concentration en sel de 0.05 M avec le modèle PS (lignes bleus) et le modèle ZB (lignes rouges). Les barres verticales montrent les frontières entre exons (lignes pleines) et les frontières entre la partie codante et les régions non-traduites (lignes pointillés).

TABLE 2.2 – Composition des génomes des organismes procaryotes (†) et eucaryotes (*) étudiés.

Espèces (abréviation)	N(Mbp)	%CDS	f_{GC}^{tot}	f_{GC}^{cds}	f_{GC}^{rest}	$\Delta\tau_{max}$
*P.falciparum** (PF)	22.86	52.5	0.190	0.237	0.146	0.33
*S.pombe** (SP)	12.57	56.4	0.36	0.397	0.314	0.30
*S.cerevisiae** (SC)	12.07	72.4	0.383	0.396	0.348	0.14
*D.melanogaster** (DM)	120.38	18.6	0.424	0.534	0.403	0.17
*C.elegans** (CE)	100.26	25.2	0.354	0.426	0.330	0.16
T.whipplei† (TW)	0.93	84.4	0.463	0.464	0.458	0.03
E.coli† (EC)	4.64	88.0	0.508	0.519	0.43	0.10

2.3 Application à l'analyse des génomes

Dans la section précédente, nous avons démontré l'utilité du modèle ZB en termes d'efficacité et de précision numérique. Dans cette section, on étend l'analyse proposée par Yeramian dans [106, 113, 49, 50] aux génomes complets de 7 organismes (*Plasmodium falciparum, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, Tropheryma whipplei, Escherichia Coli*, voir la table 2.2 pour plus de détail et pour les abréviations utilisées) qui ont déjà été partiellement étudiés précédemment (PF, TW, SC, SP, DM) ou qui sont des organismes modèles (SC, EC, CE). Ces exemples incluent des procaryotes (TW, EC) ainsi que des eucaryotes (PF, SP, SC, DM, CE) ayant des densités en parties codantes variées (de 18% à 88%) et couvrant une large gamme de contenus GC (entre 19% et 51%). Dans un premier temps, nous définissons et utilisons des outils et des méthodes nous permettant d'étudier de manière systématique les corrélations entre propriétés de dénaturation et codantes aussi bien au niveau des paires de bases qu'au niveau des domaines. Dans un deuxième temps, nous discutons de la possibilité d'utiliser notre approche pour faire des prédictions et identifier des gènes.

2.3.1 Corrélations au niveau des paires de bases

On commence par une comparaison plus quantitative que celle abordée dans la section 2.2.2 des propriétés codantes et dénaturantes. Pour chaque température étudiée, on attribue à chaque paire de bases i une prédiction sur son état : si $T_m(i) > T$ (c'est à dire, si i est fermée à T), son état est prédit comme codant sinon, la paire est ouverte et son état est prédit comme non-codant. De plus, pour chaque génome, on connaît les parties qui ont été annotées comme codantes (le reste étant considéré comme non-codant) via des bases de données comme celle du NCBI ou celle du EMBL-Bank (European Molecular Biology Laboratory). Ainsi, pour chaque organisme et en fonction de la température, on peut évaluer parmi nos prédictions le nombre de vrais-positifs N_{VP} (i est correctement prédite comme étant une base codante), de vrais-négatifs N_{VN}, de faux-positifs N_{FP} et de faux-négatifs N_{FN}. A partir de ces nombres, on en déduit des indicateurs statistiques :

- la sensibilité $\beta \equiv N_{VP}/N_{cod}$ qui mesure le taux de paires codantes ayant été effectivement identifiées comme fermées avec $N_{cod} = (N_{VP} + N_{FN})$ le nombre de paires codantes dans le génome ;
- la spécificité $\alpha \equiv N_{VN}/N_{non-cod}$ qui mesure le taux de paires non-codantes ayant été effectivement identifiées comme ouvertes avec $N_{non-cod} = (N_{VN}+N_{FP})$ le nombre de paires non-codantes dans le génome ;
- le taux de corrélation $\tau \equiv (N_{VP}+N_{VN})/N$ qui mesure la fraction de paires ayant été effectivement bien identifiées.

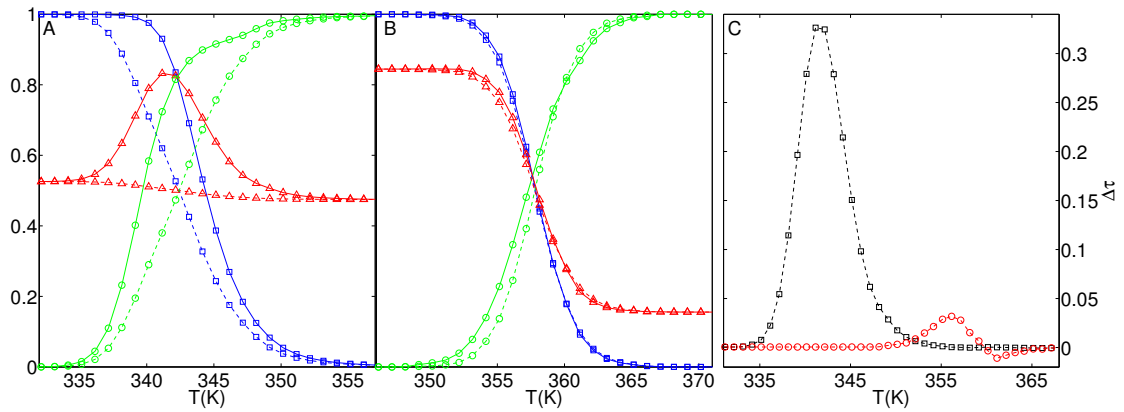

Figure 2.7 – (A,B) Sensibilité (carrés bleus), spécificité (cercles verts) et taux de corrélation (triangles rouges) pour PF (A) et TW (B) calculés à partir de l'annotation réelle du génome (lignes pleines) ou à partir d'une annotation aléatoire à $\%CDS$ constant (lignes pointillées). (C) Différence $\Delta\tau = \tau - \tau^r$ par rapport au cas aléatoire pour PF (carrés noirs) ou pour TW (cercles rouges). Les fluctuations des variables aléatoires α^r, β^r et τ^r sont plus petites que la taille des symboles utilisés, par exemple $\sigma_{\tau^r} = 10^{-3}$ pour PF et 5×10^{-3} pour TW.

La correspondance entre le caractère codant et fermé est d'autant meilleure que ces trois indicateurs ont des valeurs proches de 1 simultanément. Les figures 2.7 A et B représentent α, β et τ pour PF et TW en fonction de la température et les comparent avec les résultats α^r, β^r et τ^r que l'on obtiendrait pour une distribution aléatoire des paires codantes le long de la séquence (voir annexe 6.4). A basse températures, toutes les paires de bases sont fermées et donc prédites comme codantes. Ainsi, la sensibilité tend vers 1 (toutes les paires codantes sont reconnues), la spécificité tend vers 0 (aucune paire non-codante n'est reconnue) et le taux de corrélation tend vers $\%CDS$, le pourcentage de séquences codantes dans le génome (voir table 2.2). Dans la limite des hautes températures, les observations sont inversées, $\beta \to 0$, $\alpha \to 1$ et $\tau \to (1 - \%CDS)$.

La figure 2.7 C présente la dépendance en température de l'accroissement relatif du taux de corrélation par rapport au cas aléatoire $\Delta\tau = \tau - \tau^r$. Comme τ^r correspond au cas aléatoire moyen, les fluctuations σ_{τ^r} autour de la valeur moyenne sont évaluées en générant plusieurs annotations aléatoires du génome. Alors que pour TW, aucune différence significative n'est observée avec le cas aléatoire, pour PF, il existe un large intervalle de température sur lequel cette différence est importante et les corrélations entre caractère codant et fermé sont fortes. Concernant la sensibilité et la spécificité, on peut montrer (voir annexe 6.4) que

$$\Delta\beta = \frac{1}{2 \times \%CDS}\Delta\tau \tag{2.7}$$

$$\Delta\alpha = \frac{1}{2 \times (1 - \%CDS)}\Delta\tau \tag{2.8}$$

Ces deux équations impliquent directement que pour une espèce, $\Delta\tau$, $\Delta\beta$ et $\Delta\alpha$ sont proportionnelles et sont donc maximales pour la même température (définie comme la température optimale T_{opt}). La figure 2.8 A indique que les corrélations sont maximales pour une température proche de $T_{moy} = (\langle T_m^{cod}\rangle + \langle T_m^{reste}\rangle)/2$, c'est à dire, entre la température de dénaturation moyenne des régions codantes $\langle T_m^{cod}\rangle$ et celle du reste du génome $\langle T_m^{reste}\rangle$. Les valeurs maximales de $\Delta\tau$ sont données dans la table 2.2 et varie fortement entre les espèces indiquant que les corrélations observées sont plus ou moins fortes selon les organismes.

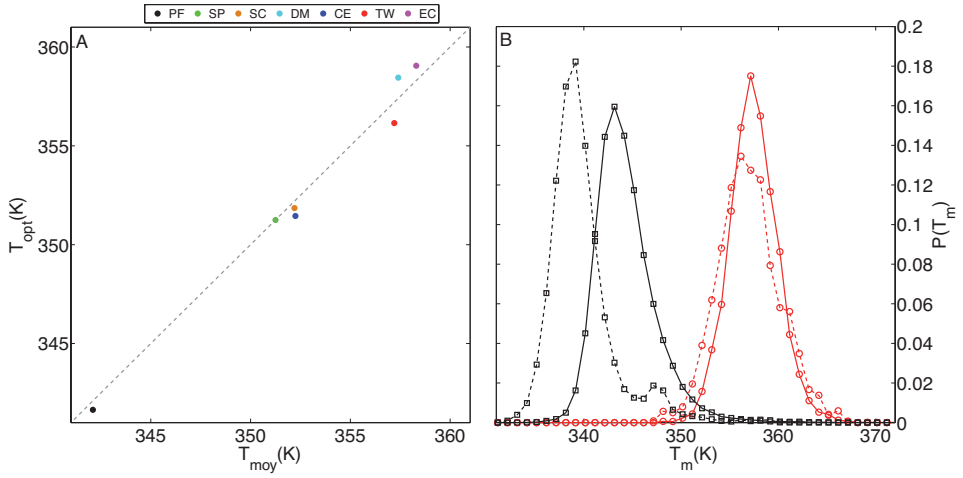

FIGURE 2.8 – (A) Température T_{opt} pour laquelle $\Delta\tau$ est maximale, tracée en fonction de T_{moy} la moyenne entre la température de dénaturation moyenne des régions codantes et celle des régions non-codantes, et calculée pour différentes espèces (voir légende). (B) Distribution normalisée $P(T_m)$ des températures de dénaturation des paires de bases codantes (lignes pleines) ou non-codantes (lignes pointillées) pour PF (carrés noirs) et TW (cercles rouges).

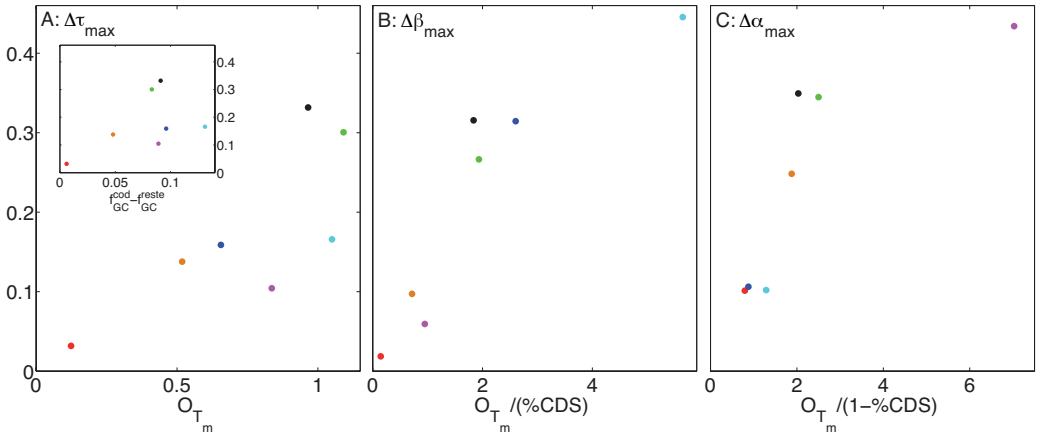

FIGURE 2.9 – Différences maximales $\Delta\tau_{max}$ (A), $\Delta\beta_{max}$ (B) et $\Delta\alpha_{max}$ (C) pour différentes espèces (même couleur que dans la figure 2.8 A) en fonction du paramètre de chevauchement O_{T_m}. L'encart dans (A) représente $\Delta\tau_{max}$ en fonction de la différence entre le contenu GC des régions codantes f_{GC}^{cod} et celui du reste du génome f_{GC}^{reste}.

La raison pour laquelle on obtient ces différents niveaux de succès pour l'analyse faite précédemment devient apparente dans la figure 2.8 B. Elle montre la distribution normalisée des températures de dénaturation des paires de bases codantes et non-codantes pour un exemple où la correspondance codant/fermé est forte (PF) et pour un autre où elle est faible (TW). Le chevauchement des deux distributions dépend de l'organisme étudié et peut être évalué via le paramètre $O_{T_m} \equiv \Delta T/\sqrt{w_{cod}^2 + w_{reste}^2}$ où $\Delta T = \langle T_m^{cod}\rangle - \langle T_m^{reste}\rangle$ et où w_{cod} et w_{reste} sont respectivement les écart-types des distributions pour les régions codantes et non-codantes (voir annexe 6.4). Dans le cas de TW (peu de correspondance), ΔT est faible (≈ 0.5 K) et $O_{T_m} = 0.124$. Au contraire pour PF (forte correspondance), les deux distributions sont bien séparées ($\Delta T \approx 4.5K$) et le chevauchement est plus faible ($O_{T_m} = 0.965$).

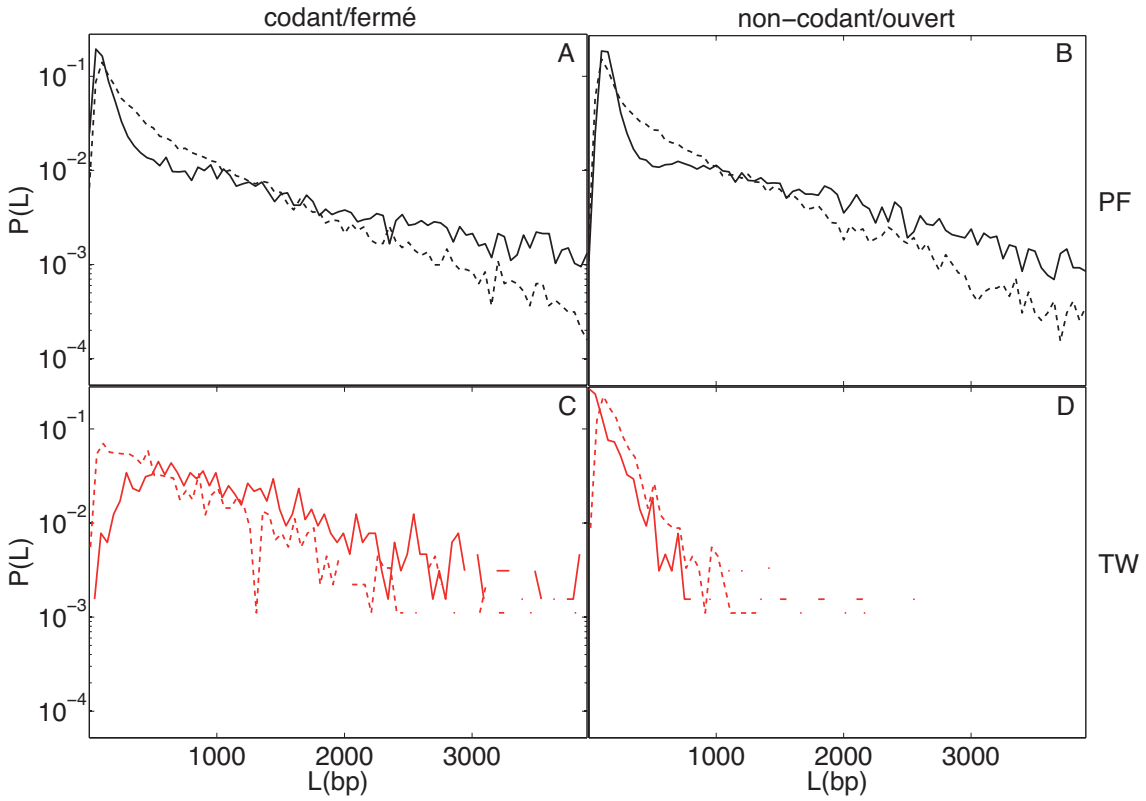

Figure 2.10 – (A,C) Distributions normalisées de la taille des domaines codants (lignes pleines) et des domaines fermés (lignes pointillées) pour PF (A) et TW (C). (B,D) Distributions normalisées de la taille des domaines non-codants (lignes pleines) et des domaines ouverts (lignes pointillées) pour PF (B) et TW (D).

La figure 2.9 nous montre comment le chevauchement des deux distributions de températures de dénaturation est connecté au succès de l'analyse des corrélations entre propriétés de dénaturation et propriétés codantes à travers les valeurs maximales de $\Delta\tau$, $\Delta\beta$ et $\Delta\alpha$. Dans la figure 2.9 A, on trace $\Delta\tau_{max} \equiv \Delta\tau(T_{opt})$ en fonction de O_{T_m}. Plus le chevauchement est important, plus le pouvoir de prédiction de l'analyse des génomes devient faible. Si l'on compare ces résultats avec les données décrivant la composition des différents génomes étudiés (voir table 2.2), on observe que la correspondance entre propriétés biologiques et thermodynamiques marchera d'autant mieux que les contenus GC des régions codantes et non-codantes seront significativement différents (voir encart figure 2.9 A). Cela souligne clairement la relation entre le contenu GC et le comportement de dénaturation. Les figures 2.9 B et C montrent $\Delta\beta_{max}$ et $\Delta\alpha_{max}$ en fonction respectivement de $O_{T_m}/\%CDS$ et de $O_{T_m}/(1-\%CDS)$ (ces paramètres sont inspirés par les équations 2.7 et 2.8). On remarque que les génomes pauvres en régions codantes (DM, CE) ont tendance à avoir une forte différence avec le cas aléatoire pour la sensibilité et une plus faible pour la spécificité. L'observation inverse peut être faite pour les génomes riches en régions codantes (SC, EC). Ainsi, l'analyse des génomes tend plutôt à mieux reconnaître les régions minoritaires (non-codantes pour les génomes riches en gènes et codantes pour les génomes pauvres en gènes).

2.3.2 Corrélations au niveau des domaines

Dans la section précédente, nous avons étudié la relation entre propriétés de dénaturation et codante au niveau des paires de bases. Dans la suite, nous regardons si l'analyse des génomes permet

de reconnaître des domaines biologiquement fonctionnels. Pour enlever toute dépendance en la température de nos résultats, on étudie maintenant chaque espèce à sa température optimale T_{opt} définie précédemment.

La figure 2.10 montre les distributions normalisées de la taille des différents types de domaines (codant/non-codant, fermé/ouvert) pour PF et TW. Pour PF, les distributions correspondantes aux domaines fermés et codants, ainsi que celle pour les domaines ouverts et non-codants, sont piquées autour de la même taille de domaine, cependant les distributions pour les domaines de dénaturation (fermé/ouvert) sont plus concentrés autour des petits domaines. Pour TW, les distributions s'étendent sur des intervalles équivalents, mais elles différent au niveau des petits domaines.

Afin d'être plus précis sur l'identification de domaine par l'analyse des génomes, on définit pour chaque domaine codant, une sensibilité locale β_{loc} (nombre de paires de bases dans ce domaine étant fermées), et pour chaque domaine non-codant une spécificité locale α_{loc} (nombre de paires de bases dans ce domaines étant ouvertes). La figure 2.11 montre l'évolution de la valeur moyenne de β_{loc} et α_{loc} en fonction respectivement de la taille des domaines codants et non-codants.

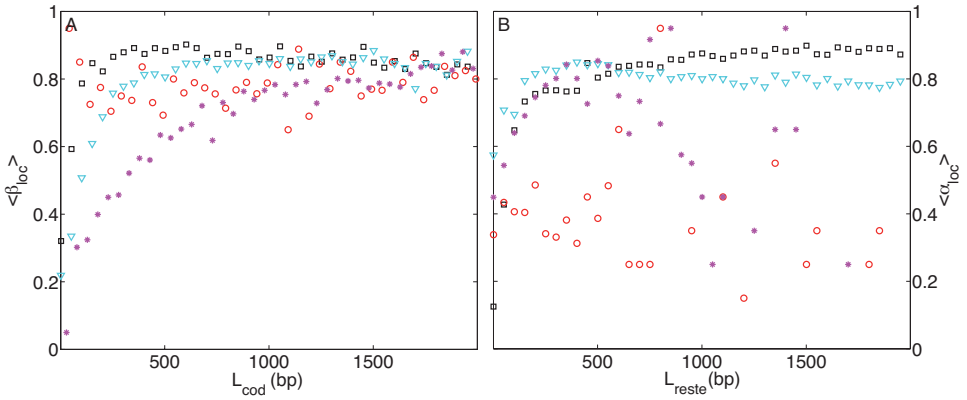

FIGURE 2.11 – (A) Valeur moyenne $\langle\beta_{loc}\rangle$ de la sensibilité locale en fonction de la taille L_{cod} des domaines codants pour PF (carrés noirs), TW (cercles rouges), DM (triangles cyans) et EC (étoiles mauves). (B) Valeur moyenne $\langle\alpha_{loc}\rangle$ de la spécificité locale en fonction de la taille L_{reste} des domaines non-codants.

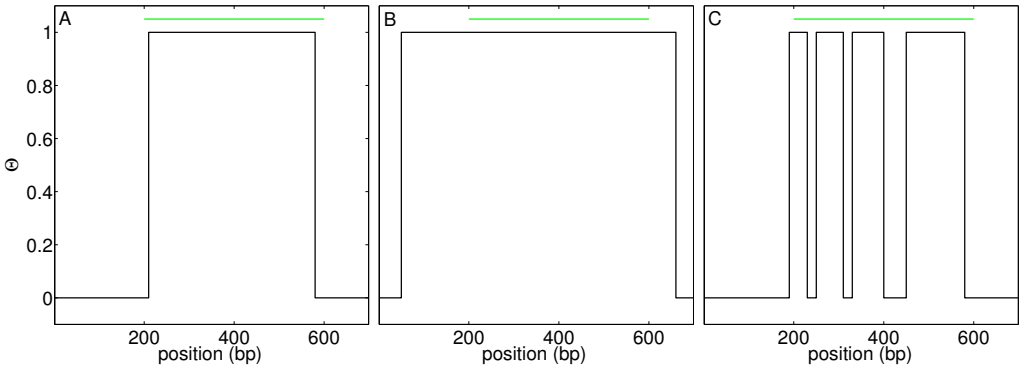

FIGURE 2.12 – Exemples de répartitions possibles de domaines fermés ($\Theta = 1$) et ouvert ($\Theta = 0$) (lignes noires) autour d'une région codante de 400 bp (lignes vertes). La sensibilité locale vaut $\beta_{loc} = 0.93$ (A), 1 (B) et 0.73 (C).

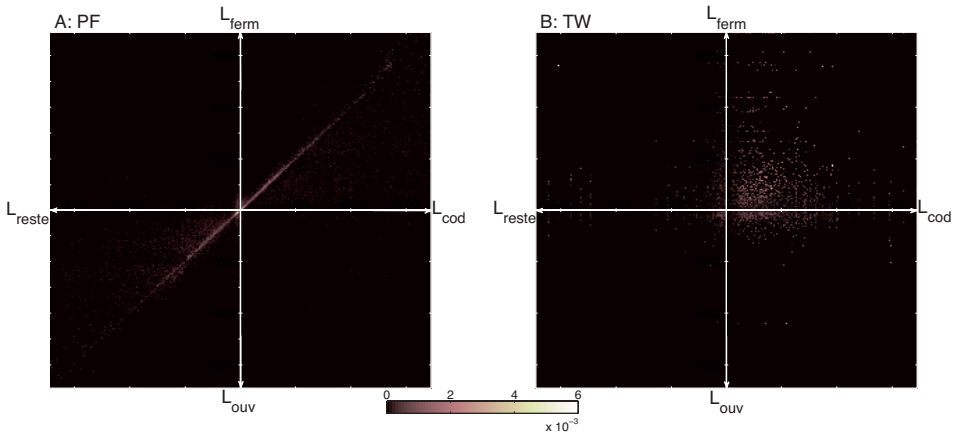

Figure 2.13 – Distributions de probabilité jointes pour une paire de base d'appartenir simultanément à un domaine codant ou non-codant (de taille L_{cod} ou L_{reste}) et à un domaine fermé ou ouvert (de taille L_{ferm} ou L_{ouv}) pour PF (A) et TW (B).

On remarque qu'à la température optimale, les petits domaines (< 200 bp) sont généralement faiblement identifiés. Cela vient du fait que la coopérativité est dominante pour des tailles inférieures à la taille typique $\bar{n} = 170$ bp du modèle ZB. Pour les domaines plus grands (entre 500 et 2000 bp), les valeurs de $\langle\beta_{loc}\rangle$ et $\langle\alpha_{loc}\rangle$ sont approximativement constantes autour des valeurs (moyennées sur la séquence) $\beta(T_{opt})$ et $\alpha(T_{opt})$. Cela indique que pour cette gamme de longueur, il n'y a pas de taille privilégiée pour le succès de l'analyse des génomes. De plus, on remarque localement, même pour TW (dont les valeurs de $\Delta\tau_{max}$, $\Delta\beta_{max}$ et $\Delta\alpha_{max}$ étaient très faibles), que $\langle\beta_{loc}\rangle$ et $\langle\alpha_{loc}\rangle$ ne sont pas négligeables. Cependant, une bonne valeur pour $\langle\beta_{loc}\rangle$ et $\langle\alpha_{loc}\rangle$ n'est pas forcément synonyme d'une bonne reconnaissance de domaine. Par exemple, sur la figure 2.12, on a représenté trois répartitions possibles de domaines ouverts et fermés autour d'une région codante. Dans tous les cas, β_{loc} reste élevée mais c'est seulement dans le premier cas que le domaine codant est correctement identifié.

Une inspection superficielle des indicateurs locaux définies précédemment ne révèle pas de différences frappantes pour le pouvoir de prédiction de l'analyse des domaines entre PF et TW par exemple, alors que l'analyse au niveau des paires avait révélé un contraste saisissant entre les 2 espèces. Pour aller plus en avant dans l'analyse au niveau des domaines, on s'intéresse dans la figure 2.13 à la distribution de probabilité jointe pour une paire de base d'appartenir simultanément à un domaine codant ou non-codant (de taille L_{cod} ou L_{reste}) et à un domaine fermé ou ouvert (L_{ferm} ou L_{ouv}). Pour TW (figure 2.13 B), aucune corrélation significative entre domaines fermés et codants et entre domaines ouverts et non-codants n'est visible : les points sont distribués de manière diffuse et souligne la caractère aléatoire de la reconnaissance de domaine pour TW. Pour PF (figure 2.13 A), les bonnes corrélations entre L_{ferm} et L_{cod} et entre L_{ouvert} et L_{reste} prouvent le succès de l'identification des domaines codants et non-codants. Pour les longs domaines codants (ou non-codants), on observe une perte des corrélations avec les régions fermées (ou ouvertes) et la distribution de points est plus ou moins homogène pour $L_{ferm} < L_{cod}$ (ou $L_{ouv} < L_{reste}$) et quasi nulle pour $L_{ferm} > L_{cod}$ (ou $L_{ouv} > L_{reste}$). Cela implique que pour PF, les grands domaines biologiques ont tendance à être diviser en plus petites régions thermodynamiques (comme dans la figure 2.12 C) indiquant des cas où la correspondance thermodynamique/génomique n'est pas valide ou des gènes multi-exons [50] qui n'auraient pas été identifiés par les méthodes standard d'annotation des génomes [55, 56, 57, 58, 59, 60, 61, 62, 63] (voir section suivante). Pour les autres organismes étudiés, la même représentation montre qu'une bonne corrélation entre domaines codants et fermés est observée pour les espèces avec une forte valeur pour $\Delta\beta_{max}$ et une bonne corrélation entre domaines non-codants et ouverts pour les espèces avec une forte valeur de $\Delta\alpha_{max}$.

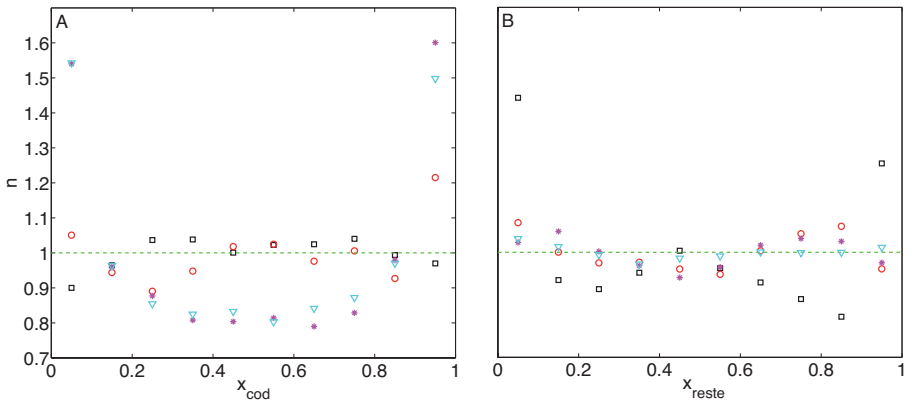

FIGURE 2.14 – Histogrammes normalisés $n_{cod}(x)$ et $n_{reste}(x)$ de la distance relative x entre une erreur située dans un domaine codant (A) ou non-codant (B) et le bord gauche de ce domaine, calculés pour plusieurs espèces (même légende que pour la figure 2.12). Les lignes pointillées vertes représentent le cas où les erreurs sont aléatoirement distribuées le long des domaines.

La dernière étape de notre analyse des domaines est d'étudier la location des paires de bases incorrectement identifiées dans les domaines codants ou non-codants. Suivant l'idée développé dans [71], pour chaque erreur positionné en i, on calcule la distance normalisée $x \equiv (i - i_1)/(i_2 - i_1)$ où $i_1 < i_2$ sont les paires de bases délimitant le domaine codant ou non-codant auquel appartient i. $x \sim 0$ ou $x \sim 1$ indique que l'erreur est située près des bords. Sur la figure 2.14, on trace les histogrammes normalisés $n_{cod}(x)$ et $n_{rest}(x)$ des distances relatives x observées dans les domaines codants et non-codants. Aucune règle ou dépendance n'apparaît clairement. On peut juste remarquer qu'en général, si $n_{cod}(x)$ a une distribution piquée autour des bords, $n_{reste}(x)$ demeure relativement plat et aléatoire, et vice et versa. Si les erreurs sont plus ou moins localisées autour des bords d'un domaine, on peut cependant considérer que l'identification du domaine en entier est assez bonne.

2.3.3 Identification de gènes et exons

Dans les deux sections précédentes, on a estimé si pour certaines espèces il existait une corrélation entre les propriétés thermodynamiques et génomiques de l'ADN. Dans cette section, on évalue si l'on peut se servir de l'analyse des propriétés de dénaturation d'un génome pour faire des prédictions fiables sur l'existence de gènes ou d'exons en utilisant les annotations des génomes déjà présentes dans les bases de données. Ces dernières sont pour la plupart réalisées en combinant plusieurs méthodes bioinformatiques de recherche de gènes. Des programmes comme GlimmerM [60] ou phat [62] sont basés sur la recherche de motifs pré-définis et de modèles de Markov [63]. Leurs résultats ne constituent pas l'absolue vérité, ainsi les incohérences relevées dans les sections précédentes pourraient bien être des gènes ou des exons non-identifiés comme tels dans les bases de données. Il est clair cependant qu'il y a très peu de raison de faire confiance à l'annotation issue de l'étude des domaines thermodynamiques pour TW, par contre pour PF, on peut imaginer que certaines des prédictions peuvent être pertinentes.

Par exemple, sur la figure 2.15 A, on a représenté une partie du chromosome 3 de PF où notre analyse des génomes identifie un gène putatif composé de 4 exons entre les gènes PFC0905c et PFC0910w. De même, PFC0910w, décrit dans la base de données comme formé d'un seul exon, pourrait être composé de 3 exons. L'inspection du même segment de chromosome avec le modèle de PS complet montre que les domaines proposés ne sont pas des artefacts du modèle ZB (voir figure 2.15 A). D'ailleurs, si au cours de l'analyse des génomes avec la méthode rapide ZB, des doutes apparaissaient sur une partie

de la segmentation proposée, on peut très bien envisager de refaire la même analyse sur des sections spécifiques mais avec le modèle PS pour vérifier ou ajuster les prédictions.

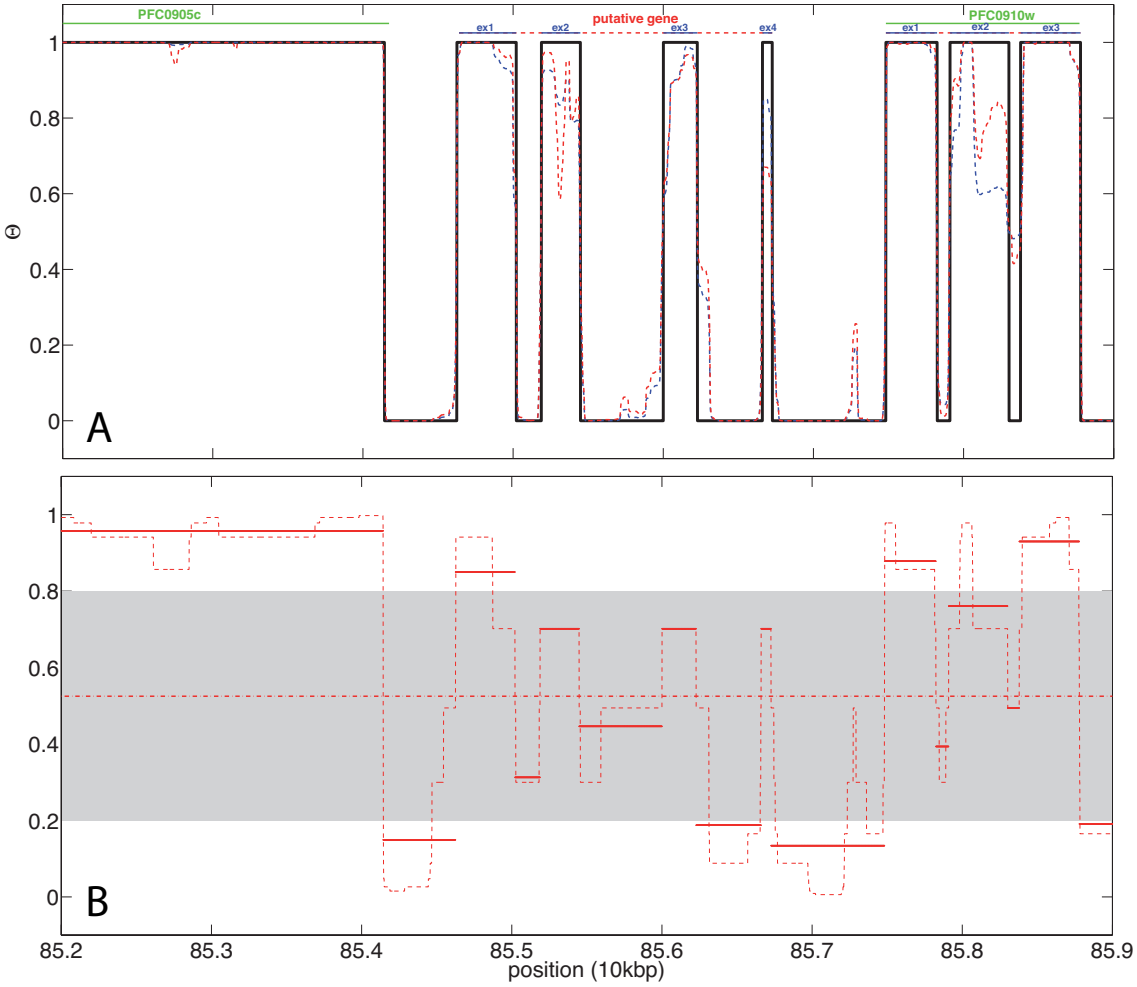

Figure 2.15 – (A) Domaines codants (vert) et fermés/ouverts (noir) à T_{opt} pour une partie du chromosome 3 de PF. On identifie un gène putatif composé de 4 exons et une division putative de PFC0910w en 3 exons. On trace également la probabilité Θ calculée à T_{opt} avec le modèle ZB (ligne pointillée bleue) et le modèle PS (ligne pointillée rouge). (B) Évolution de $P(\text{codant}|T_m)$ (ligne pointillée rouge) ou de sa valeur moyennée sur un domaine thermodynamique (lignes pleines rouges) pour un segment du chromosome 3 de PF (le même que précédemment). La ligne points-et-traits rouge représente la probabilité $\%CDS$ de prédire aléatoirement une paire de base comme codante. En dehors de la zone grisée ($0.2 < P(\text{codant}|T_m) < 0.8$), les prédictions sont considérées comme étant sures.

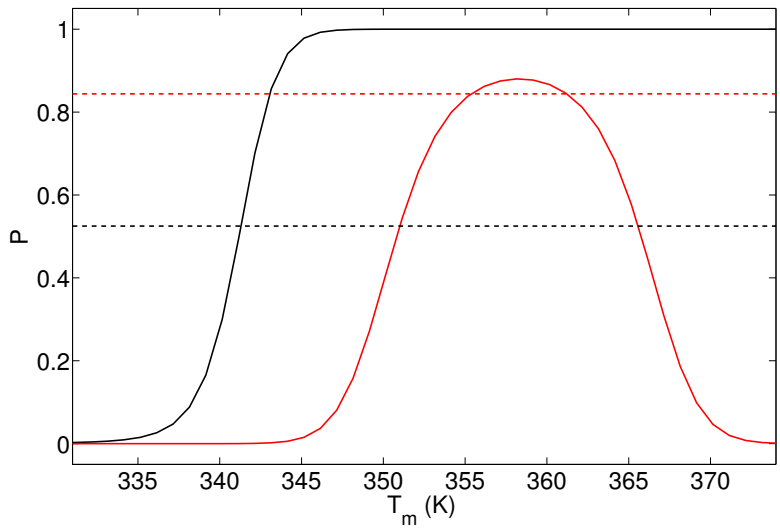

FIGURE 2.16 – Probabilité $P(\text{codant}|T_m)$ pour une paire de bases d'appartenir à une région codante en fonction de sa température de dénaturation pour PF (ligne noire) et TW (ligne rouge). Le pourcentage de paires codantes $\%CDS$ est tracé en ligne pointillée.

Dans ce qui suit, on veut déterminer un moyen pour quantifier pour chaque paire de bases ou chaque domaine le niveau de confiance des prédictions faite par l'analyse thermodynamique des génomes. On base notre estimation sur les distributions des températures de dénaturation locales, $P(T_m|\text{codant})$ (voir lignes pleines de la figure 2.8 B) et $P(T_m|\text{non}-\text{codant})$ (voir lignes pointillées de la figure 2.8 B), dans les parties du génomes identifiées comme codante ou non-codante par d'autres méthodes d'annotation. A partir de celles-ci, on peut les inverser, avec le théorème de Bayes (voir annexe 6.4), pour obtenir la probabilité qu'une paire de bases ayant une température de dénaturation T_m soit codante ou non-codante :

$$\begin{aligned} P(\text{codant}|T_m) &= \frac{P(\text{codant})}{P(T_m)} P(T_m|\text{codant}) \\ &= [\%CDS \times P(T_m|\text{codant})] \,/\, [\%CDS \times P(T_m|\text{codant}) + \\ &\quad (1-\%CDS) \times P(T_m|\text{non}-\text{codant})] \qquad (2.9) \\ P(\text{non}-\text{codant}|T_m) &= 1 - P(\text{codant}|T_m) \qquad (2.10) \end{aligned}$$

Ces quantités nous donne un excellent moyen d'évaluer localement la confiance en nos prédictions. Par souci de simplicité et de robustesse, on les détermine à partir des modélisations par des gaussiennes de $P(T_m|\text{codant})$ et $P(T_m|\text{non}-\text{codant})$. Si nos prédictions étaient faites via un processus d'annotation aléatoire, on aurait $P(\text{codant}|T_m) = P(\text{codant}) = \%CDS$.

Considérons tout d'abord le cas de PF pour lequel on s'attend à ce que l'analyse marche plutôt bien. Sur la figure 2.16, on observe que $P(\text{codant}|T_m)$ a une forme sigmoïdale. Cela correspond à l'argument original de Yeramian posant que les bases stables (instables) thermodynamiquement (donc ayant une température de mélange plus élevée (basse) que la moyenne) ont une forte probabilité d'être codantes (non-codantes). La figure 2.15 B représente le même segment du chromosome 3 de PF que la figure 2.15 A pour lequel on a tracé $P(\text{codant}|T_m)$ (pour chaque paire de bases ou moyennée sur un domaine). On remarque que pour le gène putatif introduit précédemment, seule la prédiction pour le premier exon atteint un bon niveau de confiance ($P(\text{codant}|T_m) > 0.8$). Pour la division putative de

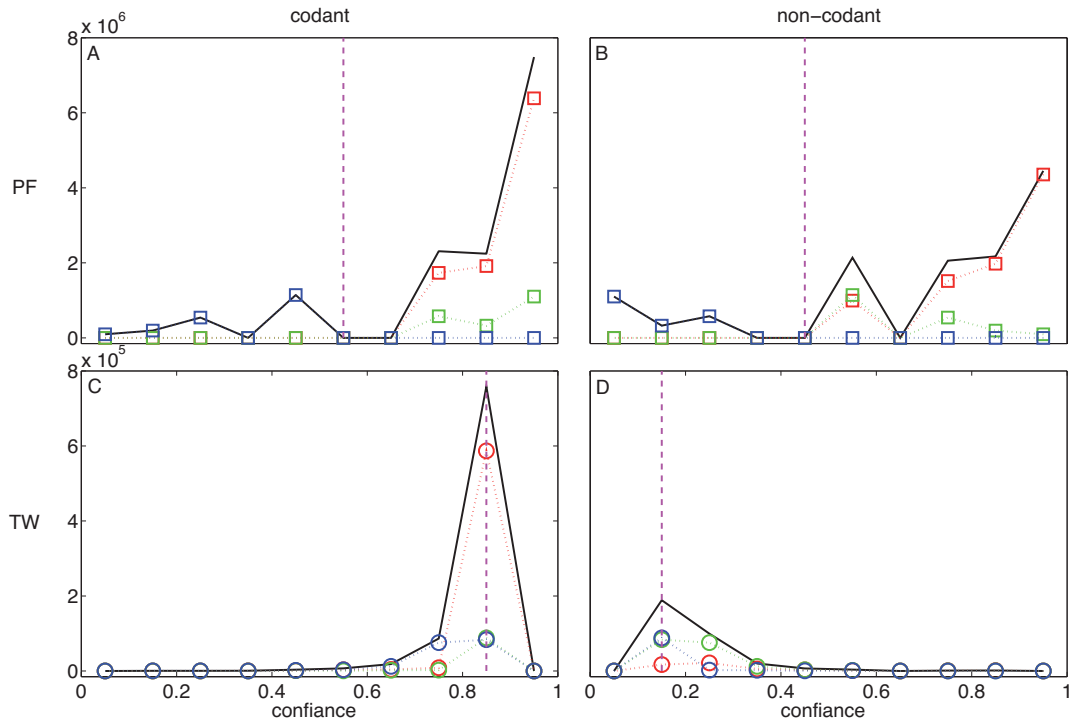

Figure 2.17 – Nombre de paires de bases prédites comme codantes (A,C) ou non-codantes (B,D) par l'analyse thermodynamique ou par d'autres méthodes d'annotations (lignes pleines noires), par l'analyse thermodynamique et par d'autres méthodes d'annotations (lignes pointillées rouges), par l'analyse thermodynamique uniquement (lignes pointillées vertes) et par d'autres méthodes uniquement (lignes pointillées bleues), en fonction du niveau de confiance ($P(\text{codant}|T_m)$ pour les prédictions codantes et $P(\text{non}-\text{codant}|T_m)$ pour les prédictions non-codantes), pour PF (A,B) et TW (C,D). Les lignes mauves représentent le niveau de confiance pour une annotation aléatoire.

PFC0910w en 3 exons, le niveau de confiance des introns prédits est relativement faible.

Sur les figures 2.17 A et B, on compare les annotations issues des méthodes standard [55, 56, 57, 58, 59, 60, 61, 62, 63] et de l'analyse thermodynamique pour le génome complet de PF en fonction du niveau de confiance que nous avons envers nos résultats. Comme prévu par nos résultats des précédentes sections, les deux annotations coincident largement. La majorité des domaines codants putatifs sont détectés par les deux méthodes avec un haut niveau de confiance en ce qui concerne l'analyse thermodynamique. Les déviations peuvent clairement être groupées en deux catégories :

1. les échecs, où l'analyse thermodynamique propose des annotations différentes ayant des faibles niveaux de confiance ;
2. les prédictions qui sont réalisées à un haut niveau de confiance.

La deuxième catégorie correspond principalement au petit pic présent dans $P(T_m|\text{non}-\text{coding})$ autour de la température de dénaturation moyenne des régions codantes, et qui est visible sur la figure 2.8 B. Ces domaines putatifs à haut niveau de confiance uniquement détectés par l'analyse thermodynamique sont listés dans le CD annexe (fichier *annotationPF.pdf*).

Les résultats pour TW sont complètement différents. La figures 2.17 C montrent que le haut de confiance des prédictions pour les régions codantes ne fait que refléter la forte proportion de séquences codantes dans le génome. En effet, $P(\text{codant}|T_m) \sim \%CDS$ correspond au niveau de confiance d'une annotation aléatoire qui reproduit la densité moyenne en séquences codantes. Ce niveau peut être assez

élevé ($\%CDS(TW) = 84\%$), mais, bien sûr, on n'apprendra rien de significatif avec cette analyse. Sur la figure 2.17 D, on observe que l'analyse thermodynamique donne des résultats similaires au cas aléatoire pour les régions non-codantes de TW. En particulier, notre analyse ne peut prédire aucune de ces régions avec un niveau de confiance raisonnable. Dans la section 2.3.1, on précisait que le problème venait du chevauchement important des deux distributions de températures de dénaturation pour les paires de bases codantes et non-codantes (figure 2.8 B). Une inspection plus fine de $P(\text{codant}|T_m)$ dans la figure 2.16 montre que la situation est même pire. Le génome de TW représente un contre-exemple pour l'hypothèse de travail de l'analyse thermodynamique que nous faisons : à cause de l'étalement des températures de dénaturation des régions non-codantes, une paire de bases avec une température de dénaturation élevée a plus de chance d'être non-codante que codante ! Notons cependant que cela ne représente pas un problème conceptuel en soit : tant que les domaines codants et non-codants diffèrent de part leurs caractéristiques thermodynamiques, il est possible d'exploiter ces différences pour construire un schéma d'annotation en suivant les principes développés dans cette section.

2.4 Conclusion

2.4.1 Bilan

Dans les sections précédentes, nous nous sommes intéressés à deux questions indépendantes concernant l'identification de séquences génomiques sur la base d'une analyse thermodynamique de la dénaturation des génomes :

1. quel modèle doit-on utiliser pour calculer les profils de dénaturation de long génomes ?
2. comment quantifier la fiabilité de nos prédictions ?

Ainsi, dans une première partie (sections 2.1 et 2.2), nous avons montré que le "vieux" modèle ZB permettait d'estimer fidèlement les températures de dénaturation locales même si les termes d'entropie de bulles n'étaient traités qu'incorrectement comparés au modèle PS. Cela souligne l'importance de l'hétérogénéité de la séquence dans les propriétés physiques des ADN génomiques. La méthode de calcul du modèle ZB le rend en plus très attractif d'un point de vue du coût numérique, permettant d'étudier les profils de dénaturation de génomes complets avec un PC ordinaire ($\sim 10^8$ bp/heure).

Dans la partie principale (section 2.2.2), nous avons évalué les possibles corrélations existantes entre les propriétés codantes et thermodynamiques de dénaturation pour des séquences ADN à l'échelle de génomes entiers. En particulier, nous avons développé une méthode pour estimer le niveau de confiance de notre annotation thermodynamique qui est basée sur la comparaison statistique avec d'autres annotations faites indépendamment. Cette analyse thermodynamique peut être résumée par le schéma suivant :

1. Calculer $T_m(i)$ pour toutes les paires de base du génome à étudier.
2. En se basant sur une annotation existante du génome ou d'une de ces parties, évaluer les distributions $P(T_m|\text{prop})$ et $P(T_m|\text{non}-\text{prop})$ où "prop" est une propriété génomique (le caractère codant pour notre étude) et T_{opt} la température optimale où la différence de taux de corrélation par rapport au cas aléatoire est maximale .
3. Inverser ces distributions pour obtenir les probabilités $P(\text{prop}|T_m)$ et $P(\text{non}-\text{prop}|T_m)$ d'avoir ou non la propriété voulue (voir équations 2.9 et 2.10).
4. Pour tous les domaines formés par les paires de bases adjacentes fermées ou ouvertes à T_{opt}, estimer la fiabilité de la prédiction : domaine fermé (ouvert) = domaine ayant (ou non) la propriété génomique étudiée.
5. En déduire l'existence de régions putatives pour les prédictions à haut niveau de confiance.

Pour certaines espèces, les corrélations sont telles qu'elles nous permettent d'identifier de nouveaux gènes ou exons avec fiabilité, suggérant qu'un couplage entre une approche thermodynamique (ou

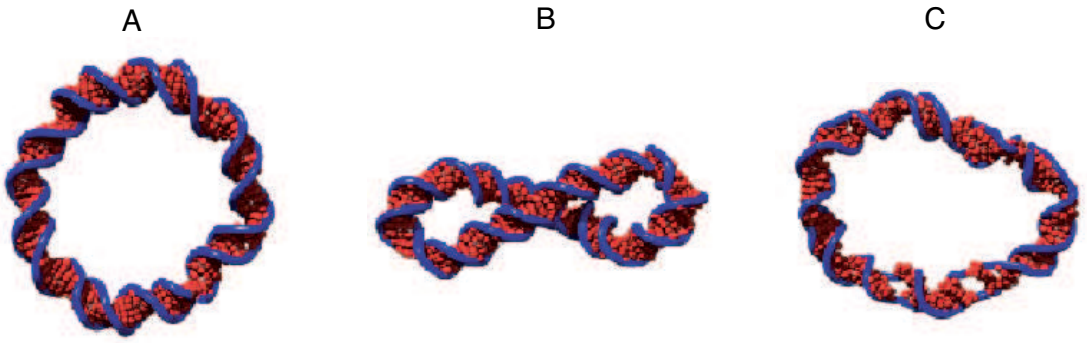

Figure 2.18 – Différentes conformations d'un ADN circulaire : (A) molécule totalement fermée et relaxée, (B) molécule totalement fermée et surenroulée, (C) molécule partiellement ouverte et faiblement surenroulée. Les figures sont tirées de [119].

physique) et d'autres méthodes d'identification des gènes pourrait améliorer le processus d'annotation des génomes. Dans d'autres cas, l'analyse thermodynamique ne donne pas de résultats probants. Le facteur clé semble être la différence en contenu GC entre les parties codantes et non-codantes.

Qualitativement, ce succès mitigé avait déjà été noté par Yeramian [49] et avait été interprété comme la signature d'une influence forte des propriétés physiques sur l'organisation des génomes archaiques, qui se serait partiellement estompée au cours des différentes étapes de l'évolution. Il pourrait être intéressant d'utiliser notre méthode pour analyser plus précisément cette hypothèse. Nos résultats actuels ne montrent cependant pas de différences significatives entre procaryotes et eucaryotes ainsi qu'entre des espèces appartenant à la même classe phylogénétique (SP et SC).

Ce ne sera qu'à travers une comparaison extensive avec des résultats issus de schémas d'annotation indépendants que l'on pourra juger de la pertinence de l'analyse thermodynamique et de la fiabilité de ses prédictions. Malheureusement, cette validation ne peut être faite une fois pour toute et nécessite d'être répétée pour chaque génome (ou parties de génome) à étudier. L'utilisation du modèle ZB n'est pas essentiel pour l'analyse des résultats que nous avons faite dans la partie principale de notre étude. Bien entendu, les profils de dénaturation calculés par le modèle PS pourraient être analyser exactement de la même manière. Nous n'avons vu aucun indice traduisant que cela conduirait à des résultats significativement différents. Cependant, pour des ressources de calcul fixées et limitées, nous préconisons un élargissement du nombre de génomes étudiés plutôt qu'une amélioration marginale de la description physique sous-jacente.
Ce travail a abouti à la publication d'un article dans *Journal of Physics : Condensed Matter* [118].

2.4.2 Perspective : inclure les effets de superhélicité

Précédemment, on a étudié la dénaturation de l'ADN en utilisant des modèles basés uniquement sur une description thermodynamique de l'ouverture des paires de bases. Ces approches prédisent, pour des longues séquences, une dénaturation de l'ADN à des températures non-biologiques ($\sim 80^o$ C à une concentration en sel de 0.1 M) et la probabilité d'observer l'ouverture de paires de bases à 37° C est quasiment nulle. Le lien éventuel entre propriétés thermodynamiques et génomiques entrevu ici est plus le reflet de corrélations croisées entre contenu GC et stabilité thermodynamique d'une part et entre contenu GC et composition des gènes d'autre part, mais ne reflète pas un processus biophysique réel car in vivo la dénaturation locale des paires de bases ne peut pas être réalisée thermiquement. Un moyen commun pour contrôler cette ouverture est d'imposer un stress superhélical à l'ADN (voir figures 2.18 B et C). Dans les organismes vivants, l'ADN est hautement contraint topologiquement en domaines circulaires (molécules circulaires ou boucles dans la chromatine) [120] (voir figure 2.18).

Chaque domaine est caractérisé par un invariant topologique L appelé le nombre de raccordement. L représente le nombre algébrique de tours que fait chaque brin d'ADN autour de l'axe centrale de la double-hélice. Faire varier L est possible uniquement sous l'action externe d'enzymes ou de protéines qui coupent puis recollent les deux brins d'ADN ou qui exercent une contrainte mécanique sur la double-hélice [121]. Pour la majorité des organismes, un surenroulement négatif est imposé à l'ADN, ce qui signifie que L est plus faible que le nombre de raccordement L_0 qu'aurait le même ADN si il n'était pas contraint (la différence $\alpha = L - L_0$ est alors négative). Seuls les organismes thermophiles sont connus pour avoir une valeur positive pour α.

Pour décrire l'effet du surenroulement sur la dénaturation des molécules circulaires d'ADN, Benham a introduit un modèle où la description thermodynamique standard de l'appariement des paires de bases est couplée à une modélisation des énergies dues au stress superhélical [122, 123, 124] (voir annexe 6.5). Avec son modèle, Benham trouve que des sites spécifiques du génome comme les sites de début de transcription ou les origines de réplication, sont préférentiellement déstabilisés par un surenroulement négatif [74, 125, 126]. La résolution exacte du modèle nécessite un algorithme en $\mathcal{O}(N^2)$ [124] ce qui est trop coûteux pour l'étude de génome entier, des algorithmes simplifiés ont été développés [127, 128] réduisant le temps de calcul.

Pour accélérer encore plus ce temps de calcul (sans une perte importante de précision bien évidemment) et pour avoir accès à des quantités difficiles à évaluer avec les approches utilisées par Benham (telles que la probabilité d'ouverture d'une bulle entre deux paires de bases données), nous avons développé une résolution auto-consistante d'une approximation de champ moyen du modèle de Benham (voir annexe 6.5). Deux étudiants de Master travaillent actuellement sur la validation (comparaison avec la solution exacte) et les applications d'une telle approche. Par exemple, nous aimerions quantifier la probabilité d'ouverture moyenne d'une bulle d'une taille donnée en fonction du contenu GC de la séquence, de la température et du stress superhélical, ou encore nous pourrions reprendre les idées développées dans le cadre de l'étude thermodynamique des génomes pour estimer l'existence de corrélations entre ces probabilités d'ouverture de bulle et les propriétés génomiques.

Deuxième partie

Repliement de l'ARN

Introduction

État de l'art

Depuis plus de trente ans, des efforts considérables ont été réalisés afin de prédire les structures secondaires et tertiaires ainsi que les propriétés thermodynamiques des molécules d'ARN. Ces efforts se sont principalement basés sur des approches bioinformatiques ou physico-chimique (pour un large passage en revue de ces différentes méthodes voir [129, 130]).

La plupart des méthodes étudiant la structure secondaire sont basées sur le modèle de Turner [85] (voir section 3.1.1). Ce modèle est la version ARN du modèle de Poland-Scheraga pour l'ADN (voir chapitre 1) : une structure secondaire y est décrite par un agencement de boucles et de parties double-brins. Pour traiter le modèle de Turner, plusieurs techniques ont été développées parmi lesquelles :

1. des algorithmes récursifs [87, 88, 86, 131, 132, 133] consistant à résoudre des relations de récurrence sur des fonctions de partition partielles (généralisation aux structures secondaires plus complexes de l'ARN de la méthode utilisée pour le modèle de Poland-Scheraga, section 1.1.3) ; l'ajout de contraintes structurelles déterminées expérimentallement améliorant l'efficacité des prédictions [134, 135] ;
2. des algorithmes génétiques [136, 137] consistant à faire évoluer (au sens biologique) une population de structures secondaires en ne gardant que les structures les plus stables thermodynamiquement : la génération de nouvelles structures secondaires est réalisée en recombinant des stuctures existantes et en y intégrant des mutations aléatoires (ouverture/fermeture de paires de bases).
3. des schémas de Monte-Carlo cinétiques [138, 139, 140] consistant à faire évoluer (au sens cinétique) une structure secondaire à l'aide de mouvements élémentaires (ouverture/fermeture de paires de bases).

Une description plus détaillée de certaines versions du modèle de Turner que l'on utilisera par la suite (appartenant notamment à la catégorie 1 et 3) est donnée en annexe 6.6.
Prédire la structure tertiaire de l'ARN est une affaire encore plus complexe car il nécessite une description plus fine de la molécule et la prise en compte d'interactions non-canoniques qui sont en grande partie responsables de la compactification de la structure tertiaire. Cependant, plusieurs stratégies existent parmi lesquelles :

1. des modélisations 3D de la structure secondaire prédite par une analyse phylogénétique de la séquence [141, 144] consistant à construire la structure tertiaire correspondante à partir de coordonnées atomiques typiques issues de données cristallographique (voir figure 2.19 A) ;
2. des simulations de dynamique moléculaire de modèles gros-grains utilisant des champs de forces dont la parametrisation est basée sur des données thermodynamiques [34, 39, 145, 142, 35] (voir figure 2.19 B) ;
3. des méthodes d'assemblage de fragments d'ARN guidé par la minimisation d'une fonction énergie empirique tirée de données cristallographiques [146, 143] (voir figure 2.19 C) ; l'ajout de contraintes structurelles déterminées expérimentallement améliore l'efficacité des prédictions [147].

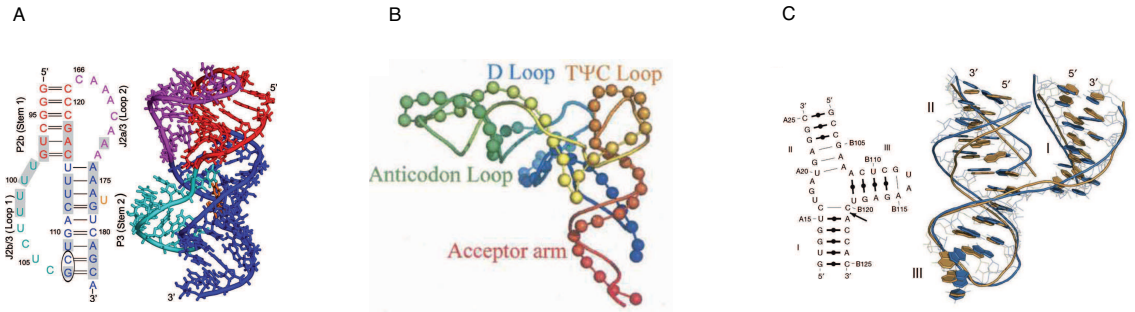

FIGURE 2.19 – Exemples de modèles tridimensionnelle de l'ARN : (A) Reconstruction 3D à partir de la structure secondaire par Yingling et Shapiro [141] ; (B) Dynamique moléculaire d'un modèle gros-grain de l'ARN par Ding et collaborateurs [142] ; (C) Assemblage de fragments d'ARN par Parisien et Major [143].

Objectifs

L'évaluation de la différence d'énergie libre $\Delta G = G_f - G_u$ entre une structure repliée (caractérisée par son énergie libre G_f) et la structure dénaturée correspondante (G_u) est primordiale pour prédire les propriétés thermodynamiques ainsi que la structure native des acides nucléiques. La plupart des modèles existants suivent l'idée de Poland et Scheraga [26] de considérer les polynucléotides (ADN/ARN) comme des polymères dont les contributions énergétiques sont associées à des intéractions locales (association de paires de bases, empilement au niveau des fourches, capping,etc.) et à des effects génériques associés à la nature polymériques des acides nucléiques (entropies des boucles, interaction à longue portée de volume exclu). ΔG peut alors être décomposée en une partie due aux intéractions locales spécifiques (dépendantes de la séquence) et en une partie comptant pour l'énergie de conformation de la structure et pour la nucléation des boucles. Dans le cas de l'ARN, les intéractions spécifiques sont principalement fournit par le modèle de Turner dont plus de 650 paramètres indépendants ont été déterminés expérimentalement.

Les contributions génériques ne doivent pas être négligée car elles jouent un rôle prépondérant dans la thermodynamique du repliement de l'ARN [133, 148]. Dans la plupart des modèles standard dévoués à l'étude de la structure secondaire, la partie de ΔG décrivant l'énergie de conformation et de nucléation des boucles est donnée par la somme des énergies de formation $\Delta g_{loop}(i)$ propres à chaque boucle i présente dans la structure considérée. Pour décrire les petites boucles, les modèles courants utilisent essentiellement des données expérimentales tabulées [85]. Pour des plus grandes boucles, une modélisation est nécessaire pour prendre en compte la dépendance en la taille de la boucle.

Pour des structures simples comme les boucles en épingle ou internes (voir figures 3.1 c et d), Δg_{loop} est en général bien décrite par une équation de Jacobson-Stockmayer [149] de la forme

$$\Delta g_{loop}^{JS}(n) = -T(\Delta s_{loop} - k_B c \log n) \tag{2.11}$$

où n est la taille de la boucle, $-T\Delta s_{loop}$ est l'énergie de nucléation et c est un exposant caractéristique de la nature polymérique des boucles. Le modèle de Turner ne tient compte que des interactions de volume exclu internes à la boucle et prend $c = 1.75$, l'exposant universelle des chemins fermés auto-évitant [77, 81]. Notons, cependant, que dans le cas de petites boucles connectées à de longues parties double-brins, les interactions avec le reste de la chaîne ne sont pas négligeables et un exposant $c \sim 2.1$ serait dans ce cas mieux adapté [78, 150].
Pour des structures plus complexes comme les boucles multiples (figure 5.2 C) ou les pseudo-noeuds (figure 5.11 B), dont les sous-structures peuvent interagir stériquement, il n'existe pas de description standard. La plupart du temps, Δg_{loop} est approximée par une équation de Jacobson-Stockmayer

généralisée

$$\Delta g_{loop}^{genJS}(n,h) = k_B T(a + b \times n + c \log n + d \times h) \tag{2.12}$$

avec n le nombre de nucléotides non-appariés dans la boucle, h le nombre de parties double-brins connectées à la boucle, et a, b, c et d sont des paramètres qui dépendent de la structure considérée. Changer (heuristiquement par exemple) la valeur de c dans les deux équations précédentes peut avoir des conséquences importantes sur les prédictions faites avec ce genre de modèle [133] si l'on ne prend pas soin de réajuster la pénalité de nucléation [80, 33, 118]. Remarquons que des relations de Jacobson-Stockmayer généralisées peuvent être requises pour chaque sous-classe de boucles, comme par exemple pour les pseudo-noeuds avec une relation décrivant les H-pseudo-noeuds sans lieur [151], une autre quand il y a un lieur [152], etc. Alternativement, Vernizzi, Orland et Zee [153, 154], proposent de modéliser les topologies complexes en incluant dans l'équation 2.12 un terme comptant pour le genus topologique de la structure secondaire considérée.

Se basant sur la nature polymérique des chaînes d'acides nucléiques, de nombreuses approches utilisant des modèles simples ont été développées pour étudier ces énergies de formation de boucles. En particulier, une importante partie de ces descriptions utilise une modélisation sur réseau de la molécule car cela permet un accès direct à l'entropie de conformation. Par exemple, l'énumération exacte de sous-structures comme des boucles ou des pseudo-noeuds sur des réseaux carrés ou cubiques a été employée pour estimer la dépendance en n de Δg_{loop} [155, 156, 157]. Ou encore, des simulations basées également sur des modèles sur réseau [158, 159, 160, 161, 33] ont été utilisées pour étudier les propriétés de dénaturation des acides nucléiques, comme l'ordre de la transition de phase du processus de dénaturation de l'ADN. Cependant, pour le moment, aucun modèle sur réseau n'a pu, de manière convaincante (par comparaison avec l'expérience), prédire précisément les propriétés thermodynamiques du repliement d'ARN hétéropolymères.

Notre objectif principal dans cette partie va être de développer un modèle sur réseau qui réussisse à capter les aspects polymériques génériques des chaînes d'ARN comme la connectivité, le volume exclu ou l'entropie de conformation grâce à sa définition sur un réseau, tout en décrivant la spécificité des molécules d'ARN par l'intermédiaire de paramètres dépendants de la séquence et similaires à ceux du modèle de Turner. Notre modèle doit être capable de reproduire de manière quantitative les données expérimentales et doit nous permettre de mieux comprendre certains aspects du repliement comme l'importance des interactions stériques internes à la molécule ou avec l'extérieur.

Plan

Dans un premier chapitre, nous définissons le modèle sur réseau. Après avoir introduit les différentes interactions considérées dans le modèle en comparaison à celles du modèle de Turner et avoir discuté en détail de certaines subtilités du modèle de Turner, nous expliquons comment paramétrer le modèle sur réseau pour qu'il reproduise le comportement de structures simples prédit par le modèle de Turner.

Dans un deuxième chapitre, nous développons les différentes méthodes numériques et algorithmiques que nous avons utilisées pour étudier le modèle sur réseau. En particulier, nous décrivons les deux schémas dynamiques et statiques employés respectivement pour prédire la thermodynamique du repliement d'une séquence donnée et pour corriger les énergies libres issues du modèle de Turner.

Dans un troisième chapitre, nous présentons les résultats obtenus avec le modèle sur réseau. Tout d'abord, on s'attelle à valider l'approche sur des exemples simples. Puis on estime le pouvoir de prédiction du modèle pour des structures complexes comprenant des boucles multiples ou des pseudo-noeuds. Ensuite, on évalue l'impact des interactions de volume exclu sur les prédictions et on explique comment corriger les approches standard avec un champ moyen pour qu'elles tiennent bien compte des effets stériques. Enfin, on exploite la définition tridimensionnelle du modèle pour étudier la prise en compte d'interactions tertiaires dans le modèle.

Finalement, nous concluons sur les avantages et les inconvénients de notre approche aux vues des résultats présentés et nous présentons de possibles perspectives à notre travail.

Chapitre 3

Modèle sur réseau

Dans une première partie, nous présentons le modèle sur réseau en le comparant au modèle de Turner [85, 162] et en l'illustrant d'exemples typiques. Puis, dans une deuxième partie, nous analysons en détail les paramètres du modèle de Turner. Enfin, dans une troisième partie, nous expliquons comment le modèle sur réseau est paramétré.

3.1 Définition du modèle

Notre approche suit la logique introduite par mon directeur de thèse, Ralf Everaers, avec S. Kumar et C. Simm [33] de considérer des intéractions locales spécifiques dépendantes de la séquence dans un modèle sur réseau générique inspiré par le travail de Causo, Coluzzi et Grassberger [158]. En particulier, nous faisons le rapprochement systématique entre les paramètres du modèle sur réseau et ceux des descriptions standard au niveau structure secondaire qui ont été paramétrés expérimentalement durant les 30 dernières années. En plus, notre modèle tient compte des effets de connectivité de la chaîne et des interactions polymériques génériques en modélisant chaque structure secondaire par un ensemble de conformations gros-grain tridimensionnelles. Dans l'esprit, notre approche est similaire à Kinefold [138] (voir section 6.6) mais va bien au-delà de la simple prise en compte de la connectivité en considérant explicitement et systématiquement les interactions stériques.

Le bien-fondé du modèle sur réseau repose sur la nature hiérarchique du repliement de l'ARN et peut ainsi être perçu comme une amélioration systématique des descriptions basées sur la structure secondaire. Cependant, notre modèle reste gros-grain, en particulier, nous ne décrivons pas la structure interne des éléments de la structure secondaire comme la structure en double-hélice des parties double-brins, ou nous n'essayons pas d'expliquer la dépendance des paramètres locaux en la séquence. Bien que ces derniers points constituent des problèmes extrêmement intéressants, nous pensons que le pouvoir de prédiction d'approches tertiaires plus microscopiques [34, 39, 145, 142, 35] pour des grandes molécules d'ARN est rapidement perdu si la thermodynamique des petits ARN n'est pas reproduite avec une précision équivalente à celle des modèles plus proches voisins standard.

3.1.1 Interactions dans le modèle de Turner et dans le modèle sur réseau

Dans le modèle de Turner, un structure secondaire, $\mathcal{S}$, est modélisée comme une succession de parties double-brins et de boucles (voir figures 3.1). Le modèle sur réseau quant à lui fournit une description gros-grain d'une conformation tridimensionnelle $\mathcal{C}$. Un brin d'ARN est alors modélisé par un chemin auto-évitant (SAW) sur un réseau régulier (voir figures 3.2 b et c) [158, 33]. Les positions possibles pour les nucléotides sont les noeuds du réseau. Deux nucléotides sont autorisés à se chevaucher si et seulement si ils peuvent s'apparier formant une paire de base de Watson-Crick (orientation antiparallèle) A/U ou G/C, ou la paire de base antiparallèle dite bancale G/U. La structure secondaire $\mathcal{S}(\mathcal{C})$ d'une conformation sur le réseau est facilement définie par ses contacts. Inversement, une structure secondaire

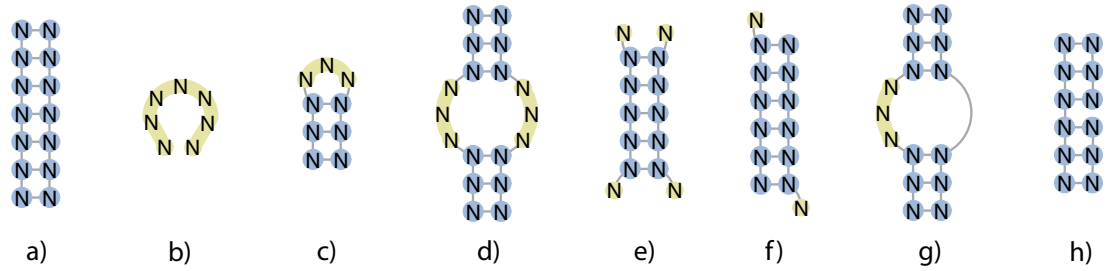

Figure 3.1 – Exemples de structures secondaires simples (de la gauche vers la droite) : double-brin, simple brin, épingle ("hairpin"), boucle interne, double-brin avec fourches externes, double-brin avec des brins pendants ou dangles, boucle latérale ("bulge"), double-brin entaillé ("nick"). Ces structures ont été réalisées à l'aide de l'application internet Pseudoviewer [163]. Les nucléotides colorés en bleu sont appariés, ceux en jaune ne le sont pas.

donnée est représentée par un ensemble $\{\mathcal{C}\}_{\mathcal{S}}$ de conformations. En général, le cardinal $\Omega(\mathcal{S})$ de cet ensemble est grand.

Le modèle de Turner et le modèle sur réseau sont définies à la même échelle et ont donc des paramètres similaires. L'énergie libre d'une structure secondaire $\mathcal{S}$ dans le modèle de Turner et d'une conformation $\mathcal{C}$ dans le modèle sur réseau peuvent être décomposées en une somme de termes à courte portée dit de plus proche voisin (NN) et de termes non-locaux (voir figures 3.2 a et b) :

- L'énergie libre d'association (notée $\Delta g^{st}_{NN} = \Delta h^{st}_{NN} - T\Delta s^{st}_{NN}$ dans le formalisme de Turner et $\epsilon(T) = \epsilon_H - T\epsilon_S$ dans celui du réseau) d'un segment de paires de bases tient compte de l'empilement entre deux paires voisines et dépend des 21 différentes possibilités d'empiler deux paires de bases côte à côte [162].
- L'énergie libre terminale ou de capping (Turner : $\Delta g_{term} = \Delta h_{term} - T\Delta s_{term}$; réseau : $\omega(T) = \omega_H - T\omega_S$) tient compte des paires terminales et dépend de la nature de la paire de base A/U, G/C et G/U.
- L'énergie libre de fourche (Turner : $\Delta g^{tm}_{NN} = \Delta h^{tm}_{NN} - T\Delta s^{tm}_{NN}$; réseau : $\gamma(T) = \gamma_S - T\gamma_H$) tient compte des interfaces entre une section double-brin et deux simples brins. Elle dépend des quatre nucléotides formant la fourche.
- L'énergie libre de dangle (Turner : $\Delta g^{dg}_{NN} = \Delta h^{dg}_{NN} - T\Delta s^{dg}_{NN}$; réseau : $\lambda(T) = \lambda_H - T\lambda_S$) tient compte des interfaces entre une section double-brin et un unique simple brin. Elle dépend des 3 nucléotides voisins.
- L'énergie libre non-locale de nucléation d'une boucle (Turner : $-T\Delta s^{\mathcal{T}}_{loop}$; réseau : $\sigma(T) = \sigma_H - T\sigma_S$). On la suppose indépendante de la composition de la boucle. Pour des raisons stériques, les boucles en épingle avec moins de deux nucléotides sont exclues [20].
- L'énergie libre d'empilement coaxial (Turner : Δg_{coax} ; réseau : $\chi(T) = \chi_H - T\chi_S$) tient compte de l'intéraction favorable entre deux double-brins empilés tête bêche. Elle dépend des 2 paires terminales empilées.
- L'énergie intermoléculaire de mélange ou d'initiation (Turner : Δg_{init} ; réseau : G_{mix}) tient compte de la perte d'entropie de translation et de rotation lors de l'association de plusieurs brins.
- Dans le modèle sur réseau uniquement, pour tenir compte de la rigidité du double-brin, on introduit une énergie libre de pliage $\kappa(T, \Psi) = \kappa_H(\Psi) - T\kappa_S(\Psi)$. Ψ représente l'angle de pliage de la double-hélice. Pour un réseau CFC, il existe quatre valeurs possibles pour Ψ (voir figure 3.7) : 0° (1 possibilité), 60° (4), 90° (2) and 120° (4). La possibilité $\Psi = 180^o$ est exclue. Dans l'exemple de la figure 3.2 b, $\Psi = 60^o$.

Les nucléotides non appariés adjacents à plusieurs double-brins ne peuvent pas participer à plus d'une

intéraction NN (fourche ou dangling). Les intéractions non-canoniques telles que les intéractions entre 3 nucléotides ou les paires de base *trans* de Watson-Crick, peuvent être en principe inclues de manières analogues. Par souci de simplicité et en l'absence de données fiables, on se restreint aux intéractions décrites ci-dessus. Dans le modèle de Turner, les énergies de fourche et de nucléation dépendent de la nature de la nulle adjacente (boucle interne, à épingle ,etc.). En fait, chaque conformation ou structure secondaire représente un grand nombre de microétats du système ARN/solvant. C'est pourquoi, dans le modèle sur réseau ainsi que dans celui de Turner, l'Hamiltonien est défini par une énergie libre dépendante de la température et non par une énergie. Chaque intéraction peut alors être décomposée en une contribution enthalpique et en une contribution entropique que l'on supposera chacune indépendante de la température.

La liste de paramètres ci-dessus illustre le lien intime entre les deux modèles. Il existe cependant une exception notable concernant les exposants universelles décrivant la nature polymérique des acides nucléiques [133, 77, 78, 80, 165]. En particulier, les relations de Jacobson-Stockmayer (equations 2.11 et 2.12), qui tiennent compte des énergies de formation d'une boucle (incluant l'énergie de nucléation), n'ont pas d'équivalent dans le modèle sur réseau. Au contraire, les effets de volume exclu y sont traités explicitement. Cependant, notre modèle ne décrit pas finement la structure 3D de la molécule. En particulier, la structure en double-hélice des brins n'est pas décrite.

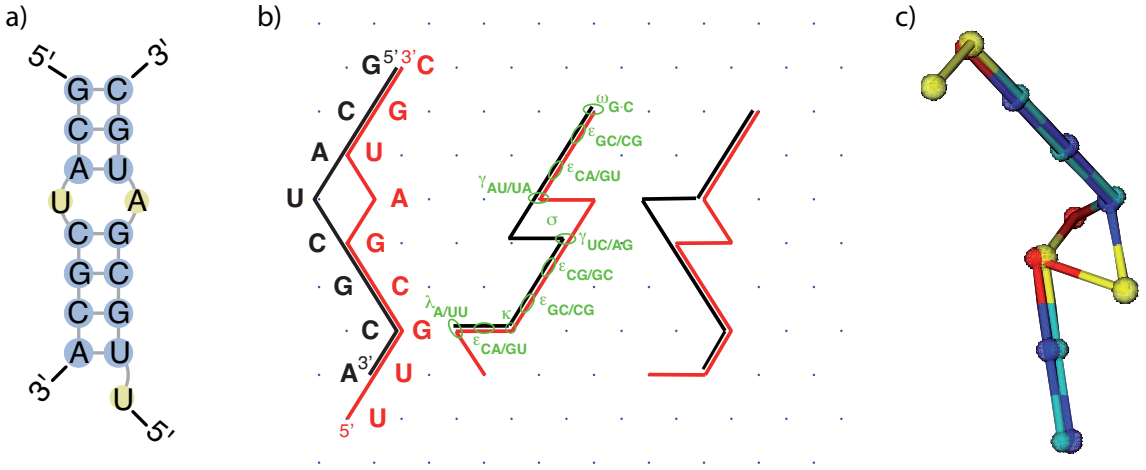

FIGURE 3.2 – Différentes représentations du complexe $(GCAUCGCA)\cdot(UUGCGAUGC)$. (a) Possible structure secondaire. (b) Projection 2D de conformations correspondantes à la même structure secondaire que dans (a) et définition des différentes contributions énergétiques sur le réseau. (c) Exemple de conformation 3D sur un réseau cubique à face centrée. Chaque sphère représente un nucléotide. Deux nucléotides appariés occupent la même position. On visualise les conformations sur réseau à l'aide du programme VMD [164]. Code couleur : les adénosines A en rouge, les guanines G en cyan, les uracyles U en jaune et les cytosines C en bleu.

3.1.2 Entropies des structures secondaires dans le modèle sur réseau

Dans le modèle sur réseau, l'énergie libre d'une structure secondaire $\mathcal{S}$ est donc donnée par la somme des intéractions décrites dans la section précédente et d'une contribution entropique, l'énergie libre de conformation $-k_BT\log\Omega(\mathcal{S})$. $\Omega(\mathcal{S})$ représente le nombre de conformations sur le réseau correspondantes à la même structure secondaire $\mathcal{S}$, la position du premier nucléotide étant fixée. Illustrons ceci en explicitant les énergies libres des structures secondaires typiques de la figure 3.1. Notons que dans le modèle de Turner, l'énergie libre d'une structure est donnée relativement à son état dénaturé (c'est donc en fait une différence d'énergie libre). Dans le modèle sur réseau, l'état dénaturé (simple brin) est traitée comme un SAW et le calcul de l'énergie libre d'une conformation quelconque se fait avec cette convention.

- Pour un double-brin composé de $N+1$ paires de bases (figure 3.1 a) : Turner :$\Delta G_{ds} = N\Delta g^{st}_{NN} + 2\Delta g_{term} + \Delta g_{init}$; réseau : $G_{ds} = N\epsilon + 2\omega - k_BT\log(z)$ avec z est le nombre de coordinence, c'est à dire, le nombre de plus proche voisins pour un noeud du réseau.
- Pour un simple-brin composé de $N+1$ nucléotides (figure 3.1 b) : Turner : $\Delta G_{ss} = 0$; réseau : $G_{ss} = -k_BT\log(f_s\mu^N N^{c'})$ où $f_s\mu^N N^{c'}$ est le nombre total de SAW pour un polymère de taille $N+1$. μ représente le nombre effectif de plus proche voisin et $c' = 1/6$ est un exposant universelle [81, 82].
- Pour une épingle composée par un double-brin avec $N+1$ paires de bases et une boucle avec $M-1$ nucléotides (figure 3.1 c) : Turner : $\Delta G_{hp} = N\Delta g^{st}_{NN} + \Delta g_{term} + \Delta g^{tm}_{NN} + \Delta g^{JS}_{loop}(M)$; réseau : $G_{hp} = N\epsilon+\omega+\gamma+\sigma-k_BT\log((z-2)f_l\mu^M M^{-c})$. Le nombre de conformations est égale au produit du nombre de polygones auto-évitants pour un polymère de taille $M+1$, $f_l\mu^M M^{-c}$, et du nombre de possibilités d'attacher une tige à une boucle, donné approximativement par $z-2$. $c \approx 1.76$ est un exposant universel [81, 77].
- Pour un double-brin comprenant une boucle interne avec $M-2$ nucléotides (figure 3.1 d) : Turner : $\Delta G_{int} = N\Delta g^{st}_{NN} + 2\Delta g_{term} + 2\Delta g^{tm}_{NN} + \Delta g^{JS}_{loop}(M) + \Delta g_{init}$; réseau $G_{int} = N\epsilon + 2\omega + 2\gamma + \sigma - k_BT\log((z-2)^2 f_l\mu^M(M)^{-c})$. On a appliqué la même logique que pour l'épingle afin d'évaluer $\Omega(\mathcal{S})$.
- Pour un double-brin avec deux fourches externes (figure 3.1 e) : Turner $\Delta G_{ext} = N\Delta g^{st}_{NN} + 2\Delta g^{tm}_{NN} + \Delta g_{init}$; réseau : $G_{ext} = N\epsilon+2\gamma-k_BT\log(z(z-1)^2(z-2)^2)$, où le facteur $(z-1)(z-2)$ tient compte du nombre de possibilités d'attacher les deux nucléotides non-appariés à l'interface avec le double-brin.
- Pour un double-brin avec deux dangles (figure 3.1 f) : Turner : $\Delta G_{dg} = N\Delta g^{st}_{NN} + 2\Delta g^{dg}_{NN} + \Delta g_{init}$; réseau : $G_{dg} = N\epsilon + 2\lambda - k_BT\log(z(z-1)^2)$. On a appliqué la même logique que pour les fourches externes afin d'évaluer $\Omega(\mathcal{S})$.
- Pou un double-brin avec une boucle latérale composée de $M-2$ nucléotides (figure 3.1 g) : Turner : $\Delta G_{lat} = N\Delta g^{st}_{NN} + 2\Delta g_{term} + \Delta g^{JS}_{loop}(M) + \Delta g_{init}$ (l'énergie coaxiale est négligée pour les grandes boucles) ; réseau : $G_{lat} = N\epsilon + 2\omega + \sigma + \chi - k_BT\log[f_l\mu^{M+1}(M+1)^{-c}(z-2)^2]$. On a appliqué la même logique que pour l'épingle afin d'évaluer $\Omega(\mathcal{S})$.
- Pour un double-brin entaillé au milieu (figure 3.1 h) : Turner : $\Delta G_{nick} = 2N\Delta g^{st}_{NN} + 2\Delta g_{term} + \Delta g_{coax} + 2\Delta g_{init}$; réseau : $G_{nick} = 2N\epsilon+2\omega+\chi-k_BT\log[z(z-1)^2]$ où le facteur $z(z-1)^2$ tient compte approximativement du nombre de possibilités d'attacher trois tiges rigides consécutivement.

Alors que les paramètres géométriques z, μ, f_s et f_l dépendent de la nature du réseau sous-jacent (cubique à face centrée, cubique,etc.), les exposants c et c' sont universels et sont charactéristiques de la nature polymérique des acides nucléiques.

Par la suite, nous travaillerons avec un réseau cubique à face centrée (CFC) pour lequel $z = 12$, $\mu = 10.035$, $f_s = 1.14$, $f_l = 0.25$ [166]. Le choix du réseau CFC au lieu du réseau cubique simple original [33] se justifie par la plus grande symétrie du réseau et par la possibilité de décrire toutes tailles de boucles (alors que seules les boucles de taille $(2n+2)$ avec n entier sont autorisées sur un

réseau cubique).

3.2 Paramètres dans le modèle de Turner

Avant de décrire comment nous paramétrons le modèle sur réseau, discutons quelques points subtiles du modèle de Turner. Après avoir brièvement expliqué comment Turner paramétrise son modèle (section 3.2.1), nous proposons dans les sections 3.2.2, 3.2.3 et 3.2.4 une unification des paramètres de fourches, coaxiaux et de défauts d'appariement ("mismatchs") au niveau structure secondaire. Ceci réduit de manière drastique le nombre de paramètres indépendants du modèle de Turner ($\sim 650 \rightarrow \sim 350$). Puis dans la section 3.2.5, nous montrons que le modèle de Turner est en fait défini modulo une constante qui bien sur n'influe en rien sur les prédictions du modèle. Enfin dans la section 3.2.6, nous étudierons la sensibilité de prédictions du modèle de Turner par rapport aux incertitudes sur les paramètres.

3.2.1 Paramétrisation du modèle de Turner

La paramétrisation du modèle de Turner se base sur des expériences de dénaturation réalisées pour des structures secondaires simples (comme celles explicitées dans la figure 3.1) qui permettent de mesurer la différence d'enthalpie totale ΔH et d'entropie totale ΔS (et donc ΔG).

Tout d'abord, les données sur un grand nombre d'oligomères courts d'ARN permettent d'avoir accès aux paramètres d'association, de capping et d'initiation [162] (même méthode que celle utilisée à la section 1.2.1 pour l'ADN). Cette étape est très importante puisqu'elle conditionne en grande partie la paramétrisation des autres paramètres. En effet, pour évaluer la contribution d'un terme dans l'énergie libre totale d'une structure secondaire donnée, la première étape est de soustraire à ΔG toutes les contributions issues des paramètres d'association, de capping et d'initiation.

Par exemple, pour la paramétrisation des dangles, de nombreuses expériences sur des structures comme celles décrites à la figure 3.1 f ont été réalisées pour plusieurs combinaisons de nucléotides au niveau du dangle. La contribution dûe au dangle est alors estimée par $\Delta g^{dg}_{NN} = (\Delta G_{tot} - \Delta G_{stem})/2$ où ΔG_{tot} est l'énergie libre totale mesurée et ΔG_{stem} est l'énergie libre de la partie double-brin calculée à l'aide des paramètres précédemment estimés (association des paires de bases contenu dans la partie double-brin et initiation).

Pour les paramètres de boucles (fourches et nucléation), la démarche précédente donne accès à l'énergie totale de la boucle incluant les énergies de fourches et l'énergie de formation de la boucle mais ne permet pas de faire de distinction entre les deux. Pour pouvoir séparer les deux, une modélisation est proposée : l'énergie de formation est décrite par une relation de Jacobson-Stockmayer (comme Eq.2.11) et les termes de fourches par une somme de bonus et de malus dépendant des nucléotides composant la fourche.

Deux versions des paramètres existent : la version 2.3 [167] et la version 3.0 [85]. Même si la version la plus récente (3.0) a été paramétrée à l'aide de beaucoup plus de données et reste la version la plus utilisée, quelques modèles comme Kinefold [140] et Vfold [132] utilisent toujours la version 2.3.

3.2.2 Unification des paramètres de fourches et de nucléations de boucle

Contrairement aux enthalpies de fourches, les entropies de fourches $\{\Delta s^{tm}_{NN}\}$ et de nucléation de boucle $\{\Delta s^{\mathcal{T}}_{loop}\}$ fournies par Turner dépendent de la nature de la boucle (épingle, interne,externe). D'une part, ceci n'est pas pratique dans notre formulation sur réseau puisque le calcul d'une variation de l'énergie libre associée à un réarrangement structural local nécessiterait une analyse globale de la structure secondaire pour connaître l'eventuel nouveau type de boucle associée. D'autre part, cette dépendance est surprenante d'un point de vue physique : comment des énergies locales comme les énergie de fourches pourraient être différentes selon la nature de la boucle connectée ? Dans la suite,

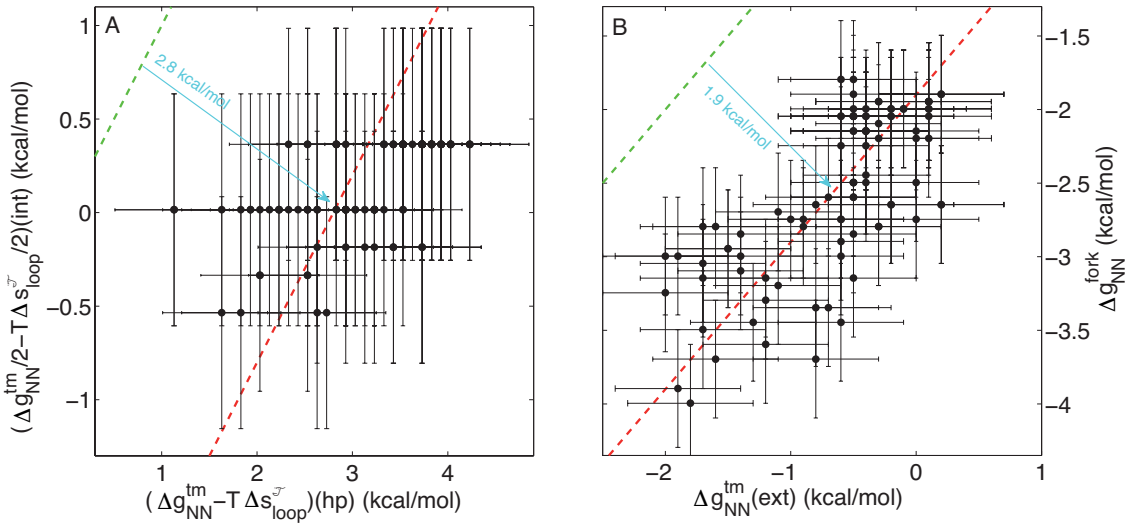

Figure 3.3 – (A) Corrélation entre les paramètres de Turner pour des boucles internes et en épingle. Le décalage observée réprésente la moitié de la pénalité unifiée de nucléation de boucle $-T_{37}\Delta s_{loop}$. (B) Corrélation entre les paramètres de Turner pour les fourches externes et les paramètres de fourches unifiés. On attribue la différence moyenne observée de 1.9 kcal/mol à des effets de bord.

nous allons montrer que les paramètres de nucléation de boucle et de fourche peuvent s'écrire de manière non-spécifique.

Regardons tout d'abord l'entropie totale liée à une boucle en épingle (hp) de taille N, $\Delta s_{hp} = \Delta s^{tm}_{NN}(\text{hp}) + \Delta s^{\mathcal{T}}_{loop}(\text{hp}) - k_B c \log N$. La prise en compte d'une jauge arbitraire ϕ_1 ne modifie en rien l'évaluation de Δs_{hp}, $(\Delta s^{\mathcal{T}}_{loop}(\text{hp}) - \phi_1) + (\Delta s^{tm}_{NN}(\text{hp}) + \phi_1) - k_B c \log N$. De même, on peut introduire un décalage arbitraire entre l'entropie de nucléation et celle de fourche pour les boucles internes (int) sans changer la mesure de l'entropie totale Δs_{int} : $(\Delta s^{\mathcal{T}}_{loop}(\text{int}) - \phi_2) + 2(\Delta s^{tm}_{NN}(\text{int}) + \phi_2/2) - k_B c \log N$.

La non-spécificité des paramètres impose

$$\Delta s^{\mathcal{T}}_{loop}(\text{hp}) - \phi_1 = \Delta s^{\mathcal{T}}_{loop}(\text{int}) - \phi_2 \equiv \Delta s_{loop} \tag{3.1}$$

$$\Delta s^{tm}_{NN}(\text{hp}) + \phi_1 = \Delta s^{tm}_{NN}(\text{int}) + \phi_2/2 \equiv \Delta s^{fork}_{NN} \tag{3.2}$$

où Δs_{loop} et Δs^{fork}_{NN} sont les paramètres non-spécifiques (ou unifiés) correspondants. Notre hypothèse de travail implique bien entendu l'existence d'une corrélation entre $\Delta s^{\mathcal{T}}_{loop}(\text{hp}) + \Delta s^{tm}_{NN}(\text{hp}) = \Delta s_{loop} + \Delta s^{fork}_{NN}$ et $\Delta s^{\mathcal{T}}_{loop}(\text{int})/2 + \Delta s^{tm}_{NN}(\text{int}) = \Delta s_{loop}/2 + \Delta s^{fork}_{NN}$. La figure 3.3 A montre clairement une telle corrélation. Ceci nous permet d'évaluer Δs^{fork}_{NN} et Δs_{loop} en modélisant les données par les équations 3.1 et 3.2 à l'aide de la méthode des moindres carrés déjà explicitée à la section 1.2.1. Pour évaluer la pertinence de la modélisation, on calcule la valeur correspondante de la fonction gamma incomplète Q (voir annexe 6.3). On trouve une valeur proche de 1, signature d'une bonne modélisation.
Finalement, on obtient comme paramètres unifiés

$$\Delta s_{loop} = -9.2 \pm 2\, k_B = -18.3 \pm 4 \text{ cal/mol/K} \tag{3.3}$$

Concernant les enthalpies de fourches, elles sont indépendantes du type de boucle dans le modèle de Turner, donc $\Delta h^{fork}_{NN} = \Delta h^{tm}_{NN}$ (voir table 3.1). Traditionnellement, on ne donne pas les entropies mais les énergies libres à 37° C qui sont présentées dans la table 3.2.

TABLE 3.1 – Enthalpies unifiées de fourches Δh_{NN}^{fork} en kcal/mol (1 kcal/mol $\approx 1.6 k_B T_{37^oC}$).

	$5'-AX-3'$ $3'-UY-5'$				$5'-CX-3'$ $3'-GY-5'$				$5'-GX-3'$ $3'-CY-5'$			
$X^{\ Y}$	A	C	G	U	A	C	G	U	A	C	G	U
A	−4.3	−6.0	−6.0	−6.0	−10.3	−9.5	−10.3	−10.3	−5.2	−8.8	−5.6	−8.8
C	−2.6	−2.4	−2.4	−2.4	−5.2	−4.5	−5.2	−6.7	−7.2	−3.1	−3.1	−3.9
G	−3.4	−6.9	−6.9	−6.9	−9.4	−9.4	−9.4	−9.4	−7.1	−7.4	−6.2	−7.4
U	−3.3	−3.3	−3.3	−3.3	−8.1	−7.4	−8.1	−8.6	−5.0	−5.0	−5.0	−5.7
	$5'-GX-3'$ $3'-UY-5'$				$5'-UX-3'$ $3'-AY-5'$				$5'-UX-3'$ $3'-GY-5'$			
$X^{\ Y}$	A	C	G	U	A	C	G	U	A	C	G	U
A	−4.3	−6.0	−6.0	−6.0	−4.0	−6.3	−8.9	−5.9	−7.2	−7.9	−9.6	−8.1
C	−2.6	−2.4	−2.4	−2.4	−4.3	−5.1	−2.0	−1.8	−4.8	−4.8	−3.6	−4.8
G	−3.4	−6.9	−6.9	−6.9	−3.8	−6.8	−8.9	−6.8	−6.6	−8.1	−9.2	−8.1
U	−3.3	−3.3	−3.3	−3.3	−2.8	−1.4	−2.8	−1.4	−5.5	−4.4	−5.5	−3.6

TABLE 3.2 – Énergies libres unifiées de fourches Δg_{NN}^{fork} en kcal/mol à 37° C (±0.4 kcal/mol).

	$5'-AX-3'$ $3'-UY-5'$				$5'-CX-3'$ $3'-GY-5'$				$5'-GX-3'$ $3'-CY-5'$			
$X^{\ Y}$	A	C	G	U	A	C	G	U	A	C	G	U
A	−2.1	−2.2	−2.6	−2.1	−3.0	−3.0	−3.5	−3.2	−2.8	−3.0	−3.5	−3.3
C	−2.0	−2.0	−2.7	−2.0	−2.8	−2.7	−3.7	−2.7	−2.8	−2.6	−3.5	−2.5
G	−3.0	−2.5	−2.0	−1.8	−3.9	−3.3	−3.1	−2.8	−4.0	−3.7	−3.0	−2.9
U	−2.1	−2.1	−2.2	−2.8	−3.1	−3.0	−3.2	−3.6	−3.2	−2.8	−3.4	−3.4
	$5'-GX-3'$ $3'-UY-5'$				$5'-UX-3'$ $3'-AY-5'$				$5'-UX-3'$ $3'-GY-5'$			
$X^{\ Y}$	A	C	G	U	A	C	G	U	A	C	G	U
A	−1.8	−2.2	−2.6	−2.1	−2.2	−2.1	−2.8	−2.2	−2.2	−2.1	−2.8	−2.2
C	−2.0	−2.0	−2.7	−2.0	−2.0	−2.0	−2.5	−1.9	−2.0	−2.0	−2.8	−1.9
G	−2.9	−2.5	−2.1	−1.9	−3.2	−2.5	−2.3	−2.0	−2.9	−2.5	−2.1	−2.3
U	−2.1	−2.1	−2.1	−2.8	−2.1	−2.0	−2.2	−2.7	−2.2	−2.0	−2.2	−2.7

3.2.3 Différence entre petites et grandes fourches

Selon les arguments de la partie précédente, l'entropie d'une fourche externe devrait être donnée par le paramètre unifié Δs_{NN}^{fork}. La figure 3.3 B montre certes une corrélation assez forte entre les deux mais également un décalage inattendu de près de $3k_BT$. Pour comprendre un éventuel échec de notre processus d'unification, il faut se pencher sur les expériences effectivement réalisées par Turner pour en tirer ses paramètres.

Contrairement aux paramètres de fourches des boucles internes ou en épingles, ceux pour les fourches externes ont été paramétrés grâce à des expériences sur des petites fourches externes composées uniquement de deux nucléotides non-appariés (comme sur la figure 3.1 e). Turner suppose donc implicitement que dans une boucle externe (partie dénaturée terminale non bouclée), tous les nucléotides, excepté la paire de bases adjacente à la jonction avec le double-brin, ont un environnement identique comme dans un simple brin dénaturé (une telle hypothèse n'est pourtant pas supposée pour les petites boucles internes ou en épingle pour lesquelles les couts entropiques sont tabulés et non calculés à l'aide de l'équation 2.11 et des paramètres de fourches correspondants). Il semblerait plus probable que cet effet s'étende sur une petite distance le long de la chaîne avec une correction en énergie libre Δg_{fork}^i par rapport au simple brin qui décroîtrait rapidement quand la distance i à la fourche grandirait. Dans ce cas, l'énergie d'une fourche attachée à de longs simples brins serait donnée par $\Delta g_{fork} = \sum_{i=1}^{\infty} \Delta g_{fork}^i$ avec $\Delta g_{fork} \equiv \Delta g_{NN}^{fork}$ le paramètre non-spécifique explicitée dans la partie précédente. En particulier, nous affirmons que $\Delta g_{NN}^{tm}(ext) = \Delta g_{fork}^1$. Puisque les expériences ne permettent pas de décrire précisément la dépendance à la distance, nous fixons $\Delta g_{fork}^2 = \Delta g_{fork} - \Delta g_{fork}^1$ and $\Delta g_{fork}^i \equiv 0$ pour $i > 2$. La figure 3.3 B suggère $\Delta g_{fork}^2 = -3.1 \pm 0.8\, k_BT$ ($Q = 0.98$). Concernant les enthalpies, une analyse similaire donne $\Delta h_{fork}^2 = -6.8 \pm 6.5\, k_BT$ ($Q = 0.77$). En incluant ces corrections, les énergies de fourches du modèle de Turner deviennent bien indépendantes de la nature de la boucle adjacente.

Des arguments semblables s'applique également aux fins libres avec dangles (voir figure 3.1 f). On peut donc définir $\Delta g_{dangle}^2 = \Delta g_{dangle} - \Delta g_{dangle}^1$ et $\Delta g_{dangle}^i \equiv 0$ pour $i > 2$. De plus, on peut raisonnablement supposer que la correction Δg_{fork}^2 pour les fourches vaut le double de la correction Δg_{dangle}^2 pour un dangle. Ainsi, on fixe $\Delta g_{dangle}^2 = -1.6k_BT$. Cette valeur est consistante avec de précédentes observations réalisées par Ohmichi et al [168].

3.2.4 Unification des paramètres d'empilement coaxial et de défaut d'appariement

Des expériences [169, 170, 171] ont montré que l'empilement coaxial (empilement des paires de bases terminales de deux doubles brins voisins) stabilise les complexes. Des données sont disponibles pour des structures avec ou sans un défaut d'appariement intercalé ("intervening mismatch") entre les deux doubles brins. À partir de ces énergies libres, on peut facilement extraire la contribution due à l'empilement coaxial en soustrayant à l'énergie totale toutes les autres contributions (association de paires de base, fourches, capping,etc.) [85].
À l'aide de nos paramètres unifiés de fourches et de nucléation de boucle, nous trouvons que l'énergie libre Δg_{coax} d'un empilement coaxial sans défaut d'appariement intercalé et sans extension de brin (figure 3.4 a) est fortement corrélée à l'énergie libre d'association Δg_{NN}^{st} correspondante, $\Delta g_{coax} - \Delta g_{NN}^{st} = -2.9 \pm 0.6k_BT \equiv -T\Delta s_c^0$ ($Q = 0.99$). Ainsi, une entaille dans une double brin le stabilise d'environ $3k_BT$. Ceci pourrait s'expliquer par une augmentation du nombre de degrés de liberté internes au squelette de la double-hélice due à sa relaxation au niveau de l'entaille. Pour une interface suivie d'une seule extension de brin (figure 3.4 b), on trouve $\Delta g_{coax} - \Delta g_{NN}^{st} = -2.3 \pm 0.6k_BT \equiv -T\Delta s_c^1$ ($Q = 0.99$), et pour une interface suivie de deux extensions de brin (Fig.3.4 c), $\Delta g_{coax} - \Delta g_{NN}^{st} = -0.8 \pm 0.6k_BT \equiv -T\Delta s_c^2$ ($Q = 0.96$). On peut noter que, originellement, Turner ne distingue pas les deux situations précédentes (une ou deux extensions). Cette hypothèse mène d'ailleurs à une plus petite valeur de Q (0.4).
Pour les empilements coaxiaux avec un défaut d'appariement intercalé (figure 3.4 d), on observe une

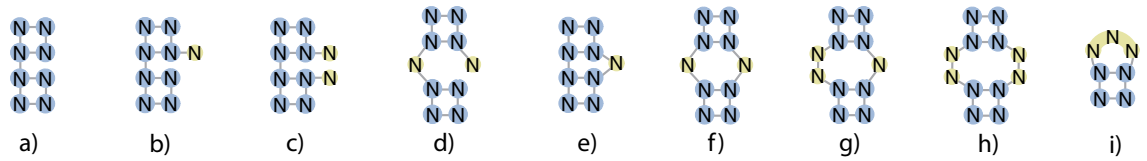

FIGURE 3.4 – Exemples de situations avec empilement coaxial et avec petites boucles : (de la gauche vers la droite) empilement coaxial sans défaut d'appariement intercalé et sans (a) ou avec une (b) ou deux (c) extensions de brin ; empilement coaxial avec un défaut d'appariement intercalé (d) ; boucle latérale avec un nucléotide (e) ; boucle interne 1nt×1nt (f) ; boucle interne 1nt×2nt (g) ; boucle interne 2nt×2nt (h) ; boucle en épingle comprenant 3 nucléotides (i).

corrélation entre Δg_{coax} et les énergies de fourches entre les paires de bases fermant les doubles brins et les nucléotides du défaut, $\Delta g_{coax} - 2\Delta g_{NN}^{fork} = 2.9 \pm 0.4 k_B T \equiv -T\Delta s_c^{im}$ ($Q = 0.97$). Dans ce cas, par manque de donnée, on ne fait pas de distinction entre de possibles extensions de brin.

Pour les boucles internes, latérales ou en épingle contenant un petit nombre de nucléotides (voir figures 3.4 e-i), les paramètres diffèrent de ceux décrivant une boucle plus grande [85] (décrit par les paramètres unifiés définis à la section 3.2.2). La raison principale est la non-perturbation de la structure hélicoidale par le défaut d'appariement : les deux doubles brins voisins appartiennent à la même double-hélice [172, 173]. Pour une boucle latérale avec un nucléotide, on trouve $\Delta g - \Delta g_{NN}^{st} = 6.2 k_B T \equiv -T\Delta s_b^1$ [85]. À l'aide des données de Turner, pour les défauts d'appariements internes, on observe une corrélation entre l'énergie libre totale de la boucle Δg (nucléation et fourches) et les énergies de fourches correspondantes ($Q \sim 1$ dans tous les cas). On trouve alors que l'énergie de nucléation vaut $\Delta g - 2\Delta g_{NN}^{fork} = 10.2 \pm 1.4 k_B T \equiv -T\Delta s_{loop}^{1,1}$ pour 1nt × 1nt (figure 3.4 f), $15 \pm 2 k_B T \equiv -T\Delta s_{loop}^{1,2}$ pour 1nt × 2nt (figure 3.4 g) et $11.1 \pm 2 k_B T \equiv -T\Delta s_{loop}^{2,2}$ pour 2nt × 2nt (figure 3.4 h). Pour les boucles en épingle composée de 3 nucléotides (figure 3.4 i), l'énergie libre de boucle ne comprend qu'un terme de nucléation (les fourches ne sont pas considérées) $-T\Delta s_{loop}^3 = 6.8 k_B T$.

Dans le modèle de Turner, le nombre total de paramètres pour les défauts d'appariement est de l'ordre de 10^4. Leur estimation est tirée d'un ensemble de paramètres indépendants mesurés expérimentalement (~ 280) complétés par des hypothèses sur l'égalité en énergie libre de tel ou tel défaut d'appariement [85]. Les relations précédentes permettent donc de réduire de manière drastique les nombre de paramètres indépendants (de 280 à 4).

3.2.5 Liberté de jauge pour les termes de bords et d'initiation dans le modèle de Turner

Dans le modèle de Turner, la paramétrisation des énergies libres d'association (Δg_{NN}^{st}), de capping (Δg_{term}) et d'initiation (Δg_{init}) est faite à partir de données expérimentales sur des oligomères cours [162] (voir section 3.2.1). Cependant, pour les mêmes raisons que celles évoquées dans la section 1.2.1 sur la paramétrisation des énergies pour l'ADN, il n'est pas possible de déterminer de manière unique (c'est à dire indépendamment les uns des autres) les trois termes de capping Δg_{term} ($A/U, G/C, G/U$) et le terme d'initiation Δg_{init}. Il en va de même pour la paramétrisation des autres termes de bords (fourches, dangles) et de nucléation. Dans la jauge fixée par Turner, $\Delta g_{term}(G/C) = 0$. Est-ce que cela a-t-il une influence quelconque sur les prédictions du modèle ?

On se demande donc si il existe un jeu de constantes Φ_{init}, Φ_{nuc}, Φ_{bound}, et Φ_{term} qu'on peut ajouter aux énergies libres d'initiation, de nucléation, de bord internes (fourches de boucles internes, en épingle, latérales) et terminales (capping,fourches externes, dangles) sans modifier les predictions du modèle de Turner. Tout d'abord, l'équilibre d'association entre des oligomères complémentaires doit

être invariant par changement de jauge, donc il vient immédiatement $\Phi_{term} = -\Phi_{init}/2$. De même, le choix de la jauge ne peut pas affecter le coût de formation d'une boucle interne dans un double-brin ($\Phi_{nuc} = -2\Phi_{bound}$) ou d'une boucle en épingle pour un simple-brin ($\Phi_{nuc} + \Phi_{term} + \Phi_{bound} = 0$). Ces différentes conditions impliquent que

$$\Phi_{init} = \Phi_{nuc} = \Phi \tag{3.4}$$

$$\Phi_{bound} = \Phi_{term} = -\Phi/2 \tag{3.5}$$

Prenons maintenant une structure secondaire quelconque comprenant n_s brins. Elle est formée de n_{loop} boucles et comprend n_{term} termes de bord externes (dangles, fourches externes, capping) et n_{bound} termes de bord internes (fourches de boucles interne, en épingle ou latérales). Changer de jauge (c'est à dire ajouter les constantes définies dans les équations 3.4 et 3.5) revient à modifier l'énergie libre de la structure par

$$\Delta G_\Phi = (n_s - 1)\Phi_{init} + n_{bound}\Phi_{bound} + n_{term}\Phi_{term} + n_{loop}\Phi_{nuc} \tag{3.6}$$

$$= (n_s + n_{loop} - 1 - (n_{bound} + n_{term})/2)\,\Phi \tag{3.7}$$

Or, on peut prouver par récurrence (voir annexe 6.7) que $n_{loop} = n_{stem} - (n_s - 1)$. De plus, comme le nombre de termes de bord (internes et externes) vaut deux fois le nombre de parties en double-hélice, il vient facilement $\Delta G_\Phi = 0$ quelque soit Φ. Ainsi, le choix de la jauge n'influe pas sur l'estimation de l'énergie libre d'une structure arbitraire. On peut donc modifier tous les termes d'initiation, de nucléation et de bord dans le modèle de Turner par respectivement un offset de Φ et $-\Phi/2$ sans changer les predictions pour des quantités mesurables. Dans la jauge de Turner, c'est utilisé pour imposer $\Delta g_{term}(G/C) = 0$, mais on pourrait très bien prendre $\Phi = T\Delta s_{loop}$ afin d'éliminer l'entropie non-locale de nucléation unifiée introduite dans la section 3.2.2.

3.2.6 Sensibilité des résultats

3.2.6.1 Méthode d'étude

En ce qui concerne la dénaturation de l'ADN (voir section 1), on a vu que l'incertitude sur les paramètres se traduisait parfois par une forte incertitude sur les prédictions faites par le modèle PS unifié. Quant est-il pour les résultats donnés par les modèles décrivant le repliement de l'ARN ? Pour quantifier cette propagation des erreurs, nous nous intéressons au modèle de Turner via le programme RNAfold [86]. Le nombre de paramètres étant beaucoup plus important dans le modèle de Turner que dans le modèle PS (plus de 600 pour Turner contre une vingtaine pour PS), nous n'avons pas (faute de données) réitéré le processus de paramétrisation décrit dans [85] pour avoir accès à la matrice de covariance des paramètres comme nous l'avions fait dans la section 1.2.4. A la place, nous supposons les paramètres indépendants entre eux (matrice de covariance diagonale) et, en consistence avec les données expérimentales [85], nous considérons une erreur de 0.1 kcal/mol pour les énergies libres locales (association, forking, capping, dangling) et une erreur de 0.8 kcal/mol pour les pénalités de nucléation. Avec cette hypothèse de paramètres non-corrélées, on surestime certainement la propagation des erreurs, mais les conclusions que l'on en tire ne devraient pas être significativement différentes que pour le cas avec une matrice de covariance non diagonale.

On génère dans un premier temps des séquences aléatoires de différentes tailles (100 séquences pour chaque taille avec $N = 100, 200, ..., 1000$). Puis, pour le jeu de paramètres standard et pour 250 autres jeux de paramètres aléatoirement tirés (chaque paramètre est choisi suivi une distribution gausienne autour de la valeur standard et avec l'écart-type défini ci-dessus) avec RNAfold, on calcule la structure la plus stable $\mathcal{S}_m$, l'énergie libre G_{ens} de l'ensemble thermodynamique, la probabilité Θ qu'un nucléotide soit appararié et la carte des contacts $c(i, j)$ qui donne les probabilités que les nucléotides i et j soient appariés ensemble, à 37° C, la température courante où sont étudiées les molécules d'ARN.

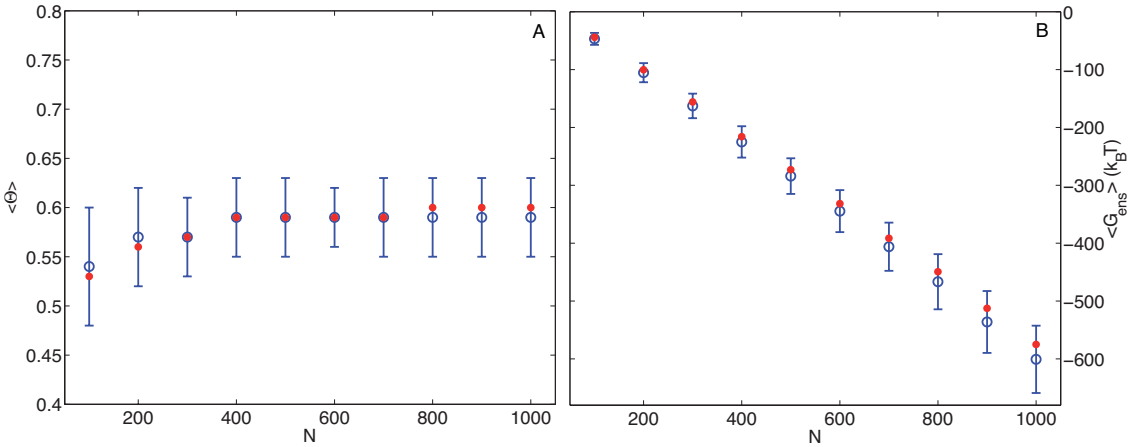

FIGURE 3.5 – Evolution de Θ (A) et de G_{ens} (B) moyennées sur des séquences aléatoires en fonction de leur taille N et calculées avec les paramètres standard (points rouges), ou en moyennant également sur les différents jeux de paramètres aléatoires générés (cercles bleus).

3.2.6.2 Propriétés thermodynamiques globales

La figure 3.5 A montre l'évolution de la valeur moyenne de Θ prise sur les séquences aléatoires et sur les différents jeux de paramètres générés. On remarque qu'à partir de $N \sim 400$, $\langle\Theta\rangle$ ne dépend plus de la taille des chaînes est vaut 0.59 ± 0.04. Ainsi, dans une longue séquence quelconque, en moyenne près de 60% des nucléotides seront appariés à 37° C. La faible variabilité des résultats par rapport aux jeux de paramètres indique que cette affirmation est assez robuste.

Sur la figure 3.5 B, on observe la décroissance linéaire de la valeur moyenne de G_{ens} en fonction de N avec une pente de l'ordre de $-60k_BT$ pour 100 bp. L'étude de la propagation des erreurs indique un écart-type qui croit également linéairement avec N, l'erreur relative étant d'environ 10%. Cela semble assez logique vu que plus la séquence sera grande, plus le nombre de contacts et de boucles le sera également (et ceci de manière linéaire vu que $\langle\Theta\rangle$ est indépendant de N), et donc plus l'erreur sera importante. Ce qui l'est moins (du moins aux premiers abords), est de trouver que l'énergie libre obtenue en moyennant sur les différents jeux de paramètres est plus petite que celle obtenue avec le jeu de paramètres standard (alors que pour Θ, aucune différence notable n'était à signaler). En fait, cela se comprend bien si l'on écrit précisément ce que vaut la valeur moyenne de G_{ens} sur les paramètres :

$$\langle G_{ens}\rangle = \int \mathrm{d}\xi_1...\mathrm{d}\xi_n P(\xi_1)...P(\xi_n) G_{ens}\left(\mathcal{H}[p_1+\xi_1,...,p_n+\xi_n]\right) \tag{3.8}$$

où $\{p_1,...,p_n\}$ sont les n paramètres standard (association,forking,nucléation,...), $\mathcal{H}$ est l'Hamiltonien du système (défini par le modèle de Turner), $\{\xi_i\}$ sont les déviations par rapport aux paramètres standard et $P(\xi_i)$ est une distribution de probabilité gaussienne centré en 0 et d'écart-type égal à celui de p_i. Puis, on applique l'inégalité de Gibbs-Bogoliubov [1] à $\mathcal{H}[p_1+\xi_1,...,p_n+\xi_n]$ et $\mathcal{H}[p_1,...,p_n]$, soit

$$\begin{aligned}\langle G_{ens}\rangle \quad \leq \quad & \int \mathrm{d}\xi_1...\mathrm{d}\xi_n P(\xi_1)...P(\xi_n)\left\{G_{ens}\left(\mathcal{H}[p_1,...,p_n]\right)\right.\\ & \left.+\langle\mathcal{H}[p_1+\xi_1,...,p_n+\xi_n]-\mathcal{H}[p_1,...,p_n]\rangle_{\{p_i\}}\right\}\end{aligned} \tag{3.9}$$

Or $\mathcal{H}$ dépend linéairement de ces paramètres, soit $\mathcal{H}[p_1+\xi_1,...,p_n+\xi_n]-\mathcal{H}[p_1,...,p_n] = \sum_{i=1}^{p}\xi_i(\partial\mathcal{H}/\partial p_i)$.

1. Soit deux Hamiltoniens $\mathcal{H}_1$ et $\mathcal{H}_2$, alors l'inégalité de Gibbs-Bogoliubov stipule que $G_1 \leq G_2 + \langle\mathcal{H}_1-\mathcal{H}_2\rangle_2$ avec G_i l'énergie libre du système i et $\langle X\rangle_2$ la valeur moyenne de X par rapport à l'Hamiltonien 2.

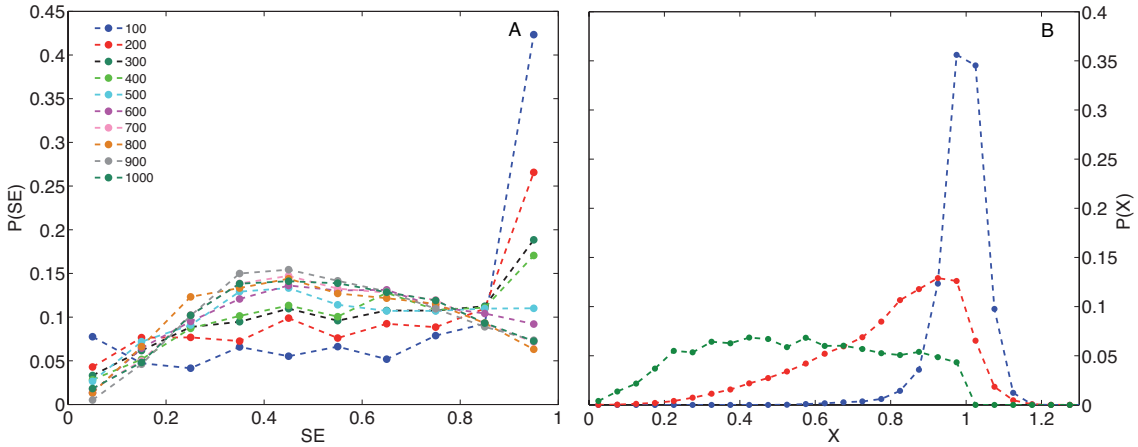

Figure 3.6 – (A) Distribution de probabilité pour SE pour différents jeux de paramètres aléatoires en fonction de la taille des séquences (voir légende des couleurs). (B) Distribution de probabilité pour SE (vert foncé), SE_{thermo} (rouge) et Θ^*/Θ^0 (bleu) pour $N = 1000$.

On obtient donc

$$\langle G_{ens} \rangle \leq G_{ens}\left(\mathcal{H}[p_1, ..., p_n]\right) + \int \mathrm{d}\xi_1 ... \mathrm{d}\xi_n P(\xi_1) ... P(\xi_n) \left[\sum_{i=1}^{p} \xi_i \langle \frac{\partial \mathcal{H}}{\partial p_i} \rangle_{\{p_i\}} \right] \tag{3.10}$$

et comme $\int \mathrm{d}\xi_i \xi_i P(\xi_i) = 0$, on a finalement

$$\langle G_{ens} \rangle \leq G_{ens}\left(\mathcal{H}[p_1, ..., p_n]\right) \tag{3.11}$$

L'équation précédente justifie ainsi la différence observée entre les deux énergies libres.

3.2.6.3 Propriétés de la structure la plus stable

Intéressons nous maintenant à la sensibilité des résultats en ce qui concerne la prédiction de la structure la plus stable. Pour cela on calcule, la sensibilité SE et la spécificité SP de la structure la plus stable prédite par un jeu de paramètres aléatoires $\mathcal{S}_m^*$ par rapport à la structure la plus stable prédite avec le jeu de paramètres standard $\mathcal{S}_m^0$ (* fait référence aux paramètres aléatoires et ^{0} à ceux standard). SE est définie comme le quotient entre le nombre de paires de bases en commun dans $\mathcal{S}_m^*$ et $\mathcal{S}_m^0$ et le nombre total de paires de bases dans $\mathcal{S}_m^0$ (soit dans le langage employé à la section 2.3.1, le nombre de vrais positifs sur la somme du nombre de vrais positifs et du nombre de faux négatifs [2]). SP quant à elle est définie comme le quotient entre le nombre de paires de bases en commun dans $\mathcal{S}_m^*$ et $\mathcal{S}_m^0$ et le nombre total de paires de bases dans $\mathcal{S}_m^*$ (soit le nombre de vrais positifs sur le nombre de positifs). $SE \to 1$ signifie que toutes les paires de bases présentes dans $\mathcal{S}_m^0$ le sont également dans $\mathcal{S}_m^*$, et $SP \to 1$ signifie que toutes les paires de bases présentes dans $\mathcal{S}_m^*$ le sont également dans $\mathcal{S}_m^0$. La figure 3.6 A montre la distribution des SE pour les différents jeux de paramètres générés en fonction de la taille des séquences (nous n'avons pas tracé les distributions des SP car elles sont similaires à celles des SE). On remarque que pour les petites séquences la distribution est piquée autour de 1, signe que la détermination de la structure la plus stable est assez fiable pour $N \leq 100$. Quand la taille des séquences augmente, la courbe tend vers une distribution assez large centrée autour de 0.5. Elle traduit la forte incertitude sur $\mathcal{S}_m$ quand la séquence est suffisamment longue. En moyenne, 50% des paires

2. Attention, les définitions de la sensibilité et de la spécificité sont ici légèrement différentes de celles données à la section 2.3.1.

de bases seront conservées, indiquant tout de même l'existence de parties très stables (par exemple de longues parties double-brins) dans $\mathcal{S}_m^0$. Le fait que la distribution des SE devienne indépendante de N indique que malgré la croissance exponentielle de l'espace des configurations [20, 174], la composition de l'ensemble des structures secondaires d'énergies minimales reste plus ou moins constante. Alors qu'on pourrait s'attendre à ce que plus N augmente, plus les configurations minimales en énergie soient différentes en composition (plus la distribution des SE soit piquée autour de 0).

3.2.6.4 Propriétés thermodynamiques locales

Dans le paragraphe précédent, on a vu que la fiabilité des prédictions pour la structure la plus stable était assez limitée pour des séquences longues. Une meilleure méthode pour estimer l'existence de contact entre deux nucléotides est de regarder la probabilité thermodynamique de contact (via la carte des contacts) calculée sur l'ensemble des configurations (et non plus sur une seule comme précédemment). Afin de visualiser les incertitudes découlant des barres d'erreurs sur les paramètres au niveau de la carte de contact, on évalue, pour chaque séquence et chaque jeu de paramètres aléatoires, le produit scalaire entre $c^*(i,j)$ et $c^0(i,j)$ normalisé par la norme de $c^0(i,j)$:

$$SE_{thermo} = \frac{\sum_{i,j} c^*(i,j) \times c^0(i,j)}{\sum_{i,j} [c^0(i,j)]^2} \tag{3.12}$$

On appelle cette quantité la sensibilité thermodynamique (en référence à SE défini plus haut), car si on remplace $c(i,j)$ par $\mathcal{S}_m$ et que l'on pose $\mathcal{S}_m(i,j) = 1$ (0) si i et j sont (ne sont pas) appariés, on a exactement la définition de SE. Comme pour SE, on observe que pour des séquences longues SE_{thermo} tend vers une distribution représentée sur la figure 3.6 B. Cependant, comparée à celle de SE, la distribution de la sensibilité thermodynamique reste maximale autour de 1 et est moins étendue, signe que l'on peut avoir une plus grande confiance dans les prédictions pour la carte des contacts.

3.2.6.5 Bilan

Sur la figure 3.6, en plus de la distribution de SE et SE_{thermo}, nous avons également tracé la distribution de Θ^*/Θ^0 qui, quant à elle, est piquée autour de 1 pour $N = 1000$. Les allures totalement différentes de ces trois courbes incitent donc à émettre quelques recommandations quant à l'utilisation brute des résultats prédits par le modèle de Turner (ou par tout autre modèle basé sur les paramètres de Turner comme le modèle sur réseau par exemple).

Pour les séquences courtes ($N \leq 100$), les prédictions pour les propriétés thermodynamiques locales ou globales ainsi que pour la structure la plus stable peuvent être considérées comme assez robustes. Précisons tout de même que l'étude a été faite à 37° C, une température où les structures sont compactes (Θ grand). Pour des températures proches de la température de dénaturation du complexe ($\sim 80^o$ C), là où les fluctuations sont très importantes, même pour les petites séquences, l'incertitude sur toutes les propriétés étudiées n'est plus forcément faible (voir section 1.4.4).

Pour les séquences longues, on distingue 3 niveaux de confiance :
- pour les propriétés thermodynamiques globales comme Θ, la propagation des erreurs est assez faible (avec près de 98% des résultats pour Θ^*/Θ^0 compris entre 0.8 et 1.2.) et le niveau de confiance est élevé ;
- pour les propriétés thermodynamiques locales comme $c(i,j)$, le niveau de confiance est moyen, avec environ 55% des résultats qui ont $0.8 \leq SE_{thermo} \leq 1.2$;
- pour la prédiction de la structure la plus stable (ou de tout autre propriété calculée sur un nombre fini de structure), le niveau de confiance est assez faible, avec une distribution assez large des résultats.

Ainsi, quant on a affaire à une séquence longue et qu'on veut pouvoir comparer les prédictions faites par le modèle de Turner avec une possible expérience ou que l'on veut les utiliser pour interpréter ou étudier un phénomène quelconque, il est préférable de ne pas s'en tenir uniquement aux résultats obtenus avec les paramètres standard, mais aussi d'effectuer une analyse des erreurs sur les propriétés que l'on veut utiliser pour quantifier le niveau de confiance de telles prédictions [175].

3.3 Paramétrisation du modèle sur réseau

Avant d'aborder la paramétrisation du modèle sur réseau proprement dite (sections 3.3.2, 3.3.3 et 3.3.4), intéressons nous aux implications qu'un changement du réseau sous-jacent 3.3.1 aurait sur les paramètres.

3.3.1 Influence du réseau

Considérons deux versions du modèle sur réseau (L_1 et L_2) sur deux réseaux différents (cubique simple, cubique à face centrée,etc.). Les chemins sur ces réseaux sont caractérisés par deux jeux de constantes $\{z_i, \mu_i, f_{s_i}, f_{l_i}\}$ $(i = 1, 2)$ définies dans la section 3.1.2. Pour obtenir des résultats indépendants du réseau, on force les différences d'énergie libre entre deux structures secondaires quelconques $\mathcal{S}_a$ et $\mathcal{S}_b$ à être invariantes pour un changement du réseau sous-jacent

$$G_a(1) - G_b(1) - k_B T \log\left[\frac{\Omega_1(\mathcal{S}_a)}{\Omega_1(\mathcal{S}_b)}\right] = G_a(2) - G_b(2) - k_B T \log\left[\frac{\Omega_2(\mathcal{S}_a)}{\Omega_2(\mathcal{S}_b)}\right] \tag{3.13}$$

où $G_a(i)$ et $G_b(i)$ représente la somme des intéractions locales et non-locales définies pour le modèle sur réseau dans la section 3.1.1 et paramétrées pour le réseau i. Par exemple, pour l'équilibre à deux-états entre un double-brin de taille N et les deux simples brins correspondants, cela donne

$$G_{ds}(1) - 2\,G_{ss}(1) + G_{mix_1} = G_{ds}(2) - 2\,G_{ss}(2) + G_{mix_2} \tag{3.14}$$

En remplacant dans l'équation 3.14 G_{ds} et G_{ss} par leur expression donnée à la section 3.1.2, on trouve

$$\begin{aligned} & N(\epsilon_1 + 2k_B T \log \mu_1) + 2\omega_1 + G_{mix_1} + k_B T \log\left(\frac{f_{s_1}^2 N^{2c'}}{z_1}\right) \\ = \; & N(\epsilon_2 + 2k_B T \log \mu_2) + 2\omega_2 + G_{mix_2} + k_B T \log\left(\frac{f_{s_2}^2 N^{2c'}}{z_2}\right) \end{aligned} \tag{3.15}$$

Enfin, en séparant les termes qui dépendent de linéairement N et ceux qui n'en dépendent pas, on obtient la relation entre paramètres

$$\epsilon_2 = \epsilon_1 + 2k_B T \log(\mu_1/\mu_2) \tag{3.16}$$

$$\omega_2 = \omega_1 + k_B T \log\left[\frac{f_{s_1}}{f_{s_2}}\left(\frac{z_2}{z_1}\right)^{1/2}\right] + \frac{\Phi_{12}}{2} \tag{3.17}$$

avec $\Phi_{12} = G_{mix_1} - G_{mix_2}$. En appliquant la même procédure aux exemples simples de la section 3.1.2, il vient

$$\gamma_2 = \gamma_1 + k_B T \log\left[\frac{z_2 - 2}{z_1 - 2}\left(\frac{z_1}{z_2}\right)^{1/2}\right] + \frac{\Phi_{12}}{2} \tag{3.18}$$

$$\sigma_2 = \sigma_1 + k_B T \log(f_{l_2}/f_{l_1}) - \Phi_{12} \tag{3.19}$$

$$\lambda_2 = \lambda_1 + k_B T \log\left[\frac{z_2 - 1}{z_1 - 1}\left(\frac{z_1}{z_2}\right)^{1/2}\right] + \frac{\Phi_{12}}{2} \tag{3.20}$$

$$\chi_2 = \chi_1 + k_B T \log\left[\frac{z_1}{z_2}\left(\frac{z_2 - 1}{z_1 - 1}\right)^{2}\right] + \Phi_{12} \tag{3.21}$$

Ainsi les deux jeux de paramètres sont reliés entre eux par les constantes du réseau $\{z_i, \mu_i, f_{s_i}, f_{l_i}\}$ et $\Phi_{12} = G_{mix_1} - G_{mix_2}$.

L'équation 3.19 montre qu'on ne peut pas supprimer le paramètre de nucléation σ du modèle ($\sigma = 0$ pour tous les réseaux) sans être incohérent quand l'on passe d'un réseau à un autre, puisque f_l est une quantité non-universelle dépendante du réseau. C'est un point subtile qui ne doit pas être négligé. En effet, il montre l'importance de tenir compte proprement de l'énergie libre non-locale de nucléation de boucle si l'on veut décrire au niveau tertiaire les acides nucléiques. Cependant, en pratique, pour un réseau fixé, on préfère travailler avec $\sigma = 0$ pour n'avoir à compter que les énergies locales et éviter ainsi de devoir analyser la structure globale du complexe après chaque changement local. Heureusement, l'analyse précédente faite sur les paramètres du réseau implique également une liberté de jauge (comme pour le modèle de Turner) via la constante Φ_{12}. Ainsi pour un réseau i fixé, on peut toujours exploiter cette jauge pour imposer $\sigma_i \equiv 0$. Dans la suite on appelera ce choix, la jauge "$\sigma = 0$".

3.3.2 Paramétrisation naïve

Une première paramétrisation naive du modèle sur réseau serait d'égaler directement tous les paramètres du réseau avec leurs homologues du modèle de Turner ($\epsilon = \Delta g^{st}_{NN}$, $\omega = \Delta g_{term}$, $\gamma = \Delta g^{fork}_{NN}$,etc. ; voir section 3.1.1). Cependant, comme, pour une structure secondaire quelconque $\mathcal{S}$, le nombre d'états $\Omega(\mathcal{S})$ dépend de la nature du réseau utilisée, les résultats seraient aussi dépendant du type de réseau choisi, et ceci n'est pas satisfaisant d'un point de vue physique. De plus, les résultats que prédirait alors le modèle sur réseau seraient aberrant en comparaison avec l'expérience. Par exemple, pour des petites structures en épingle, la paramétrisation naive ne reproduit pas les transitions à deux-états observées expérimentalement et prédit des températures de fusion 45K plus petites que les résultats expérimentaux (voir figure 5.1).

Une deuxième idée similaire et utilisée par Kinefold [138] et Vfold [132] serait de prendre directement les paramètres de Turner pour les intéractions d'empilement (association, capping, fourche, dangle, coaxial) et de considérer que l'énergie de conformation $-k_B T \log \Omega(\mathcal{S})$ d'une structure secondaire $\mathcal{S}$ calculée sur le réseau inclut les termes de nucléation. Ainsi, la différence d'énergie libre entre $\mathcal{S}$ et l'état dénaturé correspondant $\mathcal{S}_0$, serait donnée par

$$\Delta G(\mathcal{S}) = \Delta G^{\mathcal{T}}_{stack}(\mathcal{S}) - k_B T \log \left[\frac{\Omega_i(\mathcal{S})}{\Omega_i(\mathcal{S}_0)} \right] \tag{3.22}$$

avec $\Delta G^{\mathcal{T}}_{stack}(\mathcal{S})$ comptant pour les intéractions d'empilement de Turner définies ci-dessus. L'introduction de différence d'énergie libre issue d'un calcul sur réseau dans le modèle de Turner est délicate pour deux raisons. Premièrement, cela implique que les résultats vont dépendre de la nature du réseau, or on veut définir un cadre général qui fait des prédictions fiables quelque soit le réseau sous-jacent. Deuxièmement, supplanter les termes de nucléation (Δs_{loop}) qui dépendent de la jauge utilisée dans le modèle de Turner par une partie de $k_B \log[\Omega(\mathcal{S})/\Omega(\mathcal{S}_0)]$ qui est une mesure indépendante de la jauge de Turner, rend l'évaluation de $\Delta G(\mathcal{S})$ tributaire de la jauge de Turner, ce qui n'est pas acceptable.

3.3.3 Paramétrisation correcte

De bien meilleurs et plus cohérents résultats peuvent être obtenus en imposant que le modèle sur réseau ait le même comportement au niveau structure secondaire que le modèle de Turner. Ceci implique que les paramètres du modèle sur réseau soient obtenus en égalant, pour des structures secondaires simples, les différences d'énergie libre calculées à partir du modèle de Turner et du modèle sur réseau (voir section 3.1.2) [33]. Par exemple, pour l'équilibre à deux-états entre un double-brin de taille N et les deux simples brins correspondants, cela donne

$$G_{ds} - 2\, G_{ss} + G_{mix} = N \Delta g^{st}_{NN} + 2 \Delta g_{term} + \Delta g_{init} \tag{3.23}$$

Comme dans la section 3.3.1, on groupe les termes qui dépendent linéairement de N et on obtient

$$\epsilon = \Delta g^{st}_{NN} - 2k_BT\log\mu \tag{3.24}$$

$$\omega = \Delta g_{term} + \frac{\Delta g_{init} - G_{mix}}{2} - k_BT\log(f_s\bar{N}^{c'}/z^{1/2}) \tag{3.25}$$

où $\bar{N} \approx 10$ est la taille typique des brins utilisés dans les expériences qui ont servi à paramétrer le modèle de Turner [162].

Dans la section 3.3.1, nous avons signalé qu'une liberté de jauge existait également dans le modèle sur réseau pour les termes de bord, de nucléation et de mélange. Par la suite, nous travaillerons dans la jauge "$\sigma = 0$" introduite Sec.3.3.1. Elle a l'avantage de rendre le modèle strictement local et d'éviter l'analyse global de la conformation pour l'évaluation de son énergie. On rappelle que le choix de cette jauge n'influe pas sur le calcul des propriétés thermodynamiques. En appliquant la même procédure que précédemment aux exemples simples de la section 3.1.2, on obtient la paramétrisation du modèle su réseau

$$\sigma = -T\Delta s_{loop} + T\Delta s_{loop} \equiv 0 \tag{3.26}$$

$$G_{mix}(T) = \Delta g_{init}(T) + T\Delta s_{loop} - k_BT\log f_l \tag{3.27}$$

$$\epsilon(T) = \Delta g^{st}_{NN}(T) - 2k_BT\log\mu \tag{3.28}$$

$$\omega(T) = \Delta g_{term}(T) - \frac{T\Delta s_{loop}}{2} + k_BT\log\left[\frac{(f_lz)^{1/2}}{f_s\bar{N}^{c'}}\right] \tag{3.29}$$

$$\gamma(T) = \Delta g_{fork}(T) - \frac{T\Delta s_{loop}}{2} + k_BT\log\left[\left(\frac{f_l}{z}\right)^{1/2}(z-2)\right] \tag{3.30}$$

$$\lambda(T) = \Delta g_{dangle}(T) - \frac{T\Delta s_{loop}}{2} + k_BT\log\left[\left(\frac{f_l}{z}\right)^{1/2}\left(\frac{z-1}{\bar{N}^{c'}}\right)\right] \tag{3.31}$$

$$\chi(T) = \Delta g_{coax}(T) - T\Delta s_{loop} + k_BT\log\left(\frac{f_l(z-1)^2}{z\bar{N}^{c'}}\right) \tag{3.32}$$

où Δg_{fork} (Δg_{dangle}) dépend de la taille de la fourche (dangle) comme précisé dans la section 3.2.3 et Δg_{coax} dépend de la nature de l'empilement coaxial (avec ou sans défaut d'appariement intercalé ou extension de brin ; voir section 3.2.4). De manière générale, les paramètres sur réseau sont définis par la somme du paramètre de Turner correspondant et d'une correction entropique dépendante des constantes du réseau (voir table 3.3).

Table 3.3 – Corrections entropiques Δs_{corr} des paramètres du modèle sur réseau dans la jauge "$\sigma = 0$" pour le réseau CFC. La valeur d'un paramètre sur réseau vaut le terme de Turner correspondant moins $T\Delta s_{corr}$.

Paramètre sur réseau	Correction entropique (en unité k_B)
σ	9.2
G_{mix}	7.8
ϵ	4.6
ω	-4.6
γ	-5.0
λ	-4.7
χ	-9.8

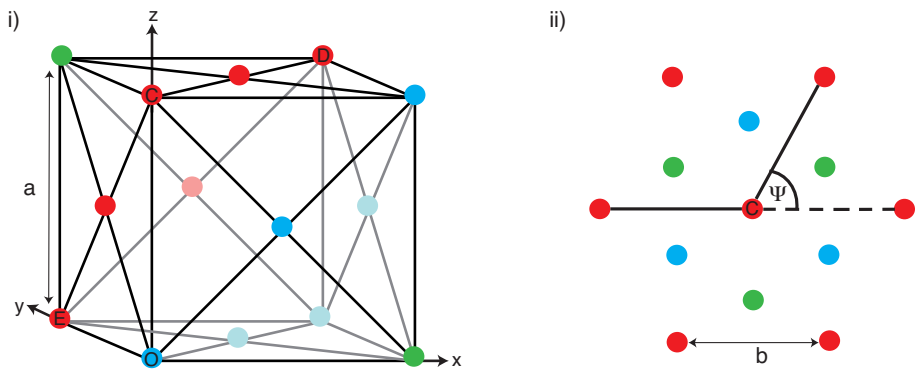

FIGURE 3.7 – Réseau cubique à faces centrées (CFC). (i) Représentation tridimensionnelle. a représente la taille de la maille. (ii) Coupe dans le plan (CDE) et définition de l'angle Ψ entre deux segments consécutifs. Les noeuds appartenant au même plan de coupe sont de la même couleur. b représente la distance entre deux noeuds voisins, $b = a/\sqrt{2}$. Les vecteurs de bases du réseau sont $\vec{a}_1 = b/\sqrt{(}2)(\hat{x}+\hat{y})$, $\vec{a}_2 = b/\sqrt{(}2)(\hat{y}+\hat{z})$ et $\vec{a}_3 = b/\sqrt{(}2)(\hat{z}+\hat{x})$.

En ce qui concerne, les paramètres pour les cas particuliers des boucles internes, en épingle et latérales étudiées dans la section 3.2.4, on trouve

$$\sigma_b^1(T) = -T(\Delta s_b^1 - \Delta s_c^2) + T\Delta s_{loop} + k_B T \log\left[\left(\frac{z-2}{z-1}\right)^2 \bar{N}^{c'} 3^{-c}\right] \tag{3.33}$$

$$\sigma_{1,1}(T) = -T\Delta s_{loop}^{1,1} + T\Delta s_{loop} \tag{3.34}$$

$$\sigma_{1,2}(T) = -T\Delta s_{loop}^{1,2} + T\Delta s_{loop} \tag{3.35}$$

$$\sigma_{2,2}(T) = -T\Delta s_{loop}^{2,2} + T\Delta s_{loop} \tag{3.36}$$

$$\sigma_3(T) = -T\Delta s_{loop}^{3} + \frac{T\Delta s_{loop}}{2} + k_B T \log\left(\frac{z-2}{z^{1/2}}\right) \tag{3.37}$$

où σ_b^1 est l'énergie de nucléation pour une boucle latérale avec un nucléotide, $\sigma_{1,1}$, $\sigma_{1,2}$ et $\sigma_{2,2}$, celles pour les défauts d'appariement internes et σ_3 celle pour une boucle en épingle comprenant 3 nucléotides.

Δg_{NN}^{st}, Δg_{init}, Δg_{term} et Δg_{dangle}^1 sont calculées avec la version 3.0 [85] des paramètres de Turner. Les données pour Δg_{NN}^{fork} sont fournies par les paramètres de Turner unifiés des tables 3.1 et 3.2. Δs_{loop} est donnée par l'équation 3.3. Δg_{fork}^2, Δg_{dangle}^2 et les contributions pour les petites boucles, les défauts d'appariement et les empilements coaxiaux sont estimées grâce aux différentes unifications et réductions de paramètres abordées dans les sections 3.2.3 et 3.2.4. Tous ces paramètres sont évalués pour une concentration en sel de 1 M, et, à notre connaissance, il n'existe pas de correction efficace pour tenir compte d'une dépendance en la concentration en sel (comme c'est le cas pour l'ADN).

3.3.4 Énergie de pliage

Le seul paramètre sur réseau qui n'a pas d'équivalent dans le modèle de Turner est l'énergie de pliage κ car elle rend compte d'aspects tertiaires absents du schéma de Turner. Il faut donc trouver un autre moyen pour paramétrer κ.

Tout d'abord, on modélise l'énergie de pliage de la double-hélice par une formule classique

$$\kappa(\Psi, T) = \bar{\kappa}(T)(1 - \cos\Psi) \tag{3.38}$$

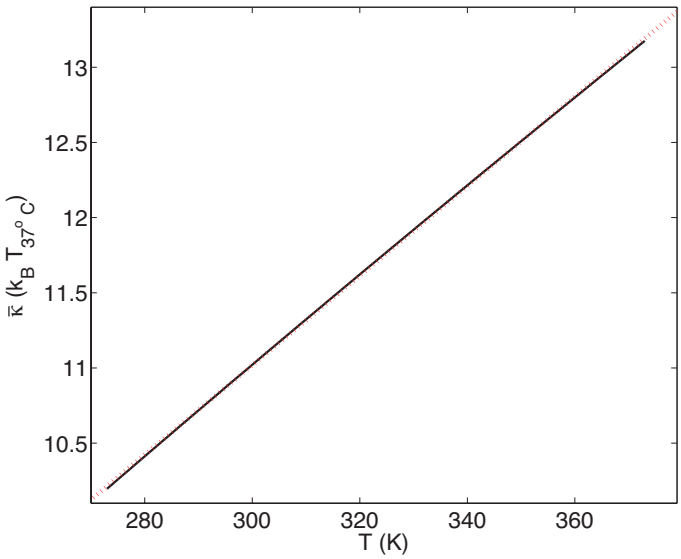

Figure 3.8 – Evolution de $\bar{\kappa}(T)$ donnée par la résolution de l'équation 3.42 (ligne noire) et modélisation par $\bar{\kappa}(T) = \bar{\kappa}_h - T\bar{\kappa}_s$ (pointillé rouge).

avec Ψ l'angle de pliage entre deux segments consécutifs du double-brin (voir figure 3.7 ii). Pour un double-brin ARN de taille N (qui peut être considéré comme un polymère semi-rigide), si on néglige les effets stériques (voir justification en annexe 6.8) , la valeur moyenne de la distance bout-à-bout carrée vaut (voir annexe 6.8)

$$\langle R_e^2 \rangle = Nb^2 \frac{1 + \langle \cos \Psi \rangle}{1 - \langle \cos \Psi \rangle} \equiv N_K l_K^2 \tag{3.39}$$

avec N_K et l_K le nombre de segments et la longueur de Kuhn [176] tels que $N_K l_K = Nb$ (conservation de la taille du polymère), et b la distance entre deux noeuds du réseau. L'équation 3.39 donne

$$\frac{l_K}{b} = \frac{1 + \langle \cos \Psi \rangle}{1 - \langle \cos \Psi \rangle} \equiv c_\infty \tag{3.40}$$

Pour un réseau CFC, il y a 11 angles possibles (on interdit le retour en arrière $\Psi = 180^o$) : 1 possibilité pour $\Psi = 0^o$, 4 possibilités pour $\Psi = 60^o$, 2 possibilités pour $\Psi = 90^o$ et 4 possibilités pour $\Psi = 120^o$ (voir figure 3.7). Ainsi, d'après la définition de κ (équation 3.38), on a

$$\langle \cos \Psi \rangle = \frac{1 + 2e^{-\beta\bar{\kappa}(T)/2} - 2e^{-3\beta\bar{\kappa}(T)/2}}{1 + 4e^{-\beta\bar{\kappa}(T)/2} + 2e^{-\beta\bar{\kappa}(T)} + 4e^{-3\beta\bar{\kappa}(T)/2}} \tag{3.41}$$

Finalement, en replaçant dans l'équation 3.40, $x \equiv \exp(-\beta\bar{\kappa}/2)$ doit vérifier

$$(3c_\infty - 1)x^3 + (c_\infty - 1)x^2 + (c_\infty - 3)x - 1 = 0 \tag{3.42}$$

Pour les acides nucléiques, à $T = 300$ K, $c_\infty \approx 300$, soit une longueur de persistance d'environ 150 paires de bases [177]. Si on suppose que c_∞ suit la loi

$$c_\infty(T) = \frac{300\text{K}}{T} c_\infty(300\text{K}) \tag{3.43}$$

pour chaque température comprise entre 273 K et 373 K, on résout l'équation 3.42 en x avec la fonction *fzero* de Matlab. La figure 3.8 représente l'évolution de $\bar{\kappa}(T)$ dans cette gamme de température. On modélise alors $\bar{\kappa}(T) = \bar{\kappa}_h - T\bar{\kappa}_s$. Une simple régression linéaire faite avec Matlab, nous donne

$$\bar{\kappa}_h = 646 \pm 3k_B\text{K} \quad \text{et} \quad \bar{\kappa}_s = -9.23 \pm 0.01k_B \tag{3.44}$$

3.3.5 Énergie libre totale d'une conformation sur réseau

L'énergie libre totale d'une conformation $\mathcal{C}$ est calculée à partir de son enthalpie $h_{latt}(\mathcal{C})$ et de son entropie $s_{latt}(\mathcal{C})$, données par la somme des termes sur réseau décrits dans la section 3.3.3. Par exemple, pour la structure en épingle définie à la figure 3.1 c, on aurait $h_{latt} = \omega_h + \epsilon_h + \epsilon_h + \kappa_h + \gamma_h + \sigma_h$ (idem pour s_{latt}). En dehors des contributions de pliage, les autres termes dépendent exclusivement de la structure secondaire $\mathcal{S}(\mathcal{C})$ correspondante à $\mathcal{C}$. On peut donc décomposer $h_{latt}(\mathcal{C}) = h_{latt}(\mathcal{S}(\mathcal{C})) + \kappa_h^{tot}(\mathcal{C})$ (idem pour l'entropie) où $\kappa^{tot}(\mathcal{C})$ est la somme de toutes les contributions de pliage. Pour l'exemple décrit ci-dessus, on aurait $h_{latt}(\mathcal{S}) = \omega_h + 2\epsilon_h + \gamma_h + \sigma_h$ et $\kappa_h^{tot}(\mathcal{C}) = \kappa_h$.

Pour le réseau CFC considéré, comme $\Psi = 0$, 60, 90 et 120°, on a $\kappa \equiv \bar{\kappa}(1 - \cos\Psi) = n\bar{\kappa}/2$ avec respectivement $n = 0$, 1, 2 et 3. La contribution totale des énergies de pliage peut donc se mettre sous la forme $\kappa^{tot} = n_{tot}\bar{\kappa}/2$ avec $n_{tot} \in \mathbb{N}$. Ainsi, l'énergie libre totale d'un conformation $\mathcal{C}$ est totalement caractérisée par la donnée de $\mathcal{S}(\mathcal{C})$ (dont on peut calculer directement $h_{latt}(\mathcal{S})$ et $s_{latt}(\mathcal{S})$) et de $n_{tot}(\mathcal{C})$.

Chapitre 4

Méthodes numériques

Afin d'utiliser le modèle sur réseau décrit dans la partie précédente pour calculer les propriétés thermodynamiques du repliement d'une molécule d'ARN, nous avons développé plusieurs méthodes numériques basées essentiellement sur des techniques de Monte-Carlo [178]. Nous expliquons tout d'abord comment nous avons modélisé le système en machine (section 4.1), puis nous introduisons quelques notions sur les simulations de Monte-Carlo (section 4.2), et enfin nous décrivons ces différentes méthodes (sections 4.3 et 4.4). Toutes ces méthodes ont été développées et codées pour n'étudier le repliement que d'un unique brin d'ARN. Cependant, la généralisation pour des molécules contenant plusieurs brins en découle directement.

4.1 Modélisation *in silico*

4.1.1 Représentation du brin d'ARN

Un brin d'ARN est représenté en machine par un vecteur Seq de taille N dont les éléments codent pour les nucléotides constituant le brin. A est codé par 1, G par 2, U par 3 et C par 4. Ainsi par exemple, pour le brin $5'AUGCGCUC3'$, on aurait $Seq = [1, 3, 2, 4, 2, 4, 3, 4]$.

La conformation du brin est quant à elle conservée dans un vecteur $Chain$ de taille $N \times 3$. La position du nucléotide n est alors donnée par le triplet $Chain(n)$ qui précise le noeud du réseau occupé par le nucléotide dans le système de coordonnées défini par les vecteurs de bases du réseau. Pour l'exemple de la figure 3.7 ii, on aurait $Chain = [(0, 0, 0); (1, 0, 0); (1, 1, 0)]$.

Pour décrire la connectivité entre nucléotides (soit la structure secondaire), on dispose d'un vecteur $Link$ tel que $Link(n) = m$ si n et m sont appariés dans la conformation courante ou $Link(n) = 0$ sinon. Par exemple pour une conformation ayant comme structure secondaire la figure 3.1 c, on aurait $Link = [9, 8, 7, 0, 0, 0, 3, 2, 1]$. Ainsi pour une conformation quelconque, $Link(n) = m \neq 0$ est équivalent à $Chain(n) = Chain(m)$ (2 nucléotides appariés occupent le même noeud) et $\{|Seq(n) - Seq(m)| = 2\} \vee \{Seq(n) \times Seq(m) = 6\}$ (appariement du type $A \cdot U$, $G \cdot C$ ou $G \cdot U$) et $\{Link(n+1) = m-1\} \vee \{Link(n-1) = m+1\}$ (association anti-parallèle).

L'enthalpie $h_{latt}(\mathcal{C})$ et l'entropie $s_{latt}(\mathcal{C})$ d'une conformation $\mathcal{C}$ sont alors calculées en analysant la structure secondaire associée via $Link$. En partant du nucléotide 1, on parcoure le vecteur $Link$: si n est à une interface (soit $[Link(n) \neq 0] \wedge [\{Link(n-1) = 0\} \vee \{Link(n+1) = 0\}]$), on ajoute à l'enthalpie et à l'entropie les termes interfacial (capping, fourche, dangle, empilement coaxial) et d'association correspondants ; si n appartient à l'intérieur d'une partie double-brin (soit $[Link(n) \neq 0] \wedge [Link(n-1) \neq 0] \wedge [Link(n+1) \neq 0]$), en plus des termes d'association, on ajoute l'éventuelle contribution du pliage entre les segments $(n-1, n)$ et $(n, n+1)$ en analysant l'angle formé par $Chain(n-1)$, $Chain(n)$ et $Chain(n+1)$.

4.1.2 Occupation du réseau

Pour pouvoir construire le vecteur de liaison $Link$ et pour décider si la conformation courante est compatible avec le fait que le chevauchement de deux nucléotides sur le même noeud est possible si et seulement si il y a formation d'une paire anti-parallèle de Watson-Crick ou bancale, on a besoin d'avoir accès à l'occupation du réseau. La méthode naïve pour vérifier que le nucléotide n est le seul à occuper le noeud $Chain(n)$ est de comparer sa position avec celle de tous les autres nucléotides. La complexité de cette méthode est $\mathcal{O}(N)$ par position à vérifier. Ainsi, pour confirmer la validité d'une certaine conformation, il faudrait $\mathcal{O}(N^2)$ opérations. Pratiquement, plusieurs méthodes sont utilisées pour diminuer significativement cette complexité [179]. Nous en avons utilisé principalement deux : une méthode utilisant une table de bits et une autre utilisant une table de hachage.

4.1.2.1 Méthode de la table de bits

On définit un tenseur B où l'on stocke l'état de tous les noeuds du réseau accessibles par le système. Soit, pour un noeud (i, j, k)

$$B(i,j,k) = (0,0) \quad \text{si } (i,j,k) \text{ n'est pas occupé} \tag{4.1}$$

$$= (n,0) \quad \text{si } \exists!\ n \text{ tel que } (i,j,k) = Chain(n) \tag{4.2}$$

$$= (n,m) \quad \text{si } \exists\ n,m \text{ tel que } (i,j,k) = Chain(n) = Chain(m) \tag{4.3}$$

Ainsi, vérifier l'occupation d'un noeud occupé par un nucléotide n se fait en $\mathcal{O}(1)$ opérations en regardant directement $B(Chain(n))$. La validation d'une conformation entière a donc une complexité en $\mathcal{O}(N)$. La contrepartie de cette méthode est l'énorme place mémoire ($\mathcal{O}(N^3)$) pris par B. Cela devient handicapant quand $N \sim 70 - 100$.

4.1.2.2 Méthode de la table de hachage

Soit un vecteur H de taille M (ou table de hachage), à chaque noeud $x = (i, j, k)$ du réseau, on attribue une adresse primaire $h(x)$ dans ce vecteur, via une fonction dite de hachage h que l'on définira plus tard. Le noeud $x = (i, j, k)$ est occupé si il existe un élément l de H tel que $H(l) = x$. Comme en général M ($\propto N$) est beaucoup plus petit que le nombre total de noeuds accessibles au système ($\propto N^3$), plusieurs noeuds partagent la même adresse primaire. On ne peut donc pas simplement définir $H(h(x)) = x$ sans potentiellement rencontrer des problèmes de chevauchement d'adresses. Pour éviter ce risque de collision, on effectue le schéma suivant pour vérifier l'occupation du noeud x : si $h(x)$ est déjà occupée, l'algorithme cherche successivement dans les adresses $h(x) + 1$, $h(x) + 2$,... (modulo M) jusqu'à ce qu'il trouve soit la position x déjà placée, soit un emplacement vide. Dans le pire des cas, une simple requête requiert $\mathcal{O}(N)$ opérations. Cependant, si h et M ont été bien choisis (voir plus bas), le nombre moyen d'opérations d'une requête est en $\mathcal{O}(1)$. Pour obtenir une telle efficacité, il faut que la table de hachage ne se remplisse pas trop rapidement (M ne doit pas être trop proche de N), $M = 5N$ est suffisant. De plus, la fonction de hachage h doit être éparse, c'est à dire : pour x et y deux noeuds proches sur le réseau, leurs adresses primaires $h(x)$ et $h(y)$ doivent être assez éloignées l'une de l'autre. Un bon choix est une fonction de hachage de la forme $h(i, j, k) = (c_1 i + c_2 j + c_3 k)$ modulo M avec c_1, c_2, c_3 et M premiers entre eux et $c_k \approx M^{k/4} > 1$ [179]. De ce fait, les $\{c_l\}$ ont des ordres de grandeur différent ce qui assure le comportement désiré pour h [1]. Ainsi, si on choisit bien ses paramètres, cette méthode permet en $\mathcal{O}(N)$ opérations et en occupant $\mathcal{O}(N)$ espaces mémoires de vérifier la compatibilité d'une conformation. Ses performances sont de loin les meilleures par rapport à la méthode naïve et à la table de bits. C'est d'ailleurs cette méthode que l'on utilisera concrètement.

1. Pour comprendre cela, regardons la fonction de hachage $h(i, j, k) = (i + j + k)$ modulo M. Deux noeuds voisins vont avoir des adresses primaires voisines. Or un chemin auto-évitant est un objet compacte ($R \sim N^{0.588}$) où les noeuds occupés sont proches les uns des autres. Ainsi avec une telle fonction de hachage, la vérification de l'occupation d'un site risque de prendre beaucoup trop de temps.

4.2 Notions sur les simulations de Monte-Carlo

4.2.1 Définitions

Prenons un système ergodique quelconque $\mathbb{S}$ caractérisé par un certain nombre n de paramètres d'ordre $\{V_j\}$ (par exemple, l'énergie, la magnétisation, le nombre de paires de bases, etc.). Et plaçons nous dans un ensemble thermodynamique quelconque (micro-canonique, canonique, grand-canonique, etc.) où le système à l'équilibre est défini par une distribution de probabilité $\rho(V_1(S_i), ..., V_n(S_i))$ d'observer le micro-état S_i [2]. Notons $\Omega(v_1, ..., v_n)$ le nombre de micro-états S_i du système tels que $\{V_1(S_i) = v_1, ..., V_n(S_i) = v_n\}$. Ainsi, la valeur moyenne à l'équilibre thermodynamique d'une observable quelconque $A(V_1, ..., V_n)$ est donnée par

$$\langle A \rangle = \frac{\sum_{S_i} A(V_1(S_i), ..., V_n(S_i))\rho(V_1(S_i), ..., V_n(S_i))}{\sum_{S_i} \rho(V_1(S_i), ..., V_n(S_i))} \tag{4.4}$$

$$= \frac{\sum_{\{V_j\}} A(\{V_j\})\Omega(\{V_j\})\rho(\{V_j\})}{\sum_{(\{V_j\})} \Omega(\{V_j\})\rho(\{V_j\})} \tag{4.5}$$

Cette dernière équation nous renseigne sur l'importance de la fonction nombre d'état $\Omega(\{V_j\})$. En effet, si par un moyen quelconque (une simulation par exemple) on a accès à cette fonction, il est alors possible de calculer la valeur moyenne d'une observable quelconque.

Dans le cadre du modèle sur réseau, les paramètres d'ordre que l'on choisit doivent permettre une description précise des propriétés que l'on désire étudier : la structure (via la connectivité) et la thermodynamique (via l'enthalpie et l'entropie). On a vu dans la section 3.3.5 que l'énergie libre d'une conformation $\mathcal{C}$ était caractérisée par la donnée de $\mathcal{S}(\mathcal{C})$, sa structure secondaire correspondante, et du nombre $n_{tot}(\mathcal{C})$ de contributions élémentaire $\bar{\kappa}/2$ dans son énergie de pliage. Ainsi, un bon jeu de paramètres d'ordre va être donnée par $\mathcal{S}(\mathcal{C})$ (dont on peut évaluer $h_{latt}(\mathcal{S})$ et $s_{latt}(\mathcal{S})$ et qui décrit aussi la connectivité) et $n_{tot}(\mathcal{C})$. En particulier, si on a accès à $\Omega(\mathcal{S}, n_{tot})$ (par exemple avec la méthode multi-histogramme, section 4.3.3 ; ou avec la méthode de la contrainte maximale, section 4.4.2), la différence d'énergie libre ΔG entre une structure secondaire arbitraire $\mathcal{S}$ et l'état dénaturé $\mathcal{S}_0$, à une température T quelconque, sera donnée par

$$\exp\{-\beta\Delta G(\mathcal{S}, T)\} = \exp\{-\beta\left[h_{latt}(\mathcal{S}) - Ts_{latt}(\mathcal{S})\right]\} \sum_{n_{tot}=0}^{\infty} \frac{\Omega(\mathcal{S}, n_{tot})}{\Omega_0} \exp\{-\beta n_{tot}(\bar{\kappa}_h - T\bar{\kappa}_s)/2\} \tag{4.6}$$

avec $\Omega_0 = \sum_{n_{tot}} \Omega(\mathcal{S}_0, n_{tot}) = \Omega(\mathcal{S}_0, 0)$. Et vu que $\bar{\kappa} \sim 10k_BT$, à l'ordre 0 en $\exp(-\beta\bar{\kappa}/2)$, on a

$$\Delta G(\mathcal{S}, T) \approx h_{latt}(\mathcal{S}) - T\left[s_{latt}(\mathcal{S}) + k_B \log\left(\Omega(\mathcal{S}, 0)/\Omega_0\right)\right] \tag{4.7}$$

La correction due aux ordres supérieures en $\exp(-\beta\bar{\kappa}/2)$ vaut environ $0.01k_BT$ pour chaque angle de pliage existant entre deux segments double-brins consécutifs (voir section 5.3.1). Cette correction peut être non négligeable quand la structure contient beaucoup de parties double-brins.

4.2.2 Évaluation pratique des valeurs moyennes

Supposons que l'on ait à notre disposition un algorithme qui génère aléatoirement une suite de micro-états $\{S_1, .., S_N\}$ selon une distribution $\pi(\{V_j\})$ connue. Soit $\mathcal{N}_\pi(\{V_j\})$ l'histogramme des paramètres d'ordre visités, on a

$$\mathcal{N}_\pi(\{V_j\}) \propto \Omega(\{V_j\}) \times \pi(\{V_j\}) \tag{4.8}$$

2. Par exemple, pour un système défini uniquement par son énergie E, dans l'ensemble canonique à une température fixée T $(= 1/(k_B\beta))$, $\rho(E(S_i)) = \exp(-\beta E(S_i))/Z$ avec $Z = \sum_{S_i} \exp(-\beta E(S_i)) = \sum_E \Omega(E)\exp(-\beta E)$, la fonction de partition.

En inversant l'équation précédent on a accès de manière relative au nombre d'état $\Omega \propto \mathcal{N}_\pi/\pi$.
De plus, en introduisant pour chaque micro-état généré un poids $W_i = \rho(\{V_j(S_i)\})/\pi(\{V_j(S_i)\})$, on obtient pour N grand

$$\begin{aligned}\bar{A} &\equiv \frac{\sum_i A(\{V_j(S_i)\})W_i}{\sum_i W_i} = \frac{\sum_{\{V_j\}} \mathcal{N}_\pi(\{V_j\})A(\{V_j\})\rho(\{V_j\})/\pi(\{V_j\})}{\sum_{\{V_j\}} \mathcal{N}_\pi(\{V_j\})\rho(\{V_j\})/\pi(\{V_j\})} \\ &\approx \frac{\sum_{\{V_j\}} \Omega(\{V_j\})A(\{V_j\})\rho(\{V_j\})}{\sum_{\{V_j\}} \Omega(\{V_j\})\rho(\{V_j\})} = \langle A \rangle \end{aligned} \tag{4.9}$$

Si l'on suppose que A et W sont décorrélés, l'erreur commise dans l'évaluation de $\langle A \rangle$ par l'équation 4.9 est bornée par (voir annexe 6.9)

$$\sigma_{\bar{A}}^2 \leq \frac{1}{N_B} \frac{\langle \widetilde{W}^2 \rangle}{\langle \widetilde{W} \rangle^2} \sigma_A^2 \tag{4.10}$$

avec N_B le nombre de blocks indépendants (décorrélés entre eux) dans la suite $\{S_1, .., S_N\}$ et $\widetilde{W}$ la valeur moyenne de W dans chaque block. Ainsi, plus la série de données est grande, plus l'erreur commise en évaluant $\langle A \rangle$ sera faible. De plus, via $\widetilde{W}$, on voit que le choix de la distribution π a une importance cruciale sur l'erreur commise. Comme $\langle \widetilde{W}^2 \rangle \geq \langle \widetilde{W} \rangle^2$, à nombre de blocks (donc de mesures indépendantes) constant, $\sigma_{\bar{A}}^2$ va être minimale si W est constant car dans ce cas $\langle \widetilde{W}^2 \rangle = \langle \widetilde{W} \rangle^2$. Deux cas particuliers sont communément étudiés

- L'echantillonage simple ("simple sampling") avec $\pi = 1$. Cela revient à générer aléatoirement et de manière équiprobable des micro-états. Dans ce cas, $W = \rho$ et l'erreur $\sigma_{\bar{A}}$ peut être relativement importante, notamment si les états à fort ρ (les états importants dans l'ensemble choisi) ont une densité d'état relativement faible et sont donc générés avec une très faible probabilité.
- L'échantillonage préférentiel ("importance sampling") avec $\pi = \rho$. Cela revient à générer aléatoirement des micro-états selon la même distribution que l'ensemble étudié. Dans ce cas, $W = 1$ et l'erreur $\sigma_{\bar{A}}$ est minimale.

D'autres types d'échantillonages existent comme l'échantillonage de Rosenbluth pour les chaînes polymériques que nous étudierons plus en détail dans la section 4.4.1.

4.2.3 Choix de l'algorithme

Nous avons supposé dans la partie précédente que nous avions à notre disposition un algorithme qui nous générait des micro-états selon une distribution donnée π. Le problème reste de construire un tel algorithme. Plusieurs stratégies existent. Le choix de telle ou telle méthode dépend essentiellement du système considéré et de l'ensemble thermodynamique à traiter. On peut cependant distinguer deux grandes familles :

- Les schémas dynamiques où la génération de micro-états se fait séquentiellement : le nouvel état S_i est construit à partir du précédent S_{i-1}.
- Les schémas statiques où chaque block indépendant de micro-états est généré séparément des autres.

Dans le cas précis du modèle sur réseau, nous avons développé des algorithmes dans les deux familles. L'approche dynamique va nous permettre à partir d'un état initial de générer une série de conformations et ainsi de pouvoir prédire et calculer les propriétés thermodynamiques des molécules étudiées. Elle est limitée pour le moment à l'étude de petits ARN (< 80 nts). L'approche statique quant à elle va nous permettre d'estimer l'entropie de conformation d'une structure secondaire donnée afin de corriger son énergie libre donnée par le modèle de Turner pour tenir compte complètement des effets de volume exclu. Elle autorise l'étude de plus grands ARN (jusqu'à environ 5000 nts).

4.3 Schémas dynamiques

Avant d'entrer plus en détail dans le cas particulier du modèle sur réseau (sections 4.3.2 et 4.3.3), intéressons nous de manière générale aux règles régissant la construction et l'acceptation de micro-états dans un schéma dynamique (section 4.3.1).

4.3.1 Règles d'acceptance

À partir d'un état S_o, on génère de manière aléatoire un nouvel état test S_n en modifiant S_o d'une certaine façon qui dépend du problème (pour le modèle sur réseau voir section 4.3.2). On doit décider maintenant si on rejette ou accepte ce nouvel état dans la suite d'états que notre algorithme construit selon la distribution d'équilibre π fixée. Le fait que nous nous plaçons à l'équilibre pour un système ergodique impose que dans la suite générée, le nombre de transition de S_o vers S_n doit être le même que le nombre de transitions de S_n vers S_o. Cette balance détaillée implique

$$\pi(S_o)\mathcal{P}(S_o \to S_n) = \pi(S_n)\mathcal{P}(S_n \to S_o) \tag{4.11}$$

avec $\mathcal{P}(S_o \to S_n)$ la probabilité de transition de S_o vers S_n qu'on réécrit sous la forme

$$\mathcal{P}(S_o \to S_n) \equiv \alpha(S_o \to S_n) \times \mathrm{acc}(S_o \to S_n) \tag{4.12}$$

avec $\alpha(S_o \to S_n)$ la probabilité de construire un mouvement S_o vers S_n et $\mathrm{acc}(S_o \to S_n)$ la probabilité d'accepter ce mouvement. Si l'on suppose que α est symmétrique (c'est à dire, $\alpha(S_o \to S_n) = \alpha(S_n \to S_o)$, l'équation 4.11 donne

$$\frac{\mathrm{acc}(S_o \to S_n)}{\mathrm{acc}(S_n \to S_o)} = \frac{\pi(S_n)}{\pi(S_o)} \tag{4.13}$$

Beaucoup de choix pour acc satisfont la condition 4.13. Un choix classique est celui dit de Metropolis [180] (originellement donné dans le cas d'un échantillonage préférentiel dans l'ensemble canonique)

$$\mathrm{acc}(S_o \to S_n) = \pi(S_n)/\pi(S_o) \quad \text{si} \quad \pi(S_n) < \pi(S_o) \tag{4.14}$$

$$= 1 \quad \text{sinon} \tag{4.15}$$

L'algorithme pour un schéma dynamique a alors la forme générale :

1. Génére aléatoirement un état initial S_0.
2. Construit à partir de S_0 un état S_1 selon la probabilité $\alpha(S_0 \to S_1)$.
3. Accepte le nouvel état avec une probabilité $\mathrm{acc}(S_0 \to S_1)$.
4. Si l'état est accepté, on rajoute S_1 à la liste des états générés, sinon on remet S_0 dans cette liste.
5. On réitère les étapes 2-3-4 le plus grand nombre de fois possibles.
6. À partir de la liste finale d'états générés, on calcule les valeurs moyennes des observables qui nous intéresse, dans l'ensemble thermodynamique choisi, grâce à la formule 4.9.

Dans le cadre du modèle sur réseau, on se place dans l'ensemble canonique et on réalise un échantillonage préférentiel suivant le schéma de Metropolis, ainsi pour une conformation $\mathcal{C}$, π est choisi proportionnel à son poids de Boltzmann

$$\pi(\mathcal{C}) \propto \exp\{-\beta[h_{latt}(\mathcal{C}) - Ts_{latt}(\mathcal{C})]\} \tag{4.16}$$

où $h_{latt}(\mathcal{C})$ et $s_{latt}(\mathcal{C})$ sont l'enthalpie et l'entropie de la conformation $\mathcal{C}$ calculées avec les paramètres du réseau (section 3.3.3). Dans la section suivante, nous décrivons comment sont générés les nouveaux états.

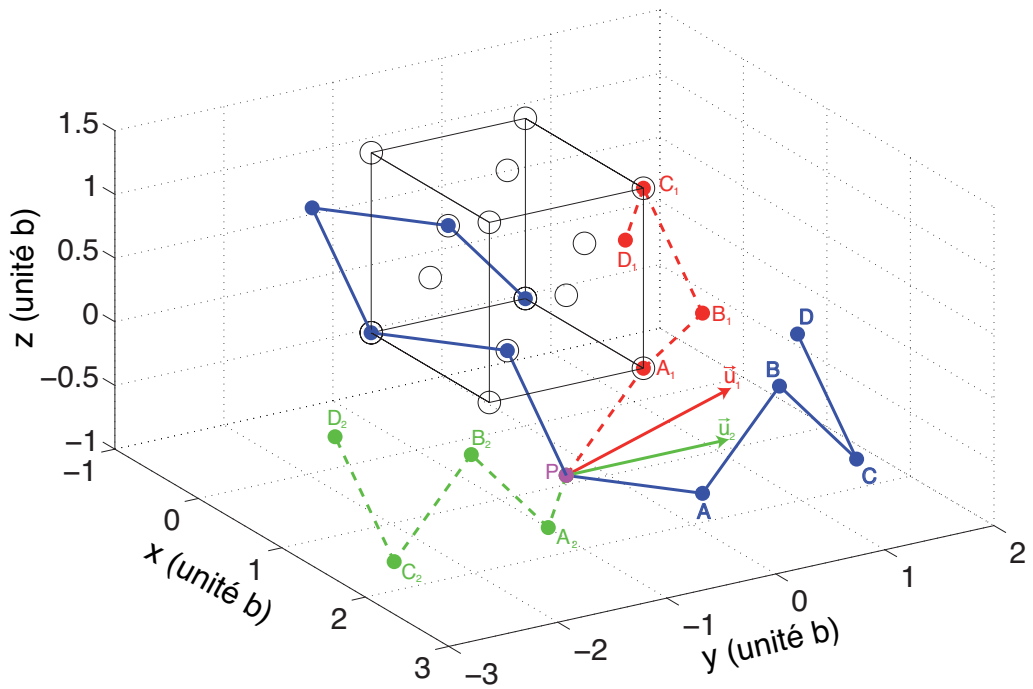

Figure 4.1 – Exemple de deux mouvements pivots appliqués au segment de la chaîne bleu $(P-A-B-C-D)$: une rotation d'axe $\vec{u}_1 = (\hat{x}+\hat{y}+\hat{z})/\sqrt{3}$ et d'angle $2\pi/3$ (pointillé rouge) et une réflexion de vecteur orthogonal $\vec{u}_2 = \hat{y}$ (pointillé vert). On a représenté également une maille du réseau CFC (ligne noire).

4.3.2 Mouvements élémentaires

Afin d'échantillonner l'ensemble canonique à une température T fixée, on part d'une conformation initiale aléatoire (par exemple un chemin auto-évitant sur le réseau), puis on génère successivement des conformations tests à l'aide de mouvements élémentaires de la chaîne. D'une part, ces mouvements doivent permettre de passer d'une conformation quelconque à une autre en un nombre fini d'étapes. On dit alors que l'algorithme est ergodique. D'autre part, ils doivent être efficaces dans le sens où ils doivent permettre un échantillonage rapide et complet de l'espace des conformations. Dans le cas des polymères, deux grandes familles de mouvements vérifiant les propriétés requises sont couramment utilisées : les mouvements non-locaux (comme l'algorithme du pivot [179] ou les transformations de segments internes [181]) et les mouvements locaux (comme la reptation ou le flip [182]). Dans le cadre du modèle sur réseau, en plus de ces mouvements "classiques", nous avons ajouté des mouvements spécifiques à notre système.

4.3.2.1 Mouvements non-locaux

Algorithme du pivot

Dans l'algorithme du pivot [179], un mouvement consiste à choisir aléatoirement un nucléotide de la chaîne, appelé pivot, et d'appliquer une opération de symétrie du réseau (réflexions ou rotations) au segment compris entre le pivot et la fin de la chaîne. Dans le cas du réseau CFC, les symétries sont les mêmes que celles du cube (voir figure 3.7)

- 3 plans de symétrie médiateurs de vecteurs orthogonaux $\hat{x}$, $\hat{y}$ et $\hat{z}$;
- 6 plans de symétrie diagonaux de vecteurs orthogonaux $(\hat{x}+\hat{y})/\sqrt{2}$, $(\hat{x}-\hat{y})/\sqrt{2}$, $(\hat{y}+\hat{z})/\sqrt{2}$, $(\hat{y}-\hat{z})/\sqrt{2}$, $(\hat{z}+\hat{x})/\sqrt{2}$ et $(\hat{z}-\hat{x})/\sqrt{2}$;
- 3 axes de rotation d'ordre 4, $\hat{x}$, $\hat{y}$ et $\hat{z}$, et d'angles 0, $\pi/2$, π et $3\pi/2$;
- 4 axes de rotation d'ordre 3, $(\hat{x}+\hat{y}+\hat{z})/\sqrt{3}$, $(\hat{x}+\hat{y}-\hat{z})/\sqrt{3}$, $(\hat{x}-\hat{y}+\hat{z})/\sqrt{3}$ et $(\hat{x}-\hat{y}-\hat{z})/\sqrt{3}$, et d'angles 0, $2\pi/3$ et $4\pi/3$;
- 6 axes de rotation d'ordre 2, $(\hat{x}+\hat{y})/\sqrt{2}$, $(\hat{x}-\hat{y})/\sqrt{2}$, $(\hat{y}+\hat{z})/\sqrt{2}$, $(\hat{y}-\hat{z})/\sqrt{2}$, $(\hat{z}+\hat{x})/\sqrt{2}$ et $(\hat{z}-\hat{x})/\sqrt{2}$, et d'angles 0 et π.

Il y a donc, en plus de l'identité, 9 réflexions et 23 rotations.
De manière générale, l'image d'un vecteur $\vec{v}$ par une rotation d'axe $\vec{u}$ (de coordonnée cartésienne (u_x, u_y, u_z)) et d'angle θ est donnée par

$$\mathrm{Rot}_{\vec{u}}(\vec{v}) = Q \times \vec{v} \tag{4.17}$$

avec

$$Q_{i,j} = u_i u_j (1-\cos\theta)\epsilon_{i,j,k} u_k \sin\theta \quad \text{pour } (i \neq j) \in \{x,y,z\} \tag{4.18}$$
$$Q_{i,i} = u_i^2(1-\cos\theta) + \cos\theta \quad \text{pour } i \in \{x,y,z\} \tag{4.19}$$

où $\epsilon_{i,j,k}$ est le symbole anti-symétrique de Levi-Civita [3].
L'image par une réflexion d'axe orthogonal $\vec{u}$ est quant à elle donnée par

$$\mathrm{Ref}_{\vec{u}}(\vec{v}) = \vec{v} - 2\left(\frac{\vec{v}\cdot\vec{u}}{|\vec{u}|^2}\right)\vec{u} \tag{4.20}$$

Ainsi, soit P, le pivot choisi, on a $\vec{OP} = i_p\vec{a}_1 + j_p\vec{a}_2 + k_p\vec{a}_3$ où $\{\vec{a}_i\}$ sont les vecteurs générateurs du réseau CFC (voir figure 3.7). L'image d'un nucléotide M (de coordonnée (i_m, j_m, k_m) dans la base des $\{\vec{a}_i\}$) appartenant au segment entre P et la fin de la chaîne sera donc situé sur le noeud de coordonnée

$$i'_m = i_p + (x'+y'-z')/\sqrt{2} \tag{4.21}$$
$$j'_m = j_p + (-x'+y'+z')/\sqrt{2} \tag{4.22}$$
$$k'_m = k_p + (x'-y'+z')/\sqrt{2} \tag{4.23}$$

avec, pour une rotation d'axe $\vec{u}$ et d'angle θ (voir figures 4.1 et 4.2 a),

$$\begin{aligned} x' &= \frac{1}{\sqrt{2}}\big\{[u_x^2(1-\cos\theta)+\cos\theta]d_1 + [u_x u_y(1-\cos\theta) - u_z\sin\theta]d_2 \\ &\quad + [u_x u_z(1-\cos\theta) + u_y\sin\theta]d_3\big\} \end{aligned} \tag{4.24}$$
$$\begin{aligned} y' &= \frac{1}{\sqrt{2}}\big\{[u_x u_y(1-\cos\theta) + u_z\sin\theta]d_1 + [u_y^2(1-\cos\theta)+\cos\theta]d_2 \\ &\quad + [u_y u_z(1-\cos\theta) - u_x\sin\theta]d_3\big\} \end{aligned} \tag{4.25}$$
$$\begin{aligned} z' &= \frac{1}{\sqrt{2}}\big\{[u_x u_z(1-\cos\theta) - u_y\sin\theta]d_1 + [u_y u_z(1-\cos\theta) + u_x\sin\theta]d_2 \\ &\quad + [u_z^2(1-\cos\theta)+\cos\theta]d_3\big\} \end{aligned} \tag{4.26}$$

où $d_1 = i_m - i_p + k_m - k_p$, $d_2 = i_m - i_p + j_m - j_p$ et $d_3 = j_m - j_p + k_m - k_p$;

3. $\epsilon_{i,j,k} = +1$ si $(i,j,k) = (x,y,z)$, (y,z,x) ou (z,x,y), $\epsilon_{i,j,k} = -1$ si $(i,j,k) = (x,z,y)$, (y,x,z) ou (z,y,x), et $\epsilon_{i,j,k} = 0$ si $i = j$, $j = k$ ou $k = i$.

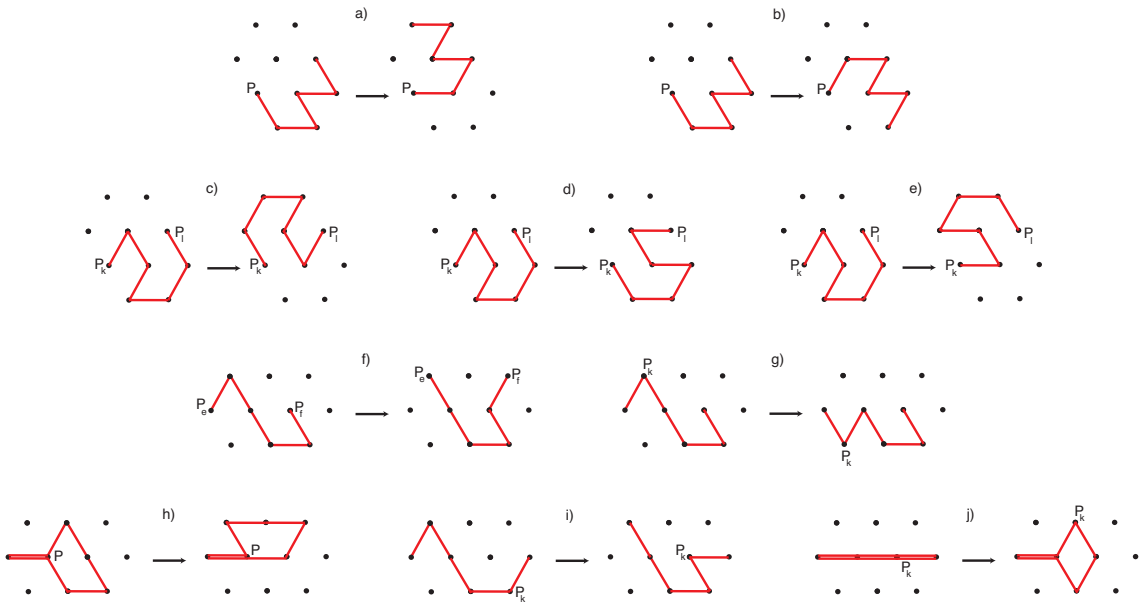

Figure 4.2 – Exemples de mouvements élémentaires pour une projection 2D du réseau CFC : (a) pivot-rotation, (b) pivot-réflexion, (c) inversion, (d) réflexion de segments internes, (e) échange, (f) reptation, (g) flip, (h) pivot spécifique, (i) glissement et (j) double-glissement.

et pour une réflexion de vecteur orthogonal $\vec{u}$ (voir figures 4.1 et 4.2 b)

$$x' = \frac{1}{\sqrt{2}}\left[d_1 - 2(d_1u_x + d_2u_y + d_3u_z)u_x\right] \tag{4.27}$$

$$y' = \frac{1}{\sqrt{2}}\left[d_2 - 2(d_1u_x + d_2u_y + d_3u_z)u_y\right] \tag{4.28}$$

$$z' = \frac{1}{\sqrt{2}}\left[d_3 - 2(d_1u_x + d_2u_y + d_3u_z)u_z\right] \tag{4.29}$$

Par construction, l'algorithme du pivot est symétrique.

Transformations de segments internes de chaînes

Ces transformations modifient des segments de la chaîne compris entre deux nucléotides aléotoirement choisis. Il existe trois différentes classes de transformations : l'inversion de segments et la réflexion ou l'échange de coordonnées [181]. On se place dans le système de coordonnée (i, j, k) défini par les vecteurs générateurs $\{\vec{a}_i\}$ du réseau CFC. Si l'on veut repasser en coordonnées cartésiennes, il faut effectuer la transformation

$$x = (i + k)/\sqrt{2} \tag{4.30}$$

$$y = (i + j)/\sqrt{2} \tag{4.31}$$

$$z = (j + k)/\sqrt{2} \tag{4.32}$$

Soit une chaîne composée de N nucléotides $(P_1 - P_2 - ... - P_N)$, tirons au hasard deux nucléotides P_k et P_l $(k \leq l)$.

L'inversion consiste à inverser le chemin $(P_k - ... - P_l)$ par rapport au point $(P_k + P_l)/2$ (voir figure 4.2 c), soit

$$P'_n = P_k + P_l - P_{k+l-n} \quad \text{pour } k \leq n \leq l \tag{4.33}$$
$$= P_n \quad \text{sinon} \tag{4.34}$$

La réflection consiste à réfléchir le chemin $(P_k - ... - P_l)$ par rapport au plan bissecteur de la ligne $[P_k, P_l]$ (voir figure 4.2 d). Plus précisément, soit $m \in \{-1, +1\}$ et $\alpha \neq \beta \in \{i, j, k\}$, si $P_l^\alpha - P_k^\alpha \neq m(P_l^\beta - P_k^\beta)$ ou si $P_l^\gamma \neq P_k^\gamma$ pour $\gamma \neq \alpha, \beta$, alors la réflexion est l'identité ; sinon on définit la réflexion par

$$P'^\delta_n = P_k^\delta - m(P_{k+l-n}^{\alpha+\beta-\delta} - P_l^{\alpha+\beta-\delta}) \quad \text{pour } k \leq n \leq l \text{ et } \delta = \alpha \text{ ou } \beta \tag{4.35}$$
$$= P_n^\delta \quad \text{sinon} \tag{4.36}$$

L'échange consiste à échanger deux coordonnées du chemin $(P_k - ... - P_l)$ (voir figure 4.2 e). Plus précisément, soit $m \in \{-1, +1\}$ et $\alpha \neq \beta \in \{i, j, k\}$, si $P_l^\alpha - P_k^\alpha \neq m(P_l^\beta - P_k^\beta)$, alors l'échange est l'identité ; sinon on définit l'échange par

$$P'^\delta_n - P'^\delta_{n-1} = m(P_n^\beta - P_{n-1}^\beta) \quad \text{pour } k < n \leq l \text{ et } \delta = \alpha \tag{4.37}$$
$$= m(P_n^\alpha - P_{n-1}^\alpha) \quad \text{pour } k < n \leq l \text{ et } \delta = \beta \tag{4.38}$$
$$= P_n^\delta - P_{n-1}^\delta \quad \text{sinon} \tag{4.39}$$

On peut vérifier de manière triviale que l'application successive de deux fois la même transformation de segments est l'application identité [4]. Chaque transformation de segments internes à la chaîne est donc symétrique.

Pivot spécifique

Dans l'algorithme du pivot défini ci-dessus, on applique l'opération de symétrie entre le pivot P et la fin de la chaîne. Dans la cas où P appartient à une partie double-brin, cela peut entraîner une ouverture du double-brin. Le but du pivot spécifique est de modifier la conformation initiale sans trop changer la connectivité dans la structure secondaire correspondante. Pour cela, on choisit un pivot P le long de la chaîne. Si P n'est pas apparié avec un autre nucléotide alors on effectue une pivot "classique" entre P et le bout de la chaîne. Par contre si P est apparié à Q (c'est à dire, si P et Q occupe le même noeud du réseau), on applique alors une opération de symétrie aux nucléotides compris entre P et Q(voir figure 4.2 h). Ce mouvement est trivialement symétrique.

4.3.2.2 Mouvements locaux

Les deux catégories précédentes mettent en jeu essentiellement des mouvements à grande échelle de la chaîne. Afin d'échantilonner efficacement sur des petites échelles, il est utile de définir également des mouvements locaux.

Reptation et flip

Soit une chaîne composée de N nucléotides $(P_1 - P_2 - ... - P_N)$, la reptation consiste à effacer le segment terminale à une extrémité de la chaîne et d'en rajouter un nouveau segment à l'autre

4. Par exemple, pour l'inversion, pour $k \leq n \leq l$, $(P'_n)' = P'_k + P'_l - P'_{k+l-n} = P_k + P_l - (P_k + P_l - P_{k+l-(k+l-n)}) = P_n$.

extrémité (voir figure 4.2 f). Plus précisément, soit une extrémité $e \in \{1, N\}$ (notons f l'autre) et $V \in \{\text{voisins de } P_f\}$, si on note $\delta = +1$ (-1) si $e = 1$ (N), la reptation est définie par

$$\begin{aligned} P'_n &= V \quad \text{pour } n = f & (4.40)\\ &= P_{n+\delta} \quad \text{sinon} & (4.41) \end{aligned}$$

Ainsi, la probabilité de construire un chemin entre deux conformations successives par reptation est donnée par la probabilité de choisir le bout correspondant multipliée par celle de choisir le voisin correspondant, soit $\alpha(o \to n) = (1/2) \times (1/z)$ qui est indépendante des conformations o et n. La reptation est donc un mouvement symétrique.

Le flip consiste à modifier uniquement la position d'un nucléotide dans la chaîne (voir figure 4.2 g). Plus précisément, soit $1 \leq k \leq N$ et $V \in \{\text{voisins communs à } P_{k-1} \text{ et } P_{k+1}\}$ si $2 \leq k \leq N-1$ ou $V \in \{\text{voisins de } P_2(P_N)\}$ si $k = 1$ (N), le flip est défini par

$$\begin{aligned} P'_n &= V \quad \text{pour } n = k & (4.42)\\ &= P_n \quad \text{sinon} & (4.43) \end{aligned}$$

Dans ce cas, la probabilité de construire un chemin entre deux conformations successives par flip est donnée par le probabilité de choisir le nucléotide k multipliée par celle de choisir le voisin correspondant, soit $\alpha(o \to n) = (1/N) \times (1/(\#\{\text{voisins communs à } P_{k-1} \text{ et } P_{k+1}\})$. Or le flip ne change la position que de P_k (et non de P_{k-1} et P_{k+1}) ainsi, le mouvement $n \to o$ aura la même probabilité de construction que le mouvement inverse (puisque $\#\{\text{voisins communs à } P_{k-1} \text{ et } P_{k+1}\}$ est conservé). Le flip est donc un mouvement symétrique.

Mouvements spécifiques de glissement

Dans la reptation, le mouvement d'une extrémité de la chaîne fait glisser la position des autres nucléotides sur celle de leur voisin dans le sens de la reptation (figure 4.2 f). Cependant ce mouvement devient rapidement inefficace quand la conformation comporte des parties double-brins. Par exemple, prenons une conformation qui comporte dans sa structure secondaire correspondante, la partie $^{AAGCU}_{UCGAA}$, après reptation, elle deviendra $^{AAGCU}_{UCGAA}$ et sera rejetée puisque les nucléotides se chevauchant ne forment plus des paires autorisées (voir section 3.1.1). Pour améliorer l'acceptation de mouvement de glissement, nous définissons une reptation interne où seule une partie de la chaîne glisse (voir figure 4.2 i).

Soit P_k un nucléotide et $P_{k+\delta}$ un de ses voisins ($\delta = \pm 1$), on note
$\mathcal{V} = \{\text{voisins communs à } P_k \text{ et } P_{k+\delta}\}$, $e = 1(N)$ si $\delta = -1(+1)$ et $V \in \mathcal{V}$.
Si $V \neq P_{k+2\delta}$, on définit le glissement par

$$\begin{aligned} P'_n &= V \quad \text{pour } n = k + \delta & (4.44)\\ &= P_{n-\delta} \quad \text{pour } \min(k+2\delta, e) \leq n \leq \max(k+2\delta, e) & (4.45)\\ &= P_n \quad \text{sinon} & (4.46) \end{aligned}$$

Sinon, si $V = P_{k+2\delta}$, soit $V_e \in \{\text{voisins de } P_e\}$,

$$\begin{aligned} P'_n &= V_e \quad \text{pour } n = e & (4.47)\\ &= P_{n+\delta} \quad \text{pour } \min(k+\delta, e-\delta) \leq n \leq \max(k+\delta, e-\delta) & (4.48)\\ &= P_n \quad \text{sinon} & (4.49) \end{aligned}$$

Le mouvement par construction n'est pas symétrique (puisqu'il dépend de la nature de V), on doit donc par un moyen détourné le rendre symétrique. Notons π_0 la probabilité de choisir un voisin V différent

de $P_{k+2\delta}$ et π_1 celle de choisir $P_{k+2\delta}$. Suivant les situations où $P_{k+2\delta}$ appartient ou n'appartient pas à $\mathcal{V}$, cela force

$$\begin{aligned} \#\mathcal{V} \times \pi_0 &\leq 1 && (4.50) \\ (\#\mathcal{V} - 1) \times \pi_0 + \pi_1 &\leq 1 && (4.51) \end{aligned}$$

De plus, la probabilité de construire un chemin entre deux conformations vaut $(1/N) \times (1/2) \times \pi_0$ si $P_{k+2\delta} \notin \mathcal{V}$ et $(1/N) \times (1/2) \times (\pi_1/z)$ si $P_{k+2\delta} \in \mathcal{V}$ (avec z le nombre de coordinence du réseau). Donc la symétrie que l'on veut imposer force $\pi_0 = \pi_1/z$. Ceci, conjugué aux inégalités 4.50 et 4.51, donne un choix optimal de

$$\begin{aligned} \pi_0 &= 1/[\#\mathcal{V} + (z-1)] && (4.52) \\ \pi_1 &= z/[\#\mathcal{V} + (z-1)] && (4.53) \end{aligned}$$

Pour le réseau CFC, cela donne $\pi_0 = 1/15$ et $\pi_1 = 4/5$. Dans ces conditions le mouvement est symétrique.

De plus, dans l'optique d'avoir un mouvement qui permet l'ouverture de boucle dans un double-brin, nous définissons un mouvement élémentaire de double-glissement qui consiste à effectuer deux glissements d'affilé (voir figure 4.2 j).

4.3.2.3 Efficacité et validation

Un pas de la simulation consiste donc à appliquer un mouvement élémentaire et à l'accepter ou non suivant la règle de Metropolis (Eq.4.14). On définit un pas de Monte-Carlo comme la succession d'un mouvement pivot, d'une transformation de segments, d'un pivot spécifique, d'une reptation, d'un flip, d'un glissement et d'un double glissement. La figure 4.3 A montre l'évolution de la structure secondaire courante lors de plusieurs simulations de Monte-Carlo pour une séquence pseudo-noeud : à $T = 290$ K, la structure pseudo-noeud est prédominante ; à $T = 330$ K, on est proche de la température de fusion et la structure courante fluctue énormément ; à $T = 370$ K, le pseudo-noeud n'est plus stable et l'état dénaturé prédomine. Pour $T = 330$ K, l'évolution de la conformation courante sur le réseau est représentée sur la figure 4.4. La figure 4.3 B montre l'évolution de la fonction d'autocorrélation du rayon de giration définie par

$$\text{AutoCorrel}_{R_G}(N_{MC}) = \frac{\langle R_G(N_{MC}) R_G(0) \rangle - \langle R_G \rangle^2}{\langle R_G^2 \rangle - \langle R_G \rangle^2} \tag{4.54}$$

avec N_{MC} le nombre de pas de Monte-Carlo. Cette fonction peut-être modélisée par une exponentielle décroissante $e^{-N_{MC}/\tau}$ où τ est le temps de corrélation. τ peut-être vu comme le temps moyen que met le système pour oublier sa configuration initiale. Plus il est petit, plus le nombre de mesures indépendantes (c'est à dire, décorrélées) sera grand (à nombre de pas de Monte-Carlo fixé). Typiquement, pour les tailles de séquences étudiées avec les schémas dynamiques, $\tau \sim 10^2 - 10^5$ pas de Monte-Carlo et dépend de la composition et de la taille de la séquence [179] et de la température. De plus, le taux d'acceptation des mouvements élémentaires considérés doit aussi être le plus grand possible afin d'éviter de "perdre" son temps à générer des nouvelles conformations tests qui ne seront pas acceptées. Dans l'exemple du pseudo-noeud étudié dans la figure 4.3, le taux moyen d'acceptation varie entre 0.27 à 290 K et 0.53 à 370 K. Bien entendu, en plus de la température, ce taux dépend aussi de la taille et de la composition de la chaîne [179].

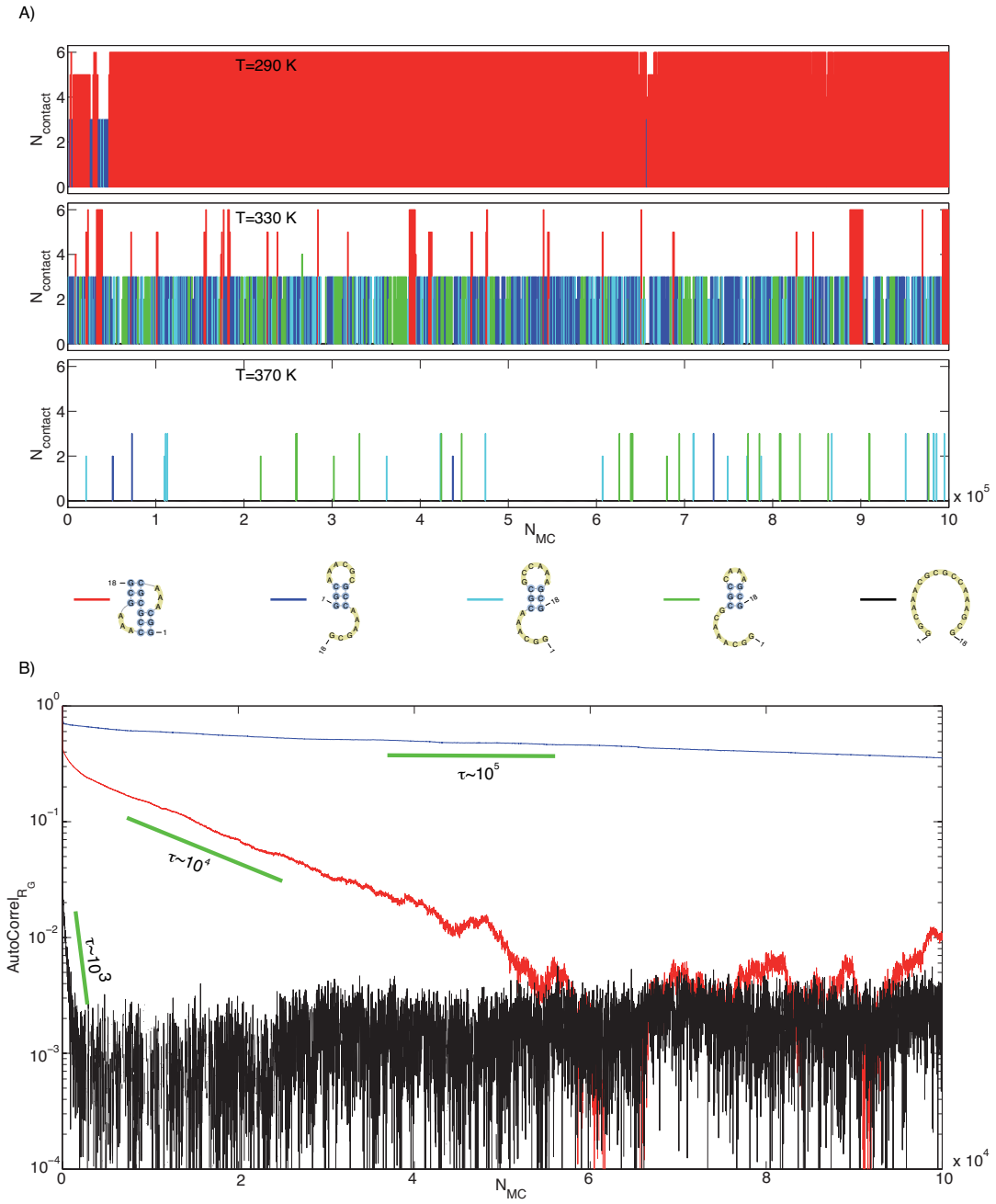

Figure 4.3 – (A) Evolution de la structure courante durant plusieurs simulations de Monte-Carlo à des températures différentes pour la séquence pseudo-noeud $GGCAAACGCGCCAAAGCC$ (voir aussi figure 5.4), en fonction du nombre de pas de Monte-Carlo N_{MC}. Pour chaque conformation, on trace son nombre de paires de bases $N_{contact}$ ainsi que son type de structure secondaire (voir légende). (B) Evolution de la fonction d'autocorrélation du rayon de giration $\text{AutoCorrel}_{R_G}(N_{MC})$ (équation 4.54) en fonction de N_{MC} pour la même séquence pseudo-noeud que dans (A) à $T = 290$ K (ligne bleu), 330 K (ligne rouge) et 370 K (ligne noire).

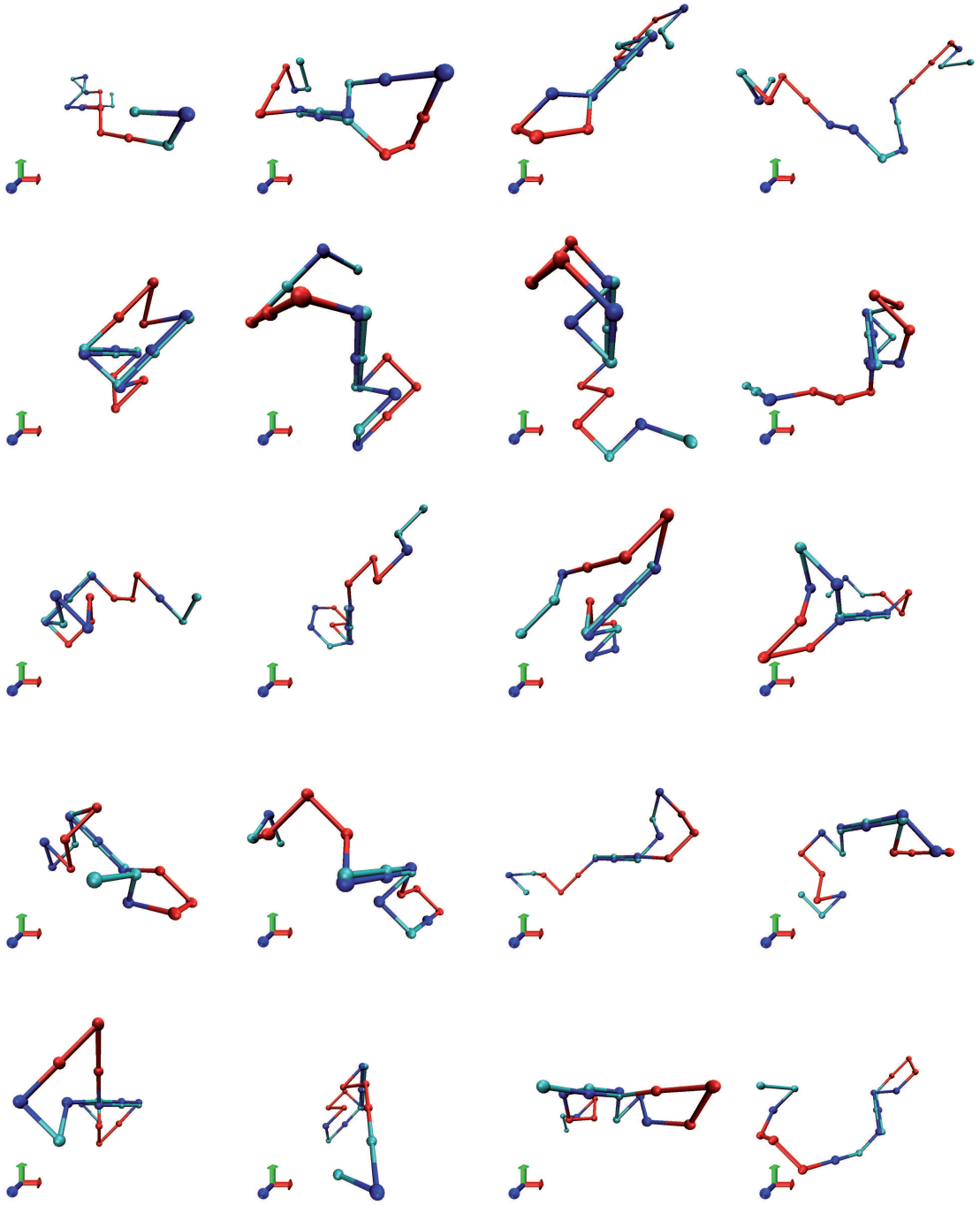

FIGURE 4.4 – Evolution de la conformation courante sur le réseau pour une simulation de Monte-Carlo à $T = 330$ K pour la séquence pseudo-noeud $GGCAAACGCGCCAAAGCC$. Les clichés sont pris tous les 5000 pas de Monte-Carlo.

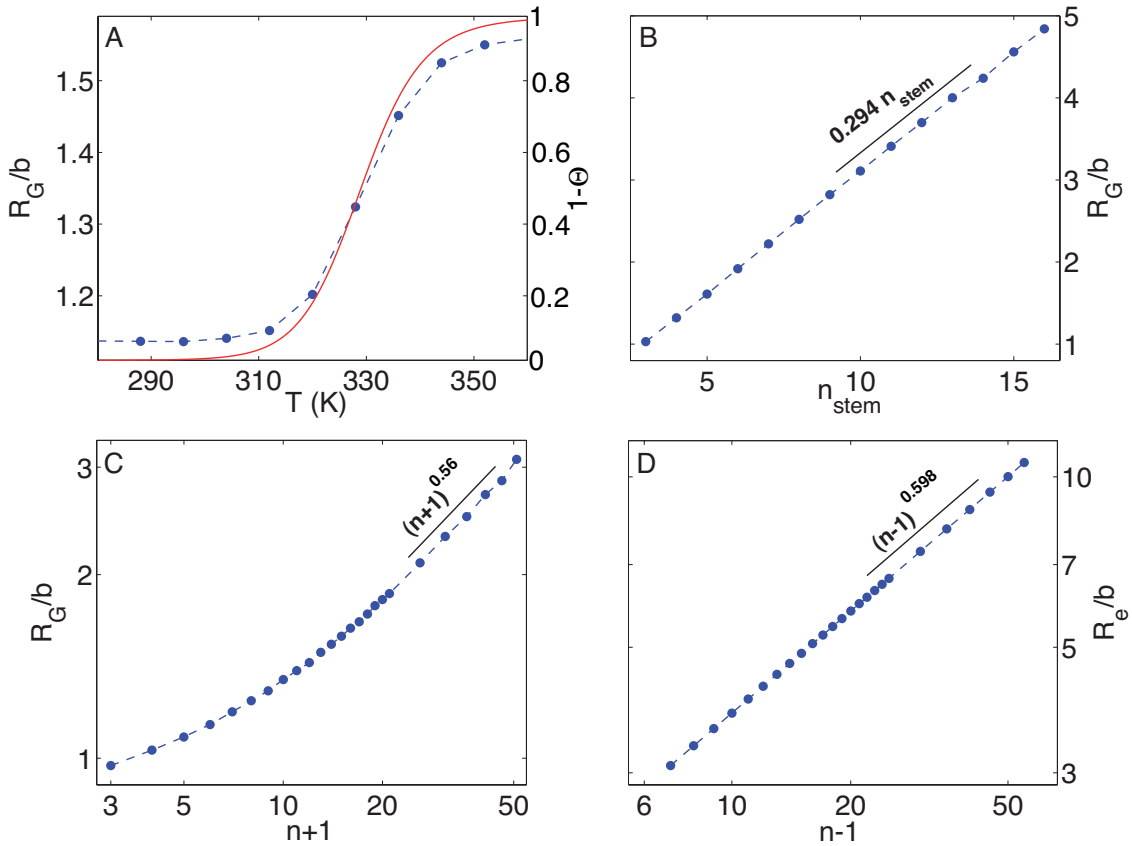

Figure 4.5 – (A) Evolution du rayon de giration R_G (points bleus) et de la probabilité d'ouverture $1-\Theta$ (ligne rouge) en fonction de la température T pour la séquence $GGCA_5GCC$. Pendant la dénaturation, les conformations passent d'une structure compact à une pelote aléatoire. (B) Evolution du rayon de giration R_G en fonction de la taille de la partie double-brin n_{stem} de la structure en épingle ayant pour séquence $GC_{n_{stem}-1}GCA_3GUG_{n_{stem}-1}C$. (C) Evolution du rayon de giration R_G en fonction de la taille $n+1$ de la boucle pour la structure en épingle $GGCA_nGCC$. (D) Evolution de la distance bout-à-bout moyenne en fonction du nombre de segments $n-1$ pour l'état dénaturé d'une chaîne contenant n nucléotides.

La figure 4.5 montre que les mouvements élémentaires choisis permettent de reproduire le comportement polymérique attendu du rayon de giration ou de la distance bout-à-bout En particulier, la figure 4.5 B montre l'influence de la taille d'un double-brin n_{stem} sur la rayon de giration pour une structure en épingle. On voit que R_G dépend linéairement de n_{stem} avec une pente de $0.294b$. Ce comportement est caractéristique des tiges rigides pour lesquelles $R_G \approx n_{stem}b/\sqrt{12} \approx 0.289bn_{stem}$ (voir annexe 6.8).La figure 4.5 C représente l'évolution de R_G pour une structure en épingle en fonction de la taille de la boucle. On trouve que $R_G \sim bN^{0.56}$. L'exposant que l'on trouve (0.56) est plus petit que l'exposant universelle des polygones auto-évitants (0.593) pour un réseau CFC [183] à cause de la relative compactification imposée par la petite partie double-brin connectée à la boucle en épingle. Enfin, la figure 4.5 D montre que l'on retrouve bien le comportement attendu pour un chemin auto-évitant sur un réseau CFC.

4.3.3 Méthode multi-histogramme

Dans le modèle sur réseau, les paramètres d'ordre utilisés sont la structure secondaire $\mathcal{S}$ et le nombre n_{tot} de contributions élémentaires dans l'énergie de pliage (voir section 4.2.1). Lors d'une simulation à une température T fixée, on enregistre l'histogramme $\mathcal{N}_T(\mathcal{S}, n_{tot})$. Comme précisé dans la section 4.2.2, on peut à partir de $\mathcal{N}_T$ évaluer de manière relative le nombre d'état par renormalisation

$$\Omega(\mathcal{S}, n_{tot}) \propto \frac{\mathcal{N}_T(\mathcal{S}, n_{tot})}{\exp\left\{-\beta\left[(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2) - T\left(s_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_s/2\right)\right]\right\}} \tag{4.55}$$

La figure 4.6 montre les courbes de dénaturation d'une structure en épingle calculées à partir des nombres d'état obtenus avec la formule précédente pour 3 températures différentes. Si l'on compare les courbes calculées à partir des histogrammes à $T = 280$ K et 360 K aux résultats issus directement des simulations de Metropolis (qui sont les résultats référents), on observe une bonne correspondance uniquement pour une plage température se trouvant autour de la température d'étude d'où l'on a extrait $\mathcal{N}_T$ et donc Ω. En effet, à une température d'étude donnée T_i, l'algorithme de Metropolis ne va générer avec une bonne statistique que les états ayant un fort poids de Boltzmann à T_i. Si l'on désire mesurer, à partir du nombre d'état estimé avec $\mathcal{N}_{T_i}$, une observable A à une température T éloignée de T_i, comme les états importants à T vont être différents de ceux à T_i, ils n'auront pas été visités avec une statistique suffisante lors de la simulation initiale, et l'évaluation de A aura de grande chance d'être erronée. C'est pourquoi, la meilleure des 3 courbes est obtenue pour la température de dénaturation où le système fluctue beaucoup et visite un grand nombre d'états. Ainsi, pour que l'utilisation du nombre d'états évalués à partir d'un unique simulation soit assez fiable, il faudrait connaître la température de dénaturation de la séquence ce qui nécessiterait de faire au préalable des simulations de Monte-Carlo. De plus, pour des séquences ayant des transitions à plusieurs paliers, quelles températures de fusion choisir ?

Le but de la méthode multi-histogramme [178, 184] est de donner une évaluation plus précise de Ω en combinant plusieurs histogrammes réalisés pour différentes températures. Pour chaque température T_i, on enregistre l'histogramme $\mathcal{N}_i(\mathcal{S}, n_{tot})$ des états visités. La probabilité d'observer $(\mathcal{S}, n_{tot})$ est alors donnée par

$$p_i(\mathcal{S}, n_{tot}) \equiv \frac{\mathcal{N}_i(\mathcal{S}, n_{tot})}{\mathcal{N}_i^{tot}} = \frac{\Omega(\mathcal{S}, n_{tot}) \exp\left\{-\beta_i\left[(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2) - T_i\left(s_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_s/2\right)\right]\right\}}{Z_i} \tag{4.56}$$

avec $\mathcal{N}_i^{tot}$ le nombre total d'entrées dans l'histogramme et Z_i la fonction de partition du système évaluée à T_i. En inversant l'équation 4.56, on trouve une estimation de Ω (similaire à l'équation 4.55). Ainsi, pour une température arbitraire T, la probabilité d'observer $(\mathcal{S}, n_{tot})$ estimée à partir de p_i vaut

$$p_i^{est,T}(\mathcal{S}, n_{tot}) = p_i(\mathcal{S}, n_{tot})\frac{Z_i}{Z}\exp\left\{(\beta_i - \beta)\left(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2\right)\right\} \tag{4.57}$$

On veut combiner tous les $p_i^{est,T}$ pour donner une meilleure estimation $p^{est,T}$ de la probabilité d'observer $(\mathcal{S}, n_{tot})$ à T :

$$p^{est,T} = \sum_{i=1}^{r} w_i p_i^{est,T} \tag{4.58}$$

où r est le nombre de températures d'étude et les w_i sont des poids statistiques tels que $\sum_i w_i = 1$ et $w_i \geq 0$. Le meilleur choix pour les poids est celui qui minimise l'erreur faite sur $p^{est,T}$ (voir annexe 6.9) soit

$$w_i = \frac{1/\sigma^2_{p_i^{est,T}}}{\sum_{j=1}^{r} 1/\sigma^2_{p_j^{est,T}}} \tag{4.59}$$

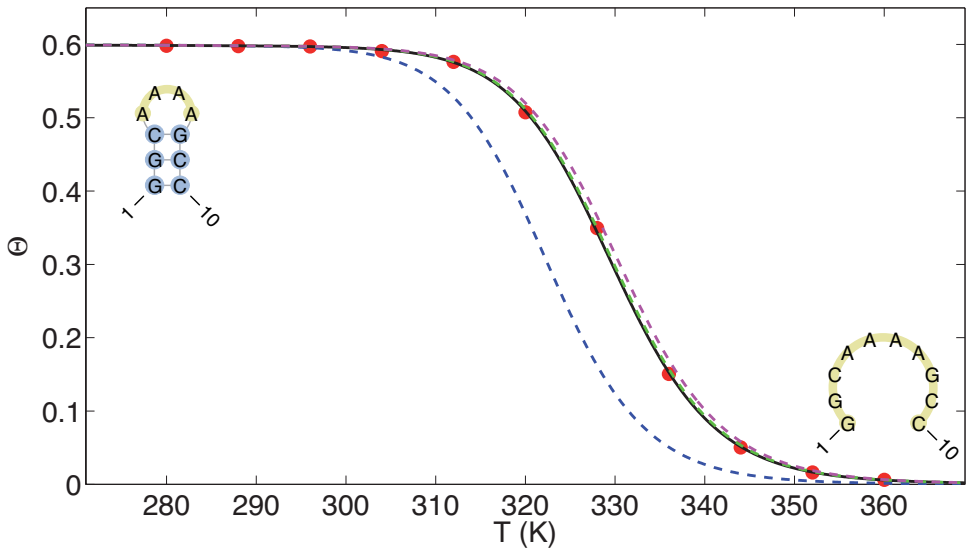

Figure 4.6 – Probabilité Θ pour un nucléotide d'être apparié en fonction de la température T pour la séquence $GGCAAAAGCC$; calculée avec différentes méthodes : simulations directes de Monte-Carlo (points rouges), renormalisations à partir des histogrammes obtenus à $T = 260$ K (pointillé bleu), à $T = T_m = 329.7$ K (pointillé vert) et à $T = 360$ K (pointillé violet), méthode multi-histogramme (ligne noire).

Evaluons $\sigma^2_{p_i^{est,T}}$. Chaque élément de $\mathcal{N}_i$ suit une loi binomiale (voir annexe 6.9), donc $\sigma^2_{\mathcal{N}_i} \approx \mathcal{N}_i$, cela donne $\sigma^2_{p_i} = p_i/\mathcal{N}_i^{tot}$ et finalement

$$\sigma^2_{p_i^{est,T}} = \frac{p_i^{est,T}}{\mathcal{N}_i^{tot}} \frac{Z_i}{Z} \exp\left\{(\beta_i - \beta)\left(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2\right)\right\} \tag{4.60}$$

En combinant les équations 4.58, 4.59 et 4.60, et en supposant que $p_i^{est,T} \approx p^{est,T}$ dans l'équation 4.60, on trouve

$$p^{est,T}(\mathcal{S}, n_{tot}) = \Omega(\mathcal{S}, n_{tot}) \frac{\exp\left\{-\beta\left[(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2) - T\left(s_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_s/2\right)\right]\right\}}{Z} \tag{4.61}$$

avec

$$\Omega(\mathcal{S}, n_{tot}) = \frac{\sum_{i=1}^{r} \mathcal{N}_i(\mathcal{S}, n_{tot})}{\sum_{i=1}^{r} \mathcal{N}_i^{tot} \exp\left\{-\beta_i\left[(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2) - T_i\left(s_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_s/2\right)\right]\right\}/Z_i} \tag{4.62}$$

Reste à déterminer les Z_i. Par définition de la fonction de partition, ils doivent vérifier le système d'équations implicites, pour $1 \leq i \leq r$

$$Z_i = \sum_{(\mathcal{S}, n_{tot})} \left(\frac{\sum_{j=1}^{r} \mathcal{N}_j(\mathcal{S}, n_{tot})}{\sum_{j=1}^{r} \mathcal{N}_j^{tot} \exp\left\{-\beta_j\left(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2\right)\right\}/Z_j} \right) \exp\left\{-\beta_i\left(h_{latt}(\mathcal{S}) + n_{tot}\bar{\kappa}_h/2\right)\right\} \tag{4.63}$$

La méthode multi-histogramme consiste d'abord à évaluer plusieurs histogrammes à des températures régulièrement espacées, puis à résoudre le système d'équations implicites définies par l'équation 4.63 et enfin à calculer le nombre d'état avec l'équation 4.62 et de s'en servir pour estimer les valeurs moyennes des observables désirées à des températures arbitraires. Ainsi, à l'aide d'un petit nombre de simulations (typiquement une dizaine), on reconstruit fidèlement le nombre d'état. La figure 4.6 confirme

l'exactitude des résultats obtenus sur toute la gamme de température étudiée. Traditionnellement, la méthode multi-histogramme est appliquée à des systèmes caractérisés par des potentiels énergétiques et décrits par un petit nombre de paramètres d'ordre. L'originalité de notre approche est de construire les histogrammes en partie par rapport à la structure secondaire (et non par rapport à l'énergie) et de travailler avec un Hamiltonien dépendant de la température.

D'un point de numérique, la résolution du système d'équations implicites se fait par la méthode du point fixe. Réécrivons tout d'abord le système sous la forme vectorielle $\vec{Z} = \vec{f}(\vec{Z})$ avec $\vec{Z} = (Z_1, ..., Z_r)$ et $\vec{f}$ une fonction vectorielle. On peut remarquer que si $\vec{Z}^*$ est solution alors $\alpha\vec{Z}^*$ (avec $\alpha \in \mathbb{R}$) aussi. Ce qui implique que la solution du système sera connue à une constante multiplicative près, dans ce cas fixons $Z_1 = 1$. Résoudre le système revient donc à trouver le point fixe de la suite $\vec{Z}_{n+1} = \vec{f}(\vec{Z}_n)$, ce qui revient, en supposant l'unicité du point fixe (ce qui est le cas en pratique), à chercher la limite de cette même suite. L'idée est donc de partir d'une condition initiale quelconque $\vec{Z}_0$ est d'appliquer successivement $\vec{f}$ jusqu'à obtenir une converge à 10^{-3} près. Cela nécessite typiquement $\sim 10^2 - 10^3$ applications de la fonction.

Ainsi, Ω est a priori connu à une constante multiplicative près. Ce qui n'est pas dérangeant puisque lorsque l'on calcule une valeur moyenne, la normalisation élimine cette constante (voir équation 4.5). Cependant, on peut estimer de manière absolue Ω dans le cas particulier du modèle sur réseau, car on connaît sa valeur pour l'état dénaturé $\mathcal{S}_0$ qui est le nombre de SAW sur le réseau soit $\Omega_0 \equiv \Omega(\mathcal{S}_0) = f_s \mu^N N^{c'}$. Si on note $\Omega(\mathcal{S}, n_{tot})'$ le résultat donné par l'équation 4.62, le nombre absolu d'état sera alors donné par

$$\Omega(\mathcal{S}, n_{tot}) = \Omega(\mathcal{S}, n_{tot})' \times \frac{\Omega_0}{\Omega_0'} \tag{4.64}$$

Typiquement, dans nos simulations, les températures T_i étudiées appartiennent à l'intervalle 273-380 K espacées de 5 à 8 K. Pour chaque température, on enregistre l'histogramme cumulé de 5 à 10 différentes simulations contenant chacune environ 10^7 pas de Monte-Carlo. Cela correspond au minimum à environ 10^3 mesures indépendantes (en comparaison avec les temps de corrélation, voir section 4.3.2). Pour une séquence composée de 20 nucléotides, la méthode multi-histogramme met 2 heures à tourner sur une ordinateur 2.4 GHz pour 13 températures d'étude et 5×10^7 pas de Monte-Carlo à chaque température. Avec de tels paramètres, on obtient une faible erreur statistique sur nos mesures. Par exemple, l'erreur réalisé sur l'évaluation de l'énergie libre d'une structure secondaire est typiquement de $0.05 k_B T$, se traduisant par une erreur d'environ 0.2 K sur la détermination de la température de mélange.

Néanmoins, la méthode multi-histogramme devient inefficace pour des séquences de longueur $\gtrsim 80$ nucléotides où la chaîne reste longtemps piégée dans des minima secondaires du paysage énergétique [20]. Avec nos mouvements élémentaires, il n'est pas possible de s'échapper de ses états sans franchir des hautes barrières d'énergie dues à des états de transition instables thermodynamiquement.

4.4 Schémas statiques

Le schéma statique développé pour le modèle sur réseau étant inspirer d'algorithmes de croissance de chaînes, nous nous intéressons dans un premier temps à plusieurs algorithmes qui permettent de générer des conformations de polymères sur réseau (section 4.4.1) ; puis nous introduirons la méthode de la contrainte maximale utilisée pour échantillonner l'ensemble des conformations sur réseau décrivant une même structure secondaire (section 4.4.2).

4.4.1 Algorithmes de croissance de chaînes

Considérons une chaîne polymérique auto-évitante composée de $N+1$ monomères . Chaque conformation $\mathcal{C}$ du polymère est caractérisée par une énergie $E(\mathcal{C})$ et on se place dans l'ensemble canonique à

température T, soit $\rho(\mathcal{C}) \propto \exp[-\beta E(\mathcal{C})]$. Dans la suite, on décrit brièvement des méthodes statiques pour calculer les propriétés thermodynamiques d'un tel système.

4.4.1.1 Énumération exacte et échantillonnage simple

Pour une chaîne idéale sans volume exclu sur un réseau, on connaît exactement le nombre total de conformations $z(z-1)^{N-1}$ ($z-1$ car on interdit le retour en arrière). Une première méthode est donc d'énumérer exactement toutes les conformations possibles sans volume exclu et de repérer celles qui sont auto-évitantes. Sur le réseau CFC, pour $N = 10$ cela revient à générer environ 10^9 conformations, pour $N = 100$, plus de 8×10^{21} conformations. Comme le nombre de conformations croît exponentiellement avec N, l'énumération exacte ne sera donc efficace et envisageable que pour de très petites ($N \sim 10$) chaînes.

Si notre système est décrit par p paramètres d'ordre $\{V_j\}$, une deuxième méthode est d'évaluer le nombre d'état $\Omega(V_1, ..., V_p)$ à l'aide d'un échantillonnage simple. Elle consiste à faire croître itérativement la chaîne du premier au dernier monomère suivant le schéma suivant (voir figure 4.7 A) :

1. Insérer le premier monomère à une position initiale.
2. Pour l'insertion du n^{ieme} monomère ($2 \leq n \leq N+1$), on choisit sa position aléatoirement parmi les $z-1$ voisins du monomère $n-1$.
3. L'étape 2 est répétée jusqu'à ce que le chaîne soit entièrement construite, on évalue alors les paramètres d'ordre pour la conformation générée et on met à jour l'histogramme des micro-états visités. Si durant la génération de la conformation, un des monomère est inséré à un noeud déjà occupé, on arrête le processus et on recommence à l'étape 1.

On répète cet algorithme un grand nombre de fois pour avoir une statistique suffisante sur le nombre d'état Ω avec lequel on peut calculer la valeur moyenne d'une observable quelconque à l'aide de l'équation 4.5. La probabilité de générer une conformation quelconque $\mathcal{C}$ par l'échantillonnage simple est de $\pi = 1/[z(z-1)^{N-1}]$. Or il y a $\mathcal{N}_{SAW}(N) = f_s \mu^N N^{c'}$ conformations auto-évitantes au total (voir section 3.1.2). Donc la probabilité de générer une conformation auto-évitante est donnée par $\mathcal{P} = \pi \times \mathcal{N}_{SAW}(N)$. L'inverse de cette probabilité est appelé l'attrition et notée $Att(N) = \mathcal{N}_{SAW}(N)/[z(z-1)^{N-1}]$ et représente le nombre moyen de départ de construction de chaîne à effectuer avant d'obtenir une conformation auto-évitante. La figure 4.8 montre l'évolution de Att en fonction de N pour un réseau CFC. On observe une croissance exponentielle pour l'attrition. Pour $N = 10$, plus de la moitié des essais de construction aboutissent alors que pour $N = 100$, seul un essai sur 5000 en moyenne aboutit à une conformation auto-évitante (pour $N = 1000$, $\mathcal{P} \sim 4 \times 10^{-40}$). Ainsi pour des longueurs $\gtrsim 100$, l'algorithme peut mettre beaucoup de temps avant de construire une bonne conformation et est fortement inefficace. De plus, même pour les longueurs où $\mathcal{P}$ reste raisonnable (typiquement $N \sim 10 - 30$), l'échantillonnage simple a tendance à générer très peu de conformations compactes qui sont souvent thermodynamiquement dominantes à faibles températures. Pour ces températures, la statistique n'est donc pas suffisante et cela peut donner des erreurs conséquentes sur l'évaluation des moyennes. Notons également que l'échantillonnage simple ne permet pas un accès direct au nombre d'état exacte mais nous le fournit à une constante multiplicative près inconnue (que l'on peut éventuellement évaluer si l'on connaît la valeur d'un des éléments de Ω, voir équation 4.64).

4.4.1.2 Échantillonage de Rosenbluth

Principe

Afin d'éviter de perdre du temps à générer des conformations qui seront soit rejetées car non auto-évitantes soit non pertinentes thermodynamiquement, l'idée de Rosenbluth et Rosenbluth [185] est de

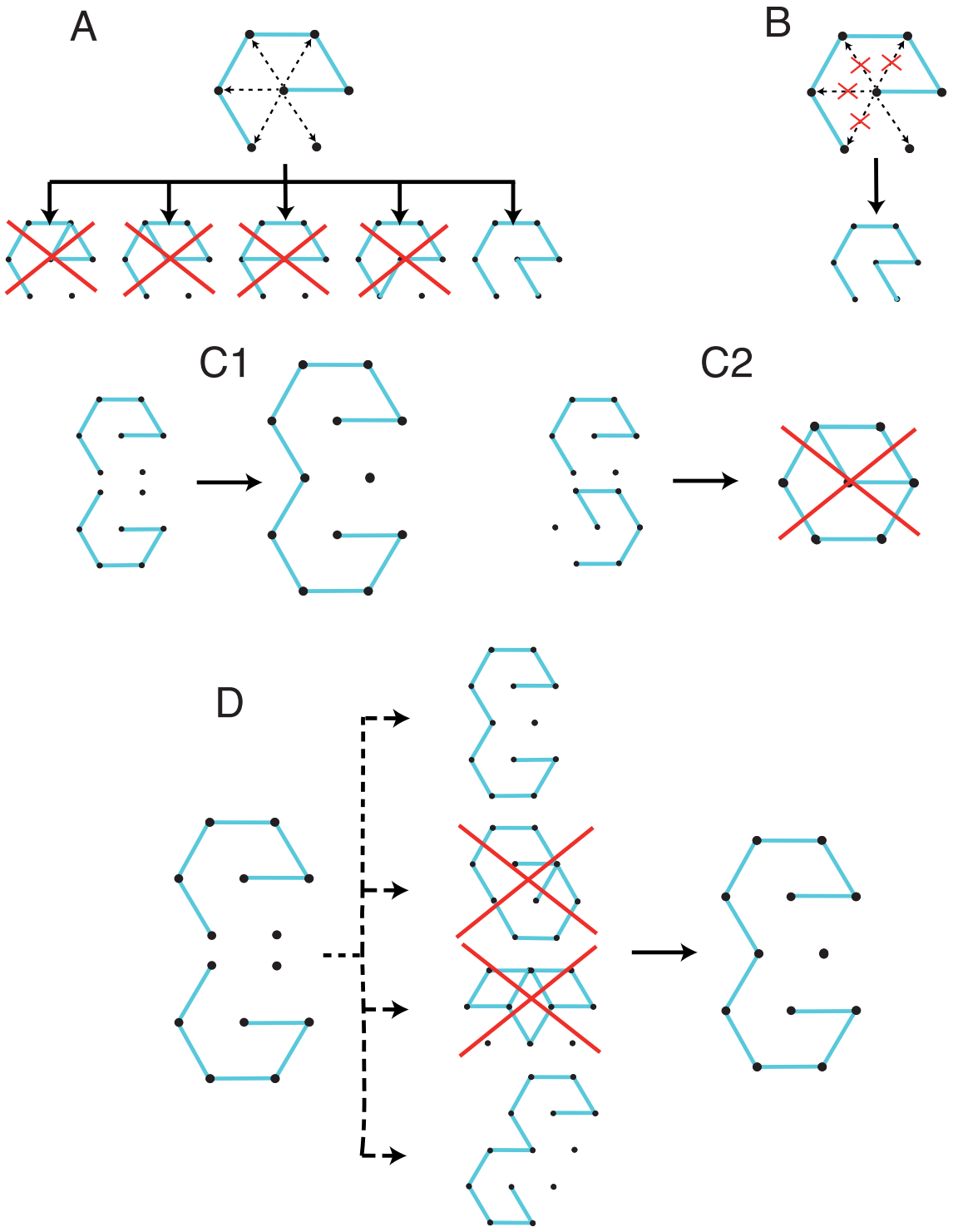

FIGURE 4.7 – (A) Échantillonage simple : on choisit au hasard parmi les $z-1$ directions possibles. (B) Échantillonnage de Rosenbluth : on choisit au hasard parmi les k_l voisins libres. (C) Dimérisation simple : on colle bout-à-bout les deux demi-conformations. (D) Dimérisation avec biais : on génére plusieurs conformations à l'aide de symétries sur les demi-conformations et on en choisit une au hasard parmi les k_l auto-évitantes.

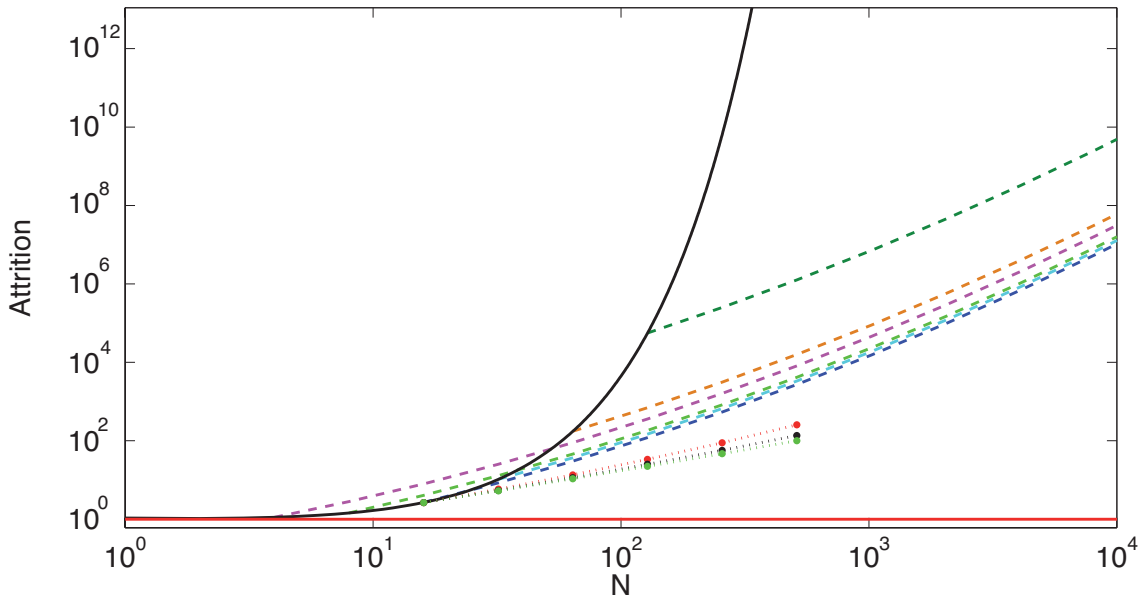

Figure 4.8 – Évolution de l'attrition en fonction de la taille N de la conformation à constuire pour différentes méthodes : échantillonnage simple (ligne noire), échantillonnage de Rosenbluth et méthodes PERM (ligne rouge), dimérisation simple (pointillés) avec $N_0 = 4$ (violet), 8 (vertclair), 16 (bleu), 32 (cyan), 64 (orange) et 128 (vert foncé), et la dimérisation avec biais (points) avec $N_0 = 16$ et $N_{trans} = 3$ (rouge), 8 (noir) et 16 (vert).

biaiser la croissance itérative de la chaîne en favorisant à chaque étape de construction la direction avec le plus grand poids de Boltzmann. Plus précisément, pour une chaîne contenant $N+1$ monomères, la génération d'une conformation suit l'algorithme suivant (voir figure 4.7 B) :

1. Insérer le premier monomère à une position initiale $\vec{r}_0$ avec une probabilité

$$p_1(\vec{r}_0) = \frac{\exp[-\beta E_1(\vec{r}_0)]}{w_1} \quad \text{avec } w_1 = \sum_{\vec{r}_i} \exp[-\beta E_1(\vec{r}_i)] \tag{4.65}$$

   avec $E_1(\vec{r}_i)$ l'énergie du monomère 1 à la position $\vec{r}_i$ dans le potentiel extérieure à la chaîne.

2. Pour l'insertion du n^{ieme} monomère ($2 \leq n \leq N+1$), on considère toutes les z positions-tests adjacentes au monomère $n-1$. Si l'on note $E_n(\alpha)$ l'énergie de la position-test α dans le potentiel crée par tous les monomères précédemment insérés plus le potentiel extérieur (si le site α est déjà occupé, on a $E_n(\alpha) = +\infty$), parmi les z possibilités, on en choisit une avec une probabilité

$$p_n(\alpha) = \frac{\exp[-\beta E_n(\alpha)]}{w_n} \quad \text{avec } w_n = \sum_{\alpha'=1}^{z} \exp[-\beta E_n(\alpha')] \tag{4.66}$$

3. L'étape 2 est répétée jusqu'à ce que le chaîne soit entièrement construite ou que l'on soit dans une impasse (tous les sites voisins sont déjà occupés).

L'énergie totale de la conformation $\mathcal{C}$ générée est alors donnée par $E(\mathcal{C}) = \sum_{n=1}^{N} E_n(\alpha)$. La probabilité d'avoir construit la conformation $\mathcal{C}$ vaut

$$\pi(\mathcal{C}) = \prod_{n=1}^{N+1} p_n = \frac{\exp[-\beta E(\mathcal{C})]}{W(\mathcal{C})} \quad \text{avec } W(\mathcal{C}) = \prod_{n=1}^{N+1} w_n \tag{4.67}$$

Comme $\rho(\mathcal{C}) \propto \exp[-\beta E(\mathcal{C})]$, W, appelé poids de Rosenbluth, est exactement le poids introduit dans la section 4.2.2 pour calculer une valeur moyenne quelconque ($W = \rho/\pi$). Ainsi la moyenne thermodynamique d'une observable A quelconque peut être calculée avec

$$\langle A \rangle = \frac{\sum_{i=1}^{N_{essai}} A(\mathcal{C}_i) W(\mathcal{C}_i)}{\sum_{i=1}^{N_{essai}} W(\mathcal{C}_i)} \tag{4.68}$$

où N_{essai} est le nombre d'essais de construction de conformations (incluant les cas où l'on tombe sur une impasse pour lesquelles on adopte la convention $W = 0$).

Une des avantages de la méthode de Rosenbluth est de pouvoir estimer simplement l'énergie libre du système en calculant la valeur moyenne arithmétique du poids de Rosenbluth sur les différentes réalisations [5] :

$$G = -k_B T \log Z \quad \text{avec } Z = \frac{1}{N_{essai}} \sum_{i=1}^{N_{essai}} W(\mathcal{C}_i) \tag{4.69}$$

Un autre avantage est que sans simulations supplémentaires, on a accès à la thermodynamique de toutes les sous-chaîne $[1, ..., n]$ avec $1 \leq n \leq N+1$. En effet, à chaque étape de construction, après l'insertion du monomère n, on peut définir un poids de Rosenbluth partiel $W_n = \prod_{n'=1}^{n} w_{n'}$ qui nous permet de calculer les propriétés thermodynamiques de ces sous-chaînes en utilisant également les formules 4.68 et 4.69.

Alors qu'avec l'échantillonnage simple la probabilité de générer un chemin auto-évitant était de l'ordre de 0.02% pour $N = 100$, elle est de presque 100% avec l'échantillonnage de Rosenbluth et reste même importante ($> 98\%$) pour $N \sim 1000$. L'attrition reste longtemps d'ordre 1. Cependant, la figure 4.9 illustre la tendance de l'algorithme de Rosenbluth à générer des structures plus compactes que la distribution de Boltzmann correspondante. En effet, pour une conformation compacte, le nombre de voisins libres à chaque étape de construction est en moyenne plus faible que pour une conformation non compacte, donc les $\{w_i\}$ vont avoir tendance à être plus petits et donc la probabilité π plus grande (voir équation 4.67). Bartoulis et Kremer [186] montrent que la différence entre la distribution de Rosenbluth et celle de Boltzmann croît exponentiellement avec N. Ainsi à partir d'une certaine gamme de longueur (typiquement $N \gtrsim 200$), les deux distributions vont faiblement se recouvrir, les fluctuations de W ($\langle W^2 \rangle / \langle W \rangle^2$) vont donc devenir grandes et l'erreur commise sur l'évaluation des moyennes et de l'énergie libre va alors être importante (voir équation 4.10).

Croissance de chaînes entre deux points fixes

Pour les cas où l'on veut générer des conformations dont la position des extrémités est fixées, Dijkstra, Frenkel et Hansen [187] proposent de modifier le biais de Rosenbluth en choisissant la prochaine direction avec une probabilité proportionnelle au poids de Boltzmann et au nombre de chemins idéaux sans volume exclu entre la position-test et la position de l'extrémité à atteindre. Plus précisément, soit $\vec{a}_0$ et $\vec{b}_0$ les positions fixes des deux extrémités, le premier monomère est toujours inséré en $\vec{a}_0$, puis dans l'étape 2 de l'échantillonnage de Rosenbluth, pour chaque position-test α du monomère n, on note c_α le nombre de chemins idéaux sans volume exclu entre $\vec{r}(\alpha)$ et $\vec{b}_0$, alors on choisit la position α avec une probabilité

$$p_n(\alpha) = \frac{\exp[-\beta E_n(\alpha)]}{w_n} \quad \text{avec } w_n = \frac{\sum_{\alpha'=1}^{z} c_{\alpha'} \exp[-\beta E_n(\alpha')]}{c_\alpha} \tag{4.70}$$

La définition du poids de Rosenbluth et le calcul des valeurs moyennes et de l'énergie libre restent inchangés par rapport au cas classique non-contraint. Dans la publication originale, Dijkstra, Frenkel et Hansen dérivent une formule pour $c(\alpha)$ dans un réseau cubique. En annexe 6.8, nous généralisons la formule pour un réseau CFC.

5. En effet, $\langle W \rangle = (1/N_{essai}) \sum_{i=1}^{N_{essai}} W(\mathcal{C}_i) = \sum_{\mathcal{C}} \pi(\mathcal{C}) W(\mathcal{C})$. Soit avec l'équation 4.67, $\langle W \rangle = \sum_{\mathcal{C}} \exp[-\beta E(\mathcal{C})] = Z$.

4.4.1.3 Méthodes PERM

Stratégie "éliminer-enrichir" pour la méthode de Rosenbluth

Afin de supprimer les grandes fluctuations observées avec la méthode de Rosenbluth quand les tailles des chaînes deviennent importantes, Grassberger [188] développe une stratégie, que l'on nommera méthode PERM (pour "Pruned-Enriched Rosenbluth Method"), où l'on élimine les conformations à faible poids de Rosenbluth et où l'on enrichit l'échantillon avec des copies de conformations à fort poids de Rosenbluth. Plus précisément, si l'on reprend le schéma de Rosenbluth décrit dans la section précédente, à chaque étape de la construction, après l'insertion du monomère n, si le poids de Rosenbluth partiel W_n est plus grand qu'un certain seuil supérieur $W_n^>$, on clone la conformation partielle courante et on continue le processus de croissance indépendamment pour les deux copies. Le poids partiel de Rosenbluth pour chaque clone vaut alors la moitié du poids original. Par contre, si W_n est plus petit qu'un certain seuil inférieur $W_n^<$, on arrête le processus de croissance de la conformation courante avec une probabilité 1/2. Si la conformation "survit", son poids partiel est doublé. Le calcul des moyennes thermodynamiques et des énergies libres suit les mêmes règles que dans l'échantillonnage de Rosenbluth simple. La figure 4.9 illustre bien le fait que la distribution PERM et celle de Boltzmann sont maintenant très similaires même pour des grandes chaînes.

A priori, la justesse de l'algorithme est indépendante des valeurs des deux seuils qui peuvent également évoluer durant la simulation, cependant son efficacité va fortement en dépendre. Dans l'article original de Grassberger [188], les seuils $W_n^<$ et $W_n^>$ sont choisis proportionnels à l'estimation courante de la fonction de partition partiel Z_n, soit $W_n^< = c^< Z_n$ et $W_n^> = c^> Z_n$, avec $Z_n = (1/N_c)\sum_i W_n(i)$ (N_c est le nombre courant d'essai de construction initié), $c^> \sim 2$ et $c^>/c^< \approx 10$. Des choix plus sophistiqués existent pour étudier le repliement de protéines à faible température [189]. Cependant, à des températures très faibles, les clones évoluent souvent dans des directions similaires et les conformations issues du même départ de construction sont très fortement corrélées.

Méthode PERM avec échantillonnage préférentiel

Pour empêcher cette perte de diversité conformationnelle à très basses températures, Hsu et collaborateurs [190] introduisent une amélioration de la méthode PERM (qu'on notera nPERMis) où les clones ne sont plus exactement identiques. A chaque étape de construction, avant l'insertion du monomère n, on évalue le nombre k_l de sites libres où n peut être positionné, et on estime une prédiction W_n^{pred} du poids partiel de Rosenbluth définie par

$$W_n^{pred} = W_{n-1} \sum_{\alpha=1}^{k_l} q_\alpha \quad \text{avec } q_\alpha = \exp[-\beta E_n(\alpha)] \tag{4.71}$$

où W_{n-1} est la poids partiel de Rosenbluth de la conformation courante avant l'insertion du monomère n.

Le choix de la stratégie d'élimination a une faible incidence sur l'efficacité de l'algorithme, ainsi on conserve celle de la méthode PERM originale : si $W_n^{pred} < W_n^<$, on élimine la conformation avec une probabilité 1/2 et si elle survit on aura $W_n = 2 \times W_n^{pred}$.

Pour l'enrichissement ($W_n^{pred} > W_n^>$), si $k_l = 1$, on insère le monomère n sur le seul site libre et on continue la construction sans rajouter de clone avec $W_n = W_n^{pred}$. Si $k_l > 1$, on continue le processus de construction pour $k = \min\{k_l, \lceil W_n^{pred}/W_n^> \rceil\}$ conformations ($2 \leq k \leq k_l$) à partir de k sites différents choisis parmi les k_l sites libres. Le choix du k-uplet $A = \{\alpha_1, ..., \alpha_k\}$ des différentes directions de croissance est fait suivant la loi de probabilité

$$p(A) = \frac{k_l}{k\binom{k_l}{k}} \left(\frac{\sum_{\alpha \in A} q_\alpha}{\sum_{\alpha'=1}^{k_l} q_{\alpha'}} \right) \tag{4.72}$$

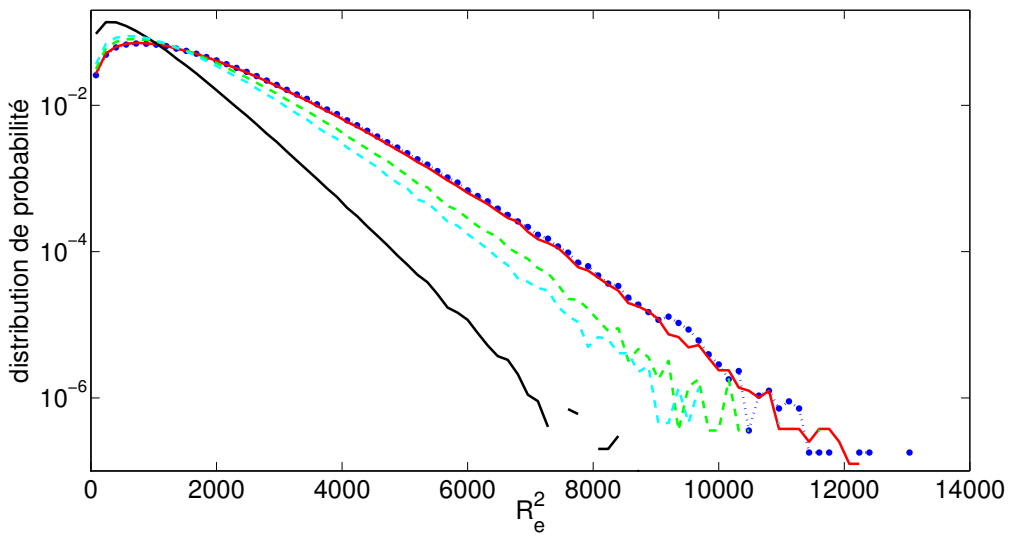

FIGURE 4.9 – Distributions de probabilité de la distance carré bout-à-bout R_e^2 pour les conformations auto-évitantes de taille $N = 512$ générées avec plusieurs méthodes : la dimérisation simple (points bleus), l'échantillonnage de Rosenbluth (ligne noire), les méthodes PERM (ligne rouge), la dimérisation avec biais (pointillé) avec $N_{trans} = 4$ (vert) et 8 (cyan). La dimérisation simple, seule méthode non biaisée, est considérée ici comme référence pour décrire la distribution de Boltzmann de chaînes auto-évitantes sans énergie.

et les poids de Rosenbluth partiels des différentes conformations sont donnés par

$$W_n(\alpha_i) = W_{n-1} \exp[-\beta E_n(\alpha_i)] \left(\frac{\sum_{\alpha'=1}^{k_l} q_{\alpha'}}{\sum_{\alpha \in A} q_\alpha} \right) \tag{4.73}$$

Ainsi cette stratégie force toutes les conformations issues d'un même départ de construction à être différentes. Concernant les seuils, ils sont définis par $W_n^> = C(Z_n/Z_1)(c_n/c_1)^2$ et $W_n^< = 0.2W_n^>$ avec Z_n l'estimation courante de la fonction de partition, c_n le nombre courant de conformations de taille n déjà construites et $C \leq 1$ une constante. L'efficacité optimale de l'algorithme (gain d'un facteur 2 par rapport au cas $C = 1$) est obtenue pour une gamme assez étroite de valeurs de C qui dépend du système étudié et qui nécessite une recherche extensive. C'est pourquoi usuellement, on fixe $C = 1$. Cependant, on peut noter que si C est très faible, l'algorithme fait essentiellement de l'énumération exacte pour les petites chaînes et devient stochastique pour des tailles plus grandes.

Notons également qu'un algorithme similaire utilisant l'échantillonnage simple a également été développé par Hsu et collaborateurs [190]. Ses performances sont légèrement moins bonnes que celles de la méthode nPERMis.

4.4.1.4 Dimérisation

Méthode simple

Pour échantillonner les chemins auto-évitants simples (sans énergie), il existe une méthode très efficace, la dimérisation [191], qui va nous permettre de comparer les résultats des algorithmes biaisés introduits précédemment avec ceux issus d'une méthode non-biaisée. L'idée d'Alexandrowicz est de remarquer qu'une conformation auto-évitante de taille N peut être en fait construite en collant bout-

à-bout deux conformations auto-évitantes de taille $N/2$. Plus précisément, soit $N_0 \in \mathbb{N}$, on définit la fonction récursive $dim(k)$ qui fournit une conformation de taille $2^k N_0$ par (voir figure 4.7 C) :

- si $k = 0$, $dim(k)$ construit une conformation par la méthode de l'échantillonnage simple ;
- sinon, on appelle deux fois $dim(k-1)$ pour avoir deux conformations de taille $2^{k-1}N_0$, on les colle bout-à-bout et on vérifie si la nouvelle conformation est auto-évitante, si ce n'est pas le cas on rappelle $dim(k)$ jusqu'à ce qu'une conformation soit validée.

Pour chaque $k \neq 0$, la probabilité $\mathcal{P}_k$ de générer une conformation auto-évitante à partir des deux demi-conformations vaut théoriquement

$$\mathcal{P}_k \equiv \frac{\mathcal{N}_{SAW}(2^k N_0)}{[\mathcal{N}_{SAW}(2^{k-1}N_0)]^2} = \frac{1}{f_s}\left(\frac{4}{2^k N_0}\right)^{c'} \tag{4.74}$$

Comme $c' \approx 1/6$, cette probabilité reste relativement grande pour des tailles de chaînes importantes. Par exemple, pour une taille de 1024 (avec $k = 6$ et $N_0 = 16$), $\mathcal{P}_6 \approx 0.35$. Pour $k = 0$, on effectue un échantillonnage simple, donc on a $\mathcal{P}_0 = \mathcal{N}_{SAW}(N_0)/(z(z-1)^{N_0-1})$.

Soit $N = 2^p N_0$ la taille maximale des conformations que l'on veut générer, la méthode consiste à appeler plusieurs fois $dim(p)$ et à compter pour chaque $0 \leq k \leq p$, le nombre total C_k de conformations de taille $2^k N_0$ réellement construites et le nombre total A_k d'appels à la fonction $dim(k)$ (c'est à dire, le nombre d'essais de construction), avec par définition de la fonction récursive dim la relation $C_{k-1} = 2A_k$. A partir de ces nombres, on évalue les probabilités $\{\mathcal{P}_k = C_k/A_k; k = 0, ..., p\}$. Or l'équation 4.74 est équivalente à la relation de récurrence $\mathcal{N}_{SAW}(2^k N_0) = \mathcal{P}_k[\mathcal{N}_{SAW}(2^{k-1}N_0)]^2$ sur les nombres de chemins auto-évitants qui a pour solution

$$\mathcal{N}_{SAW}(2^k N_0) = \left[z(z-1)^{N_0-1}\right]^{2^k} \prod_{i=0}^{k} (\mathcal{P}_i)^{2^{k-i}} \tag{4.75}$$

Cette dernière équation nous permet ainsi d'estimer le nombre de chemins auto-évitants à partir des résultats obtenus par la dimérisation simple.

De plus, pour un nombre final C_p de conformations auto-évitantes de taille N à construire, le nombre A_0 d'appel à la fonction $dim(k = 0)$ va être donné par

$$A_0 = \frac{C_0}{\mathcal{P}_0} = \frac{2A_1}{\mathcal{P}_0} = \frac{2C_1}{\mathcal{P}_1\mathcal{P}_0} = \frac{2^2A_2}{\mathcal{P}_1\mathcal{P}_0} = \frac{2^2C_2}{\mathcal{P}_2\mathcal{P}_1\mathcal{P}_0} = ... = \frac{2^pC_p}{\prod_{k=0}^{p}\mathcal{P}_k} \tag{4.76}$$

Ainsi l'attrition pour la dimérisation est donnée par

$$Att_{dim}(N) \equiv \frac{A_0}{C_p} = \frac{2^p}{\prod_{i=0}^{p}\mathcal{P}_i} \quad \text{avec } p = \log_2[N/N_0] \tag{4.77}$$

D'où, quand N_0 augmente, on a d'une part le nombre p d'étapes de dimérisation qui diminue et d'autre part les probabilités $\mathcal{P}_k$ (en particulier $\mathcal{P}_0$) qui diminuent aussi. Il va donc exister une valeur intermédiaire N_0^{opt} pour laquelle l'attrition sera minimale. La figure 4.8 montre l'évolution de $Att_{dim}(N)$ pour différentes valeurs de N_0. On remarque que $N_0^{opt} = 16$. Pour cette valeur, l'attrition est largement plus faible que celle de l'échantillonnage simple mais demeure tout de même beaucoup plus importante que celle de la méthode de Rosenbluth ou de PERM. Par exemple, pour $N = 1024$, l'attrition de la dimérisation vaut environ 1.7×10^4, contre 2.0×10^{40} pour celle de l'échantillonnage simple et 1 pour celle de l'échantillonnage de Rosenbluth.

Dimérisation avec biais

Pour essayer de diminuer encore cette attrition, on imagine une méthode de dimérisation biaisée par l'idée de l'échantillonnage de Rosenbluth : au lieu de coller bout-à-bout les deux demi-conformations, on génère plusieurs "collages" possibles et on en choisit aléatoirement un parmi ceux donnant une conformation auto-évitante. Plus précisément la fonction récursive $dim(k)$ devient (voir figure 4.7 D)

– si $k = 0$, $dim(k)$ construit une conformation par la méthode de l'échantillonnage simple, son poids de Rosenbluth est alors fixée à $W_0 = 1$;
– sinon, on appelle deux fois $dim(k-1)$ pour avoir deux demi-conformations $\mathcal{C}_A$ et $\mathcal{C}_B$, de poids respectifs $W_{k-1}(\mathcal{C}_A)$ et $W_{k-1}(\mathcal{C}_B)$. On génère N_{trans} conformations $\{\mathcal{C}_i\}$ en appliquant des mouvements pivots (voir section 4.3.2) aléatoires à $\mathcal{C}_B$, le premier élément de la chaîne étant pris comme pivot (toutes les conformations sont donc des isomères). Pour chaque $1 \leq i \leq N_{trans}$, on vérifie si la conformation $\mathcal{C}_A \cup \mathcal{C}_i$ est auto-évitante. Parmi les k_l conformations valides, on en choisit une aléatoirement. Son poids vaut alors

$$W_k = W_{k-1}(\mathcal{C}_A)W_{k-1}(\mathcal{C}_B)(k_l/N_{trans}) \tag{4.78}$$

Si $k_l = 0$, $W_k = 0$ et on rappelle $dim(k)$ jusqu'à ce qu'une conformation soit construite.
Afin de déterminer les $\{\mathcal{N}_{SAW}(2^k N_0)\}$, on aimerait pouvoir utiliser l'équation 4.75. Reste à déterminer comment estimer les $\{\mathcal{P}_k\}$. On peut montrer (voir annexe 6.10) que c'est possible en se servant de la formule

$$\mathcal{P}_k = \frac{\sum_{i=1}^{A_k} W_k(\mathcal{C}_i)}{\sum_{i=1}^{A_k} W_{k-1}(\mathcal{C}_{i,A})W_{k-1}(\mathcal{C}_{i,B})} \tag{4.79}$$

avec A_k le nombre d'appel à la fonction $dim(k)$ dans la dimérisation avec biais et $\mathcal{C}_{i,A}$ et $\mathcal{C}_{i,B}$ les deux demi-conformations utilisées pour construire $\mathcal{C}_i$.

La figure 4.8 montre la diminution significative de l'attrition pour différentes valeurs de N_{trans}. Bien évidemment, on remarque que plus N_{trans} est grand, plus l'attrition diminue. Cependant, la figure 4.9 montre également que la tendance à former des structures compactes s'accroît quand N_{trans} augmente conduisant, dans une moindre mesure, aux mêmes problèmes que l'échantillonage de Rosenbluth. Une version PERM de la dimérisation pourrait être envisagée cependant la structure récursive de la dimérisation rendrait complexe une telle approche.

4.4.1.5 Illustration pour les chemins auto-évitants sans énergie

Pour comparer quantitativement les méthodes décrites précédemment, on évalue l'évolution de l'erreur moyenne $\delta G(t)$ faite lors du calcul de l'énergie libre $G = -\log \mathcal{N}_{SAW}(N)$ en fonction du temps de simulation t. Le temps t est défini comme le nombre de fois où l'on a regardé dans la table de hachage, soit environ le nombre de fois où l'on a testé l'occupation d'un noeud du réseau. Pour calculer δG, pour chaque méthode, on effectue 10 fois la même simulation. Tous les 10^6 appels à la table de hachage, on évalue la valeur courante de l'énergie libre, puis on calcule l'écart quadratique moyen à la valeur limite G_∞ obtenue grâce à la dimérisation simple quand $t \to +\infty$, soit

$$\delta G = \left[\frac{1}{10}\sum_{i=1}^{10}(G(t) - G_\infty)^2\right]^{1/2} \tag{4.80}$$

La figure 4.10 montre que les méthodes les plus performantes sont la dimérisation simple et les méthodes PERM. L'évolution de l'erreur suit bien une loi en $t^{-1/2}$. A première vue, le fait que la dimérisation converge aussi vite que les méthodes de PERM n'est pas évident puisque son attrition est largement supérieure à celle de PERM. Cependant, il faut garder à l'esprit que toutes les conformations construites avec la dimérisation sont indépendantes les unes des autres alors que celles issues de PERM peuvent être corrélées. Ainsi, à t fixé, la dimérisation a certes généré moins de conformations que PERM, mais le nombre de mesures réellement indépendantes du système est à peu près le même pour les deux méthodes.

La figure 4.11 montre que pour le calcul de l'énergie libre (A) et la distance carrée bout-à-bout moyenne (B), les méthodes PERM et la dimérisation donnent des résultats très similaires. L'échantillonnage de Rosenbluth donne des résultats corrects jusqu'à environ $N \sim 200$. La dimérisation avec

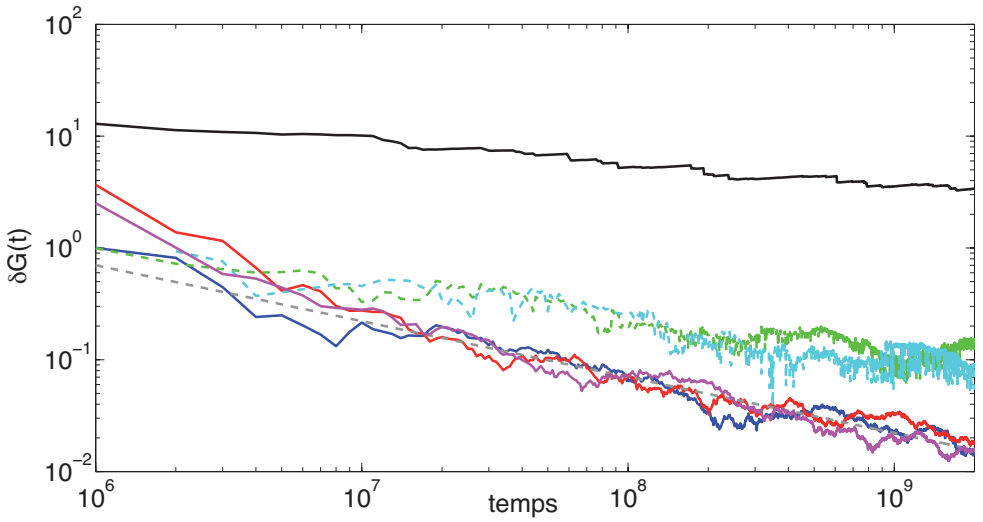

Figure 4.10 – Evolution de l'erreur moyenne δG en fonction du temps de simulation lors du calcul de l'énergie libre d'une chaîne auto-évitante sans énergie de taille $N = 1024$ pour la dimérisation simple (ligne bleue), l'échantillonnage de Rosenbluth (ligne noire), la méthode PERM (ligne rouge), la méthode nPERMis (ligne violette) et la dimérisation avec biais (pointillés) avec $N_{trans} = 4$ (vert) et 8 (cyan). La ligne pointillée grise représente la fonction $t^{-1/2}$.

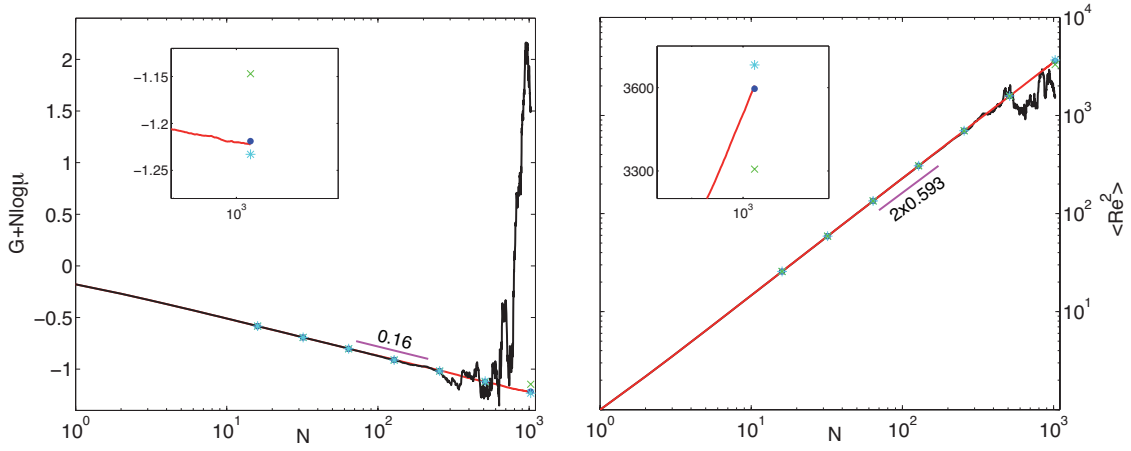

Figure 4.11 – Évolution de $G + N \log \mu$ (A) et de $\langle R_e^2 \rangle$ (B) en fonction de la taille N de la chaîne évaluée avec la dimérisation simple (points bleus), l'échantillonnage de Rosenbluth (ligne noire), les méthodes PERM (ligne rouge), la dimérisation avec biais avec $N_{trans} = 4$ (croix vertes) et 8 (étoiles cyans). Les encarts représentent des zooms autour de $N = 1000$.

biais quant à elle commence à exhiber des erreurs systématiques pour $N > 1000$.
L'énergie libre d'un chemin auto-évitant sans énergie est de la forme [81, 82]

$$G = -\log \mathcal{N}_{SAW}(N) = -\log f_s - N \log \mu - c' \log N \tag{4.81}$$

avec f_s et μ des constantes qui dépendent du réseau et c' une constante universelle. Avec les données obtenues par la méthode PERM, on trouve $f_s = 1.140 \pm 0.002$, $\mu = 10.037 \pm 0.001$ et $c' = 0.160 \pm 0.001$. La théorie de la renormalisation donne $c' = 1/6 = 0.16666$ [81]. De même, la distance carrée bout-à-bout moyenne suit la loi $\langle R_e^2 \rangle \propto N^{2\nu}$. On trouve $\nu = 0.593 \pm 0.001$. Janse van Resburg et collaborateurs trouvent également la même valeur pour une réseau CFC avec la méthode du pivot [183]. La théorie de la renormalisation donne $\nu = 0.588$ [81].

4.4.2 Échantillonnage d'une structure secondaire donnée

Dans cette partie, nous développons une méthode basée sur la méthode PERM avec échantillonnage préférentiel nous permettant de calculer, pour une structure secondaire $\mathcal{S}$ donnée, son énergie libre totale de conformation sur réseau

$$G_{conf}(\mathcal{S}) = -k_B T \log \left[\sum_{n_{tot}=0}^{\infty} \Omega(\mathcal{S}, n_{tot}) \exp(-\beta n_{tot} \bar{\kappa}/2) \right] \tag{4.82}$$

Ainsi, on pourra remonter à la différence d'énergie libre de la structure secondaire considérée grâce à (voir équation 4.6)

$$\Delta G = h_{latt}(\mathcal{S}) - T s_{latt}(\mathcal{S}) + G_{conf} + k_B T \log \Omega_0 \tag{4.83}$$

Pour pouvoir appliquer les méthodes de croissance de chaînes à la construction de conformations d'ARN sur réseau, nous définissons tout d'abord une représentation mathématique adéquate d'une structure secondaire, puis nous détaillons notre algorithme.

4.4.2.1 Graphe équivalent à une structure secondaire

Soit $\mathcal{S}$ une structure secondaire d'une séquence contenant N nucléotides, notons p le nombre de paires de bases dans $\mathcal{S}$. On représente $\mathcal{S}$ par un graphe $\mathcal{G}(\mathcal{S})$ où chaque nucléotide non-apparié et chaque paire de bases y est représentée par un noeud. Le graphe équivalent à $\mathcal{S}$ contient au total $n = N - p$ noeuds. On numérote les noeuds dans l'ordre croissant (dans le sens 5' vers 3') d'apparition du nucléotide ou de la paire de base dans la structure secondaire. On note $Id(i)$ l'identité dans $\mathcal{S}$ du noeud i de $\mathcal{G}$ et Id^{-1} la fonction réciproque qui donne pour un nucléotide, son numéro de noeud dans le graphe. Ainsi, $Id(i)$ est soit le numéro d'un nucléotide non-apparié, soit les deux numéros des nucléotides formant une paire de bases. La connectivité dans le graphe est la même que celle de la séquence originale, c'est à dire que deux noeuds i et j de $\mathcal{G}$ sont connectés si et seulement si il existe $k \in Id(i)$ et $k' \in Id(j)$ tels que $k = k' + 1$ ou $k = k' - 1$. Puisqu'on ne prend pas en compte les paires de bases isolées, un noeud du graphe a au maximum 3 voisins et au minimum 1.

Illustrons cette définition du graphe équivalent par l'exemple de la figure 4.12. La séquence contient 28 nucléotides et la structure secondaire considérée a 8 paires de bases, d'où $n = 20$ noeuds. La fonction Id correspondante vaut alors $Id(1) = (1, 12)$, $Id(2) = (2, 11)$, $Id(3) = (3, 10)$, $Id(4) = 4$, $Id(5) = 5$, $Id(6) = 6$, $Id(7) = (7, 22)$, $Id(8) = (8, 21)$, $Id(9) = 9$, $Id(10) = 13$, $Id(11) = 14$, $Id(12) = 15$, $Id(13) = (16, 27)$, $Id(14) = (17, 26)$, $Id(15) = (18, 25)$, $Id(16) = 19$, $Id(17) = 20$, $Id(18) = 23$, $Id(19) = 24$ et $Id(20) = 28$. Pour la connectivité, voir la figure 4.12 b.

Dans le graphe, un chemin $\mathcal{K}$ du noeud i au noeud j est défini par la donnée d'une série de noeuds $\mathcal{K} = \{k_0, k_1, ..., k_l\}$ allant de i à j, avec $k_0 \equiv i$, $k_l \equiv j$ et k_i est connecté à k_{i+1} $(0 \leq i \leq l-1)$ dans $\mathcal{G}$. Le nombre l, qu'on nommera la longueur du chemin $\mathcal{K}$, représente le nombre de connections entre i et j le long du chemin. Par exemple, $\{3, 4, 5, 6, 7, 18, 19, 15\}$ ou $\{3, 9, 8, 17, 16, 15\}$ sont deux chemins possibles entre les noeuds 3 et 15 du graphe de la figure 4.12 b, avec respectivement $l = 7$ ou 5.

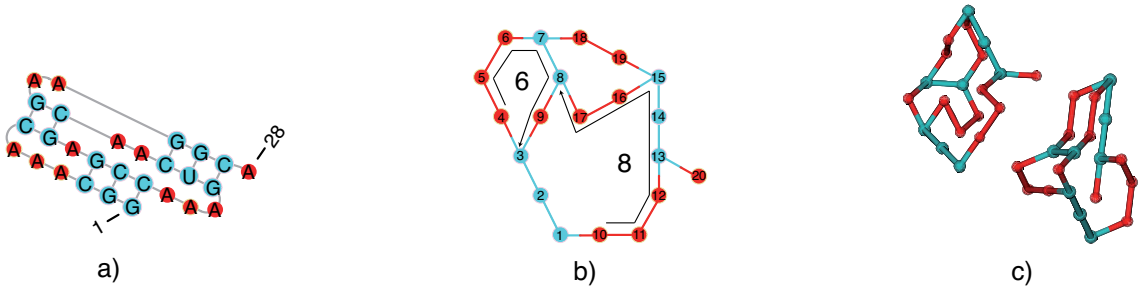

Figure 4.12 – Illustration de la construction du graphe équivalent à une structure secondaire avec pseudo-noeud. (a) La structure secondaire considérée avec en bleu les nucléotides appariés et en rouge ceux non-appariés. (b) Graphe équivalent à la structure secondaire définie à gauche. La contrainte maximale pour l'insertion du noeud 4 est le chemin $\{4, 5, 6, 7, 8, 9, 3\}$ de taille 6 aboutissant au noeud 3. Pour l'insertion du noeud 10, c'est le chemin $\{10, 11, 12, 13, 14, 15, 16, 17, 8\}$ de taille 8 aboutissant au noeud 8. (c) Exemples de conformations sur réseau construites à l'aide de la méthode de la contrainte maximale.

4.4.2.2 Méthode de la contrainte maximale

Comme une conformation $\mathcal{C}$ sur réseau ayant pour structure secondaire $\mathcal{S}$ représente également une conformation du graphe $\mathcal{G}(\mathcal{S})$, échantillonner $\{\mathcal{C}\}_{\mathcal{S}}$ est équivalent à échantillonner $\{\mathcal{C}\}_{\mathcal{G}(\mathcal{S})}$. L'idée de la méthode de la contrainte maximale décrite ici est d'utiliser la méthode PERM avec échantillonnage préférentiel pour générer des conformations sur réseau d'un graphe $\mathcal{G}$ tout en prenant soin de respecter la connectivité des noeuds de $\mathcal{G}$. Concernant ce dernier point, on va adapter la méthode vue précédemment pour échantillonner des chemins auto-évitants entre deux points fixes aux contraintes multiples imposées par le graphe.

Définition des contraintes

Considérons une séquence de taille N et une structure secondaire $\mathcal{S}$, soit $\mathcal{G}$ son graphe équivalent et n le nombre de noeuds dans $\mathcal{G}$. Supposons, lors de la construction d'une conformation, que l'on est déjà placé p noeuds sur le réseau, soit $\mathcal{E} = \{k_1, ..., k_p\}$ cet ensemble de noeuds. La connectivité du graphe impose que si il existe un chemin dans $\mathcal{G}$ entre deux noeuds déjà insérés et dont les noeuds intermédiaires n'ont pas encore été placés, l'insertion de ces derniers devra tenir compte des contraintes sur les positions déjà fixées des deux extrêmités du chemin. Ainsi, si la prochaine étape de l'algorithme est d'insérer le noeud i, qui est connecté au noeud $v_i \in \mathcal{E}$, on considère tous les chemins $\{a_0, a_1, ..., a_l\}$ entre v_i et les noeuds de $\mathcal{E}$ (y compris v_i), tels que $a_0 = v_i$ et $a_l \in \mathcal{E}$, avec $a_q \notin \mathcal{E}$ pour $1 \leq q \leq l-1$ et dont le premier noeud intermédiaire est i, soit $a_1 = i$. Parmi tous ces chemins, la plus forte contrainte sur la position de i est imposée par le chemin le plus petit, que l'on nomme contrainte maximale du noeud i. On note $l_m(i)$ la taille de ce chemin et $a_m(i)$ son extrêmité finale. Si aucun chemin ne vérifie les propriétés nécessaires, on pose par défaut $l_m(i) = +\infty$. Dans l'exemple de la figure 4.12, si $i = 4$ et $\mathcal{E} = \{1, 2, 3\}$, on a $v_i = 3$, $l_m(i) = 6$ et $a_m(i) = 3$; pour $i = 10$ et $\mathcal{E} = \{1, .., 9\}$, on a $v_i = 1$, $l_m(i) = 8$ et $a_m(i) = 8$.

En pratique, on construit une conformation du graphe toujours dans le même ordre (du noeud 1 au noeud N). Ainsi, pour l'insertion du noeud i, on aura $\mathcal{E} = \{1, ..., i-1\}$ et $v_i = Id^{-1}(\min\{Id(i)\} - 1)$. La détermination de $a_m(i)$ et de $l_m(i)$ se fait donc une fois pour toute avant le début de construction de la première conformation. L'analyse des chemins minimaux est réalisée à l'aide de l'algorithme de Dijkstra [192] qui, en théorie des graphes, est utilisé pour résoudre les problèmes de détermination du plus court chemin entre deux noeuds du graphe. L'algorithme consiste étape par étape à calculer la

distance minimale entre deux noeuds du graphe. On le modifie légèrement pour évaluer $a_m(i)$ et de $l_m(i)$. L'algorithme suit alors les étapes suivantes :

1. Pour $q \neq i$, on note $d_i(q)$ la distance courante entre q et i. On initialise $d_i(q) = +\infty$. On pose $d_i(i) = 0$.
2. Tous les noeuds de $\mathcal{G}$ sont marqués comme non-visités par l'algorithme et i est le noeud courant que l'on étudie en premier.
3. Soit j le noeud courant, pour tous ses voisins q non-visités, on met à jour leur distance à i par $d(q) = \min\{d(q)+1, d(q)\}$. On marque j comme visité. Si $j = i$, on n'applique pas la mise à jour précédente à v_i.
4. On est sure alors que la distance $d(j)$ est minimale. Le noeud courant devient alors le noeud non-visité ayant la distance courante à i la plus petite. Ceci nous assure de visiter les noeuds dans l'ordre croissant de leur distance à i.
5. Puis on recommence les étapes 3 et 4 jusqu'à ce que l'on est parcouru tous les noeuds que l'on pouvait visiter à partir de i (c'est à dire, qu'à la fin de l'étape 3, pour tous les noeuds q encore non-marqués, on ait $d(q) = +\infty$) ou que le numéro du nouveau noeud courant j soit inférieur à i (c'est à dire, qu'il sera inséré avant i lors de la construction d'une conformation de $\mathcal{G}$). Dans le premier cas, on a $l_m(i) = +\infty$ et dans le deuxième cas, $a_m(i) = j$ et $l_m(i) = d(j)$.

Algorithme

A l'aide de la contrainte maximale définie précédemment, la construction de conformations représentant le graphe $\mathcal{G}$ (et donc $\mathcal{S}$) se fait suivant l'algorithme :

1. On insére le premier noeud à une position initiale.
2. Pour l'insertion du i^{ieme} noeud ($2 \leq i \leq n$), considérons la contrainte maximale de i caractérisée par a_m et l_m. Pour une direction-test α de i, on évalue le nombre $c_\alpha = \mathcal{N}_{RW}(\vec{r}_\alpha, \vec{r}_{a_m})$ (voir annexe 6.8) de chemins idéaux de taille l_m entre la position $\vec{r}_\alpha$ de la direction-test et la position $\vec{r}_{a_m}$ de l'extrêmité finale de la contrainte maximale. Si $l_m = +\infty$, on pose par défaut $c_\alpha = 1$.
3. On applique alors la méthode PERM avec échantillonnage préférentiel avec $q_\alpha = c_\alpha \exp[-\beta E_n(\alpha)]$. E_n, ne tenant compte que du volume exclu et de la rigidité des parties double-brins (via $\bar{\kappa}$).
4. Les étapes 2 et 3 sont répétées jusqu'à ce que tous les noeuds du graphe aient été insérés. On vérifie alors que la conformation finale est valide, c'est à dire, que deux nucléotides consécutifs dans la séquence ont bien des positions voisines sur le réseau.

Pour l'estimation de c_α, son calcul à l'aide de la formule 6.92 (voir annexe 6.8) peut être assez long quand l_m devient grand et cela ralentit considérablement l'exécution pratique de la méthode de la contrainte maximale. Pour palier à ce problème, pour des petits chemins ($l_m \leq 10$), on utilise des données tabulées calculées avec l'équation 6.92. Pour des plus grands chemins, on approche c_α par la formule asymptotique décrivant la probabilité d'observer un chemin gaussien de taille de l_m entre $\vec{r}_\alpha$ et $\vec{r}_{i_m}$ [176] :

$$c_\alpha \approx z^{l_m} \left(\frac{3}{2\pi l_m}\right)^{3/2} \exp\left[-\frac{3}{2l_m b^2}(\vec{r}_\alpha - \vec{r}_{a_m})^2\right] \tag{4.84}$$

La méthode de la contrainte maximale permet donc de dynamiquement biaisé la construction d'une conformation en considérant à chaque étape la plus forte contrainte imposée par la connectivité du graphe. Elle fournit un ensemble de conformations, ainsi que leurs poids de Rosenbluth correspondants qui nous permettent d'estimer l'énergie libre de conformation de la structure secondaire $\mathcal{S}$ initiale par

$$G_{conf}(\mathcal{S}) = -k_B T \log\left[\frac{1}{N_{essai}} \sum_{i=1}^{N_{essai}} W(\mathcal{C}_i)\right] \tag{4.85}$$

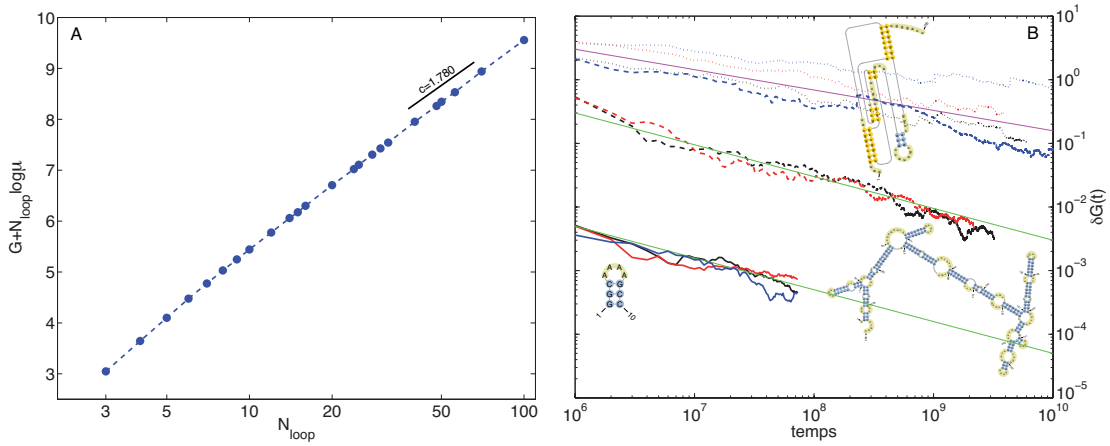

Figure 4.13 – (A) Évolution de l'énergie libre d'une boucle auto-évitante en fonction du nombre N_{loop} de segments qui la compose. (B) Évolution de l'erreur moyenne δG en fonction du temps de simulation pour la méthode de la contrainte maximale utilisée avec la méthode PERM avec échantillonnage préférentiel (noir), l'échantillonnage de Rosenbluth (bleu) ou la méthode PERM classique (rouge) ; et calculée pour une petite épingle à cheveux ($N \sim 10$) (lignes pleines), pour une structure faiblement contrainte ($N \sim 300$) (lignes tiretées) et pour une structure fortement contrainte ($N \sim 80$) (lignes pointillés). Pour chaque exemple étudié, on a tracé la structure secondaire correspondante.

avec N_{essai} le nombre de fois où l'on a commencé un essai de construction de conformation. G_{conf} pourra alors être utilisée pour évaluer la différence d'énergie libre de la structure secondaire considérée avec l'équation 4.83.

On a développé la méthode de la contrainte maximale en utilisant la méthode PERM avec échantillonnage préférentiel, mais on peut également se servir de l'échantillonnage de Rosenbluth ou de la méthode PERM classique. De plus, cette méthode, bien qu'appliquée à la génération de conformations d'ARN, pourrait facilement se généraliser à des potentiels d'interaction plus complexes.

4.4.2.3 Validation et efficacité

On teste la validité de l'algorithme en calculant l'énergie libre d'une boucle auto-évitante, soit $G = -\log \mathcal{N}_{SAP}(N)$ où $\mathcal{N}_{SAP}(N)$ est le nombre de polygones auto-évitants de taille N. En effet, une boucle contenant N segments peut être vue comme une structure secondaire particulière d'un séquence quelconque composée de $N+2$ nucléotides par laquelle seules les deux extrêmités sont appariées. La figure 4.13 A montre que G se met bien sous la forme asymptotique [81, 82]

$$G = -\log \mathcal{N}_{SAP}(N) = -\log f_l - N \log \mu + c \log N \tag{4.86}$$

où $f_l = 0.251 \pm 0.001$ et $\mu = 10.037 \pm 0.001$ sont des constantes qui dépendent du réseau, et $c = 1.780 \pm 0.005 = 3\nu = 3 \times 0.593$ une constante universelle. On retrouve ici les mêmes valeurs de μ et ν que celles obtenues dans l'étude des chemins auto-évitants avec la dimérisation et les méthodes PERM (voir section 4.4.1.5).

Pour justifier l'utilisation de la méthode PERM avec échantillonnage préférentiel dans la méthode de la contrainte maximale plûtot que l'échantillonnage de Rosenbluth ou la méthode PERM classique, on étudie l'évolution de l'erreur moyenne $\delta G(t)$ faite lors du calcul de l'énergie libre en fonction du temps de simulation t (voir équation 4.80) pour plusieurs structures secondaires. Sur la figure 4.13 B, on observe que pour des petites structures secondaires simples, les 3 méthodes convergent plus ou moins à la même

vitesse. Pour des structures secondaires plus grandes mais faiblement contraintes, les deux méthodes PERM sont équivalentes et sont beaucoup plus performantes que l'échantillonnage de Rosenbluth (plus d'un ordre de grandeur). L'utilité de se servir de PERM avec échantillonnage préférentiel ne se ressent que pour des structures fortement contraintes (avec pseudo-noeuds) où l'emploi de la méthode PERM classique entraîne une perte de diversité conformationnelle.

Pratiquement, la méthode de la contrainte maximale nous permet de calculer l'énergie libre d'un structure secondaire quelconque pour des tailles de chaînes allant jusqu'à 5000 nucléotides. Typiquement, on débute $N_{essai} = 10^6$ fois la construction d'un conformation. Pour une séquence de taille 300 et une structure secondaire faiblement contrainte (comme celle étudiée dans la figure 4.13 B), l'algorithme tourne pendant environ 8 minutes sur un ordinateur 2.4 GHz, environ 6×10^4 essais aboutissent à la construction d'une ou plusieurs conformations, générant au total près de 1.2×10^6 conformations. L'erreur statistique finale commise sur l'évaluation de G_{conf} est alors de l'ordre de $0.003 k_B T$, soit une erreur relative d'environ 10^{-5}. A N_{essai} fixé et toujours pour des structures secondaires faiblement contraintes, cette erreur statistique augmente avec la taille de la séquence : elle est de l'ordre de $0.2 k_B T$ (erreur relative $\sim 10^{-4}$) pour $N \sim 1000$ et de l'ordre de $0.5 k_B T$ (erreur relative $\sim 10^{-4}$) pour $N \sim 4000$. Pour des structures fortement contraintes comme celle étudiée dans la figure 4.13 B pour laquelle $N \sim 80$, l'erreur ($\sim 0.1 k_B T$ soit une erreur relative de $\sim 10^{-3}$) est environ une centaine de fois plus grande que celle qu'on aurait pour une séquence de taille équivalente mais sans pseudo-noeud.

Chapitre 5

Résultats

5.1 Validation du modèle pour des structures simples

Les résultats du modèle sur réseau obtenus dans cette section ont été obtenus à l'aide de la méthode multi-histogramme (voir section 4.3.3).

5.1.1 Dénaturation de petites structures en épingle

Les structures en épingle font partie des structures secondaires les plus simples chez les acides nucléiques. Elles sont composées d'un double-brin fermé par une boucle (voir figure 3.1 c). Elles interviennent dans de nombreux contextes biologiques comme la régulation de la transcription et de la réplication, la protection et la stabilisation in vivo des acides nucléiques, la facilitation de la mutagenèse ou encore l'initiation de contacts tertiaires dans les ribozymes [193, 194, 195, 196]. Elles ont donc été activement étudiées thermodynamiquement [197, 198] et mécaniquement [25, 199]. Tester notre modèle sur réseau sur ses structures simples est une étape importante dans la validation de notre approche.

Lors de leur dénaturation, les petites structures en épingle présentent une transition à deux états entre le complexe apparié et le simple brin dénaturé [162]. La figure 5.1 A compare les températures de fusion expérimentales pour des structures en épingle à celles calculées à l'aide du modèle sur réseau ou d'autres programmes qui utilisent le modèle de Turner (RNAfold [86], DINAmelt [88], Kinefold [140]). Pour estimer l'efficacité de la prédiction, on évalue la déviation standard par rapport aux expériences via

$$\sigma_{T_m} = \left\{ \langle [T_m(sim) - T_m(exp)]^2 \rangle \right\}^{1/2} \tag{5.1}$$

On trouve $\sigma_{T_m} = 5.4$ K pour le modèle sur réseau, 5.5 K pour RNAfold et 9.3 K pour Kinefold. On retrouve bien que nos prédictions ont une précision équivalente à celles du modèle de Turner complet (RNAfold), rien d'étonnant puisque notre paramétrisation est calquée sur ce modèle, en particulier pour les structures simples. La figure 5.1 B, qui montre la courbe de dénaturation de deux épingles, souligne l'excellent accord entre les résultats expérimentaux et nos prédictions. Pour justifier notre approche pour la paramétrisation du modèle sur réseau (section 3.3.3), nous avons également tracer les résultats pour une paramétrisation naïve qui consisterait à simplement identifier les paramètres sur réseau et ceux de Turner sans inclure les corrections entropiques dues à l'entropie de conformation des structures secondaires sur le réseau. Une telle paramétrisation ne permet pas de prédire la transition à deux-état usuellement observée pour les petites épingles et ses prédictions dévient de plus de 45 K par rapport aux expériences ($\sigma_{T_m} = 46.7$ K). De plus, cet écart va varier avec le réseau utilisé.

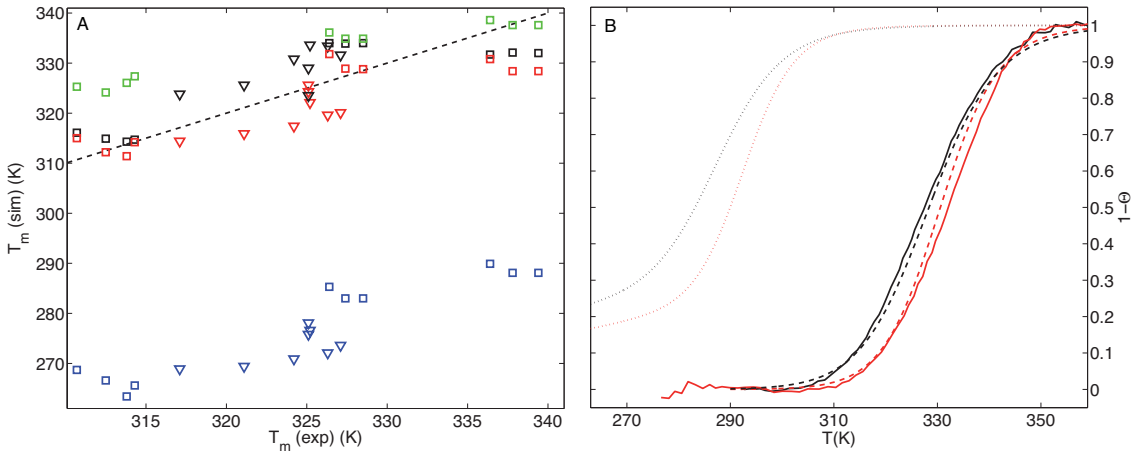

Figure 5.1 – (A) Températures de fusion calculées $T_m(sim)$ avec le modèle sur réseau (paramétrisation naïve en bleu et correcte en rouge) ou avec des programmes standard : RNAfold [86] ou DINAmelt [88] (noir) et Kinefold [140] (vert), en fonction des températures expérimentales $T_m(exp)$ pour 10 épingles ARN [198] (carré) et 8 épingles ADN [197] (triangles). (B) Probabilités $1 - \Theta$ que l'épingle soit ouverte, observée expérimentalement (lignes pleines) ou prédites par le modèle sur réseau (paramétrisation naïve en pointillé et correcte en tirets) pour les séquences $GGCAUUAGCC$ (rouge) et $GGGAUAUACCC$ (noir). Les données expérimentales ont été extraites de Ref.[198] et normalisées avec la procédure décrite dans la section 6.2 et dans Ref.[17].

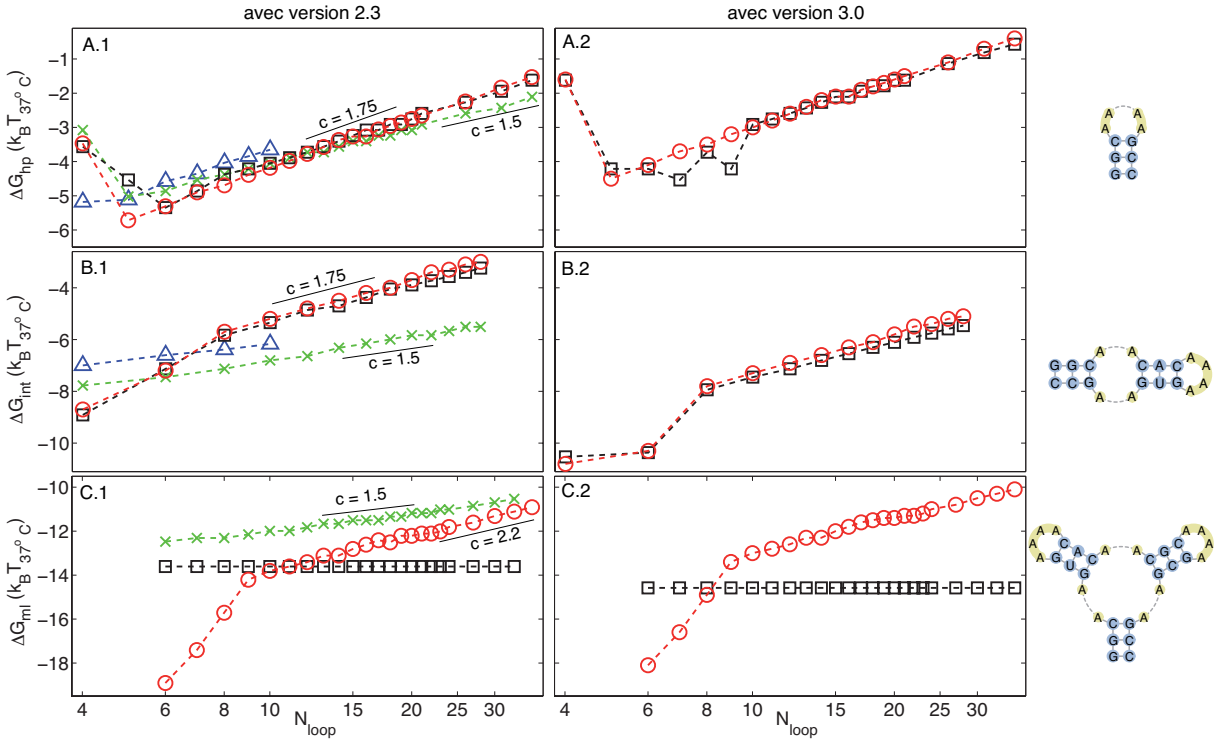

Figure 5.2 – Énergie libre ΔG en fonction de la taille de la boucle N_{loop} pour trois sortes de boucles (A : épingle , B : interne et C : multiple) et pour deux versions des paramètres de Turner, calculée avec le modèle sur réseau (cercles rouges), RNAfold [86] (carrés noirs), Kinefold [140] (croix vertes) et Vfold [132] (triangles bleus). Les données pour RNAfold et Kinefold ont été obtenues via leur application web. Les données pour Vfold ont été calculées en ajoutant aux termes d'empilement (association et fourche) l'énergie de conformation extraite de la figure 3B de [132].

5.1.2 Énergies libres de boucles internes et en épingle

Pour valider notre approche, nous comparons également les énergies libres calculées avec le modèle sur réseau et avec le modèle de Turner pour des structures simples telles que les structures avec une boucle interne ou en épingle.

Les figures 5.2 A et B montrent l'évolution de la différence d'énergie libre ΔG entre la structure native et l'état dénaturé à $T = 310$ K pour des structures avec boucles en épingle (A) ou internes (B), en fonction de la taille de la boucle. ΔG est calculée avec le modèle sur réseau, RNAfold, Kinefold et Vfold. On remarque que le modèle sur réseau et le modèle de Turner complet (via RNAfold) donnent des résultats similaires quelque soit la version des paramètres utilisés (2.3 ou 3.0). Ceci confirme notre dérivation des paramètres de fourches et de nucléation à partir des paramètres correspondants dans le modèle de Turner. Kinefold et Vfold, qui continuent à utiliser la vieille version 2.3 [167] des paramètres de Turner, donnent également des estimations raisonnables de ΔG. Cependant, notons qu'il faudrait être vigilant si ces deux modèles voulaient passer à la dernière version 3.0. En effet, dans les deux modèles, la différence d'énergie libre est donnée par $\Delta G = \Delta G_{stack} - T\Delta s_{loop}$ (voir annexe 6.6) où l'estimation de l'entropie de conformation Δs_{loop} (qui inclue la nucléation) est calculée indépendamment des paramètres de Turner et de sa jauge, et ΔG_{stack} contient des paramètres d'empilement (association, fourche, dangle) qui dépendent de la jauge. Or, dans le modèle de Turner non unifié, $\Delta s_{loop} = \Delta s^{\mathcal{T}}_{loop} - k_B c \log N_{loop}$ avec, dans la version 2.3 (3.0), $\Delta s^{\mathcal{T}}_{loop} = 6k_B$ ($0.1k_B$) pour les boucles internes et $4k_B$ ($6.5k_B$) pour les boucles en épingle. La substitution de Δs_{loop} par une contribution invariante (c'est à dire, ne dépendant pas du modèle de Turner sous-jacent et donc de sa jauge) pourrait aboutir à des déviations importantes par rapport aux prédictions de RNAfold. Il est toutefois "étonnant" que sans tenir compte d'aucune considération sur la jauge, Kinefold et Vfold donnent de si bons résultats avec la version 2.3.

Une inspection plus attentive des figures 5.2 A et B nous révèle que le modèle sur réseau supporte clairement l'idée que les grandes boucles internes et en épingle isolées peuvent être décrite par une relation de Jacobson-Stockmayer (comme supposé par le modèle de Turner) avec $c = 1.75$ [77]. Les déviations observées pour les épingles $N_{loop} = 7$ et 9 sont dûes à des corrections du modèle de Turner pour les petites boucles qui ne sont pas prises en compte dans le modèle sur réseau.

5.2 Prédiction de structures complexes

Dans cette partie, nous testons le pouvoir de prédiction du modèle sur réseau sur des structures plus complexes. À partir d'une séquence donnée et à l'aide de la méthode multi-histogramme (section 4.3.3), le modèle sur réseau est capable de prédire les propriétés thermodynamiques et la structure la plus stable pour une température quelconque. En particulier, nous comparons nos résultats avec des données expérimentales sur les structures natives.

5.2.1 Structures avec boucles multiples

Les boucles multiples (c'est à dire connectées à plus de 3 parties en double-brin) sont présentes dans beaucoup de molécules d'acides nucléiques comme les ARNs de transfert [4], et particulièrement dans les structures natives de longues séquences où la compléxité des structures secondaires impliquent la présence d'un grand nombre de sous-structures interconnectées.

5.2.1.1 Énergie libre

Pour les boucles multiples (figure 5.2 C), RNAfold suppose typiquement que ΔG est indépendant de la taille de la boucle. Ceci n'est pas confirmé par le modèle sur réseau qui prédit que ΔG peut varier par $3k_BT$ entre $N_{loop} = 9$ et $N_{loop} = 30$ avec même des déviations plus importantes pour les petites boucles où l'empilement coaxial entre deux parties double-brins voisines stabilise encore plus le

Table 5.1 – Pouvoir de prédiction (sensibilité SE/specificité SP) du modèle sur réseau et d'autres méthodes standard testée sur deux ARN de transfert à $T = 37^o$ C.

Séquence	Modèle sur réseau	RNAfold [86]	Kinefold [140]	pknotsRG [200]	Nupack [201]
ARNt-phe1 de la levure	1/1	0.95/1	0.9/0.83	0.24/0.24	0.95/0.87
ARNt-ala1 de l'homme	1/0.95	1/1	1/0.84	0.95/0.95	0.95/0.8

système. On peut remarquer également qu'avec la version 2.3 des paramètres de Turner, les résultats de Kinefold sont assez proches de ceux du modèle sur réseau. Aucune comparaison avec Vfold n'est possible puisqu'il n'existe pas (encore) de relation Jacobson-Stockmayer paramétrée par Vfold pour les boucles multiples.

En ce qui concerne la dépendance en la taille de la boucle, on observe une exposant effectif $c \sim 2.2$ qui est en accord avec les prédictions théoriques sur des chaînes avec des boucles interagissant stériquement [78, 165]. Pour Kinefold, dans les 3 cas étudiés, on trouve sans surprise que $c \sim 1.5$ qui est l'exposant pour des boucles aléatoires gaussiennes.

Notons que le modèle sur réseau (comme Kinefold) n'utilise pas de paramètre spécial pour ce type complexe de boucle (comme dans le modèle de Turner). Il donne une prédiction sur ΔG à partir des paramètres unifiés de boucles (fourches et nucléation) qui ont été paramétrés à l'aide de boucles simples.

5.2.1.2 Pouvoir de prédiction

On teste la faculté du modèle sur réseau à prédire des structures avec boucles multiples sur deux séquences d'ARN de transfert. Pour cela, on calcule la sensibilité SE et la spécificité SP de la structure prédite la plus stable que l'on compare à la structure native expérimentale (voir Table 5.1). SE est définie comme le quotient entre le nombre de paires de bases correctement prédites et le nombre total de paires de bases dans la structure native. SP quant à elle est définie comme le quotient entre le nombre de paires de bases correctement prédites et le nombre total de paires de bases prédites. On voit que le modèle sur réseau prédit de manière quantitative la structure native de ces deux séquences. Ses performances sont à peu près équivalentes à celles de RNAfold et sont meilleures que celles des autres méthodes étudiées.

5.2.1.3 Chemin de repliement à l'équilibre d'un ARN de transfert

Avec le modèle sur réseau, nous ne sommes pas limités à prédire des structures natives mais on a accès à la thermodynamique du repliement. On étudie ainsi le chemin de repliement à l'équilibre de l'ARNt-phe1 de la levure. Ce chemin représente l'ensemble des étapes à l'équilibre thermodynamique pour passer de l'état dénaturé à haute température à l'état natif à température ambiante. Par comparaison, le chemin de repliement cinétique (voir section 5.5.2.2) est l'ensemble des étapes hors-équilibre pour passer de l'état dénaturé à l'état natif à une température donnée. La figure 5.3 illustre ce chemin de repliement à l'équilibre en montrant la carte des contacts (probabilité pour deux nucléotides d'être appariés) ainsi que la structure la plus stable à différentes températures. Autour de 37° C (310 K), les quatre parties double-brins présentes dans la structure native expérimentale sont stables (probabilité de contact ≥ 0.5) et forment également la structure prédite la plus stable. Au fur et à mesure que la température augmente, on observe le dépliement de la molécule jusqu'à sa dénaturation totale. Ce processus peut être décomposé en 4 étapes caractérisées par l'ouverture successive des double-brins natifs :

1. Autour de 57° C (330 K), le double-brin I s'ouvre en premier. A première vue, ceci pourrait paraître surprenant puisque c'est le double-brin le plus long (7 bp) donc a priori le plus stable

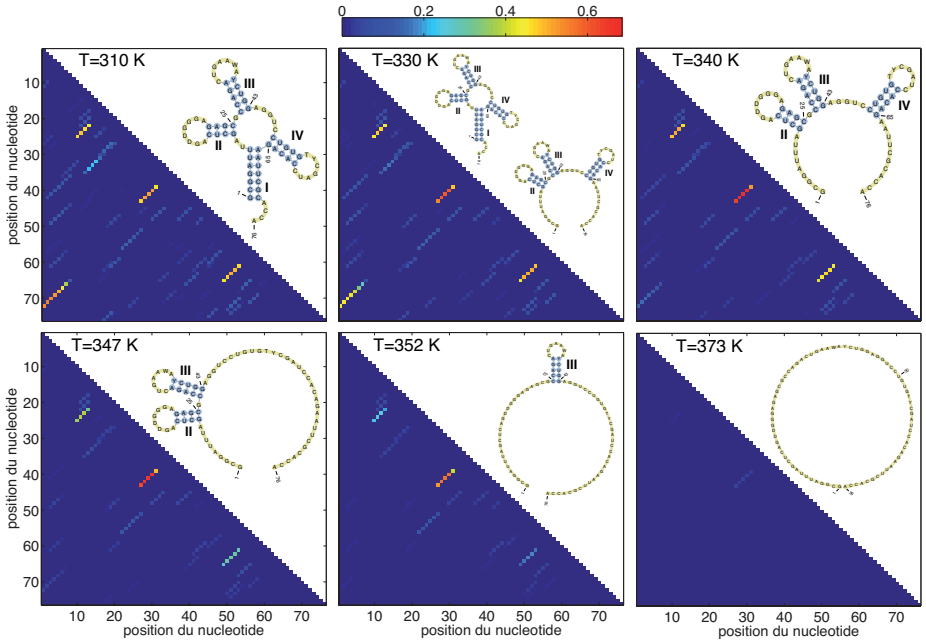

FIGURE 5.3 – Illustration du chemin de repliement à l'équilibre pour l'ARNt-phe1 de la levure. Pour différentes températures, on trace la carte des contacts et la structure la plus stable. La carte des contacts représente la probabilité pour deux nucléotides d'être appariés. La légende des couleurs est donnée en haut de la figure.

et, en plus, il est empilé coaxialement avec le double-brin IV. Mais en fait, si on analyse les contributions entropiques, on remarque que parallèlement à l'ouverture de I, la boucle multiple centrale s'ouvre également et ceci permet à la structure de se relaxer d'une importante pénalité entropique. Si au contraire, c'était un des autres double-brins qui s'était ouvert, cela aurait eu pour effet d'augmenter la taille de la boucle centrale et donc d'accroître sa pénalité entropique (voir équation 5.12).

2. Autour de 72° C (345 K), le double-brin IV se dénature, les double-brins II et III étant toujours stabilisés par empilement coaxial.
3. Autour de 76° C (349 K), le double-brin II s'ouvre alors que les nucléotides du double-brin III ont toujours une forte probabilité de contact (~ 0.5).
4. Autour de 82° C (355 K), l'état dénaturé devient la structure majoritaire.

Ainsi, lors du repliement à l'équilibre de l'ARNt, le bras anti-codon (double-brin III) est le premier formé. Puis viennent le bras T (II), le bras D (IV) et finalement le bras accepteur (I). Malheureusement aucune expérience existante ne permet de comparer nos prédictions avec le véritable chemin à l'équilibre.

5.2.2 Structures avec pseudo-noeuds

Les pseudo-noeuds sont des molécules plus complexes. Ils sont présents dans les noyaux catalytiques des ribozymes, dans les introns et les télomérases auto-épissants, ou encore dans l'initiation du décalage ribosomal du cadre ouvert de lecture [202, 203, 204, 205, 206]. La présence de paires de bases qui violent la convention gigogne (pour deux paires de bases quelconques i, j et k, l, on impose $i < k < l < j$

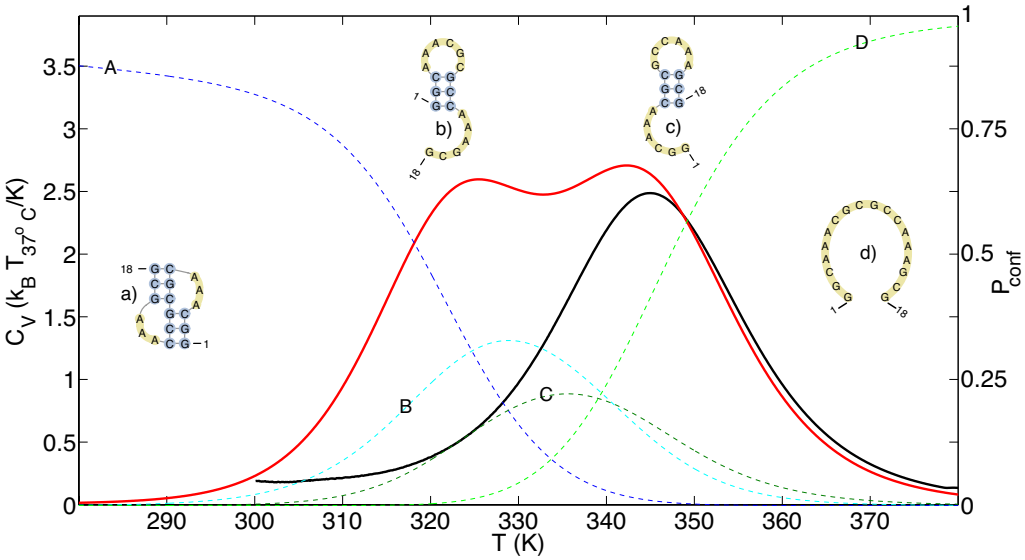

Figure 5.4 – Evolution de la capacité calorifique C_V (lignes pleines) pour la séquence H-pseudo-noeud $GGCAAACGCGCCAAAGCG$ calculée avec le modèle sur réseau (rouge) ou avec le programme RNAheat [86] (noir). Les lignes pointillées représentent la probabilité $\exp(-\beta\Delta G)/Z$ d'observer les structures secondaires majoritaires : H-pseudo-noeud A (bleu), épingle B (cyan), épingle C (vert foncé) et pelote aléatoire D (vert clair).

ou $i < j < k < l$ [20]), rend plus difficile la prédiction des structures secondaires par les méthodes standard puisqu'elle accroît fortement l'espace des configurations à considérer.

5.2.2.1 Transition à plusieurs états

Un exemple classique de pseudo-noeud est le pseudo-noeud de type H (voir figure 5.11). Il est composé de deux double-brins $S1$ et $S2$ (comprenant chacun n_1 et n_2 paires de bases), de deux boucles $L1$ et $L2$ (comprenant chacune l_1 et l_2 nucléotides) et d'un lieur L_3 (comprenant l_3 nucléotides). Les molécules d'ARN qui forment de telles structures présentent souvent une dénaturation thermique à plusieurs états [208] entre le H-pseudo-noeud, les deux épingles correspondantes et l'état dénaturé. La figure 5.4 illustre ce comportement pour une séquence ayant comme structure secondaire stable à basse température un H-pseudo-noeud. L'évolution de la capacité calorifique présente deux pics relatifs à la transition H-pseudo-noeud/épingles (autour de $T = 324$ K) et à la transition épingles/pelote (autour de $T = 340$ K). D'ailleurs, on retrouve uniquement cette dernière transition dans les prédictions du modèle de Turner (via RNAheat [86]) qui ne tient pas compte des pseudo-noeuds. Les comportements des probabilités d'observer les structures secondaires majoritaires confirment cette dénaturation à plusieurs étapes et nous renseigne, par exemple, sur quelle épingle est majoritaire (en l'occurrence l'épingle B) dans l'étape intermédiaire entre 324 et 340 K.

5.2.2.2 Énergie libre

La figure 5.5 montre la dépendance en la taille des boucles de la différence d'énergie libre ΔG_{pk} entre la structure H-pseudo-noeud et l'état dénaturé, pour $n_1 = n_2 = 3$, $l_1 = l_2$ et $l_3 = 0$. Dans la même figure, nous avons également inclu les prédictions issues d'autres modèles. Pour apprécier le bon accord général des résultats, rappelons brièvement comment les différentes méthodes étudiées ont déterminé leurs paramètres décrivant les pseudo-noeuds (voir aussi annexe 6.6). Le modèle de Gultyaev [207],

pknotsRG [200], Nupack [201] et Vfold [151] utilisent des relations de Jacobson-Stockmayer généralisées spécialement paramétrées pour les H-pseudo-noeuds. Alors que les trois premières citées ont ajusté leurs paramètres de manière à correctement prédire des pseudo-noeuds connus expérimentalement ou phylogénétiquement [1], Vfold prédit ces mêmes paramètres à l'aide d'un modèle microscopique sur réseau [151]. Kinefold [138, 139, 140] quant à lui fait des prédictions (tout comme le modèle sur réseau) sur l'entropie de conformation en modélisant les double-brins comme des tiges rigides et les boucles comme des chaînes gaussiennes, négligeant par ailleurs les interactions de volume exclu. Pour des boucles de taille $l \sim 10$, toutes les méthodes donnent environ des prédictions identiques. Pour des boucles plus petites, les prédictions du modèle sur réseau concordent avec celles de Nupack, Vfold et Kinefold et présentent une grande différence avec pknotsRG et le modèle de Gultyaev. Pour des boucles intermédiaires ($l \sim 10 - 20$), le modèle sur réseau prédit, en accord avec le modèle de Gultyaev, que le comportement asymptotique suit une équation de Jacobson-Stockmayer avec un exposant $c \approx 1.8$, très proche de l'exposant des polygones auto-évitants. Cela semble indiquer qu'il n'y a pas d'interaction stérique significative entre $L1$ et $L2$. Comme prévu, Kinefold prédit une dépendance plus faible en la taille de la boucle avec $c = 1.5$. Vfold, pknotsRG et Nupack prédisent quant à eux des énergies libres qui croîssent linéairement avec l.

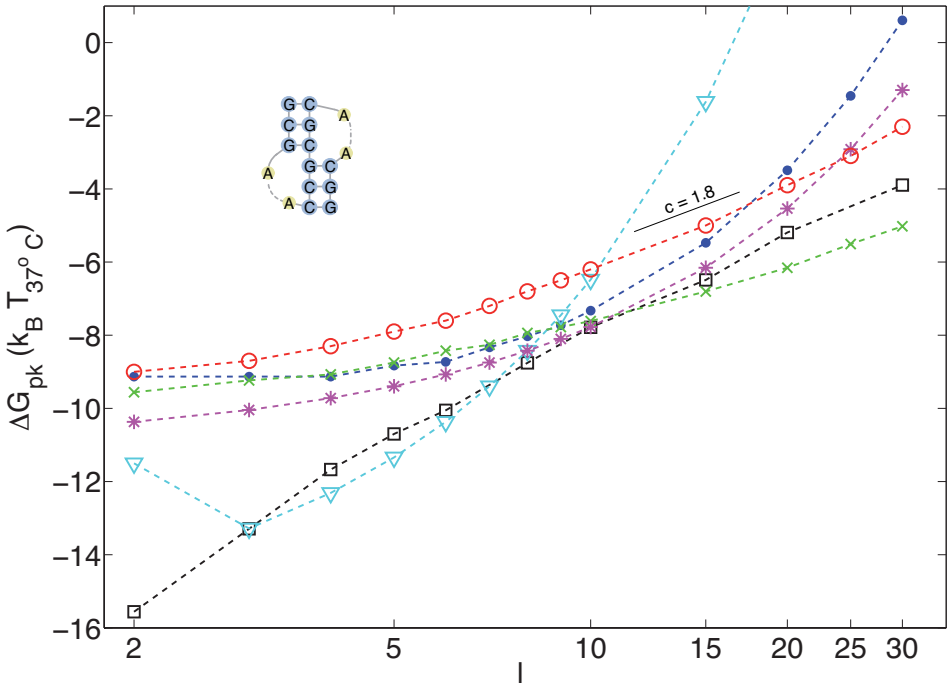

FIGURE 5.5 – Énergie libre ΔG_{pk} en fonction de la taille l des boucles du H-pseudo-noeud $GGCA_lUCGGCCA_lCGA$ calculée avec différents modèles : le modèle sur réseau (ronds rouges), le modèle de Gultyaev [207] (carrés noirs), Vfold [151] (points bleus), pknotsRG [200] (triangles cyans), Kinefold [140] (croix vertes) et Nupack [201] (étoiles mauves). L'erreur typique sur la détermination de ΔG_{pk} est de $2k_BT$ et est due principalement à l'incertitude sur les paramètres. Les données pour pknotsRG, Kinefold et Nupack ont été obtenues à l'aide de leur interface web. Celles pour le modèle de Gultyaev et Vfold ont été calculées suivant les règles décrites respectivement dans Ref.[207] et Ref.[151].

1. La détermination phylogénétique de structures secondaires consiste à supposer que des molécules ayant la même fonction dans des organismes proches au sens de l'évolution vont avoir des structures similaires. Si par un moyen expérimental quelconque (typiquement par cristallographie ou par RMN) on a accès à la structure d'une molécule, alors, on aura une prédiction pour la structure de toutes ses molécules équivalentes au sens phylogénétique. Plus la proximité évolutive sera importante, plus la prédiction sera valable.

TABLE 5.2 – Pouvoir de prédiction (sensibilité SE/spécificité SP) du modèle sur réseau (MR) et d'autres méthodes, testé sur des séquences tronquées de signaux viraux de décalage du cadre ouvert de lecture [210]. Pour le modèle sur réseau, on donne la différence d'énergie libre ΔG entre la structure prédite la plus stable et le H-pseudo-noeud natif décrit par $n_1/l_1/n_2/l_2$ (* : présence d'un nucléotide non-apparié entre les deux double-brins, soit $l_3 = 1$; pour les autres cas on a $l_3 = 0$)).

Abréviations	T (^{o}C)	Séquences tronquées	$n_1/l_1/n_2/l_2$	SE/SP (MR)	ΔG ($k_B T_{37^o}$)	RNAfold	Kinefold	Vfold	pknotsRG	Nupack
BChV	25	$G1595 - C1620$	$4/1/4/8^*$	1/1	0	0/0	0.5/ 0.57	1/1	1/1	1/1
BLV	37	$G1604 - U1630$	6/5/3/4	0.67/1	5.5	0.67/1	0.67/1	0.67/0.86	1/1	0.67/1
BWYV	25	$C1566 - G1591$	5/2/4/6	1/1	0	0.56/1	1/1	1/1	1/1	1/1
BYDV-NY-RPV	25	$G1706 - C1732$	5/2/4/7	0.33/0.33	1.9	0/0	0/0	1/1	0/0	1/1
CABYV	25	$G1494 - C1520$	$5/2/3/8^*$	0.38/0.38	5.5	0/0	0/0	0/0	1/1	1/1
EIAV	37	$G1797 - C1831$	6/3/4/12	1/1	0	0.5/0.71	1/1	1/1	1/1	0.9/1
FIV	37	$G1893 - C1927$	5/2/6/11	0.82/1	3.6	0.45/1	1/1	1/1	1/1	1/1
MMTVgag/pro	37	$G2090 - U2123$	$5/1/8/8^*$	0/0	3.8	0/0	0.92/1	1/1	1/1	0.42/0.5
PEMV	25	$U2042 - C2069$	6/2/4/6	0.9/1	0.3	0.6/1	0.9/1	0.9/1	1/1	0.9/1
PLRV-S	25	$G1781 - G1806$	4/2/4/8	1/1	0	0.5/1	0.5/0.57	1/1	1/1	1/1
PLRV-W	25	$G1676 - G1701$	$4/2/3/9^*$	0/0	2.2	0.5/1	0/0	1/0.88	1/1	1/1
SRV1gag/pro	37	$G2337 - C2373$	6/1/6/12	0.83/1	6.8	0/0	1/1	1/1	1/1	1/1

5.2.2.3 Pouvoir de prédiction

Pour tester le pouvoir de prédiction du modèle sur réseau sur les pseudo-noeuds, on étudie la structure native de petits pseudo-noeuds connus expérimentalement et qui induisent un décalage du cadre ouvert de lecture. Le décalage du cadre ouvert de lecture consiste à forcer les ribosomes à transcrire un autre cadre ouvert de lecture que le cadre habituel et ainsi à produire des protéines différentes. Ce phénomène est souvent rencontré chez les rétrovirus et peut être engendré par la présence de pseudo-noeuds en aval du ribosome [206, 209]. On a choisi le même jeu de petites séquences tronquées de signaux viraux de décalage du cadre ouvert de lecture (pris dans la base de données Pseudobase [210]) que Vfold. Pour les virus d'animaux, on prédit la structure la plus stable à 37° C et pour les virus de plantes à 25° C.

La table 5.2 compare les structures prédites et les H-pseudo-noeuds natifs en évaluant la sensibilité SE et la spécificité SP. De manière générale, on remarque que pknotsRG et Nupack font de meilleures prédictions que le modèle sur réseau, Kinefold ou Vfold. Cependant, cette comparaison est biaisée car les séquences étudiées ont été utilisées en partie pour évaluer les paramètres de ces deux modèles. On note également la moins bonne performance de pknotsRG et Nupack dans la prédiction des structures multiples (voir table 5.1). En particulier, Nupack prédit un état de plus basse énergie contenant un pseudo-noeud pour l'ARNt-phe de la levure qui n'est pas observé expérimentalement. De plus, parmi les méthodes basées sur des modèles physiques (pour décrire l'énergie de conformation), Vfold obtient de meilleurs résultats que le modèle sur réseau et que Kinefold.

En regardant de plus près les résultats du modèle sur réseau, on voit qu'il prédit parfaitement 4 structures natives (BChV, BWYV, EIAV et PLRV-S) et reproduit quasiment la bonne structure secondaire pour 3 autres séquences (FIV,PEMV et SRV1gag/pro). Dans ces 7 structures, aucune paire de bases incorrecte n'est prédite ($SP = 1$) et quelques paires sont manquantes car elles sont très contraintes géométriquement sur le modèle sur réseau. Pour les autres séquences, notre approche se trompe partiellement (BLV, BYDV-NY-RPV et CABYV) ou totalement (MMTVgag/pro et PLRV-W).

Une raison possible pour ces mauvais résultats pourrait être l'omission d'effets stabilisateurs. Pour CABYV, seuls les modèles utilisant des paramètres spécialement évalués pour bien décrire les pseudo-noeuds (pknotsRG et Nupack) prédisent correctement la structure secondaire native. De faibles interactions tertiaires (comme les bases triples) négligées dans le modèle sur réseau, Kinefold et de Vfold et qui seraient incluses de manière effective dans les paramètres de Nupack et pknotsRG, pourraient être à l'origine de ces mauvais résultats. Pour BLV, quasiment aucun modèle n'arrive à prédire la structure native. Ce pourrait être dû à l'omission d'interactions tertiaires plus fortes. Pour BYDV-NY-RPV et PLRV-W, le fait que Vfold marche bien alors que le modèle sur réseau et Kinefold se trompent, voudrait dire que des ingrédients présents dans Vfold (comme le volume exclu au niveau des fourches ou la description de la structure 3D de la double-hélice) et absents des deux autres modèles seraient importants pour la prédiction de la structure native de ces deux séquences.
Une autre raison possible serait la sous-estimation de l'énergie libre pour des petites boucles pseudo-noeuds à cause des contraintes géométriques imposées par le réseau. Cette explication semble bien fonctionner pour MMTVgag/pro dont la structure secondaire native est un H-pseudo-noeud avec $n_1 = 5$, $n_2 = 8$, $l_1 = 1$ et $l_2 = 8$. Sur le réseau CFC, construire des conformations décrivant un telle structure secondaire nécessiterait l'insertion de coudes dans les parties double-brins S_1 ou S_2. Ces coudes du double-brin sont des artefacts de la modélisation sur réseau or, ils sont très énergétiquement défavorables, donc cela va empêcher l'acceptation de telles conformations.

La figure 5.6 montre les cartes de contact pour une séquence où le modèle sur réseau prédit la bonne structure secondaire native (BWYV) et pour une où il se trompe (PLRV-W). Pour BWYV, la structure native est très largement majoritaire avec une probabilité de contact supérieure à 0.9 pour presque toutes les paires de bases. Seule la paire de base $9 - 26$ est un peu moins stable (probabilité ~ 0.7). Pour PLRV-W, bien que la structure prédite la plus stable ne soit pas la structure native, les

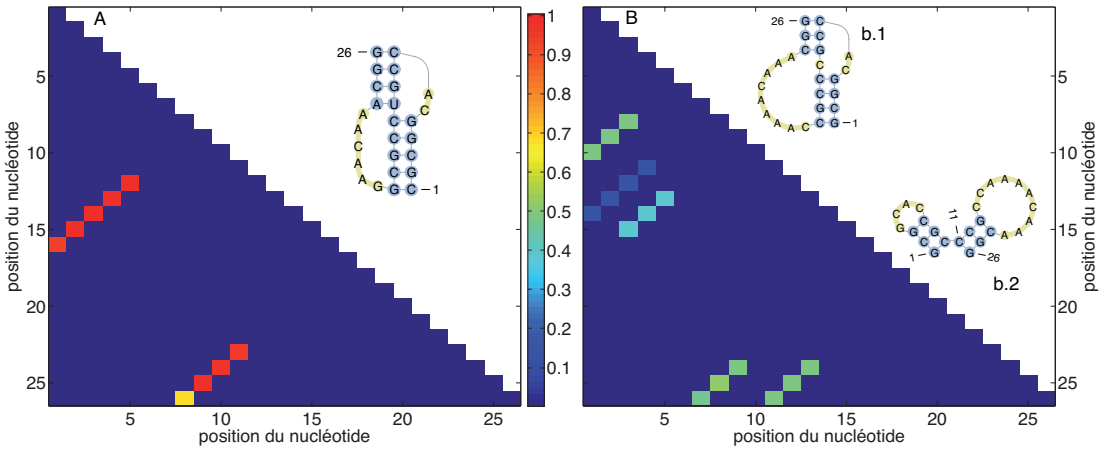

Figure 5.6 – Cartes de contacts calculées par le modèle sur réseau pour la séquence BWYV (A) et PLRV-W (B). Pour BWYV, on a tracé la structure secondaire native qui est aussi la structure prédite la plus stable ; pour PLRV-W, on a tracé la structure native (b.1) ainsi que la structure prédite la plus stable (b.2).

paires de bases composant le double-brin S_2 de la structure native ont une probabilité de contact non négligeable (0.5) ainsi que, dans une moindre mesure, ceux de S_1 (0.25). Le modèle sur réseau décrit donc partiellement la structure de cette séquence. D'ailleurs, l'écart entre la structure prédite la plus stable et la structure native dans les résultats sur réseau est faible ($\Delta G = 2.2 k_B T$) et est du même ordre de grandeur que l'incertitude sur les paramètres.

5.3 Impact du volume exclu

Dans le modèle de Turner standard, la différence d'énergie libre entre une structure secondaire $\mathcal{S}$ et l'état dénaturé est calculé en ajoutant à toutes les contributions d'empilement (association, fourche, dangle,...) la somme sur toutes les boucles des énergies libres de formation de boucle. Chacune est décrite par une relation de Jacobson-Stockmayer (voir équation 2.11), caractérisée par son exposant c. Cet exposant quantifie l'intensité des interactions stériques agissant sur la boucle. Le modèle de Turner néglige donc les possibles interactions stériques entre les différentes sous-structures (boucles et parties double- et simple-brins) qui composent $\mathcal{S}$ et ne tient compte que des intéractions stériques internes aux sous-structures via un exposant $c = 1.76$. Einert et collaborateurs [133] ont montré récemment que changer la valeur de c pour rendre compte de manière effective des interactions entre la boucle et les sous-structures voisines avait des conséquences non-négligeables sur les prédictions faites par le modèle de Turner (voir figure 5.7).

Dans cette partie, grâce au modèle sur réseau, nous étudions l'importance de tenir compte complètement des intéractions de volume exclu pour l'étude des molécules d'ARN. Tout d'abord, nous introduisons le modèle des brins fantômes (section 5.3.1) qui est une version du modèle sur réseau où l'on néglige (comme dans le modèle de Turner) les interactions stériques entre sous-structures, pour pouvoir estimer quantitativement les effets de cette approximation. Puis, nous évaluons l'impact du volume exclu sur le calcul des propriétés thermodynamiques et structurelles des ARN (sections 5.3.2, 5.3.3 et 5.3.4).

Les résultats du modèle sur réseau donnés pour cette partie ont été obtenus en utilisant la méthode de la contrainte maximale (voir section 4.4.2) pour chaque structure secondaire étudiée.

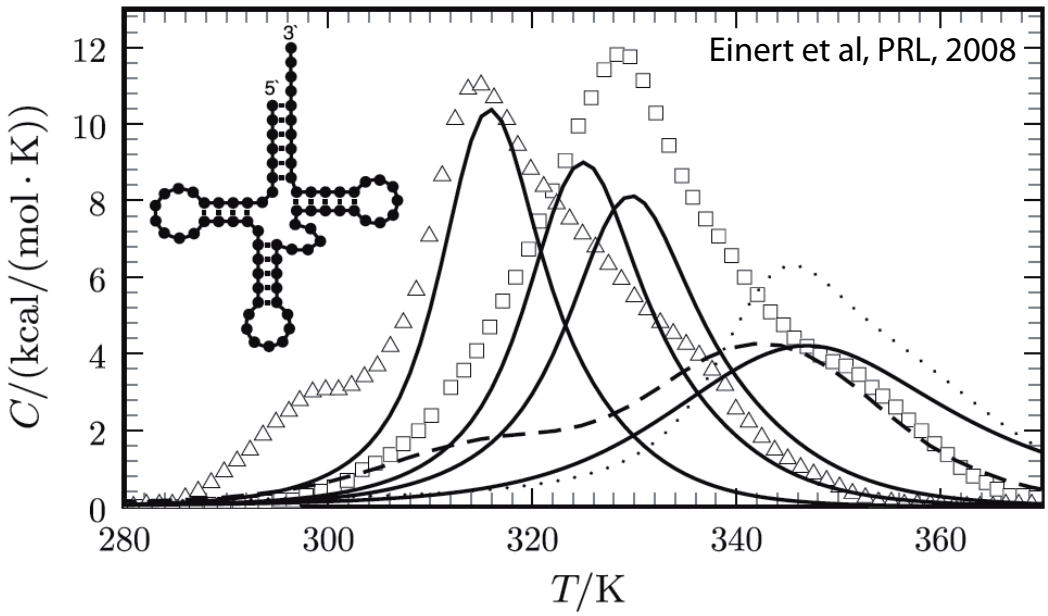

FIGURE 5.7 – Capacité calorifique C de l'ARNt-phe de la levure mesurée expérimentalement à $[Na^+] = 20$ mM (triangles) et 150 mM (carrés) ou prédite par une version de RNAfold développée par Einert et collaborateurs [133] pour différentes valeurs de l'exposant $c = 3.0$, 2.16, 1.76 et 0 (lignes pleines de gauche à droite). La courbe pointillée a été obtenue pour $c = 3$ et une énergie de nucléation nulle. La courbe tiretée a été calculée par le programme RNAheat [86] qui utilise une équation de Jacobson-Stockmayer généralisée (équation 2.12) avec $b = c = 0$ pour décrire les boucles multiples.

5.3.1 Le modèle des brins fantômes

5.3.1.1 Définition

Le modèle des brins fantômes utilise exactement le même type de paramètres (ϵ, γ, σ, λ,...) que le modèle sur réseau (voir section 3.1.1). La seule différence entre les deux modèles vient dans l'évaluation de l'entropie de conformation d'une structure secondaire $\mathcal{S}$. Alors que le modèle sur réseau tient compte complètement des effets de volume exclu en interdisant à deux nucléotides non complémentaires de se trouver sur le même noeud, le modèle des brins fantômes considère que les sous-structures (les boucles et les parties double- et simple-brins) qui composent $\mathcal{S}$ n'interagissent pas stériquement entre elles. Seul le volume exclu interne à la sous-structure est envisagé. Ainsi une sous-structure sera "invisible" (d'où le nom du modèle) pour le reste de la structure. L'entropie de conformation de $\mathcal{S}$ se calcule alors en sommant les différentes entropies de conformation individuelles de chaque sous-structure.

Pour une structure secondaire $\mathcal{S}$ sans pseudo-noeud, soit $\mathcal{B} = \{B_i\}$, $\mathcal{D} = \{D_i\}$ et $\mathcal{U} = \{U_i\}$ l'ensemble des boucles et des parties double- et simple brins de $\mathcal{S}$. On note N_i, l_i, n_i la taille de la boucle B_i, celle du double-brin D_i et celle du simple-brin U_i (en nombre de segments de chaîne). Pour les parties non-rigides (boucles et simple-brins), l'énergie libre de conformation individuelle vaut celle du chemin auto-évitant correspondant (fermé ou ouvert) :

$$G_{conf}(B_i) = -k_BT \log\left[\mathcal{N}_{SAP}(N_i)\right] = -k_BT \log\left[f_l \mu^{N_i} N_i^{-c}\right] \tag{5.2}$$

$$G_{conf}(U_i) = -k_BT \log\left[\mathcal{N}_{SAW}(n_i)\right] = -k_BT \log\left[f_s \mu^{n_i} n_i^{c'}\right] \tag{5.3}$$

Pour une partie double-brin D_i, il faut tenir compte en plus de la rigidité de la double-hélice, d'où

$$G_{conf}(D_i) = -k_B T \log \left[\sum_{n_{tot}} \omega(n_{tot}) \exp(-\beta n_{tot} \bar{\kappa}/2) \right] \tag{5.4}$$

avec $\omega(n_{tot})$ le nombre de conformations sur réseau du double-brin pour lesquelles la contribution totale des énergies de pliage vaut $n_{tot}\bar{\kappa}/2$ (voir section 3.3.5). Pour $n_{tot} = 0$, D_i peut être considérée comme une tige rigide donc $\omega(0) = z$. Pour $n_{tot} = 1$, D_i est coudée à 60°, ce coude peut se situer sur toute les $l_i - 1$ jonctions entre les segments qui composent D_i et dans chaque situation il y 4 possibilités pour un angle de 60° (voir figure 3.7), d'où $\omega(1) = (l_i - 1) \times z \times 4$. Pour $n_{tot} = 2$, D_i a soit un coude de 90° (2 possibilités à chaque fois) soit deux coudes de 60° , donc de même, $\omega(2) = (l_i - 1) \times z \times 2 + \binom{l_i-1}{2} \times z \times 4^2$. Plus généralement, on peut partitioner n_{tot} en trois contributions a_1, a_2 et a_3 qui représentent respectivement le nombre d'angles élémentaires de 60°, de 90° et de 120° dans D_i, soit $a_1 + 2a_2 + 3a_3 = n_{tot}$. Alors le nombre de conformations de D_i vérifiant cette partition vaut $\mathcal{N}(a_1, a_2, a_3) = z \times \binom{l_i-1}{a_1+a_2+a_3} \times 4^{a_1} 2^{a_2} 4^{a_3}$. D'où

$$\begin{aligned}
\omega(n_{tot}) &= \sum_{\text{partitions } \{a_1,a_2,a_3\}} \mathcal{N}(a_1, a_2, a_3) && (5.5)\\
&= z \times \sum_{a_3=0}^{\lfloor n_{tot}/3 \rfloor} \sum_{a_2=0}^{\lfloor (n_{tot}-3a_3)/2 \rfloor} \binom{l_i-1}{n_{tot}-a_2-2a_3} \times 2^{2n_{tot}-3a_2-4a_3} && (5.6)\\
&\approx z \times \binom{l_i-1}{n_{tot}} 4^{n_{tot}} && (5.7)
\end{aligned}$$

On obtient alors

$$\begin{aligned}
G_{conf}(D_i) &\approx -k_B T \log \left[z \sum_{n_{tot}=0}^{l_i-1} \binom{l_i-1}{n_{tot}} 4^{n_{tot}} \exp(-\beta n_{tot} \bar{\kappa}/2) \right] && (5.8)\\
&= -k_B T \left\{ \log z + (l_i - 1) \log \left[1 + 4 \exp(-\bar{\kappa}/2) \right] \right\} && (5.9)
\end{aligned}$$

Ainsi l'énergie de conformation de D_i est égale à l'énergie de conformation d'une tige rigide sur le réseau ($-k_B T \log z$) plus une correction qui vaut environ $-0.014(l_i - 1)k_B T$. Cette correction peut devenir non-négligeable quand l_i devient grand.

L'énergie libre de conformation totale pour $\mathcal{S}$ dans le modèle des brins fantômes vaut alors

$$G_{conf}^{fant}(\mathcal{S}) = \sum_{i=1}^{\#\mathcal{B}} G_{conf}(B_i) + \sum_{i=1}^{\#\mathcal{U}} G_{conf}(U_i) + \sum_{i=1}^{\#\mathcal{D}} G_{conf}(D_i) \tag{5.10}$$

Avec ceci, on peut paramétrer le modèle des brins fantômes exactement de la même manière que le modèle sur réseau (voir section 3.3.3) pour qu'il reproduise parfaitement le modèle de Turner pour des structures simples. On obtient que les paramètres peuvent également être définis comme la somme du paramètre de Turner correspondant et d'un correction entropique dépendante des constantes du réseau : dans les équations (3.26-3.32) pour le modèle dur réseau, il faut remplacer les termes $z - 1$ et $z - 2$ (qui traduisent le volume exclu à la jonction entre deux sous-structures) par z. La différence d'énergie libre totale de $\mathcal{S}$ est alors calculée avec

$$\Delta G^{fant}(\mathcal{S}) = [h_{fant}(\mathcal{S}) - T s_{fant}(\mathcal{S})] + G_{conf}^{fant}(\mathcal{S}) + k_B T \log \Omega_0 \tag{5.11}$$

où $\Omega_0 = f_s \mu^N N^{c'}$ est le nombre d'états sur réseau de l'état dénaturé et h_{fant} (s_{fant}) est l'équivalent de h_{latt} (s_{latt}) pour le modèle des brins fantômes.

Pour une structure secondaire $\mathcal{S}$ donnée, on pourra donc estimer l'importance des interactions stériques entre sous-structures en évaluant l'écart $\Delta\Delta G(\mathcal{S})$ entre la différence d'énergie libre ΔG^{res} de

$\mathcal{S}$ calculée avec le modèle sur réseau et celle ΔG^{fant} calculée avec le modèle des brins fantômes, soit $\Delta\Delta G(\mathcal{S}) = \Delta G^{res}(\mathcal{S}) - \Delta G^{fant}(\mathcal{S}) = -T(\Delta s^{res}_{corr} - \Delta s^{fant}_{corr}) + G^{res}_{conf}(\mathcal{S}) - G^{fant}_{conf}(\mathcal{S})$. Par définition, le modèle des brins fantômes représente en fait une généralisation du modèle de Turner. Ainsi, l'ajout de $\Delta\Delta G$ à l'énergie libre donnée par un programme standard type RNAfold ou Mfold va permettre de corriger l'estimation de ΔG faite par le modèle de Turner en prenant en compte le volume exclu entre sous-structures.

Pour des structures avec pseudo-noeuds, une généralisation du modèle des brins fantômes est envisageable, mais dans ce cas, il est plus difficile de définir clairement des sous-structures dont l'énergie libre individuelle de conformation est facilement calculable.

5.3.1.2 Calcul pratique de G^{fant}_{conf}

Considérons une séquence ayant $N+1$ nucléotides (soit N segments de chaînes) et une structure secondaire $\mathcal{S}$ modélisée par le vecteur $Link$ (voir section 4.1.1). D'après l'équation 5.10, pour calculer G^{fant}_{conf}, il faut donc évaluer le nombre de sous-structures, leur nature et leur taille. Pour cela, on parcoure itérativement $Link$ suivant l'algorithme suivant

1. Marquer tous les nucléotides comme non-visités et prendre $i = 1$ comme noeud courant.
2. Si $Link(i) \neq 0$, on a une sous-structure double-brin : a) marquer i et $Link(i)$ comme visités et initialiser la taille l de la sous-structure à 0 ; b) tant que $[Link(i+l+1) \neq 0] \wedge [Link(i+l+1) = Link(i+l) - 1]$, marquer $i+l+1$ et $Link(i+l+1)$ comme visités et mettre à jour l $(l = l+1)$; c) on a au final une partie double-brin de taille l.
3. Si $Link(i) = 0$, on a une sous-structure boucle ou simple-brin : a) marquer i comme visité et initialiser la taille l de la sous-structure à 0 ; b) prendre $j = i$ comme sous-noeud courant et initialiser à 1 le nombre n de nucléotides non-appariés consécutifs ; c) tant que $[(j+n) \leq (N+1)] \wedge \{[(Link(j+n) = 0] \vee [j \neq i-1]\}$, marquer $j+n$ comme visité et mettre à jour n $(n = n+1)$; d) si $j+n > N+1$, on a au final une partie simple-brin de taille $l+n-1$,sinon, si $Link(j+n) \neq 0$, mettre à jour l $(l = l+n)$, j $(j = Link(j+n))$ et n $(n = 1)$ et reprendre à l'étape c), sinon $(j = i-1)$, on a au final une partie boucle de taille l.
4. Parmi les nucléotides encore non-visités par l'algorithme, prendre celui ayant le plus petit numéro comme noeud courant.

L'étape 3 revient en fait à parcourir continûment la sous-structure formée de nucléotides non-appariés jusqu'à s'arrêter en bout de chaîne (dans ce cas la sous-structure est une partie simple-brin) ou jusqu'à être revenu à l'origine de la sous-structure (et dans ce cas, on a une boucle).

5.3.2 Interactions stériques entre deux boucles

Nous étudions d'abord l'impact du volume exclu sur l'exemple simple d'une structure secondaire en forme d'altère composée de deux boucles B_1 et B_2 (ayant chacune N_1 et N_2 segments) connectées par une partie double-brin D de taille L. Pour plusieurs valeurs de N_1, N_2 et L, on évalue l'énergie de conformation de la structure secondaire dans le modèle sur réseau et dans le modèle des brins fantômes.

La figure 5.8 A montre l'évolution de la différence d'énergie libre $\Delta g_{loop} = \Delta G^{res} - \sum \Delta g_{NN}$ calculée avec le modèle sur réseau, où $\sum \Delta g_{NN}$ représente la somme des contributions d'empilements du modèle de Turner. On observe, dans tous les cas, une dépendance logarithmique de Δg_{loop} en la taille N_2 de la forme $C^{ste} + c \log N_2$ avec c qui dépend de N_1 et L mais aussi de la gamme de taille N_2 que l'on regarde.
Quand la partie double-brin est petite ($L = 1$), les deux boucles peuvent interagir facilement. Pour une petite boucle B_1 ($N_1 = 4$), lorsque B_2 est petite également, on observe un exposant $c \approx 2.1$ signature d'une forte interaction stérique entre B_1 et B_2 ; par contre quand $N_2 >> N_1$, l'influence de B_1 devient négligeable et on retrouve le comportement classique d'un chemin fermé auto-évitant avec

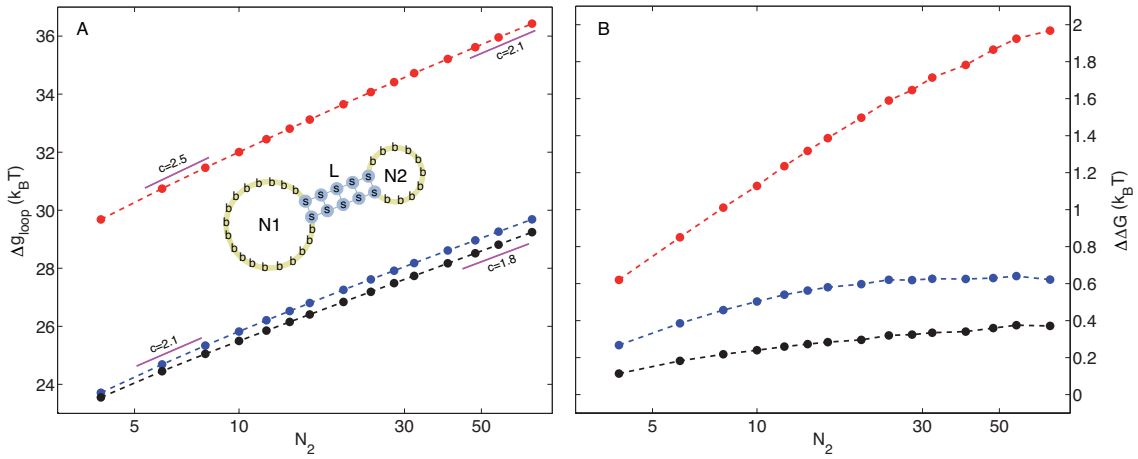

Figure 5.8 – Évolution de la différence d'énergie libre Δg_{loop} calculée avec le modèle sur réseau (A) et de l'écart $\Delta\Delta G$ au modèle des brins fantômes (B) pour la structure en forme d'altère dessinée au centre de A, en fonction de la taille N_2 de la boucle B_2 pour plusieurs valeurs des tailles N_1 et L de la boucle B_1 et la partie double-brin D : $(N_1, L) = (4, 1)$ (bleu), $(70, 1)$ (rouge) et $(4, 20)$ (noir).

$c \approx 1.8$. Quand B_1 est plus grande ($N_1 = 70$ avec toujours $L = 1$), on retrouve le comportement en $c \approx 2.1$ quand $N_2 \sim N_1$ indiquant ici une possible loi d'échelle pour c quand $N1$ et N_2 sont du même ordre dans le cas où D est petit ; l'intensité de l'interaction stérique augmente même quand $N_2 << N_1$ ($c \approx 2.5$). Pour $N_2 >> N_1$, on s'attend également à retrouver une loi asymptotique en $c \approx 1.8$.
Quand la partie double-brin devient plus grande ($L = 20$), si B_1 et B_2 sont toutes les deux petites, l'interaction mutuelle des deux boucles va être négligeable. Cependant, on observe dans ce cas, $c \approx 2.1$ ce qui signifie qu'il existe, à cette échelle de longueur, une interaction forte entre les boucles et la partie double-brin. D'ailleurs, on retrouve ici l'exposant calculé théoriquement par Kafri, Mukamel et Peliti [78] pour une boucle interne dans la limite où la taille de la boucle est petite par rapport à la taille des deux parties double-brins qui la juxtaposent. Quand $N_2 >> (N_1, L)$, on retrouve là aussi $c \approx 1.8$.

La figure 5.8 B permet de quantifier l'impact des interactions stériques entre sous-structures. On observe que $\Delta\Delta G$ augmente avec N_1 et N_2 jusqu'à saturation dans la limite $N_2 >> (N_1, L)$ où les deux modèles (sur réseau et des brins fantômes) ont le même comportement asymptotique. Plus la taille des sous-structures est grande, plus elles vont interagir entre elles et donc plus $\Delta\Delta G$ va être important. De même, plus l'exposant c dans la gamme de taille considérée est grand, plus $\Delta\Delta G$ va croître rapidement. L'intensité de $\Delta\Delta G$, dans les exemples étudiés, est de l'ordre de $k_B T$, ce qui est largement plus petit que la valeur de l'énergie de nucléation de bulle ($-T\Delta s_{loop} = 9.2 k_B T$). Ainsi, même si le volume exclu entre sous-structures modifie les lois d'évolution des énergies de conformation en fonction des tailles de boucles, en pratique, son effet peut paraître finalement négligeable devant la différence d'énergie totale. Nous verrons par la suite (section 5.3.4) que sa prise en compte peut modifier significativement les propriétés prédites pour le repliement de l'ARN.

5.3.3 Evaluation de relations de Jacobson-Stockmayer

Le modèle sur réseau incorpore explicitement les effets de volume exclu. Il peut donc être utilisé pour paramétrer des relations de Jacobson-Stockmayer ou pour calculer des énergies libres de boucles, qui pourront par la suite être utilisées dans un programme basé sur le modèle de Turner.

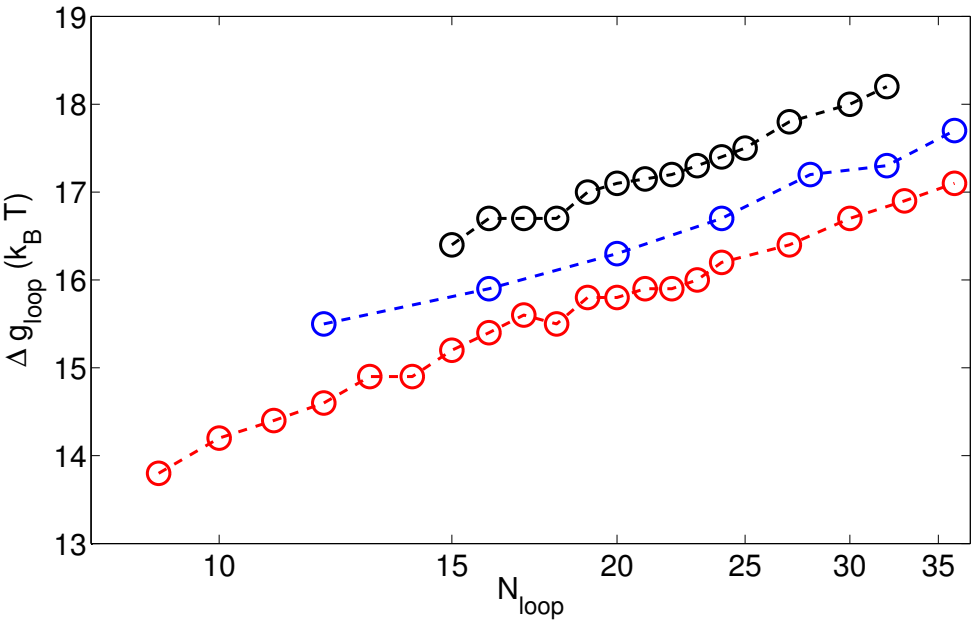

FIGURE 5.9 – Énergies libres Δg_{loop} de boucles multiples en fonction de leurs tailles N_{loop} dans le cas de 3 (rouge), 4(bleu) et 5 (noir) connexions, calculées avec le modèle sur réseau.

5.3.3.1 Boucles multiples isolées

Pour une boucle multiple isolée, on a déjà observé dans la figure 5.2 que l'exposant effectif valait $c \approx 2.2$ dans la gamme de taille de boucle étudiée. A partir de ΔG, en soustrayant toutes les contributions des parties double-brins, des fourches et des boucles en épingles, on peut remonter à l'énergie libre d'une boucle multiple Δg_{loop}. En répétant cette étude de ΔG en fonction de la taille de la boucle pour des situations avec 4 et 5 connexions (voir figure 5.9), on trouve que Δg_{loop} vérifient approximativement l'équation de Jacobson-Stockmayer généralisé

$$\Delta g_{loop}^{genJS} = \{7.2(\pm 0.1) + 0.65(\pm 0.02)h + 2.2 \log N_{loop}\} k_B T \quad (5.12)$$

$$= \{4.4 + 0.4h + 1.36 \log N_{loop}\}\text{kcal/mol à } 37^o\text{C} \quad (5.13)$$

avec N_{loop} la taille de la boucle en nombre de segments (et non en nombre de nucléotides appariés) et h le nombre de parties double-brin connectées. A titre de comparaison, RNAfold utilise $\Delta g_{loop}^{genJS} = \{3.4 + 0.4h\}$ kcal/mol. Notons que l'équation 5.12 n'est valide qu'avec le jeu unifié de paramètres de Turner définis dans la section 3.2 et que pour une boucle multiple "isolée". En effet, il n'est pas évident qu'une unique valeur asymptotique [78] ou heuristique [133] de l'exposant c décrive fidèlement une géométrie quelconque pour une boucle multiple. Par exemple, la valeur effective de c devrait dépendre de la taille relative de la boucle centrale par rapport aux sous-structures qui l'entourent, avec c qui tendrait vers l'exposant 1.76 des polygones auto-évitants pour des grandes boucles centrales (comme pour la structure en altère étudiée dans la section 5.3.2). Par contre, si la boucle fait partie d'une structure secondaire plus complexe, il faudrait tenir compte des interactions stériques avec le reste de la structure (voir ci-dessous). Cependant, dans une approche à la Turner où l'on suppose que chaque sous-structure est indépendante (pas d'interaction stérique entre sous-structures), l'équation 5.12 pourrait permettre de mieux décrire localement les boucles multiples.

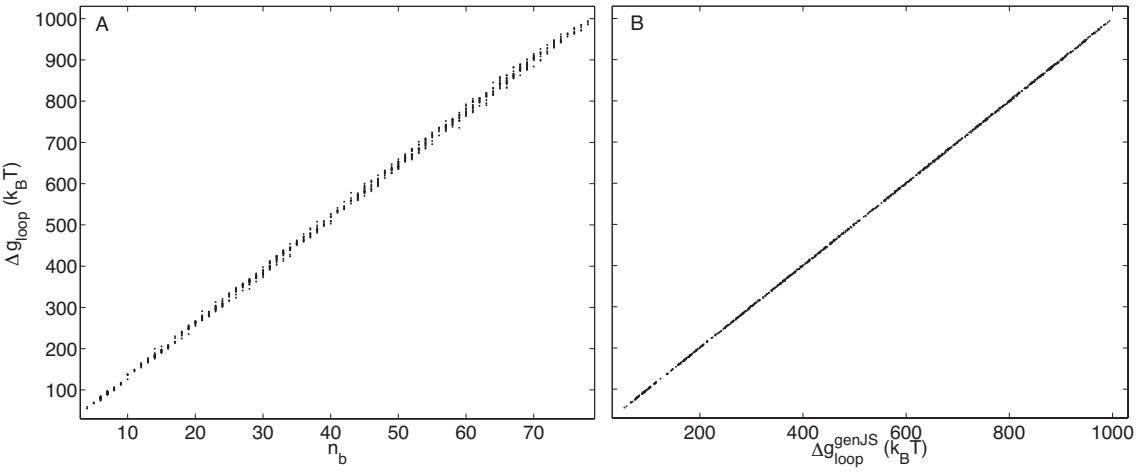

Figure 5.10 – Énergie libre totale Δg_{loop} des boucles dans la structure la plus stable prédite par RNAfold pour des séquences aléatoires de taille 100 nts à 1000 nts, en fonction du nombre n_b de boucle dans la structure (A) ou de l'énergie libre de Jacobson-Stockmayer généralisée calculée à partir de l'équation 5.14 (B).

5.3.3.2 Boucles dans structures complexes

Dans le paragraphe précédent, on regardait une boucle multiple isolée. Intéressons nous maintenant à une boucle quelconque appartenant à une struture complexe. Pour ce faire, on génère dans un premier temps des séquences aléatoires de différentes tailles (100 séquences pour chaque taille avec $N = 100, 200, ..., 1000$). Pour chacune, on considère la structure secondaire la plus stable $\mathcal{S}_m$ à 37° C calculée par RNAfold [86] pour laquelle on estime son énergie de conformation grâce à l'algorithme de la contrainte maximale et on en déduit la différence d'énergie libre totale de boucles Δg_{loop} (comprenant la nucléation des boucles et la différence d'entropie de conformation entre $\mathcal{S}_m$ et l'état dénaturé).

La figure 5.10 A montre l'évolution de Δg_{loop} en fonction du nombre n_b de boucle dans la structure secondaire considérée. On remarque bien évidemment que Δg_{loop} est une fonction quasi-linéaire de n_b ($\Delta g_{loop} \approx (12.8 n_b + 3.6) k_B T$) à cause en grande partie à la pénalité de nucléation par boucle ($-T\Delta s_{loop} = 9.2 k_B T$). Chaque boucle de la structure va donc contribuer à Δg_{loop}. Si l'on suppose, comme Turner, que l'énergie libre de boucle est la somme des énergies libres individuelles de chaque boucle, soit $\Delta g_{loop} = \sum_{i=1}^{n_b} \Delta g_{loop}(i)$, on peut modéliser chaque $\Delta g_{loop}(i)$ par une équation de Jacobson-Stockmayer généralisée $\Delta g_{loop}(i) = a + d \times h(i) + c \log N_{loop}(i)$ où $h(i)$ est le nombre de parties double-brins connectées à la boucle i et $N_{loop}(i)$ est le nombre de segment présent dans la boucle. On obtient alors

$$\Delta g_{loop}^{genJS} = \{(9.80 \pm 0.07) - (0.58 \pm 0.03)h + (2.34 \pm 0.01) \log N_{loop}\} k_B T \quad (5.14)$$

$$= \{6.05 - 0.36h + 1.44 \log N_{loop}\} \text{ kcal/mol à } 37^o\text{C} \quad (5.15)$$

La fonction gamma incomplète Q de cette modélisation vaut quasiment 1. Rajouter un terme de la forme $b \times N_{loop}$ dans la modélisation, n'améliore pas de manière significative l'erreur carré moyenne.

La figure 5.10 B montre l'excellent accord entre la modélisation et les données issues des simulations sur réseau. La relation de Jacobson-Stockmayer généralisée obtenue rend compte ainsi en moyenne de l'influence de la structure secondaire globale sur l'énergie de boucle locale, et bien sur ne décrit pas la situation d'une boucle isolée. Par exemple, pour une boucle en épingle à cheveux ($h = 1$) dans une structure complexe, l'équation 5.14 donne $\Delta g_{loop} \approx (9.2 + 2.34 \log N_{loop}) k_B T$, pour une boucle isolée, on a $\Delta g_{loop} \approx (9.2 + 1.76 \log N_{loop}) k_B T$ (voir section 5.1.2). La nucléation est similaire mais l'exposant c diffère, signature d'un influence forte de la structure globale. Autre exemple, pour une

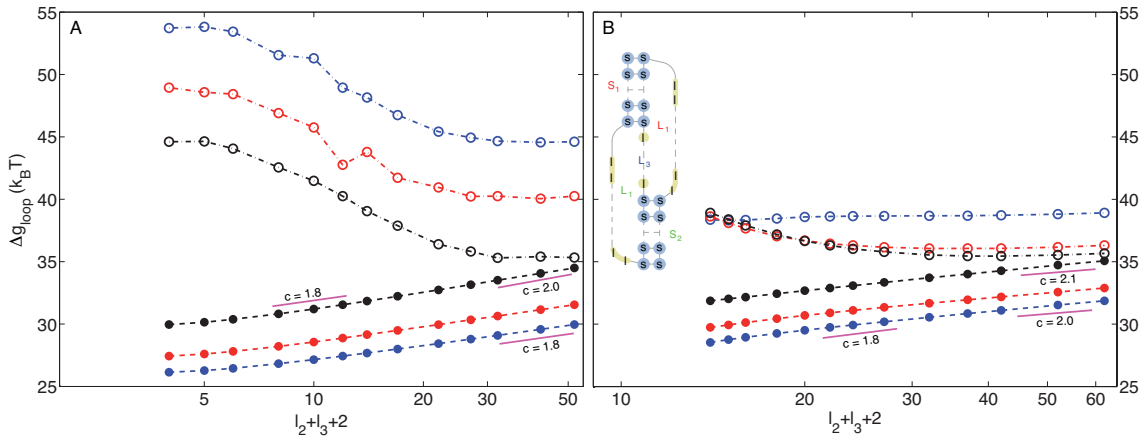

FIGURE 5.11 – Énergie libre Δg_{loop} d'un H-pseudo-noeud pour différentes tailles des parties double-brins, des boucles et du lieur, évaluée avec le modèle sur réseau, en fonction du nombre l_2+l_3+2 de segments non-appariés consécutifs formant la boucle effective L_2+L_3 : $n_1 = n_3 = 3$ (points) ou $n_1 = n_2 = 10$ (ronds) ; $l_1 = 2$ (bleu), 10 (rouge) et 50 (noir) ; $l_3 = 0$ (A) et 10 (B).

boucle multiple, les expressions 5.12 et 5.14 diffèrent également, preuve que les deux situations (boucle isolée et boucle incluse dans une structure complexe) sont bien différentes d'un point de vue de l'énergie de conformation de la boucle.

L'intérêt d'une telle relation de Jacobson-Stockmayer généralisée (équation 5.14) est de pouvoir l'incorporer dans un programme du style RNAfold ou Mfold pour l'étude du repliement de longues séquences où les effets d'interaction de volume exclu entre sous-structures commencent à être importants.

5.3.3.3 Pseudo-noeuds

De même que pour les boucles multiples, on s'intéresse ici aux énergies des pseudo-noeuds. Afin d'étudier plus précisément l'influence de la taille des différents éléments d'un pseudo-noeuds, on évalue l'énergie libre Δg_{loop} d'un H-pseudo-noeud pour différentes tailles des parties double-brins, des boucles et du lieur, à l'aide de la méthode de la contrainte maximale (voir section 4.4.2). Cette énergie comprend la nucléation des deux boucles et la différence d'énergie libre de conformation entre la structure pseudo-noeud et l'état dénaturé. Ainsi, l'énergie libre totale d'un H-pseudo-noeud sera donnée par la somme de Δg_{loop} et des énergies d'empilement (association, fourche, dangle, empilement coaxial) dans le modèle de Turner unifié. La figure 5.11 montre l'évolution de Δg_{loop} en fonction de la taille $(l_2 + l_3 + 2)$ de la boucle effective formée par L_2 et L_3 pour différentes valeurs de n_1, n_2, l_1, l_2 et l_3, les tailles respectives des deux parties double-brins S_1 et S_2, des deux boucles L_1 et L_2 et du lieur L_3 (voir schéma dans la figure 5.11 B).

Regardons d'abord le cas où les parties double-brins sont petites ($n_1 = n_2 = 3$). On observe alors que, dans la gamme de tailles de boucles étudiées, Δg_{loop} suit une loi logarithmique avec un exposant c qui varie selon la taille de L_1, L_2 et de L_3. Par exemple, pour la situation sans lieur ($l_3 = 0$) et avec une grande boucle L_1 ($l_1 = 50$), dans l'intervalle $l_2 \sim 10 - 20$, on trouve $c \approx 1.8$, signe que L_1 et L_2 interagissent faiblement. Quand l_2 devient de l'ordre de l_1, les deux commencent à interagir et $c \approx 2$. Si la taille de L_1 est plus petite ($l_1 = 2$) toujours dans le cas sans lieur, on n'observe pas de changement de comportement et $c \approx 1.8$ dans l'intervalle total étudié. La longueur du lieur L_3 joue sur la taille effective des boucles et donc sur les interactions stériques entre parties simple-brins. La présence du lieur accroît ainsi le volume exclu et les exposants c observés sont plus grands que dans le cas sans

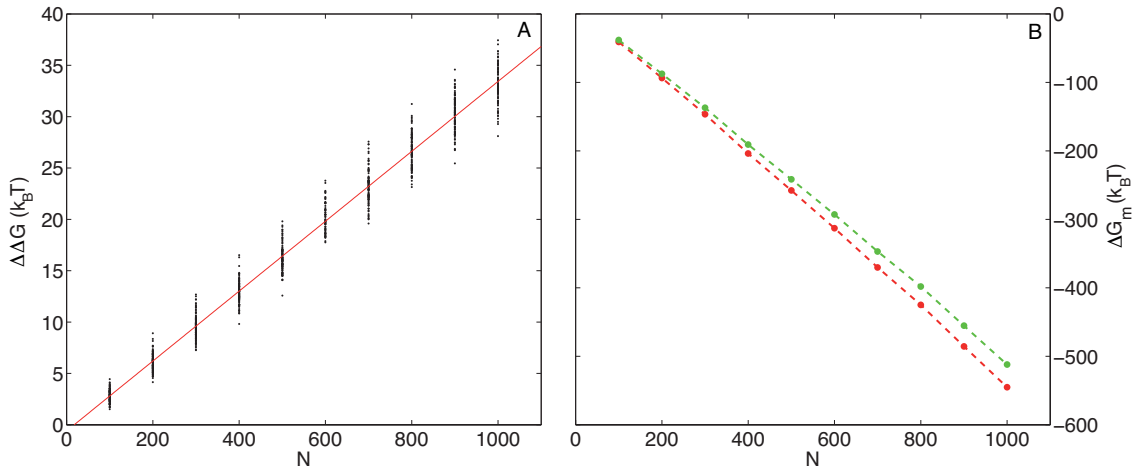

Figure 5.12 – (A) Correction $\Delta\Delta G$ en fonction de N calculée pour plusieurs séquences aléatoires de tailles différentes. La ligne rouge représente la modélisation des données par une droite affine $3.4 \times 10^{-2}N - 0.61$. (B) Énergie libre moyenne ΔG_m de la structure la plus stable à 37° C calculée par RNAfold pour des séquences aléatoires de taille N (points rouges) et cette même énergie libre mais avec la correction des interactions stériques entre sous-structures (points verts).

lieur.

Intéressons nous maintenant au cas où les parties double-brins sont plus longues ($n_1 = n_2 = 10$). Même si à plus grande échelle, on s'attend à retrouver le comportement décrit précédemment, pour les tailles de boucles étudiées, on observe un comportement complètement différent. A l_1 fixée, on remarque la décroissance initiale de Δg_{loop} en fonction de la taille de la boucle effective L_2+L_3. Pour comprendre cet effet, prenons le cas plus simple d'une chaîne gaussienne de taille N. La probabilité que sa distance bout-à-bout vaille r est donnée par la distribution gaussienne $\mathcal{P}(r) \propto (1/N)^{3/2} \exp[-3r^2/(2Nb^2)]$ [176]. L'énergie libre d'une chaîne dont la distance r est fixée vaut alors $-k_BT \log[\mathcal{P}(r) \times z^N]$ avec z le nombre de voisins accessibles sur réseau. D'où une différence d'énergie libre entre la structure à r fixé et la structure libre

$$\beta\Delta g = C^{ste} + \frac{3}{2}\log N + \frac{3r^2}{2Nb^2} \tag{5.16}$$

Δg se met donc sous la forme d'un somme d'un terme classique qui augmente logarithmiquement avec N et d'un terme qui décroît avec N mais qui augmente avec r. Retournons au cas du H-pseudo-noeud, vu que les parties double-brins sont rigides, on peut considérer que les distances bout-à-bout des boucles effectives $L_1 + L_3$ et $L_2 + L_3$ sont plus ou moins fixées avec $r \sim n_1 b$ ou $n_2 b$. Ainsi, quand n_1 et n_2 sont importants, le deuxième terme dans 5.16 devient prépondérant aux petites tailles de boucles (alors qu'il est négligeable pour des petites parties double-brins) et dicte l'évolution décroissante de Δg_{loop}. Augmenter la taille du lieur ou de L_1 revient à amoindrir cet effet.

Dans l'optique d'une utilisation de nos résultats dans un algorithme standard de prédiction de structures secondaires, nous avons tabulé Δg_{loop} pour $n_i \in \{2, 3, ..., 10, 11\}$, $l_1, l_2 \in \{2, 3, 4, 6, 8, 10, 12, 15, 20, 25, 30, 40, 50\}$ et $l_3 \in \{0, 1, 2, 4, 6, 8, 10\}$ (voir fichier *deltagloop-Hpk.dat* dans le CD annexe).

5.3.4 Impact sur les propriétés thermodynamiques et structurelles du repliement

Dans la section 5.3.2, nous avons étudié l'impact des effets stériques sur un exemple simple où seulement deux boucles interagissaient. Dans cette partie, nous regardons l'effet de tenir compte rigoureusement du volume exclu pour des structures secondaires (sans pseudo-noeud) plus grandes et plus

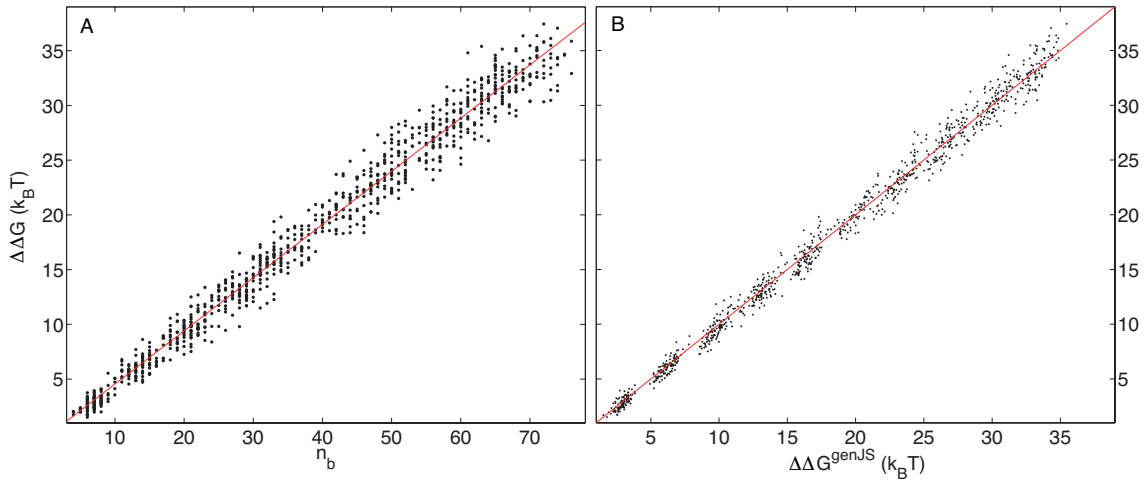

FIGURE 5.13 – Correction $\Delta\Delta G$ en fonction du nombre de boucles n_b dans la structure secondaire et sa modélisation linéaire $(0.47n_b - 0.36)k_BT$ (A) et en fonction de l'énergie libre de Jacobson-Stockmayer généralisé calculée à partir de l'équation 5.19 (B).

complexes.

5.3.4.1 Propriétés de $\Delta\Delta G$

Pour les mêmes structures secondaires que celles étudiées dans le deuxième paragraphe de la section précédente, on évalue $\Delta\Delta G$ la correction énergétique due aux interactions stériques entre sous-structures. La figure 5.12 A montre que $\Delta\Delta G$ est en moyenne une fonction croissante de la taille N. En effet, plus N est grand, plus le nombre de sous-structures présentes dans la structure secondaire va être important et donc plus l'intensité des interactions stériques va être significative. La croissance est linéaire et $\Delta\Delta G$ suit en moyenne la loi d'évolution

$$\langle \Delta\Delta G \rangle \approx \left\{3.4 \times 10^{-2} N - 0.61\right\} k_B T \tag{5.17}$$

Sur la figure 5.12 B, on représente la valeur moyenne sur toutes les séquences de même taille de l'énergie libre de $\mathcal{S}_m$ en fonction de N avec et sans la correction. On peut remarquer que la prise en compte des intéractions entre sous-structures augmente l'énergie libre de la structure d'environ 6% en moyenne quelque soit N. La relation 5.17 entre $\Delta\Delta G$ et N est cependant biaisée par le fait que l'on regarde pour chaque séquence la structure secondaire la plus stable à 37° C. Pour une autre température, on devrait avoir une autre relation. Afin de s'affranchir de ce biais, on regarde la valeur de $\Delta\Delta G$ en fonction du nombre de boucles n_b dans une structure secondaire qui est une propriété intrinsèque de la structure secondaire considérée (et non plus une propriété de N ou de la température). La figure 5.13 A montre également une évolution linéaire de la correction en fonction de n_b. Ce qui conforte l'idée que plus le nombre de sous-structures est grand et plus $\Delta\Delta G$ sera important. En modélisant les données par une droite affine on trouve

$$\langle \Delta\Delta G \rangle \approx \{0.47 n_b - 0.36\} k_B T \tag{5.18}$$

Ainsi, en moyenne, les interactions stériques entre sous-structures reviennent à corriger l'énergie libre donnée par un modèle standard (Turner ou brins fantômes) d'environ $0.5k_BT$ par boucle.

On peut être plus précis et modéliser la correction stérique par une relation de Jacobson-Stockmayer généralisée qui dépendrait de la taille de chaque boucle et du nombre de parties double-brins connectées comme dans le deuxième paragraphe de la section précédente. On trouve par boucle une correction

$$\langle \Delta\Delta G \rangle^{genJS} \approx \{-0.56 + 0.20h + 0.35 \log N_{loop}\} k_B T \tag{5.19}$$

La figure 5.13 B prouve la bonne qualité de la modélisation proposée. Pour une petite boucle en épingle, la correction restera faible ($\langle \Delta\Delta G \rangle^{genJS} \approx 0.1 k_B T$ pour $N_{loop} = 4$). L'exposant 0.35 du terme logarithmique est compatible avec les résultats observés pour une structure en altère dans la section 5.3.2 pour laquelle on observait un exposant relatif $c \approx 0.3$ pour $\langle \Delta\Delta G \rangle$ dans l'intervalle $4 - 10$ de tailles de boucles (voir figure 5.8 B). Typiquement pour une boucle interne ($h = 2$) de taille $N_{loop} \sim 6$, on trouve $\langle \Delta\Delta G \rangle^{genJS} \approx 0.5 k_B T$, ce qui est environ la correction moyenne par boucle (voir équation 5.18).

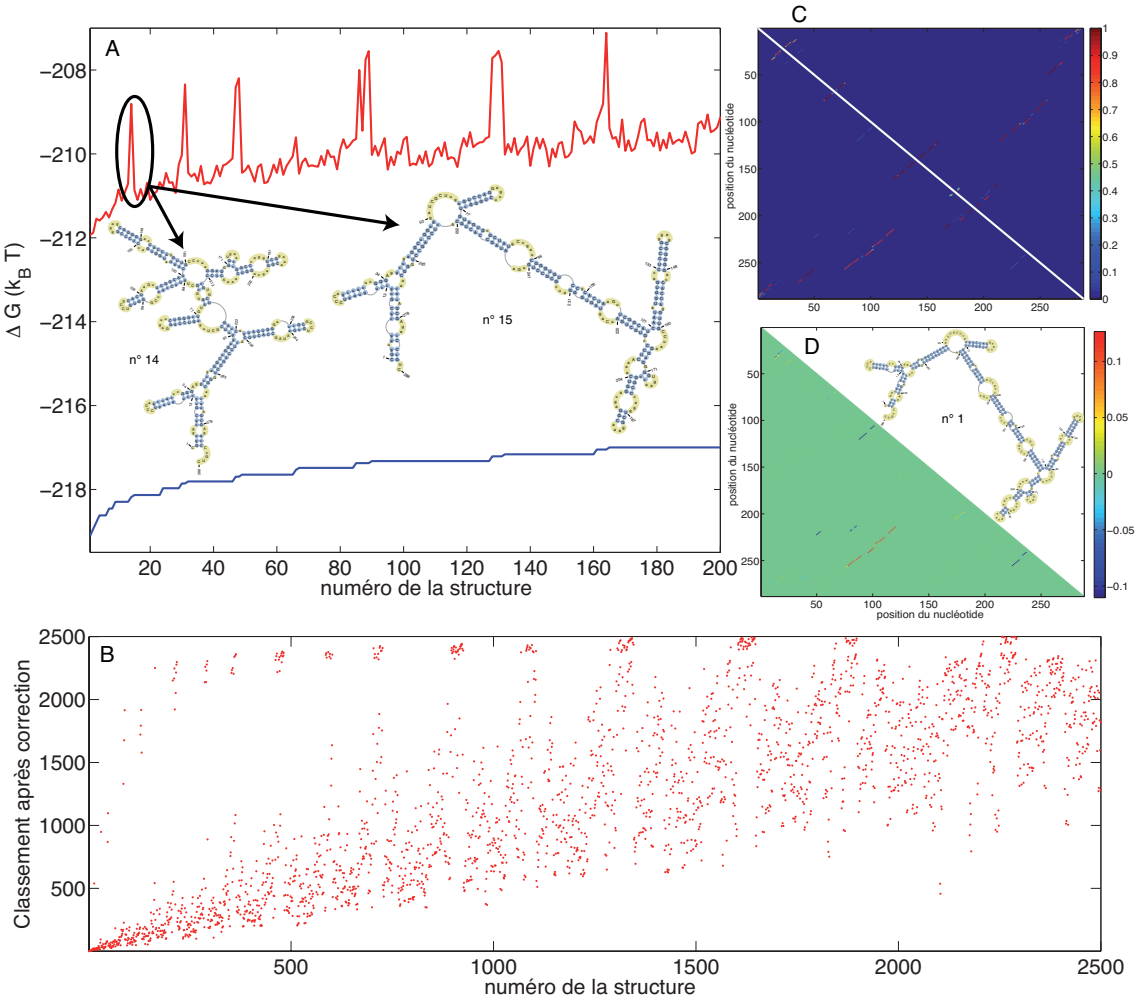

Figure 5.14 – (A) Différence d'énergie libre ΔG à 37° C pour les 200 séquences les plus stables dans le modèle de Turner et si l'on tient compte de la correction $\Delta\Delta G$ (ligne rouge). On a tracé les structures secondaires n° 14 et 15. (B) Classement des structures secondaires après correction avec $\Delta\Delta G$ en fonction de leur numéro (classement) initial. (C) Cartes de contact à 37° C calculées pour les 2500 structures les plus stables avec (haut) ou sans (bas) $\Delta\Delta G$. (D) Différence entre les deux cartes de contact de (C) et diagramme de la structure secondaire la plus stable.

5.3.4.2 Étude pour l'ARN de la particule de reconnaissance du signal humaine

Pour jauger l'impact du volume exclu sur les propriétés thermodynamiques et structurelles du repliement des molécules d'ARN, on peut, pour une séquence donnée et à température fixée, évaluer $\Delta\Delta G$ pour toutes les structures secondaires importantes et regarder comment sont modifiées ces propriétés. On étudie la séquence ARN de la particule de reconnaissance du signal humaine [2] qui est composée de 288 nucléotides. Pratiquement, à l'aide du programme RNAsubopt [86], on génère les 2500 structures secondaires les plus stables du modèle de Turner et pour chacune de ces structures, on évalue la correction $\Delta\Delta G$. La figure 5.14 A montre l'impact du volume exclu sur les 200 premières structures. On observe un $\Delta\Delta G$ moyen de l'ordre de $7k_BT$ en accord avec les résultats de la figure 5.12 A. Si l'on classe les structures en fonction de leur énergie libre après correction (voir figure 5.14 B), on remarque que des structures qui avaient des énergies similaires (classement proche) dans le modèle de Turner peuvent se retrouver assez éloigné ($> 2k_BT$) si l'on tient compte complètement des interactions stériques. Par exemple, la structure n° 14 descend au 589e rang alors que les autres structures qui avaient une énergie non-corrigée équivalente restent groupées entre la 10e et 20e position. Si l'on compare la composition des structures secondaires n° 14 et 15 (voir figure 5.14 A), on observe que les deux ont à peu près le même nombre de boucle (22) et la même correction calculée avec l'équation 5.19 ($9.9k_BT$ pour la 14 et $9.3k_BT$ pour la 15), donc cela n'explique pas la différence observée entre ces deux structures. Cependant, si on les regarde plus attentivement, on remarque que la 14 semble être plus compacte avec des parties double-brin plus courtes. Or on a vu dans la section 5.3.2 que plus les parties double-brins entre boucles étaient petites et plus elles interagissaient stériquement. Ceci pourrait expliquer la correction plus importante pour la 14 que pour la 15. Les figures 5.14 C et D évaluent l'influence du volume exclu sur la probabilité de contact entre deux nucléotides calculée sur les 2500 structures secondaires considérées. L'impact en moyenne reste limité et les plus grand écarts en probabilité sont somme toute assez faibles (~ 0.1).

5.3.4.3 Champ moyen des interactions stériques dans le modèle de Turner

Évaluer l'impact du volume exclu comme précédemment est cependant assez limitée. En effet, on ne s'intéresse qu'à une petite partie des structures secondaires possibles, et même si on choisit les plus stables, leur contribution dans l'énergie libre totale est seulement de l'ordre de 10%. Que se passe-t-il pour les autres structures participant aux 90% restants ? De plus, d'un point de vue numérique, cela va prendre beaucoup de temps pour calculer les $\Delta\Delta G$ sans toutefois être sûr de bien décrire l'ensemble thermodynamique. Ainsi, pour étudier l'impact du volume exclu sur une gamme plus large de taille de séquences de manière plus efficace, on inclut, dans le modèle de Turner, la correction dûe aux interactions stériques entre sous-structures par un champ moyen. Comme $\Delta\Delta G$ est en moyenne une fonction linéaire croissante du nombre de boucles rencontrées dans la structure secondaire (voir équation 5.18), il suffit de rajouter à tous les paramètres de nucléation de boucle de Turner une correction de $0.5k_BT$ (soit environ 0.3 kcal/mol) et d'utiliser ces nouveaux paramètres dans des programmes spécialisés tels que RNAfold [86]. On se sert de cette correction plutôt que de celle donnée par l'équation 5.19 car elle est plus facile à incorporer dans RNAfold, néanmoins, les conclusions de l'étude réalisée ci-dessous ne devraient pas changer de manière significative si l'on considérait l'autre correction.

Pour les mêmes séquences aléatoires que celles de la section 5.3.4.1, on évalue l'impact du volume exclu sur les propriétés thermodynamiques et structurelles à $T = 37^o$ C (la température d'étude la plus couramment utilisée) , à $T = 80^o$ C (une température proche de la température de dénaturation des molécules) et à $T = 130^o$ (une température où les molécules sont en grande partie dénaturées) à l'aide de RNAfold et RNAeval [86]. Pour chaque séquence, chaque température et chaque jeu de paramètres, on calcule la probabilité Θ qu'un nucléotide soit apparié, l'énergie libre ΔG_{tot} de l'ensemble

2. La particule de reconnaissance du signal est un complexe protéine-ARN qui, chez les eukaryotes, reconnaît et cible des protéines spécifiques du réticulum endoplasmique.

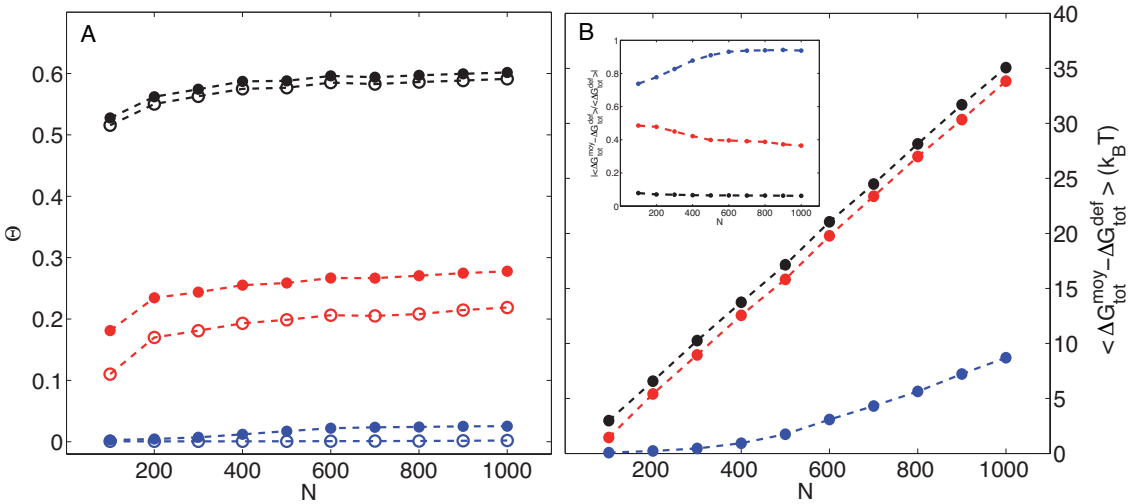

Figure 5.15 – (A) Θ moyen en fonction de la taille des séquences calculée avec le jeu de paramètres par défaut (points) ou avec le jeu de paramètres modifié (ronds) à 37° C (noir), à 80° C (rouge) ou à 130° C (bleu). (B) Écart moyen $\langle \Delta G_{tot}^{moy} - \Delta G_{tot}^{def} \rangle$ et écart relatif (encart) entre les énergies libres totales calculées avec les deux jeux de paramètres à 37° C (noir), à 80° C (rouge) et à 130° C (bleu).

thermodynamique, la structure secondaire $\mathcal{S}_m$ la plus stable, ainsi que son énergie libre $\Delta G(\mathcal{S}_m)$. Tout d'abord, on trouve qu'à 37° C, la correction moyenne $\langle \Delta\Delta G \rangle = \langle \Delta G^{moy}(\mathcal{S}_m^{def}) - \Delta G^{def}(\mathcal{S}_m^{def}) \rangle$ (où ^{def} fait référence aux résultats obtenus avec le jeu de paramètres par défaut et ^{moy} à ceux obtenus avec le jeu modifié) est égale (à $99.7 \pm 0.3\%$ près) à celle obtenue avec le traitement complet du modèle sur réseau (voir équation 5.18). Cela valide donc notre approximation de champ moyen.

On s'intéresse ensuite à l'influence de la correction stérique sur les propriétés thermodynamiques. La figure 5.15 A illustre la déstabilisation engendrée par le volume exclu entre sous-structures. A 37° C, elle reste néanmoins assez faible ($\langle \Theta^{moy} - \Theta^{def} \rangle \sim 0.01$) et quasiment indépendante de N. A 80° C, l'impact des effets stériques est nettement plus fort ($\langle \Theta^{moy} - \Theta^{def} \rangle \sim 0.06$) car on est proche de la température de dénaturation et les fluctuations sont donc plus importantes. A 130° C, l'impact redevient plus faible ($\langle \Theta^{moy} - \Theta^{def} \rangle \sim 0.02$) car les structures secondaires majoritaires sont composées de moins de sous-structures.

La figure 5.15 B représente l'écart moyen entre ΔG_{tot} calculée avec ou sans la correction stérique. Comme pour $\Delta\Delta G$, on observe une croissance linéaire avec N : plus la séquence est grande, plus les interactions stériques sont importantes. A 37 et 80° C, les corrections sont comparables. A 130° C par contre, la correction est plus faible mais, si l'on regarde l'écart relatif à la valeur moyenne par défaut de l'énergie libre (voir encart dans la figure 5.15 B), on observe qu'il est tout de même considérable.

Finalement, on étudie l'influence des interactions entre sous-structures sur la prédiction de la structure la plus stable. Pour chaque structure aléatoire étudiée, on compare la structure secondaire native prédite par le jeu de paramètres par défaut et celle prédite par le jeu de paramètres modifié. Pour cela, on évalue la sensibilité SE et la spécificité SP (voir section 5.2.1) de $\mathcal{S}_m^{moy}$ par rapport à $\mathcal{S}_m^{def}$. La figure 5.16 montre la distribution de probabilité de SE et SP pour les différentes tailles de séquences considérées à 37° et 80° C (à 130° C, dans tous les cas $\mathcal{S}_m$ est l'état dénaturé, donc $SE = SP = 1$ à chaque fois). A 37° C, pour toutes les tailles étudiées, on observe une assez bonne correspondance entre $\mathcal{S}_m^{moy}$ et $\mathcal{S}_m^{def}$ avec près de 80% des séquences qui ont une sensiblité et une spécificité supérieure à 0.8. Ainsi, en moyenne, pour cette température, les effets de volume exclu sur la prédiction de la structure la plus stable restent faibles, même si les cas où $SE \leq 0.1$ (soit une structure secondaire

totalement différente) existent ($\sim 2-3\%$). A 80° C, la situation est complètement différente. Alors que pour $N \lesssim 200$ les deux structures sont encore assez proche, pour des tailles plus grandes, la distribution des sensibilités est centrée autour de 0.5. Ainsi, en moyenne, près de 50% des paires de bases diffèrent entre $\mathcal{S}_m^{moy}$ et $\mathcal{S}_m^{def}$. La distribution des spécificités montre qu'une bonne partie ($\sim 40\%$) des structures ont une spécificité supérieure à 0.8. Cela indique que les différences observées entre $\mathcal{S}_m^{moy}$ et $\mathcal{S}_m^{def}$ sont souvent dûes à une déstabilisation de certaines parties double-brins à cause des interactions stériques. Cependant, un ensemble non négligeable de séquences ($\sim 10-20\%$) exhibe une sensibilité et une spécificité assez faible (≤ 0.3) ce qui atteste dans ce cas d'une faible correspondance entre $\mathcal{S}_m^{moy}$ et $\mathcal{S}_m^{def}$.

En conclusion, on peut dire que, comparé à l'effet de l'incertitude sur les paramètres (voir section 3.2.6), les effets du volume exclu entre sous-structures restent assez faible en moyenne sur les prédictions thermodynamiques et structurelles à 37° C. Même s'il existe une probabilité non négligeable que les prédictions faites par un programme standard (qui ne tient pas compte des interactions stériques entre sous-structures) soient faussées, surtout pour la structure la plus stable (l'effet sur les propriétés thermodynamiques comme Θ est moindre). Ces effets deviennent plus fort à des températures proche de la température de dénaturation où les fluctuations sont importantes. Pour des séquences petites ($N < 100$), on peut raisonnablement être sûr que les effets de volume exclu ne perturberont pas trop les prédictions des programmes standard (RNAfold, Mfold), mais pour des séquences plus grandes, il faut être à chaque fois très prudent par rapport à leur utilisation directe. Notons également que nous avons discuté ici d'effet moyen et que certains comportements comme celui observé dans le paragraphe précédent ne sont pas décrits par la correction moyenne.

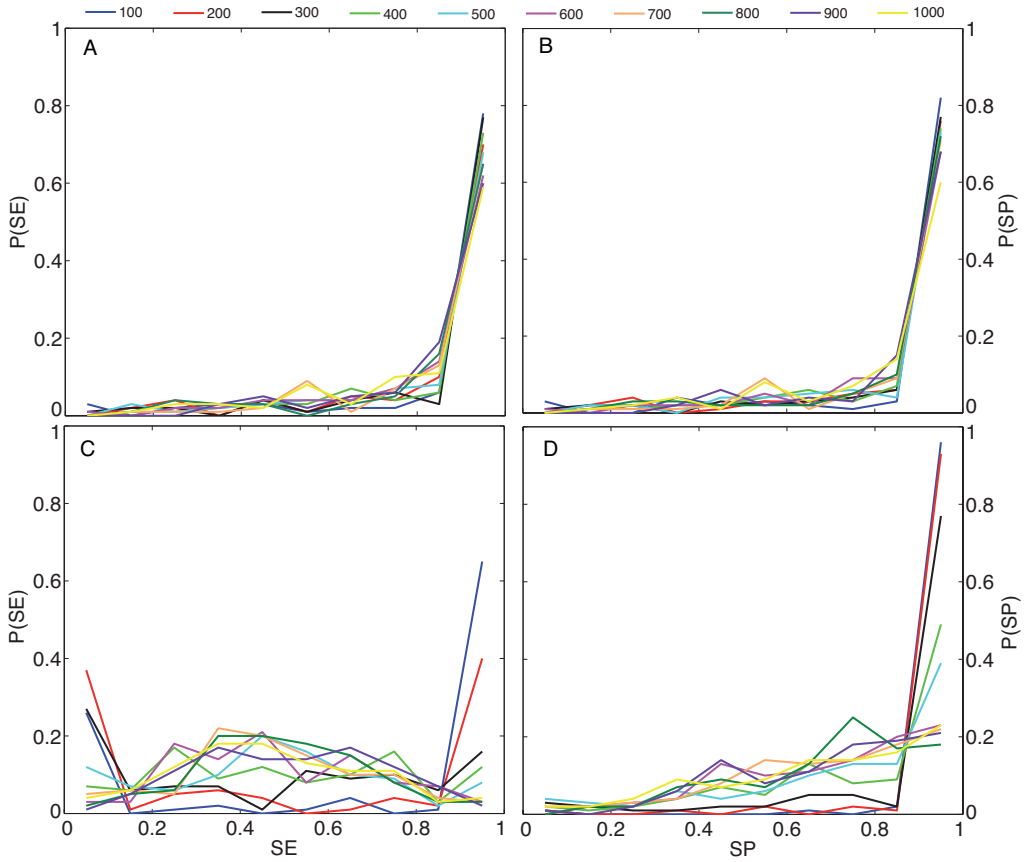

FIGURE 5.16 – Distribution de probabilité pour la sensibilité SE (A et C) et pour la spécificité SP (B et D) à 37° C (A et B) et à 80° C (C et D) pour différentes tailles de séquences (voir légende en haut de la figure).

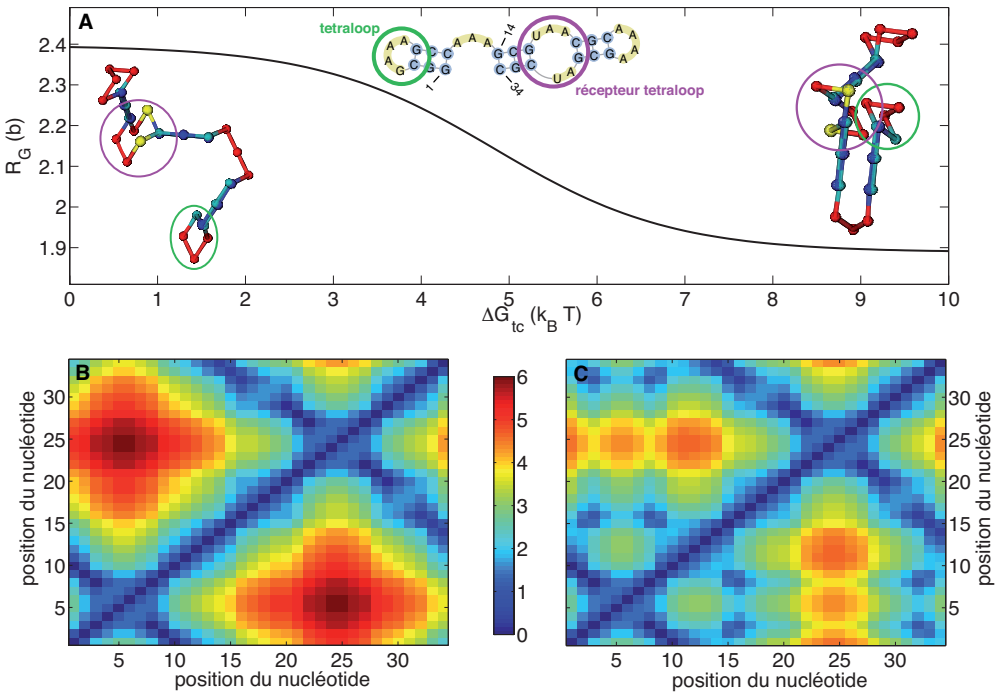

Figure 5.17 – (A) Transition à deux états pour le rayon de giration R_G (en unité b) en fonction de l'intensité du contact tertiaire ΔG_{tc}. La structure secondaire de la séquence T/Tr_3 étudiée est dessinée au centre. Des exemples de conformations tertiaires sont représentées pour $\Delta G_{tc} = 0k_BT$ (à gauche) et $-10k_BT$ (à droite). (B,C) Carte des distances (distance moyenne entre deux nucléotides) pour $\Delta G_{tc} = 0k_BT$ (B) et $-10k_BT$ (C). La légende des couleurs est donnée au milieu des deux figures (en unité b).

5.4 Interactions tertiaires

Un des avantages du modèle sur réseau par rapport aux approches standard est sa définition tridimensionnelle. Dans cette section, nous étudions l'effet d'interactions tertiaires sur le repliement d'une molécule d'ARN. En particulier, nous considérons l'action d'un contact tertiaire spécifique (section 5.4.1) ou d'interactions stériques avec l'extérieure (section 5.4.2).

5.4.1 Interactions spécifiques de contact

Le repliement de l'ARN est en partie hiérarchique : sous des conditions typiques (en température et en concentration en sel), le simple brin se replie d'abord pour former la structure secondaire. Puis, si la concentration en sel est assez forte, des contacts tertiaires entre les éléments de la structure secondaire s'établissent et donnent naissance à la structure tertiaire complètement repliée [211, 212]. Les catégories de contacts tertiaires spécifiques sont nombreuses [213] et la plupart de ces contacts sont fortement sensibles à la concentration en cations salins (spécialement Mg^{2+} [211, 214]). Un exemple de contact tertiaire est l'interaction tetraloop/recepteur tetraloop (T/Tr) présente dans le domaine P4-P6 du ribozyme groupe I de *Tetrahymena* [215, 216, 217] (voir figure 5.18 A). Elle connecte la tetraloop L5b (une boucle en épingle $GAAA$) et son récepteur J6a/b (une boucle interne UAA/AU).

Pour être le plus général possible, on étudie la séquence modèle

$$T/Tr_n = GGCGA_3GCCA_nGCGUAACGC_4GCGAUCGC \tag{5.20}$$

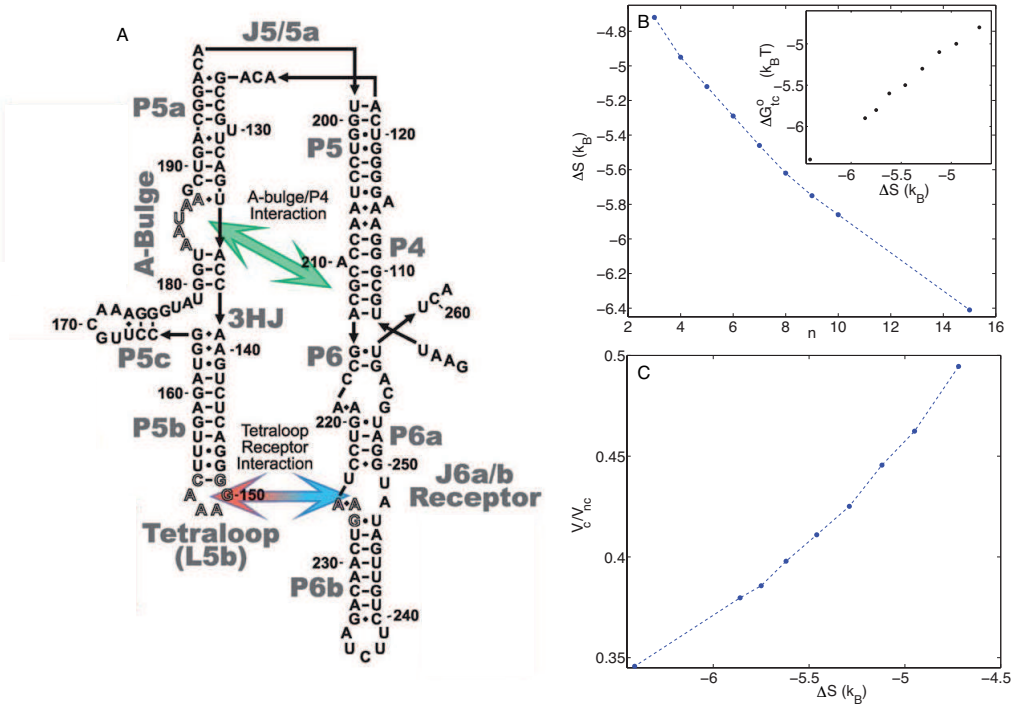

FIGURE 5.18 – (A) Structure secondaire du domaine P4-P6 du ribozyme groupe I de *Tetrahymena* tirée de Ref.[215]. (B) Différence d'entropie de conformations ΔS entre les états compactes et l'ensemble des états en fonction de la taille n du lieur. (Encart) Intensité ΔG^o_{tm} de l'interaction T/Tr à la transition de compactification (équivalent à T_m pour la dénaturation) en fonction de ΔS. (C) Taux V_c/V_{nc} de compactification des volumes en fonction de la différence d'entropie ΔS.

où n nous autorise à faire varier la taille du lieur entre la tetraloop et son récepteur (voir figure 5.17 A). On suppose qu'il y a formation d'un contact T/Tr quand les deux sous-structures sont assez proches l'une de l'autre. Arbitrairement, on choisit une distance de b entre les barycentres de la tetraloop et de son récepteur (ce choix n'affecte pas les observations générales que l'on va faire par la suite). Les simulations ont été réalisées à une température faible (273 K) où la structure secondaire native est majoritaire en utilisant les mouvements élémentaires décrits dans la section 4.3.2. La figure 5.17 A montre la compactification globale du complexe au fur et à mesure que l'on augmente l'intensité ΔG_{tc} de l'interaction T/Tr. Cette évolution du rayon de giration est caractéristique d'une transition à deux-états. Les figures 5.17 B et C représentent l'effet de la compactification au niveau des distances entre nucléotides. Quand l'intéraction tertiaire est nulle, seuls les nucléotides appartenant aux mêmes double-brins sont en moyenne proches. La distance entre les nucléotides de la tetraloop et du récepteur est de l'ordre de $5b$. Par contre, quand la compactification a lieu, on observe une diminution significative de cette distance ($\sim 2b$) et que les parties double-brins vont préférentiellement être parallèle. Ceci concorde avec les observations expérimentales faites sur le domaine P4-P6 du ribozyme groupe I de *Tetrahymena* [215, 216].

Pour chaque taille n du lieur, on estime la fraction r de conformations sur le réseau dont la distance entre les barycentres de la tetraloop et du récepteur est plus petite que b (c'est à dire, les conformations qui peuvent être impliquées dans une interaction T/Tr). A l'aide de r, on définit la différence d'entropie de conformation entre les états compacts et l'ensemble des étatss par $\Delta S = k_B \log r$. La figure 5.18 B montre l'évolution décroissante de ΔS en fonction de n. Ainsi, plus le nombre de nucléotides entre la tetraloop et le récepteur est important, plus la probabilité de former un contact T/Tr est faible. On

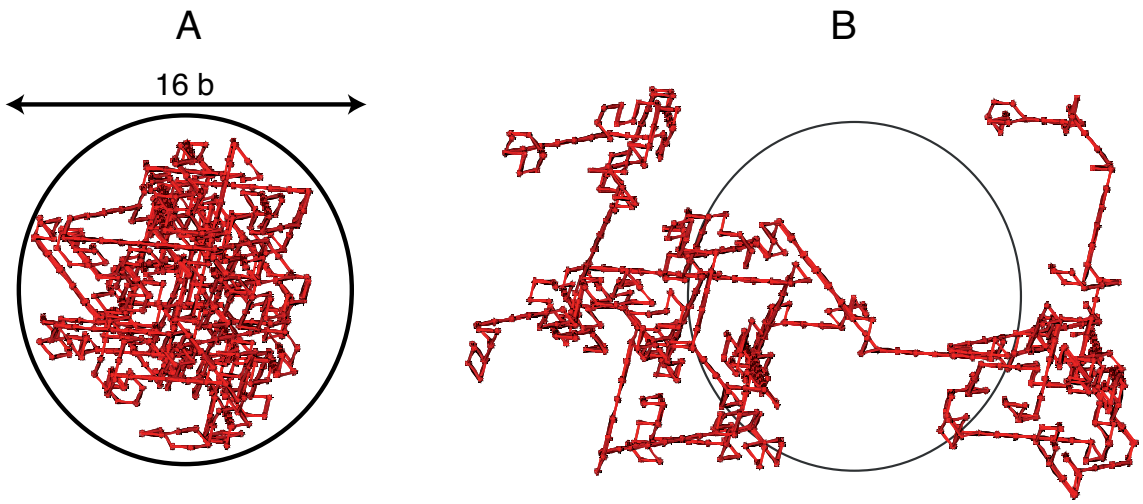

Figure 5.19 – Conformations possibles d'une structure secondaire composée de 900 nts confinée dans une sphère de rayon $8b$ équivalent à une densité $\rho = 0.3$ (A) ou non-confinée (B).

vérifie également que la valeur de ΔG_{tc} à la transition de compactification (ΔG^o_{tc}) est égale à $T\Delta S$. Autrement dit, l'énergie minimale de contact pour avoir la moitié des conformations avec un contact T/Tr vaut le coût entropique de passer d'une structure quelconque à une structure compacte.

On peut également évaluer la niveau de compactification de la molécule en calculant le quotient entre le volume moyen V_{nc} occupé quand $\Delta G_{tc} = 0$ et le volume moyen V_c occupé quand ΔG_{tc} est suffisamment intense pour avoir toutes les conformations compactes. On observe (voir figure 5.18 C) une importante compactification avec une diminution de près de 50% du volume initial, conformément aux observations expérimentales faites sur le domaine P4-P6 du ribozyme groupe I de *Tetrahymena* [215].

5.4.2 Interactions stériques avec l'extérieur

In vivo, une molécule est rarement isolée et subit des interactions constantes avec l'extérieur. Nous nous intéressons dans cette partie au cas particulier d'un confinement d'un brin d'ARN dans une cavité sphérique. Cette situation modélise les simples brins d'ARN viraux présents dans des capsides. En particulier, on veut estimer la perte d'entropie de conformation due au confinement dans une sphère de rayon R (voir figure 5.19).

Tout d'abord définissons quelques notions utiles par la suite. Soit une chaîne de taille N, on définit la densité ρ comme le quotient entre le nombre de nucléotides dans la séquence et le nombre N_s de noeuds du réseau dans la sphère. Pour le réseau CFC, dans la maille décrite dans la figure 3.7 de volume $2^{3/2}b^3$, il y a 4 noeud. D'où, dans une sphère de volume $(4/3)\pi R^3$, on aura

$$N_s = \frac{8\pi}{3 \times 2^{1/2}} \left(\frac{R}{b}\right)^3 \approx 5.924 \left(\frac{R}{b}\right)^3 \text{ noeuds} \tag{5.21}$$

et donc une densité de

$$\rho \equiv \frac{N}{N_s} = \frac{N}{5.924} \left(\frac{b}{R}\right)^3 \tag{5.22}$$

Pour estimer la perte d'entropie de conformation, on utilise les mêmes séquences aléatoires et structures secondaires que dans la section 5.3.3.2. Pour chaque structure secondaire, on évalue son énergie de conformation à l'aide d'une version légèrement modifiée de l'algorithme de la contrainte maximale,

pour 11 densités différentes $\rho \in \{10^{-5}, 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$. Cette variante consiste à choisir aléatoirement un noeud dans la sphère pour commencer la croissance d'une chaîne et d'imposer que la conformation reste dans la sphère en fixant un potentiel infini pour une direction-test sortant de la capside. On soustrait alors aux énergies mesurées les énergies de conformations des structures secondaires non-contraintes afin d'obtenir $\Delta\Delta G_{conf}(R, \mathcal{S})$ l'énergie de confinement. La figure 5.20 illustre l'évolution en fonction de R de $\Delta\Delta G_{conf}$ moyennée sur les séquences de même taille. A N fixé, on observe sa décroissance quand R augmente, l'influence des intéractions stériques avec l'extérieur diminuant rapidement quand la sphère est assez grande. A R fixé, on observe une croissance de $\Delta\Delta G_{conf}$ avec N, plus le nombre de nucléotides à confiner est important dans un volume donnée, plus on doit fournir d'énergie.

Pour deviner de quelle manière cette énergie devrait dépendre de N et de R, faisons un argument de Flory : 1) l'énergie élastique d'une chaîne gaussienne confinée dans une boîte de dimension typique R est proportionnelle à $N(b/R)^2$ si le rayon de giration de la chaîne non-confinée est grand devant R ($b\sqrt{N} >> R$) [176] ; 2) le terme issu du viriel comptant pour les interactions steriques, garde sa forme classique proportionnelle à vN/R^3 où v est le volume exclu d'un segment de chaîne. Soit une énergie totale de Flory [218]

$$\beta\Delta\Delta G_{conf}^{Flory}(N, R) = c_1 N \left(\frac{b}{R}\right)^2 + c_2 N^2 \left(\frac{b}{R}\right)^3 \tag{5.23}$$

avec c_1 un facteur géométrique et $c_2 \propto v/b^3$. En modélisant les données par l'équation 5.23, on trouve $c_1 = 1.85 \pm 0.09$ et $c_2 = 0.04 \pm 0.01$, reproduisant de manière assez fidèle le comportement observé avec le modèle sur réseau (voir figure 5.20).

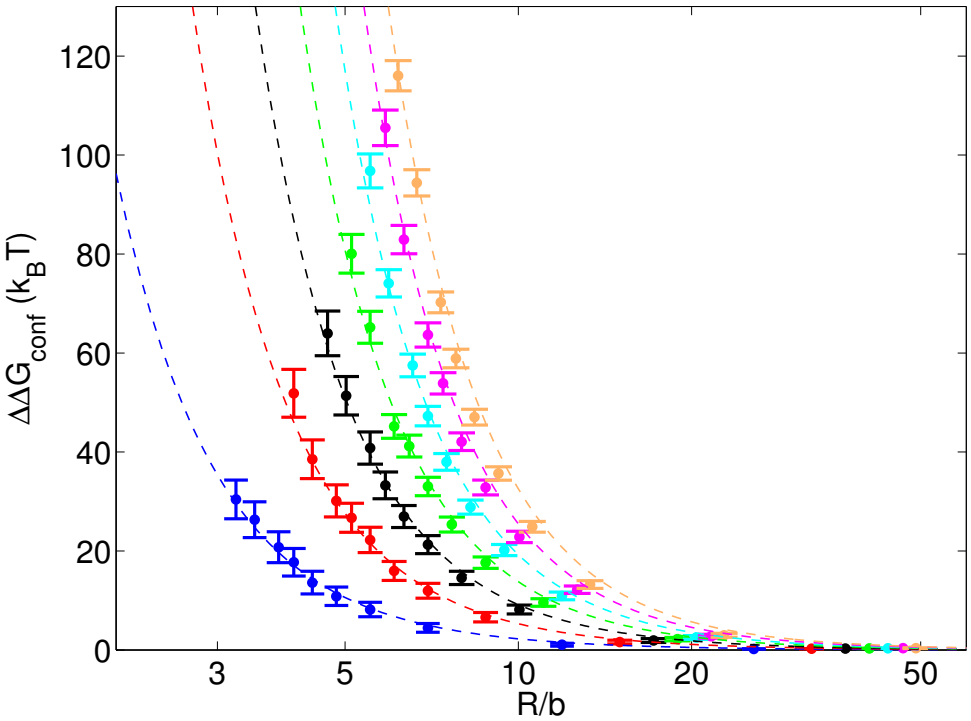

FIGURE 5.20 – Énergie de confinement moyenne en fonction du rayon de la sphère pour plusieurs tailles de brins d'ARN : $N = 100$ (bleu), 200 (rouge), 300 (noir), 400 (vert), 500 (cyan), 600 (mauve) et 700 (orange). Les barres d'erreur décrivent les fluctuations de $\Delta\Delta G_{conf}$ calculées sur 100 structures différentes de même taille. Les lignes pointillées représentent la modélisation des données par une énergie de Flory (voir équation 5.23).

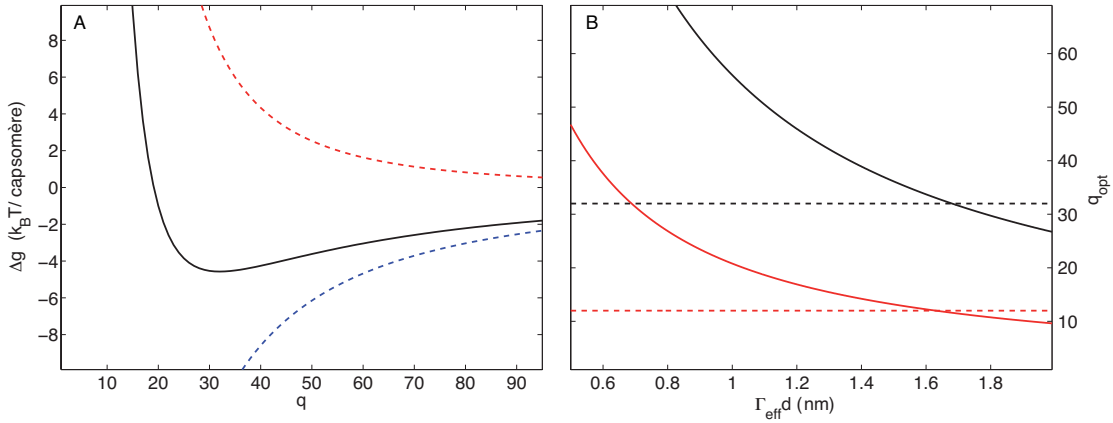

Figure 5.21 – (A) Évolution de l'énergie d'encapsidation par capsomère $\Delta G_{vir}/q$ (noir), de l'énergie de confinement par capsomère $\Delta\Delta G_{conf}/q$ (rouge) et de l'énergie d'attraction par capsomère $\Delta G_{att}/q$ (bleu) en fonction du nombre q de capsomères, pour le brin d'ARN du virus CCMV avec $\Gamma_{eff}d = 1.68$ nm afin d'avoir $q_{opt} = 32$. (B) Évolution de la valeur optimale q_{opt} du nombre de capsomères en fonction de $\Gamma_{eff}d$ pour CCMV (noir) et STNV (rouge). Les lignes pointillées représentent la valeur $q_{opt} = 32$ (noir) ou 12 (rouge).

Une application directe de cette modélisation est l'étude des virus d'ARN simple-brin. Une capside virale de rayon interne R est communément modélisée comme l'agencement de q capsomères formées chacune de 5 ou 6 protéines. Typiquement, une capside a une géométrie icosahédrale et est composée de 12 pentamères plus $10 \times (T-1)$ hexamères avec $T = h^2 + k^2 + hk = 1, 3, 4, 7, 12, 13, ...$ $(h, k \in \mathbb{N})$. Par exemple, la capside du virus chlorotique marbré de la cornille (CCMV pour "cowpea chlorotic mottle virus") contient $q = 32$ capsomères ($T = 3$), et celle du virus satellite responsable de la nécrose du tabac (STNV pour "satellite tobacco necrosis virus") en contient $q = 12$ ($T = 1$). Le rayon typique r_0 d'un capsomère est donné par la conservation de la surface interne : $q\pi r_0^2 = 4\pi R^2$, soit la relation entre R et q :

$$R = \frac{r_0}{2}\sqrt{q} \tag{5.24}$$

Pour CCMV ($N = 3171$ nts), $R = 14$nm et $q = 32$, soit $r_0 \approx 4.9$nm. Pour STNV ($N = 1239$ nts), $R = 9.2$nm et $q = 12$, soit $r_0 \approx 5.3$nm.

Soit N la taille de la molécule d'ARN encapsulée, l'énergie libre totale de la chaîne dans le virus ΔG_{vir} est approximée par la somme de l'énergie libre nécessaire $\Delta\Delta G_{conf}$ pour confiner le polymère dans la sphère et de l'énergie d'interaction ΔG_{att} attractive entre les parois internes de la capside et les nucléotides du brin d'ARN [219]. Cette dernière est la résultante des forces écrantées de Coulomb entre les résidus chargés positivement de la surface interne de la capside et la molécule d'ARN chargée négativement [220]. La portée d'une telle interaction, notée d, est typiquement la longueur de Debye, soit $d \sim 1$ nm. Pour modéliser ΔG_{latt}, on effectue également un argument de Flory : le nombre moyen de nucléotides présents dans la couronne sphérique d'épaisseur d vaut $(4\pi R^2 d) \times (N/R^3) \sim N(d/R)$. Si l'on note Γ_{eff} la valeur absolue de l'intensité effective des interactions entre la paroi et un nucléotide (en unité $k_B T$), on modélise alors l'énergie totale d'interaction par

$$\beta\Delta G_{att} = -\Gamma_{eff} N \left(\frac{d}{R}\right) \tag{5.25}$$

En remplaçant R par son expression en fonction de q (équation 5.24), on trouve une énergie d'encap-

sidation par capsomère de

$$\beta\frac{\Delta G_{vir}}{q} = c_1\left(\frac{2b}{r_0}\right)^2\frac{N}{q^2} + c_2\left(\frac{2b}{r_0}\right)^3\frac{N^2}{q^{5/2}} - \Gamma_{eff}\left(\frac{2d}{r_0}\right)\frac{N}{q^{3/2}} \tag{5.26}$$

avec c_1 et c_2 déterminés précédemment et b la constante du réseau.

Pour calculer la valeur de $\Delta G_{vir}/q$, on a besoin tout d'abord d'estimer une valeur réaliste pour b. Pour cela, considérons une structure secondaire de taille N. Soit ΘN le nombre de nucléotides appariés (avec $\Theta \sim 0.6$, voir section 3.2.6.2), elle sera alors composée d'environ $N\Theta/2$ segment en double-hélice et de $N(1-\Theta)$ segment simple-brin. Sachant que le volume d'un segment de paires de bases vaut environ 1.35 nm^3 et si l'on suppose que celui d'un segment simple-brin en vaut la moitié, le volume total occupé par la molécule sera d'environ $(0.7\text{nm}^3)N$ (pour $\Theta = 0.6$). Or sur le réseau, segments double-brin et simple-brin occupent le même volume $b^3/\sqrt{2}$, d'où un volume total occupé sur réseau de $0.5b^3N$. En égalant les deux, on trouve un taille équivalente de $b \sim 1.1$ nm. Cela correspond, pour les virus CCMV et STNV, à une densité équivalente sur réseau de $\rho \sim 0.3$.

La figure 5.21 A montre l'évolution de $\Delta G_{vir}/q$, ainsi que de $\Delta\Delta G_{conf}/q$ et $\Delta G_{att}/q$, en fonction de q, pour CCMV. La valeur de q pour laquelle $\Delta G_{vir}/q$ est minimale est fonction de $\Gamma_{eff}d$ et représente le nombre optimal de capsomères pour la taille de molécule à encapsuler. En dérivant l'équation 5.26, on trouve

$$q_{opt} = \left[\left(\frac{4c_1b^2}{3r_0\Gamma_{eff}d}\right) + \sqrt{\left(\frac{4c_1b^2}{3r_0\Gamma_{eff}d}\right)^2 + \left(\frac{20c_2b^3}{3r_0^2\Gamma_{eff}d}\right)N}\right]^2 \tag{5.27}$$

Connaissant la valeur de q_{opt} pour certains virus comme CCMV (32) ou STNV (12), on peut remonter à la valeur de $\Gamma_{eff}d$ qui correspond à ces minima. On trouve typiquement une valeur de 1.6 nm (voir figure 5.21 B), en accord en ordre de grandeur avec des estimations issues de la théorie des polyélectrolytes [220].

5.5 Conclusion

5.5.1 Bilan

Dans cette partie, nous avons introduit un modèle sur réseau semi-quantitatif du repliement de l'ARN, dont les paramètres ont été systématiquement dérivés à partir de données expérimentales portant sur des molécules courtes via le modèle de Turner. Comme ce dernier, le modèle inclut des paramètres locaux dépendants de la séquence pour décrire les énergies libres de formation d'un segment de paire de bases, d'une fourche,etc. ; mais ne résout pas la structure (hélicale ou dénaturée) interne des brins au niveau des bases, du squelette sucre-phosphate ou des atomes. Cependant nos résultats peuvent servir d'entrée à des programmes qui génèrent des conformations atomistiques à partir de structures secondaires [141, 144].

Le modèle sur réseau et celui de Turner étant définis à la même échelle, les paramètres similaires sont facilement identifiables. Nous avons montré en détail, quelles étaient les corrections à apporter aux paramètres de Turner pour tenir compte de l'entropie de conformation des structures secondaires sur le réseau, en groupant et en énumérant, pour des cas simples, le nombre de microétats (ou de conformations sur réseau) correspondants (il serait d'ailleurs intéressant d'étendre ce schéma systématique de paramétrisation à des représentations 3D plus complexes). D'un point de vue pratique, la paramétrisation et la simulation du modèle sur réseau est grandement facilité par un nombre de simplifications et d'unifications que nous avons apportées au modèle de Turner sur la base de considérations physiques. En particulier, nous avons dérivé des paramètres unifiés pour les termes de fourches et de nucléation de boucles qui sont indépendants de la nature de la boucle (interne, en épingle, etc.). Il est également possible via une transformation de jauge adéquate de s'affranchir du terme non-local de nucléation. Ces considérations devraient aussi s'appliquer à des modèles comme Kinefold ou Vfold qui mixent

paramètres de Turner et entropies de conformations estimées indépendamment. Les paramètres et le pouvoir de prédiction de ces approches devrait pouvoir évoluer avec la description standard considérée. Des tests simples pour vérifier la cohérence interne de leur paramétrisation seraient d'obtenir des comportements similaires pour différentes versions des paramètres de Turner et de faire des prédictions invariantes par transformation de jauge dans le modèle de Turner.

Grâce à sa définition tridimensionnelle, le modèle sur réseau tient compte d'effets génériques importants comme les effets de connectivité et de volume exclu entre tous les éléments d'une structure secondaire. Ce sont des ingrédients essentiels pour le traitement de structures complexes comme les pseudo-noeuds ou les boucles multiples. A titre de comparaison, Vfold, bien qu'ayant une meilleure résolution spatiale de la structure interne et donc décrivant a priori plus précisément ces effets génériques, est limité à un petit nombre croissant d'architectures, et Kinefold tient compte uniquement de la connectivité (voir annexe 6.6). Après une validation de notre approche sur des exemples simples, nous avons testé le pouvoir de prédiction du modèle sur réseau sur des structures complexes comme les ARNs de transferts et les pseudo-noeuds. Alors que les applications spécialisées (Nupack, pknotsRG) sont légèrement plus performantes sur la prédiction des états natifs de séquences avec pseudo-noeuds, le modèle sur réseau, tout en fournissant une description systématique et complète, possède une bonne fiabilité quelque soit le type de structure étudié. Actuellement, la limitation principale pour la prédiction des propriétés thermodynamiques de repliement est le temps de calcul nécessaire pour déterminer $\Omega(\mathcal{S})$ pour des séquences hétérogènes de taille supérieure à 80 nts. Plus la séquence est longue, plus l'espace des configurations devient important (croissance exponentielle avec N [20, 174]) et plus la conformation courante a tendance à rester piéger longtemps dans des minima secondaires, le schéma dynamique utilisé (mouvements de chaînes plus méthode multi-histogramme) n'étant plus adapté à un échantillonnage efficace de l'espace des configurations. Une solution envisageable est l'utilisation d'un algorithme cinétique (voir section 5.5.2.2).

En comparaison avec des approches standard comme RNAfold ou Nupack qui utilisent des relations de Jacobson-Stockmayer généralisées, il n'y a pas, dans le modèle sur réseau, de paramètres libres pour décrire l'entropie de conformation de la molécule repliée. Au contraire, notre représentation gros-grain tridimensionnelle nous permet de prédire les contributions génériques (polymériques) à l'énergie libre non-locale de formation de boucle Δg_{loop}, pour des structures secondaires arbitraires, et, grâce à l'algorithme de croissance de chaîne que nous avons développé pour échantillonner les conformations ayant la même structure secondaire (la méthode de la contrainte maximale), nous pouvons traiter des séquences jusqu'à environ 5000 nts. De plus, il nous est possible de paramétrer des équations de Jacobson-Stockmayer utilisables par les approches standard. En particulier, nous avons dérivé une expression pour Δg_{loop} qui tient compte, par un champ moyen, des interactions stériques de la boucle avec l'ensemble des éléments de la structure, alors que classiquement, les méthodes traitant le repliement au niveau structure secondaire, par manque de données ou de modélisation, ne considèrent que l'effet des interactions stériques internes à la boucle.

Nous avons également estimé l'importance de tenir compte rigoureusement du volume exclu sur la prédiction des propriétés de repliement. L'impact général est une déstabilisation de la molécule car le traitement complet des interactions stériques réduit l'espace accessible à la structure et donc l'entropie de conformation. A des températures où les structures majoritaires sont fortement repliées (à 37° C par exemple), l'effet est assez faible en moyenne car c'est l'appariement des nucléotides qui domine. Cependant il existe une partie non négligeable de séquences dont les prédictions sont fortement perturbées (notamment pour la structure la plus stable). Quand on s'approche des températures typiques de dénaturation ($\sim 80^o$ C), l'effet devient plus important (même pour l'évaluation de propriétés thermodynamiques globales comme Θ). A titre de comparaison, nous avons étudié la propagation des incertitudes sur les paramètres de Turner dans les résultats calculés par RNAfold. A 37° C, nous avons constaté une assez bonne fiabilité des prédictions pour les petites séquences (< 100 nts). Par contre, pour des séquences plus longues, le niveau d'incertitude est contrasté suivant l'observable considérée : pour les propriétés thermodynamiques globales (comme Θ), le niveau de confiance est assez bon ; pour

les propriétés thermodynamiques locales (comme la carte des contacts), il est moyen ; alors que pour propriétés mettant en jeu un petit nombre de structures (comme la structure la plus stable), il est assez faible. Cela soulève, comme pour la dénaturation de l'ADN, la question de la fiabilité des méthodes de prédiction de structures secondaires. Une amélioration de celle-ci passera forcément par un perfectionnement de la paramétrisation des données de Turner. Ainsi, pour l'étude d'une séquence donnée à l'aide des approches standard, il est préférable d'inclure la correction de volume exclu que nous avons paramétrée à l'aide du modèle sur réseau et d'effectuer une analyse des erreurs pour estimer la fiabilité que l'on peut avoir dans les prédictions réalisées.

Enfin, nous avons tiré profit de la définition tridimensionnelle du modèle sur réseau pour étudier l'impact sur le repliement d'interactions tertiaires comme l'interaction spécifique tetraloop/recepteur tetraloop ou le confinement dans une sphère. En particulier, nous avons interprété l'énergie libre moyenne nécessaire pour confiner une structure secondaire par une modélisation de Flory qui nous a permis de discuté l'énergie d'encapsidation d'une molécule d'ARN dans une capside virale et d'estimer l'intensité caractéristique des interactions attractives entre les parois internes de la capside et l'ARN. Cependant, l'équation 5.23 ne décrit pas les différences observées entre différentes structures secondaires de même taille. Il serait intéressant de quantifier ces différences par l'intermédiaire d'une énergie de Flory qui dépendrait des propriétés de la structure secondaire comme le nombre de boucles, leur taille, la longueur des parties double-brins,etc, dans la même esprit que les relations de Jacobson-Stockmayer. De telles considérations permettraient de comparer de manière systématique l'énergie de confinement de séquences virales à celle de séquences aléatoires de même composition [174] et d'étudier comment la structure secondaire évolue en réponse à un confinement. De plus, au lieu de considérer le repliement d'une molécule à l'intérieur d'une sphère, on pourrait également étudier l'énergie d'une molécule entourée de sphères afin de simuler l'environnement encombré d'une cellule.

Une partie de ce travail (présentation du modèle et test du pouvoir de prédiction) a abouti à la publication d'un article dans *Journal of Chemical Physics* [221].

5.5.2 Perspectives

5.5.2.1 Interactions non-spécifiques

Une des forces du modèle sur réseau est sa définition au niveau tertiaire qui nous a permis d'étudier l'effet d'interactions tertiaires de contact spécifiques entre éléments d'une même structure ou l'effet d'interactions stériques avec l'extérieur (voir section 5.4). Un autre type d'interactions tertiaires particulièrement intéressant est celui des interactions non-spécifiques. Par exemple, la stabilisation des capsules d'ARN viraux est en partie assurée par des interactions attractives non-spécifiques entre les nucléotides du brin d'ARN et les protéines formant la capsule, ainsi qu'entre les nucléotides eux-même [220, 219, 218]. Ou bien, la compactification des structures tertiaires est réalisée en partie par des interactions ARN-ARN dont l'intensité est pilotée par la concentration en magnésium [211, 222].

Prenons par exemple le cas des interactions non-spécifiques entre nucléotides d'un même brin. Pour une structure secondaire $\mathcal{S}$ fixée, chaque conformation $\mathcal{C}$ sur réseau représentant $\mathcal{S}$ peut être caractérisée par son rayon de giration $R_G(\mathcal{C})$ et le nombre total de contacts $N_c(\mathcal{C})$ entre nucléotides occupant des sites voisins sur le réseau. Une approche possible pour étudier l'effet de ces interactions sur le repliement est de déterminer $\Omega_{\mathcal{S}}(r_g, n_c)$ le nombre de conformations $\mathcal{C}$ pour lesquelles $R_G(\mathcal{C}) = r_g$ et $N_c(\mathcal{C}) = n_c$. Il sera alors possible d'étudier la compactification de la structure en fonction de l'énergie ϵ_c d'interaction par contact en regardant, par exemple, la quantité

$$\langle R_G \rangle_{\mathcal{S}}(\epsilon_c) = \frac{\sum_{(r_g,n_c)} r_g \Omega_{\mathcal{S}}(r_g, n_c) \exp[-\beta n_c \epsilon_c]}{\sum_{(r_g,n_c)} \Omega_{\mathcal{S}}(r_g, n_c) \exp[-\beta n_c \epsilon_c]} \tag{5.28}$$

Reste à estimer $\Omega_{\mathcal{S}}(r_g, n_c)$. Une première méthode consiste à enregistrer durant une simulation avec la

méthode de la contrainte maximale pour $\mathcal{S}$ l'histogramme

$$\mathcal{N}_{\mathcal{S}}(r_g, n_c) = \sum_{\{\mathcal{C}_i/(R_G(\mathcal{C}_i)=r_g, N_c(\mathcal{C}_i)=n_c)\}} W(\mathcal{C}_i) \tag{5.29}$$

alors on devrait avoir $\Omega_{\mathcal{S}} \propto \mathcal{N}_{\mathcal{S}}$. Cependant l'algorithme génère assez peu de conformations compactes car d'une part leur nombre est faible et d'autre part car ils ont une énergie de pliage totale élevée. Ainsi la statistique pour les éléments de $\Omega_{\mathcal{S}}$ correspondant à des faibles rayons de giration et des grands nombres de contacts sera mauvaise et ne permettra pas de bien décrire la transition de compactification. Pour améliorer cette statistique, l'idée est de biaiser la construction des conformations avec un poids proportionnel à l'inverse de $\Omega_{\mathcal{S}}(r_g, n_c)$ de manière à obtenir un histogramme $\mathcal{N}_{\mathcal{S}}(r_g, n_c)$ plat, afin de s'assurer que tous les états seront visités avec la même fréquence. Pour les schémas dynamiques, une méthode puissante pour évaluer une densité d'état en rendant plat un histogramme est l'échantillonnage multi-canonique [223]. Pour une distribution quelconque π, on a vu dans la section 4.3.1 qu'on pouvait l'échantillonner en utilisant la règle d'acceptance 4.14 pourvu qu'on ait à notre disposition un moyen symétrique de passer d'une configuration du système à une autre. L'histogramme $\mathcal{N}_\pi$ des configurations visitées résultant sera alors proportionnelle à la densité d'état Ω multipliée par la distribution π. Une estimation de Ω est alors donnée par $\mathcal{N}_\pi/\pi$. Le principe de la méthode multi-canonique est le suivant : on commence en prenant la distribution $\pi_0 = 1$ (ce qui revient à faire de l'échantillonnage simple), une simulation nous donne accès à $\mathcal{N}_{\pi_0}$. Puis on refait une simulation mais pour une distribution $\pi_1 = \pi_0/\mathcal{N}_{\pi_0}$ qui est l'estimation suite à la première simulation de l'inverse de Ω. Puis on réitère le processus $\pi_n = \pi_{n-1}/\mathcal{N}_{\pi_{n-1}}$ tant que l'histogramme courant $\mathcal{N}_{\pi_n}$ n'est pas plat. Une fois plat cela signifie que toutes les configurations ont été visitées à la même fréquence et donc que $1/\pi$ vaut bien Ω. Ainsi cette méthode est un excellent moyen d'évaluer des densités d'état à l'aide de schémas dynamiques. L'adaptation de cette idée aux schémas statiques PERM a été développée par Bachmann et Janke [224, 225]. Elle consiste à rajouter un biais proportionnel à l'inverse de Ω lors du choix des directions de croissance et de modifier le poids en conséquence. Coupler cette méthode avec celle de la contrainte maximale pour les structures secondaires permettait ainsi d'estimer $\Omega_{\mathcal{S}}$ et donc d'évaluer $\langle R_G \rangle_{\mathcal{S}}(\epsilon_c)$ avec l'équation 5.28 pour estimer l'impact d'une interaction locale non-spécifique sur le compactification de la molécule.

5.5.2.2 Cinétique de repliement

Dans toute l'étude que nous avons menée dans cette partie, nous avons supposé que le repliement de l'ARN était décrit à l'équilibre thermodynamique. Cette hypothèse est celle postulée par la plupart des programmes ou méthodes traitant du repliement et, bien sûr, facilite les calculs et les prédictions des observables (par rapport à un traitement hors-équilibre). Cependant sa validité dans les processus de repliement in vivo parait contestable. En effet, d'une part, l'espace des configurations devenant vite très complexe, même pour des "petites" séquences, la molécule va être piégée dans des minimas secondaires ; d'autre part, in vivo, la majorité des ARN (comme les pré-ARN messagers au moment de la transcription) se replient au fur et à mesure de leur création [226]. Etudier la cinétique de repliement est donc une étape importante dans le compréhension du phénomène de repliement.

Plusieurs approches basées sur le modèle de Turner traitent de la cinétique du repliement de l'ARN (pour un passage en revue de ces différentes méthodes voir [227]). Principalement, elles décrivent la cinétique avec une équation maîtresse dans l'espace des structures secondaires (voir annexe 6.11)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_{\mathcal{S}}(t) = \sum_{\mathcal{S}' \neq \mathcal{S}} \left[k(\mathcal{S}' \to \mathcal{S})\mathcal{P}_{\mathcal{S}'}(t) - k(\mathcal{S} \to \mathcal{S}')\mathcal{P}_{\mathcal{S}}(t) \right] \tag{5.30}$$

où $\mathcal{P}_{\mathcal{S}}(t)$ est la probabilité d'observer la structure secondaire $\mathcal{S}$ au temps t et $k(\mathcal{S} \to \mathcal{S}') > 0$ le taux de transition entre $\mathcal{S}$ et $\mathcal{S}'$. Généralement $k(\mathcal{S} \to \mathcal{S}') \neq 0$ uniquement si le passage de $\mathcal{S}$ à $\mathcal{S}'$ se fait

par l'ouverture ou la fermeture d'une (ou d'un petit nombre de) paire de bases. Communément, on les définit via une équation d'Arrhenius

$$k(\mathcal{S} \to \mathcal{S}') = k_0 \exp\left\{-\beta\left[\Delta G^*(\mathcal{S},\mathcal{S}') - \Delta G(\mathcal{S})\right]\right\} \tag{5.31}$$

où k_0 est l'inverse de l'échelle de temps typique du problème, déterminée par comparaison avec l'expérience et $\Delta G^*(\mathcal{S},\mathcal{S}')$ est l'énergie libre de l'état de transition entre $\mathcal{S}$ et $\mathcal{S}'$.

Le modèle sur réseau peut alors être utilisé pour évaluer les énergies de transition, par exemple, en échantillonnant l'espace des états en couplant les mouvements élémentaires décrits dans la section 4.3.2 à un algorithme de type Wang-Landau [228], ou en estimant $\Delta G^*(\mathcal{S},\mathcal{S}')$ à l'aide de la méthode de la contrainte maximale. Cependant, ces utilisations du modèle sur réseau, (qui ont un intérêt notamment pour le calcul des entropies de conformation de structures quelconques, en particulier de celles pour lesquelles il n'existe pas de relations de Jacobson-Stockmayer, comme les pseudo-noeuds), sont limitées en pratique à des petites séquences ($N < 30 - 40$ nts).

Pour essayer de repousser cette limite de taille, nous avons développé un algorithme de Monte-Carlo basé sur la méthode de biais configurationnel [229, 178] qui est une schéma mixte entre l'approche dynamique et l'approche statique. L'idée consiste à décrire le repliement d'une molécule comme un processus stochastique où le passage entre deux conformations consécutives se fait par l'ouverture ou la fermeture d'une paire de bases ou d'un segment de paires de bases : soit $\mathcal{C}_o$ la conformation courante représentant la structure secondaire $\mathcal{S}_o$ et soit W_o son poids de Rosenbluth, on génère tout d'abord aléatoirement une nouvelle structure secondaire $\mathcal{S}_n$ en ouvrant ou fermant une paire de base à partir de $\mathcal{S}_o$, puis en utilisant le méthode de la contrainte maximale avec l'échantillonnage de Rosenbluth (voir section 4.4.2), on construit une conformation $\mathcal{C}_n$ de poids W_n. La nouvelle conformation est accepter avec une probabilité (voir annexe 6.11)

$$\mathrm{acc}(\mathcal{C}_o \to \mathcal{C}_n) = \min\left\{1, \exp\left(-\beta\left[\Delta G_{latt}(\mathcal{S}_n) - \Delta G_{latt}(\mathcal{S}_o)\right]\right) \times \left(\frac{W_n}{W_o}\right) \times \left(\frac{\mathcal{N}_o}{\mathcal{N}_n}\right)\right\} \tag{5.32}$$

avec $\Delta G_{latt}(\mathcal{S}) = h_{latt}(\mathcal{S}) - Ts_{latt}(\mathcal{S})$ l'énergie libre sur réseau de $\mathcal{S}$ et $\mathcal{N}_o$ le nombre de structures secondaires voisines que l'on peut atteindre directement depuis $\mathcal{S}_o$.

Reste à valider la méthode en particulier en la comparant à des résultats expérimentaux afin d'évaluer le temps caractéristique correspondant à un pas de Monte-Carlo. Une fois notre approche cinétique entérinée nous espérons l'appliquer avec succès à des séquences longues d'ARN pour étudier l'influence de la vitesse de transcription d'un brin d'ARN sur son repliement, ou l'importance de pièges cinétiques (minima secondaires) lors du repliement de molécules complexes.

Conclusion générale

Pour conclure, dans une première partie, nous avons étudié la dénaturation de l'ADN, par l'intermédiaire de deux modèles au niveau structure secondaire, le modèle de Poland-Scheraga et celui de Zimm-Bragg, puis, dans une deuxième partie, nous nous somme intéressé au repliement de l'ARN par l'intermédiaire d'un modèle gros-grain de la structure tertiaire, le modèle sur réseau. Dans chacun des cas abordés, de nombreuses perspectives s'offrent à nous afin de poursuivre, d'approfondir ou d'appliquer les méthodes développées. Ces deux parties ont été l'occasion d'introduire de nombreux outils théoriques, numériques ou d'analyse que nous avons utilisés pour étudier les prédictions des modèles considérés.

Lors du développement ou de l'utilisation des approches introduites dans ma thèse, notre idée directrice était de faire des modèles génériques qui s'appliquent à une grande classe de séquences, et qui donnent des résultats quantitatifs par comparaison avec l'expérience. En particulier, bien qu'appliquer soit à l'ADN soit à l'ARN, les différents formalismes utilisés peuvent s'étendre à l'un comme à l'autre. De plus, dans chaque partie, nous avons toujours pris un soin particulier à discuter des incertitudes sur les résultats en partie dûes aux erreurs inhérentes à la paramétrisation de ces modèles.

Toutes les approches introduites dans ma thèse modélisent une configuration d'un acide nucléique comme l'agencement de parties double-brins et de parties dénaturées ou non-appariées (comme les boucles). La spécificité locale de l'ADN et de l'ARN y est décrite par des interactions de plus proche voisin (appariement et empilement). Par contre, les interactions non-locales (principalement les énergies de formation des boucles) sont considérées à des niveaux de modélisation différents. Le modèle sur réseau, grâce à sa définition tridimensionnelle, tient compte des interactions de volume exclu interne et externe à une boucle. Le modèle de Poland-Scheraga néglige ou approxime les effets stériques entre boucles. Alors que le modèle de Zimm-Bragg omet aussi les interactions internes à une boucle.

Pourquoi dès lors utiliser ces deux derniers modèles si on a à notre disposition une approche plus rigoureuse du volume exclu ? Une réponse pragmatique serait l'accélération significative des temps de calcul. Quels sont alors les conséquences de telles approximations sur la fiabilité des résultats ? On a vu que négliger les interactions steriques entre sous-parties pouvaient avoir un impact important sur les courbes de dénaturation d'un brin d'ARN prédites par le modèle de Turner. Comment justifier dans ce contexte que le modèle similaire de Poland-Scheraga décrive aussi bien la dénaturation de l'ADN si il ne tient pas compte rigoureusement du volume exclu ? Une raison possible serait la plus faible importance des interactions stériques dans un double-brin partiellement dénaturé que dans un simple-brin replié sur lui-même. En effet, la présence de boucles multiples dans les structures ARN rend ces structures topologiquement plus compactes et donc sujettes à plus d'interactions de volume exclu. De même, qu'en est-il pour l'approximation faite dans le modèle de Zimm-Bragg ? Le bon accord entre les prédictions de ce modèle et ceux du modèle de Poland-Scheraga, montre qu'elle se justifie totalement dans le cadre de l'étude de longues séquences génomiques, les effets d'hétérogénéité de séquence prédominant sur la mauvaise estimation de l'entropie de conformation des bulles. Ainsi, même si le modèle sur réseau reste le modèle le plus complet, les deux autres approches sont justifiées et efficaces dans leur domaine d'étude respectif.

Pour augmenter encore le degré de modélisation des effets de volume exclu, la prise en compte de la structure interne des parties double-brins et simple-brins semble être une étape décisive. Une idée

possible serait d'utiliser le canevas du modèle gros-grain de Vfold et de lui appliquer le formalisme systématique que nous avons développé pour paramétrer et simuler le modèle sur réseau. Une telle approche permettrait de traiter plus rigoureusement les interactions stériques et de mieux rendre compte de la structure tertiaire d'une molécule. Le prix à payer serait bien évidemment une augmentation sensible du temps de calcul.

Chapitre 6

Annexes

6.1 Algorithme de Fixman-Freire pour le modèle de Poland-Scheraga

Simplification des relations de récurrence

On simplifie tout d'abord les relations de récurrence 1.13, 1.14 et 1.15 en posant

$$Z_f^*(\alpha) = \exp\left(-\beta \sum_{i=1}^{\alpha-1} \Delta g_{NN}(i,i+1)\right) Z_f(\alpha) \tag{6.1}$$

$$Z_b^*(\alpha) = \exp\left(-\beta \sum_{i=\alpha}^{N-1} \Delta g_{NN}(i,i+1)\right) Z_b(\alpha) \tag{6.2}$$

$$Z_{sf}^*(\alpha) = \exp\left(-\beta \sum_{i=1}^{\alpha-1} \Delta g_{NN}(i,i+1)\right) Z_{sf}(\alpha) \tag{6.3}$$

On obtient alors les relations

$$Z_f^*(\alpha+1) = e^{-\beta\Delta g_{NN}(\alpha,\alpha+1)} Z_f^*(\alpha) + \sigma_0(\alpha) \sum_{\alpha'=2}^{\alpha-2} (\alpha-\alpha')^{-c} Z_f^*(\alpha') + \sigma_1(\alpha) \tag{6.4}$$

$$\begin{aligned} Z_b^*(\alpha) &= e^{-\beta\Delta g_{NN}(\alpha,\alpha+1)} Z_b^*(\alpha+1) + \sigma_0(\alpha-1) e^{\beta\Delta g_{NN}(\alpha-1,\alpha)} \\ &\quad \times \sum_{\alpha'=3}^{N} (\alpha'-\alpha-1)^{-c} e^{-\beta\Delta g_{NN}(\alpha'-1,\alpha')} Z_b^*(\alpha') \\ &\quad + \sigma_2(\alpha) e^{\beta\Delta g_{NN}(\alpha-1,\alpha)} \end{aligned} \tag{6.5}$$

$$Z_{sf}^*(\alpha) = \sigma \sum_{\alpha'=2}^{\alpha-2} (\alpha-\alpha')^{-c} Z_f^*(\alpha') + \bar{\sigma}(\alpha-1)^{c'} e^{\beta\omega_1} \tag{6.6}$$

avec

$$\begin{aligned} \sigma_0(\alpha) &= \sigma e^{-\beta\Delta g_{NN}(\alpha,\alpha+1)} \\ \sigma_1(\alpha) &= \bar{\sigma} e^{-\beta\Delta g_{NN}(\alpha,\alpha+1)+\beta\omega_1} (\alpha-1)^{c'} \\ \sigma_2(\alpha) &= \bar{\sigma} e^{-\beta\Delta g_{NN}(\alpha-1,\alpha)+\beta\omega_N} (N-\alpha)^{c'} \end{aligned}$$

Approximation de Fixman-Freire

On remplace alors les termes de la forme x^{-c} par $\sum_{k=1}^{I} a_k e^{-b_k\, x}$ dans les relations de récurrence précédentes et on introduit de nouvelles variables définies par $(1 \leq i \leq I)$

$$e^{b_i \alpha}\, e^{\mu_i(\alpha)} \equiv \sum_{\alpha'=2}^{\alpha-2} e^{b_i \alpha'} Z_f^*(\alpha') \tag{6.7}$$

$$e^{-b_i \alpha}\, e^{\nu_i(\alpha)} \equiv \sum_{\alpha'=\alpha}^{N} e^{-b_i \alpha'} e^{-\beta \Delta g_{NN}(\alpha'-1,\alpha')} Z_b^*(\alpha') \tag{6.8}$$

Comme on a

$$Z_f^*(\alpha) = e^{\mu_i(\alpha)} - e^{-b_i}\, e^{\mu_i(\alpha-1)} \tag{6.9}$$

$$Z_b^*(\alpha) = e^{\beta \Delta g_{NN}(\alpha-1,\alpha)} \left[e^{\nu_i(\alpha)} - e^{-b_i}\, e^{\nu_i(\alpha+1)} \right] \tag{6.10}$$

$$Z_{sf}^*(\alpha) = \sigma \sum_{k=1}^{I} a_k e^{-2b_k} e^{\mu_k(\alpha-2)} + \bar{\sigma}(\alpha-1)^{c'} e^{\beta \omega_1} \tag{6.11}$$

Les μ_i et ν_i vérifient alors les relations de récurrence

$$\mu_i(\alpha+1) = \mu_i(\alpha) + \log(A+B+C+D) \tag{6.12}$$

$$\nu_i(\alpha) = \nu_i(\alpha+1) + \log(A'+B'+C'+D') \tag{6.13}$$

avec

$$A = e^{-b_i} \qquad B = e^{-\beta \Delta g_{NN}(\alpha,\alpha+1)} \left(1 - e^{-b_i}\, e^{\mu_i(\alpha-1)-\mu_i(\alpha)}\right)$$

$$C = \sigma_0(\alpha) \sum_{k=1}^{I} a_k e^{-2b_k} e^{\mu_k(\alpha-2)-\mu_i(\alpha)} \qquad D = \sigma_1(\alpha) e^{-\mu_i(\alpha)}$$

$$A' = e^{-b_i} \qquad B' = e^{-\beta \Delta g_{NN}(\alpha-1,\alpha)} \left(1 - e^{-b_i}\, e^{\nu_i(\alpha+2)-\nu_i(\alpha+1)}\right)$$

$$C' = \sigma_0(\alpha-1) \sum_{k=1}^{I} a_k e^{-2b_k} e^{\nu_k(\alpha+3)-\nu_i(\alpha+1)} \qquad D' = \sigma_2(\alpha) e^{-\nu_i(\alpha+1)}$$

L'algorithme finale consiste donc à résoudre les relations de récurrence 6.12 et 6.13, d'en déduire les Z_f^*, Z_b^* et Z_{sf}^* grâce aux équations 6.9, 6.10 et 6.11, puis de remonter aux Z_f, Z_b et Z_{sf} en inversant les relations 6.1, 6.2 et 6.3, enfin de calculer $p(\alpha)$, Z (équations 1.16 et 1.17) pour finalement en déduire Θ_{int} (Eq.1.18), ΔG_{int} (Eq.1.19), ΔG_0 (Eq.1.8), Θ_{ext} (Eq.1.10) et Θ (Eq.1.11).

Conditions initiales des récurrences

Pour résoudre les équations 6.12 et 6.13, on définit les conditions initiales des récurrences. Comme

$$Z_f^*(2) = e^{-\beta \Delta g_{NN}(1,2)} \tag{6.14}$$

$$Z_f^*(3) = e^{-\beta[\Delta g_{NN}(2,3)+\Delta g_{NN}(1,2)]} + \bar{\sigma} e^{-\beta[\Delta g_{NN}(2,3)-\omega_1]} \tag{6.15}$$

$$Z_f^*(4) = e^{-\beta[\Delta g_{NN}(3,4)+\Delta g_{NN}(2,3)+\Delta g_{NN}(1,2)]} + \bar{\sigma} e^{-\beta[\Delta g_{NN}(3,4)+\Delta g_{NN}(2,3)-\omega_1]} + \bar{\sigma} 2^{\zeta-1} e^{-\beta[\Delta g_{NN}^0(3,4)-\omega_1]} \tag{6.16}$$

$$Z_b^*(N) = 1 \tag{6.17}$$

$$Z_b^*(N-1) = e^{-\beta \Delta g_{NN}(N-1,N)} + \bar{\sigma} e^{\beta \omega_N} \tag{6.18}$$

$$Z_b^*(N-2) = e^{-\beta[\Delta g_{NN}(N-2,N-1)+\Delta g_{NN}(N-1,N)]} + \bar{\sigma} e^{-\beta[\Delta g_{NN}(N-2,N-1)-\omega_N]} + \bar{\sigma} 2^{\zeta-1} e^{\beta \omega_N} \tag{6.19}$$

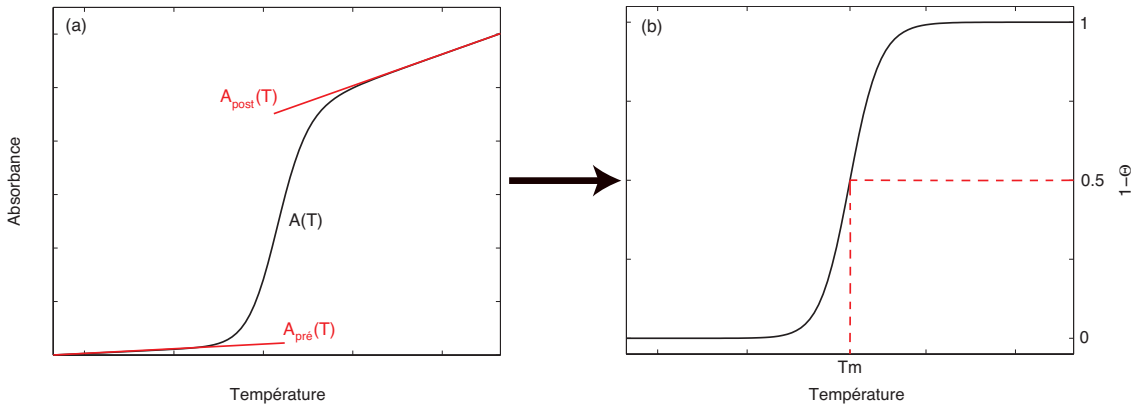

FIGURE 6.1 – (a) Exemple typique de courbe expérimentale d'absorbance en UV $A(T)$ (ligne noire). Les fonction linéaires pré- et post-transitionnelles $A_{\text{pré}}$ et A_{post} sont tracées en rouge. (b) Courbe $1-\Theta$ après traitement de la courbe expérimentale.

On trouve

$$\mu_i(2) = \log\left[Z_f^*(2)\right] \tag{6.20}$$

$$\mu_i(3) = \log\left[Z_f^*(3) + e^{-b_i} Z_f^*(2)\right] \tag{6.21}$$

$$\mu_i(4) = \log\left[Z_f^*(4) + e^{-b_i} Z_f^*(3) + e^{-2b_i} Z_f^*(2)\right] \tag{6.22}$$

$$\nu_i(N) = \log\left[Z_b^*(N) e^{-\beta\Delta g_{NN}(N-1,N)}\right] \tag{6.23}$$

$$\nu_i(N-1) = \log\left[e^{-\beta\Delta g_{NN}(N-2,N-1)} Z_b^*(N-1) + e^{-b_i} e^{-\beta\Delta g_{NN}(N-1,N)} Z_b^*(N)\right] \tag{6.24}$$

$$\begin{aligned}\nu_i(N-2) = \log\Big[&e^{-\beta\Delta g_{NN}(N-3,N-2)} Z_b^*(N-2) + e^{-b_i} e^{-\beta\Delta g_{NN}(N-2,N-1)} Z_b^*(N-1) \\ &+ e^{-2b_i} e^{-\beta\Delta g_{NN}(N-1,N)} Z_b^*(N)\Big]\end{aligned} \tag{6.25}$$

6.2 Exploitation des courbes expérimentales de dénaturation obtenues par absorbance d'UV

Dans cette annexe, nous expliquons la méthode classique [17] pour traiter les courbes expérimentales d'absorbance en UV afin d'obtenir des données comparables avec la probabilité $1-\Theta$ qu'une paire soit ouverte, calculée à partir du modèle de Poland-Scheraga, du modèle de Turner ou du modèle sur réseau. L'absorbance des UVs est très souvent utilisée pour étudier la dénaturation thermique des molécules d'acides nucléiques. Cette technique expérimentale est basée sur le principe qu'une paire fermé a une absorbance beaucoup plus faible qu'une paire ouverte. Elle décrit ainsi la probabilité $1-\Theta$. La figure 6.1 (a) montre une courbe typique d'évolution de l'absorbance $A(T)$ des UV par une solution d'oligomères courts d'ADN en fonction de la température. Dans le régime pré-transitionnelle, on observe une légère augmentation linéaire de l'absorbance (caractérisée par la fonction $A_{\text{pré}}$) à cause d'une faible perturbation de l'empilement moyen dans la double-hélice. Après la transition, on observe également une augmentation linéaire de l'absorbance (caractérisée par A_{post}) à cause du désempilement progressif des bases des simples brins [230]. Pour retrouver $1-\Theta$ à partir des données expérimentales, il faut donc éliminer ces effets linéaires à basses et hautes températures avec (voir figure 6.1 (b))

$$1-\Theta = \frac{A(T) - A_{\text{pré}}(T)}{A_{\text{post}}(T) - A_{\text{pré}}(T)} \tag{6.26}$$

6.3 Pertinence d'une modélisation

Soit $\{x_i, y_i\}$ N données (avec une déviation standard σ_i) et $y(x; a_1, ...a_M)$ leur modélisation qui dépend des M paramètres $\{a_1, ..., a_M\}$ déterminés par exemple grâce à la méthode du chi2 [93]. Quand le modèle est linéaire par rapport aux paramètres, la probabilité Q pour que le chi2 soit plus grand qu'une valeur particulière χ^2 par chance est donnée par

$$Q = \text{gammainc}(0.5\nu, 0.5\chi^2) = \frac{\int_{\chi^2/2}^{+\infty} e^{-t} t^{\nu/2-1} \mathrm{d}t}{\int_0^{+\infty} e^{-t} t^{\nu/2-1} \mathrm{d}t} \tag{6.27}$$

avec $\nu = N - M$ le nombre de degré de liberté et

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1, ..., a_M)}{\sigma_i} \right)^2 \tag{6.28}$$

Il est assez commun (et vérifié en général) de considérer que Q a aussi un sens pour les modèles non strictement linéaire par rapport aux $\{a_j\}$.

Quantifier si le modèle choisi permet de décrire fidèlement les données observées peut se faire par l'intermédiaire de Q : i)si Q est trop petit ($< 0.01 - 0.1$), la modélisation est probablement mauvaise ; (ii) par contre, Q proche de 1 est la signature d'une bonne modélisation. Si les barres d'erreurs expérimentales sont grandes, cela aura pour effet de diminuer χ^2 et donc d'augmenter Q. Ainsi, il est plus facile d'avoir un "bon" modèle qui reproduit bien les observables quand les barres d'erreurs des données sont importantes.

6.4 Distributions de probabilité dans le cadre de l'analyse thermodynamique des génomes

Distributions aléatoires des propriétés génomiques le long de la séquence

Pour une distribution aléatoire des paires de bases codantes dans la séquence, le nombre de paires de bases fermées qui correspondent effectivement à une paire codante vaut

$$N_{VP}^r = N_{ferm} \times \frac{N_{codant}}{N} = \%CDS \times N_{ferm} \tag{6.29}$$

où N_{ferm} est le nombre total de paires fermées et $N_{codant} = \%CDS \times N = N_{VP}^r + N_{FN}^r$ le nombre de paires codantes. Ainsi

$$\beta^r = \frac{N_{TP}^r}{N_{codant}} = \frac{N_{ferm}}{N} \tag{6.30}$$

De même,

$$N_{FN}^r = \%CDS \times (N - N_{ferm}) \tag{6.31}$$
$$N_{VN}^r = (1 - \%CDS) \times (N - N_{ferm}) \tag{6.32}$$
$$N_{FP}^r = (1 - \%CDS) \times N_{ferm} \tag{6.33}$$

soit

$$\alpha^r \equiv \frac{N_{VN}^r}{N_{non-codant}} = 1 - \beta^r \tag{6.34}$$
$$\tau^r \equiv \frac{N_{VP}^r + N_{VN}^r}{N} = \%CDS \times \beta^r + (1 - \%CDS)\alpha^r \tag{6.35}$$

De plus, le système

$$N = N_{VP} + N_{FP} + N_{VN} + N_{FN} \quad (6.36)$$
$$N_{ferm} = N_{VP} + N_{FP} \quad (6.37)$$
$$N_{codant} = N_{VP} + N_{FN} \quad (6.38)$$
$$N\tau = N_{VN} + N_{VP} \quad (6.39)$$

donne

$$N_{VN} = \frac{1}{2}\left(N\tau + N - N_{codant} - N_{ferm}\right) \quad (6.40)$$
$$N_{VP} = \frac{1}{2}\left(N\tau - N + N_{codant} + N_{ferm}\right) \quad (6.41)$$

d'où, en remplaçant τ par $\Delta\tau + \tau^r$ et en utilisant les équations 6.30-6.41, on trouve

$$\Delta\alpha \equiv \alpha - \alpha^r = \frac{N_{VN}}{(1 - \%CDS)N} - \alpha^r = \frac{\Delta\tau}{2(1 - \%CDS)} \quad (6.42)$$
$$\Delta\beta \equiv \beta - \beta^r = \frac{N_{VP}}{\%CDS \times N} - \beta^r = \frac{\Delta\tau}{2\%CDS} \quad (6.43)$$

Chevauchement de deux distributions gaussiennes

Soit deux distributions gaussiennes $P_1(x)$ et $P_2(x)$:

$$P_i(x) = \frac{1}{w_i\sqrt{2\pi}} \exp\left[-\frac{(x - x_i)^2}{2w_i^2}\right] \quad (6.44)$$

avec x_i et w_i la valeur moyenne et l'écart-type de la distribution P_i. Le chevauchement entre les deux distributions est quantifiable par l'intermédiaire de l'intégrale de recouvrement

$$I = \int P_1(x)P_2(x)\mathrm{d}x \quad (6.45)$$
$$= \frac{1}{\sqrt{w_1^2 + w_2^2}} \exp\left[-\frac{(x_1 - x_2)^2}{2(w_1^2 + w_2^2)}\right] \quad (6.46)$$

Dans le cadre de l'analyse thermodynamique des génomes, si on pose $1 = cod$, $2 = reste$ et $x = T_m$, on reconnaît dans l'expression 6.46 de I (argument de l'exponentielle) le paramètre de chevauchement $O_{T_m} = \Delta T/\sqrt{w_{cod}^2 + w_{reste}^2}$. Ainsi, plus O_{T_m} sera petit et plus le chevauchement I sera important.

Théorème de Bayes

Soit x et y deux variables aléatoires, la probabilité jointe $P(x, y)$ d'avoir x et y se décompose comme le produit de la probabilité conditionnelle $P(y|x)$ d'avoir y sachant x et de la probabilité a priori $P(x)$ d'avoir x, soit

$$P(x, y) = P(y|x)P(x) \quad (6.47)$$

De même, on a aussi

$$P(x, y) = P(x|y)P(y) \quad (6.48)$$

D'où, en égalant les deux équations 6.47 et 6.48, on obtient la relation entre les deux probabilités conditionnelles, appelée théorème de Bayes :

$$P(x|y) = \left[\frac{P(x)}{P(y)}\right] P(y|x) \quad (6.49)$$

6.5 Dénaturation de l'ADN surenroulé

Modèle de Benham

Dans le modèle de Benham, pour chaque état de la molécule d'ADN, l'invariant topologique α (la différence en nombre de raccordement) peut être décomposé en trois contributions :

- la dénaturation de n_o paires de bases relaxe l'hélicité par un facteur $-n_o/A$ où $A = 10.4$ bp/tour est le nombre moyen de paires de bases présents dans un tour de la double-hélice ;
- les régions dénaturées correspondantes peuvent s'enrouler sur elles-mêmes générant un twist global T ;
- le pliage et l'enroulement des parties double-brins sont compris dans la différence résiduelle d'enroulement α_r.

L'invariance de α impose

$$\alpha = -\frac{n_o}{A} + T + \alpha_r \tag{6.50}$$

La description des énergies d'association et de dénaturation des paires de bases est fournie par le modèle de Zimm-Bragg (voir section 2.1), et donc par l'Hamiltonien $\mathcal{H}_{ZB}$ défini par l'équation 2.2. Pour les régions dénaturées, la grande flexibilité des simples-brins d'ADN leurs permet de s'entortiller. L'énergie associée au twist τ_i (en radian/bp) d'une paire de bases ouvertes i vaut

$$\mathcal{H}_{tw}(\theta_i, \tau_i) = \frac{C}{2}(1-\theta_i)\tau_i^2 \tag{6.51}$$

avec $\theta_i = 0$ (1) si i est ouverte (fermée) et C la rigidité de torsion qui vaut environ $3.1 k_B T$ [123]. Les twists individuels τ_i sont reliés au twist global T par la relation

$$T = \sum_i (1-\theta_i)\frac{\tau_i}{2\pi} \tag{6.52}$$

Pour les parties double-brins, il a été trouvé expérimentalement que les déformations superhélicales induisaient une énergie quadratique en α_r

$$\mathcal{H}_r = \frac{K}{2}\alpha_r^2 = \frac{K}{2}\left(\alpha + \frac{n_o}{A} - T\right)^2 \tag{6.53}$$

avec $n_o = \sum_i (1-\theta_i)$ le nombre total de paires de bases ouvertes et $K \approx 2220 k_B T/N$ [123].
L'Hamiltonien total du système vaut alors

$$\mathcal{H}(\{\theta_i, \tau_i\}) = \mathcal{H}_{ZB}(\{\theta_i\}) + \mathcal{H}_r(n_o, \{\tau_i\}) + \sum_{i=1}^{N} \mathcal{H}_{tw}(\theta_i, \tau_i) \tag{6.54}$$

et la fonction de partition

$$Z = \sum_{\theta_1=0,1} ... \sum_{\theta_N=0,1} \int_{-\infty}^{+\infty} \mathrm{d}\tau_1 ... \int_{-\infty}^{+\infty} \mathrm{d}\tau_N \exp\left[-\beta\mathcal{H}(\{\theta_i, \tau_i\})\right] \tag{6.55}$$

$$= \sum_{\{\theta_i\}} \mathcal{Q}(n_o) \exp\left[-\beta\mathcal{H}_{ZB}(\{\theta_i\})\right] \qquad \text{avec} \tag{6.56}$$

$$\mathcal{Q}(n_o) = \left(\left[\frac{2\pi}{\beta C}\right]^{n_o} \frac{4\pi^2 C}{4\pi^2 C + K n_o}\right)^{1/2} \exp\left(\frac{-2\pi^2 \beta C K}{4\pi^2 C + K n_o}\left[\alpha + \frac{n_o}{A}\right]^2\right) \tag{6.57}$$

où le passage entre la première et la deuxième ligne a été obtenu en intégrant sur les τ_i [124]. A partir de l'équation 6.56, on définit un Hamiltonien effectif

$$\mathcal{H}_{eff} = \mathcal{H}_{ZB} + \mathcal{H}^*(n_o) \tag{6.58}$$

qui ne dépend plus que des θ_i et où l'on a posé $\mathcal{H}^* \equiv -k_B T \log \mathcal{Q}$. Ainsi, calculer la valeur moyenne thermodynamique d'une observable ne dépendant que des θ_i pour l'Hamiltonien total revient à la calculer pour $\mathcal{H}_{eff}$.

Approximation de champ moyen

Pour résoudre le système avec l'Hamiltonien $\mathcal{H}_{eff}$, on fait une approximation de champ moyen qui nous permettra d'utiliser directement la méthode rapide des matrices de transfert. On développe $\mathcal{H}^*(n_o)$ autour d'une valeur typique $\bar{n}_o$:

$$\mathcal{H}^*(n_o) \approx \mathcal{H}^*(\bar{n}_o) + (n_o - \bar{n}_o)\frac{\partial \mathcal{H}^*}{\partial n_o}(\bar{n}_o) \tag{6.59}$$

Sous cette approximation, l'Hamiltonien effectif prend la forme typique d'un modèle d'Ising

$$\mathcal{H}_{eff}(\{\theta_i\}, \bar{n}_o) \approx \mathcal{H}^*(\bar{n}_o) + (N - \bar{n}_o)h + \sum_{i=1}^{N} \left[(\Delta g_{NN}(i, i+1) - 2\gamma)\, \theta_i \theta_{i+1} + (2\gamma - h)\, \theta_i \right] \tag{6.60}$$

où $h = (\partial \mathcal{H}^*/\partial n_o)(\bar{n}_o)$ est le champ moyen. Pour une certaine valeur de $\bar{n}_o$, on peut donc estimer, avec la méthode des matrices de transfert (voir section 2.1.2), $\langle n_o \rangle$ la valeur moyenne thermodynamique de n_o correspondant à l'Hamiltonien défini par l'équation 6.60. Pour être auto-consistant, la meilleure approximation de champ moyen sera obtenue en résolvant

$$\langle n_o \rangle(\bar{n}_o) = \bar{n}_o \tag{6.61}$$

Résoudre cette dernière équation revient à recherche la racine de la fonction

$$f(h) = h - \left(\frac{\partial \mathcal{H}^*}{\partial n_o} \right) (\langle n_o \rangle(h)) \tag{6.62}$$

car $(\partial \mathcal{H}^*/\partial n_o)(n_o)$ est une fonction monotone de n_o.
Pour des valeurs fixées de la température et de la différence d'enroulement, la racine de $f(h)$ est estimée efficacement en couplant la méthode de la bissection à celle de Newton-Raphson [93]. Sachant que l'évaluation de la fonction f requière un appel à la méthode des matrices de transfert ($\mathcal{O}(N)$), l'algorithme de recherche de racine que nous avons implémenté requiert typiquement 10-20 évaluations de f pour obtenir une précision relative de 10^{-4} sur la racine, pour $N \in 10^3 - 10^6$ bp. Une fois la meilleure valeur de h déterminée, on calcule les propriétés que l'on désire étudier pour le séquence considérée.

6.6 Descriptions de quelques approches basées sur le modèle de Turner

Nous décrivons ici les approches auxquelles nous avons comparé le modèle sur réseau dans le chapitre 5 et qui sont basées également sur le modèle de Turner.

RNAfold

RNAfold fait partie du package ViennaRNA développé par Hofacker, Studler, Schuster et collaborateurs [231, 86]. Il fournit la structure la plus stable sans pseudo-noeud prédite par le modèle de Turner classique [85] (voir section 3.1.1) et mesure la carte des contacts, soit la probabilité pour chaque paire de nucléotides d'être appariée, ainsi que l'énergie libre totale. Par défaut, il utilise la version 3.0 des paramètres de Turner. On présente par la suite le principe des algorithmes utilisés par RNAfold (j'ai délibérément choisi une version simple où l'énergie des boucles multiples ne dépend pas de la taille de la boucle, ni du nombre de parties double-brins connectées).

Le calcul de la structure la plus stable est réalisé à l'aide d'un algorithme récursif développé par Zuker et Stiegler [232]. L'algorithme consiste pour chaque sous-séquence $S_{i,j}$ (formée par les nucléotides

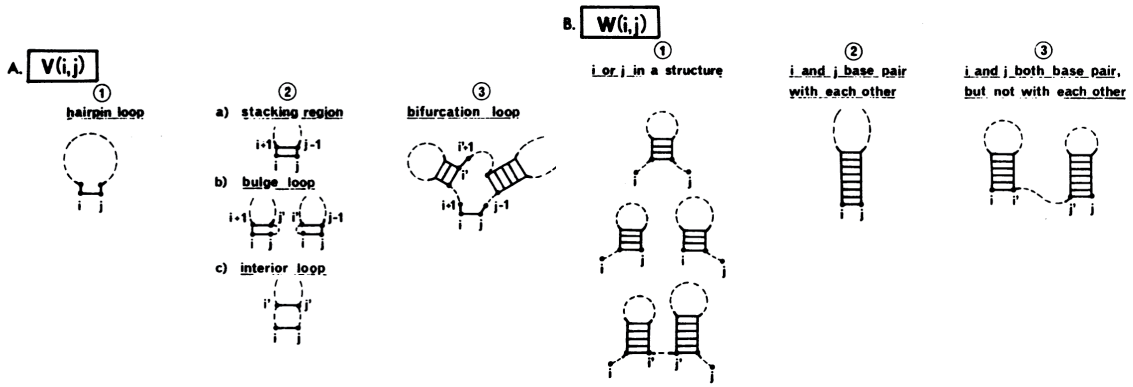

Figure 6.2 – (A) Description des 3 situations possibles pour les sous-structures de $S_{i,j}$ avec i et j appariés. (B) Pareil que A mais pour une sous-structure quelconque. Les figures sont extraites de [232].

compris entre i et j, $1 \leq i < j \leq N$) à calculer $W(i,j)$ l'énergie minimale parmi toutes les structures possibles de $S_{i,j}$ et $V(i,j)$ l'énergie minimale parmi les structures possibles de $S_{i,j}$ pour lesquelles i et j sont appariés. Si i et j ne sont pas complémentaires, $V(i,j) = +\infty$. D'après les 3 situations possibles pour $V(i,j)$ (voir figure 6.2 A), on a

$$\begin{aligned} V(i,j) &= \min\Bigg\{\Delta G_{hp}(i,j), \min_{i<i'<j'<j}\left[\Delta G_{int}(i,j;i',j') + V(i',j')\right], \\ &\qquad \min_{i+1<i'<j-2}\left[W(i+1,i') + W(i'+1,j-1) + \Delta g_{ml}\right]\Bigg\} \end{aligned} \tag{6.63}$$

avec $\Delta G_{hp}(i,j)$ l'énergie libre de la boucle en épingle fermée en i et j, $\Delta G_{int}(i,j;i',j')$ l'énergie libre du segment de paires, de la boucle interne ou de la boucle latérale compris entre les paires $i-j$ et $i'-j'$ et Δg_{ml} l'énergie libre de nucléation d'une boucle multiple. De même pour $W(i,j)$ (voir figure 6.2 B)

$$W(i,j) = \min\left\{W(i+1,j), W(i,j-1), V(i,j), \min_{i<i'<j-1}\left[W(i,i') + W(i'+1,j)\right]\right\} \tag{6.64}$$

L'énergie minimale pour la séquence entière est alors donnée par $W(1,N)$. Pour retrouver la structure la plus stable, il suffit à chaque étape du calcul de W et V de garder en mémoire quels termes étaient le plus petit et de reconstruire ainsi les paires appariés. La complexité de l'algorithme est $\mathcal{O}(N^4)$ mais en limitant la taille maximale des boucles, on peut la réduire à $\mathcal{O}(N^3)$.

Le calcul de la carte des contacts et de l'énergie libre est également fourni par un algorithme récursif développé par McCaskill [233]. Il consiste à évaluer récursivement la fonction de partition $Q(i,j)$ calculée sur toutes les structures de $S_{i,j}$ et $Q^b(i,j)$ la fonction de partition évaluée uniquement sur les structures où i et j sont appariés. Si i et j ne sont pas complémentaires, $Q^b(i,j) = 0$. En faisant le même raisonnement que pour la détermination de la structure la plus stable, on obtient

$$\begin{aligned} Q^b(i,j) &= \exp\left[-\beta\Delta G_{hp}(i,j)\right] + \sum_{i<i'<j'<j} Q^b(i',j')\exp\left[-\beta\Delta G_{int}(i,j;i',j')\right] \\ &\quad + \sum_{i+1<i'<j-2} Q(i+1,i')Q(i'+1,j-1)\exp\left[-\beta\Delta g_{ml}\right] \end{aligned} \tag{6.65}$$

$$Q(i,j) = Q(i+1,j) + Q(i,j-1) + Q^b(i,j) + \sum_{i<i'<j-1} Q(i,i')Q(i'+1,j) \tag{6.66}$$

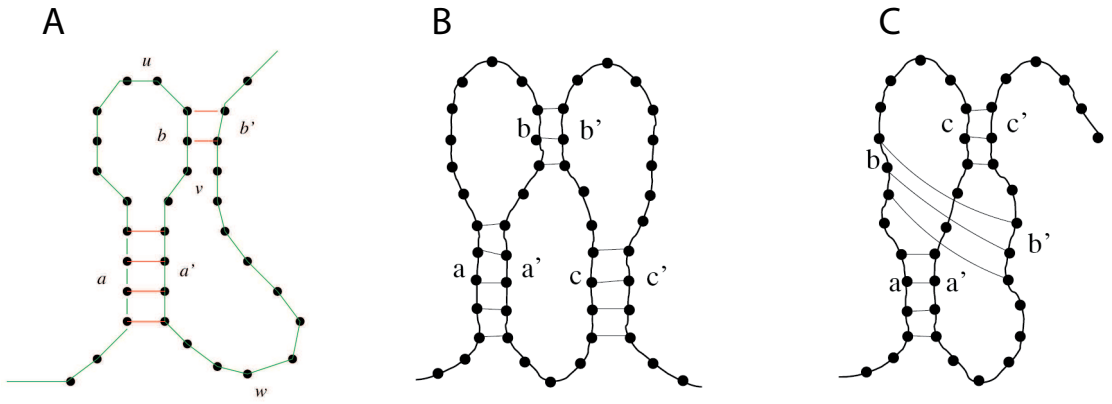

FIGURE 6.3 – (A) Pseudo-noeud simple formé de deux hélices $a-a'$ et $b-b'$ et de 3 parties intercalées u, v et w qui peuvent avoir une structure interne. (B) Appariement entre deux boucles en épingle. (C) Interaction entre 3 hélices. Les figures sont issues de [200].

On peut alors en déduire la probabilité $P(i,j)$ pour que i et j soient appariés par la relation de récurrence

$$\begin{aligned} P(i,j) &= \frac{Q(1,i-1)Q^b(i,j)Q(j+1,N)}{Q(1,N)} + \sum_{i'<i<j<j'} P(i',j')\frac{Q^b(i,j)}{Q^b(i',j')}\left\{\exp\left[-\beta\Delta G_{int}(i',j';i,j)\right] + \right. \\ &\left. \exp\left[-\beta\Delta g_{ml}\right]\left[Q(i'+1,i-1)+Q(j+1,j'-1)+Q(i'+1,i-1)Q(j+1,j'-1)\right]\right\} \end{aligned} \tag{6.67}$$

La compléxité de cet algorithme est aussi en $\mathcal{O}(N^4)$.

RNAfold ainsi que certains autres programmes du package ViennaRNA sont accessibles sur le web à l'adresse http ://rna.tbi.univie.ac.at/.

Nupack et pknotsRG

Ces deux algorithmes permettent d'étudier le repliement de l'ARN au niveau structure secondaire avec la possibilité d'avoir des pseudo-noeuds. Dans les deux cas, seules certaines classes de pseudo-noeuds sont considérées. Par exemple pour pknotsRG, on se limite aux pseudo-noeuds avec uniquement deux parties double-brins enchevêtrées (voir figure 6.3 A) où les parties u, v et w peuvent avoir une structure interne mais ne peuvent pas interagir avec d'autres parties de la structure secondaire totale (des pseudo-noeuds comme l'appariement entre deux boucles en épingles ou l'interaction à trois hélices ne sont donc pas pris en compte, voir figure 6.3 B et C). Nupack, développé par Dirks et Pierce [201], et pknotsRG, développé par Reeder et Giegerich [200], utilisent, comme RNAfold, des algorithmes récursifs pour traiter le problème du repliement. Ils introduisent, de plus, des relations de Jacobson-Stockmayer généralisées pour décrire l'énergie libre de formation d'un pseudo-noeud. Les paramètres de ces relations sont évalués de telle sorte que les algorithmes prédisent la structure native de pseudo-noeuds connus [210] comme les signaux viraux de décalage du cadre ouvert de lecture. Le reste de l'énergie d'une structure est décrit par les paramètres de la version 3.0 du modèle de Turner.

Nupack permet en $\mathcal{O}(N^5)$ de calculer la structure la plus stable et la carte des contacts, alors que pknotsRG n'évalue que la structure la plus stable en $\mathcal{O}(N^4)$.

Nupack et pknotsRG sont accessibles sur le web respectivement à http ://www.nupack.org/ et http ://bibiserv.techfak.uni-bielefeld.de/pknotsrg/.

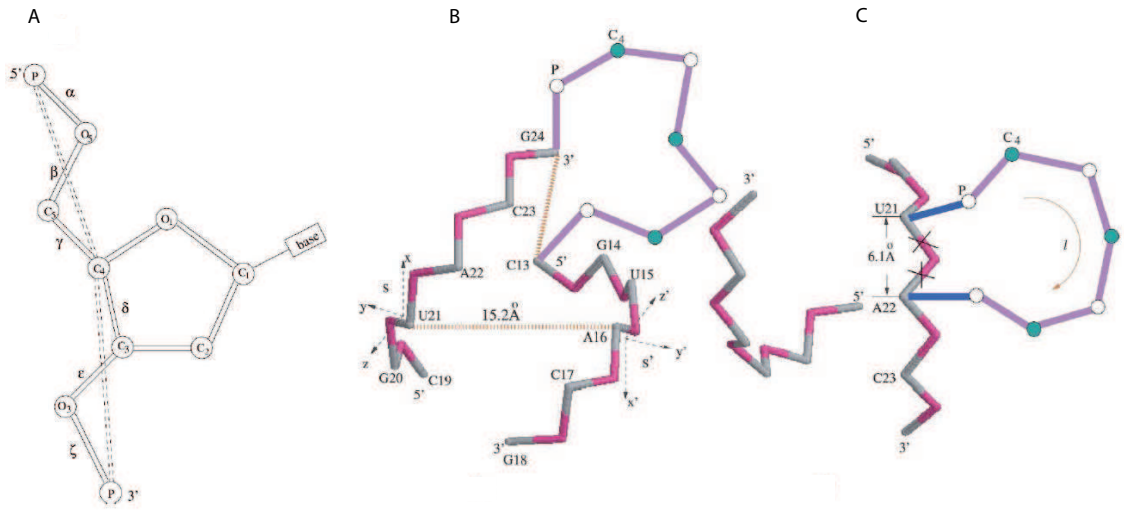

Figure 6.4 – (A) Vecteurs virtuels représentant le squelette d'un nucléotide. (B,C) Conformations possibles sur le réseau du diamant d'une boucle en épingle (B) ou d'une boucle latérale (C). Les figures sont issues de [132].

Vfold

Vfold est un modèle développé par Cao et Chen [132, 151, 152] qui permet de calculer la structure la plus stable, la carte des contacts et l'énergie libre totale en se limitant, dans sa version la plus récente, aux structures standard (sans pseudo-noeud) ou avec H-pseudo-noeuds simples (voir figure 5.11). L'algorithme consiste, comme les modèles précédents, en une récurrence sur des fonctions partielles en $\mathcal{O}(N^5)$.

L'originalité de ce modèle vient de la description de l'énergie d'une structure. Alors que les énergies locales (comme l'association de segments de paires de bases, les fourches, les dangles, l'empilement coaxial, le capping) sont fournies pas l'ancienne version 2.3 des paramètres du modèle de Turner [167], les énergies de conformation des boucles (il n'y a pas de terme pour décrire la nucléation) sont calculées à l'aide d'une modélisation gros-grain de la structure tertiaire. Pour un type de boucle donné, Cao et Chen énumèrent le nombre de conformations Ω représentant cette boucle. L'énergie de formation de la boucle est alors donnée par $\Delta g_{loop} = -k_B T \log(\Omega/\Omega_0)$ avec Ω_0 le nombre de conformation d'un simple brin dénaturé de même taille que la boucle.

La modélisation gros-grain consiste à décrire un nucléotide non apparié par deux vecteurs virtuels (voir figure 6.4 A). Une boucle est alors représentée par un chemin auto-évitant des vecteurs virtuels sur le réseau du diamant. Le choix de ce réseau est motivé par ses angles de torsion qui sont les mêmes que les états rotationnels isomériques $gauche^+$, $trans$ et $gauche^-$ des polymères. La jonction entre les parties double-brins et les boucles est modélisée en représentant la double-hélice sur le réseau du diamant en prenant les sites du réseau les plus proches des coordonnées atomiques réelles (voir figure 6.4 B et C). Ainsi Ω est calculé en comptant le nombre de chemin auto-évitant sur ce réseau.

Vfold n'est pas accessible en libre service sur le web.

Kinefold

Kinefold est un modèle au niveau structure secondaire développé par Isambert et Siggia [138, 140]. Il ne se restreint pas à certains types de structures comme les modèles précédents et peut a priori décrire n'importe quelle structure (comme le modèle sur réseau). Comme dans Vfold, les énergies locales sont données par la version 2.3 des paramètres du modèle de Turner. Pour calculer l'énergie de conformation

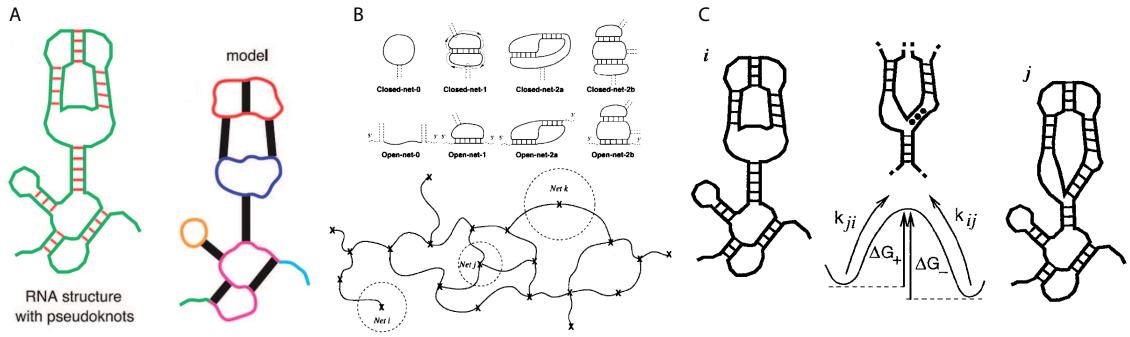

FIGURE 6.5 – (A) Modélisation par des tiges rigides (lignes noires) et des chaînes gaussiennes (lignes colorées) d'une structure secondaire. (B) Haut : les 8 sous-structures possibles dont on peut calculer exactement l'entropie de conformation. Bas : représentation sous forme d'un polymère réticulé où chaque sous-structure est modélisée par un "net" ; les contraintes entre sous-structures sont représentés par des ressorts gaussiens. (C) Schéma de la transition entre les structures i et j. Les figures sont issues de [138, 139]

d'une structure (pas de nucléation comme dans Vfold), chaque partie double-brin est modélisée par une tige rigide de longueur appropriée (2.5 Å par segment de paires) et les régions non-appariés par une chaîne gaussienne de longueur de Kuhn de 1.5 nm (2.5 bases) (voir figure 6.5 A). L'entropie de conformation est alors évalué en deux étapes (voir figure 6.5 B) : des sous-structures sont définies pour lesquelles on peut calculer exactement l'entropie, puis les contraintes entre ces sous-structures sont traitées globalement comme dans un gel réticulé par une série de $n-1$ intégrations algébriques avec n le nombre de sous-structures.

L'algorithme consiste à faire évoluer cinétiquement une structure à l'aide de mouvements élémentaires (ouverture ou fermeture d'une ou plusieurs paires de bases). Chaque mouvement est caractérisé par son taux de transition $k_\pm = k^o \exp[-\beta \Delta G_\pm]$ (+ pour une formation de paires et − pour une ouverture) avec $k^o \approx 10^8$ s^{-1} et où $\Delta G_\pm$ est la différence d'énergie entre le configuration courante et l'état de transition et se calcule à l'aide des paramètres locaux et de l'énergie de conformation décrits ci-dessus (voir figure 6.5C). A chaque étape de l'algorithme, on calcule, pour tous les mouvements élémentaires possibles, leurs taux de transition et on en choisit un aléatoirement proportionnellement à son taux de transition. Cette procédure est répétée jusqu'à atteindre l'équilibre thermodynamique.

Kinefold permet ainsi d'estimer la structure la plus stable, le chemin de repliement cinétique de la molécule et des propriétés thermodynamiques comme la carte des contacts. Il est accessible sur le web à l'adresse http ://kinefold.curie.fr/.

6.7 Preuve par récurrence d'une propriété topologique des structures secondaires

On veut prouver que, pour une structure secondaire quelconque comprenant n_s brins, n_{loop} boucles et n_{stem} parties double-brins, on a la propriété topologique

$$n_{loop} = n_{stem} - (n_s - 1) \tag{6.68}$$

Commençons par traiter le cas $n_s = 1$ et montrons par récurrence sur n_{stem} la propriété 6.68. Pour $n_s = 1$ et $n_{stem} = 0$, on vérifie trivialement $n_{loop} = 0 = 0 - (1-1)$. Pour $n_s = 1$, on suppose que $n_{loop} = n_{stem}$ soit vrai pour $n_{stem} \leq n \in \mathbb{N}$ et considérons une structure secondaire $\mathcal{S}$ quelconque avec $n_{stem} = n+1$. Si toutes les boucles de la structure ne sont connectées qu'à une partie double-brin

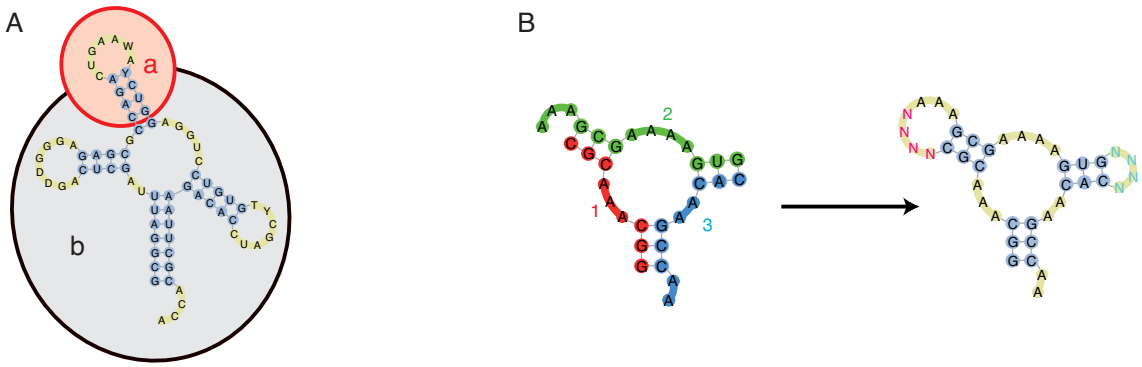

Figure 6.6 – (A) Exemple de décomposition en deux sous-structures a et b de la structure totale. (B) Transformation d'une structure composée de 3 brins (chacun repéré par une couleur différente) en une structure avec un seul brin en ajoutant un lieur simple-brin entre les brins 1 et 2 (mauve) et un autre entre les brins 2 et 3 (cyan).

chacune, on a directement $n_{stem} = n_{loop}$. Sinon il existe une boucle connectée à au moins deux parties double-brins. On peut donc décomposer la structure générale en deux sous-structures complémentaires a et b telles que $\mathcal{S} = a \cup b$ (voir figure 6.6 A) et auxquelles on applique l'hypothèse de récurrence et qui donne $n_{loop} = n^a_{loop} + n^b_{loop} = n^a_{stem} + n^b_{stem} = n_{stem}$.

Intéressons nous maintenant au cas général. Pour une structure secondaire $\mathcal{S}$ telle que $n_s > 1$ et $n_{stem} \in \mathbb{N}$, considérons la même structure à laquelle on rajoute des parties simple-brins reliant le brin 1 au brin 2, le brin 2 au brin 3,...,le brin $n_s - 1$ au brin n_s (voir figure 6.6 B). Cette structure $\mathcal{S}^*$ est alors formée d'un seul brin, a la même nombre de parties double-brin et $n_s - 1$ boucles en plus (dues aux $n_s - 1$ liens que l'on a rajoutés) que la structure initiale. Or, pour un seul brin, on a démontré ci-dessus que la propriété topologique 6.68 était vérifiée, soit

$$n_{stem} = n^*_{stem} = n^*_{loop} = n_{loop} + (n_s - 1) \tag{6.69}$$

6.8 Quelques notions sur les polymères

Négliger les effets stériques pour un polymère semi-rigide

A partir de quelle longueur les interactions stériques ne sont plus négligeables pour un polymère semi-rigide ? Faisons un argument de Flory pour répondre à la question. Pour un polymère de taille N semi-rigide défini par N_K et l_K, le nombre de segments et la longueur de Kuhn ($N_K l_K = Nb$, avec b la distance entre deux monomères), l'énergie de Flory peut approximativement s'écrire comme la somme d'un terme venant du Viriel et comptant pour les corrections de volume exclu et d'un terme entropique gaussien [176] :

$$\beta F_{flory} \approx vR^3 \left(\frac{N_K}{R^3}\right)^2 + \frac{3}{2}\left[\frac{R^2}{(N_K l_K^2)}\right] \tag{6.70}$$

où v est le volume exclu d'un segment de Kuhn et R la taille typique du polymère. Les effets stériques doivent être pris en compte quand le terme venant du Viriel devient du même ordre de grandeur que le terme entropique purement gaussien, soit

$$N^3 \sim R^5/(b^4 l_K) \tag{6.71}$$

Or $R^2 \approx N_K l_K^2 = Nbl_K$, donc le régime gaussien n'est plus valable (et donc on ne peut plus négliger les effets stériques) à partir de

$$N^* \sim (l_K/b)^3 \tag{6.72}$$

Pour un double-brin d'acide nucléique, $l_K/b \approx 300bp$ [177], soit $N^* \sim 10^7$ bp. Ainsi, négliger les effets stériques est donc largement justifié pour les molécules que nous allons étudier ($N_{max} \sim 10^3$).

Distance bout-à-bout d'un polymère semi-rigide

Soit un polymère semi-rigide de taille $N << N^*$, sa distance bout-à-bout carrée moyenne est définie par

$$R_e^2 = \langle \| \sum_{i=1}^{N} \vec{r}_i \|^2 \rangle = \sum_{i,j} \langle \vec{r}_i \cdot \vec{r}_j \rangle \tag{6.73}$$

avec $\vec{r}_i$ l'orientation du i^{eme} segment. On a donc besoin d'évaluer les facteurs de corrélation $\langle \vec{r}_i \cdot \vec{r}_j \rangle$. Supposons par exemple $i \leq j$, alors la moyenne de $\vec{r}_j$ si $\{\vec{r}_i, \vec{r}_{i+1}, ..., \vec{r}_{j-1}\}$ sont fixés vaut [176]

$$\langle \vec{r}_j \rangle_{\{\vec{r}_i, \vec{r}_{i+1}, ..., \vec{r}_{j-1}\} \text{ fixés}} = \langle \cos \Psi \rangle \vec{r}_{j-1} \tag{6.74}$$

avec Ψ l'angle entre deux segments consécutifs. En multipliant les deux côtés de l'équation précédente par $\vec{r}_i$ et en moyennant sur $\{\vec{r}_i, \vec{r}_{i+1}, ..., \vec{r}_{j-1}\}$, il vient

$$\langle \vec{r}_i \cdot \vec{r}_j \rangle = \langle \cos \Psi \rangle \langle \vec{r}_i \cdot \vec{r}_{j-1} \rangle = b^2 \langle \cos \Psi \rangle^{|j-i|} \tag{6.75}$$

Ce qui donne

$$\begin{aligned} R_e^2 &= b^2 \sum_{i,j} \langle \cos \Psi \rangle^{|j-i|} = b^2 \sum_{i=1}^{N} \sum_{k=-i+1}^{N-i} \langle \cos \Psi \rangle^k && (6.76) \\ &= Nb^2 \left(\frac{1 + \langle \cos \Psi \rangle}{1 - \langle \cos \Psi \rangle} \right) + 2 \langle \cos \Psi \rangle \left(\frac{1 - \langle \cos \Psi \rangle^N}{[1 - \langle \cos \Psi \rangle]^2} \right) && (6.77) \\ &\approx Nb^2 \left(\frac{1 + \langle \cos \Psi \rangle}{1 - \langle \cos \Psi \rangle} \right) \quad \text{pour } N \text{ grand} && (6.78) \end{aligned}$$

Rayon de giration d'une tige rigide

Soit une tige rigide composée de N monomères repérés par leur position $\vec{r}_i = (i-1)\vec{r}_0$, $i = 1, ..., N$ avec $||\vec{r}_0|| = b$. Par définition, le rayon de giration carré vaut

$$R_G^2 = \frac{1}{N} \sum_{i=1}^{N} \|\vec{r}_i - \vec{r}_m\|^2 \tag{6.79}$$

avec $\vec{r}_m = (1/N) \sum_i \vec{r}_i$ la position du barycentre de la chaîne. Donc pour une tige rigide, comme

$$\vec{r}_m = \frac{1}{N} \left(\sum_{i=1}^{N} (i-1) \right) \vec{r}_0 = \frac{N-1}{2} \vec{r}_0 \tag{6.80}$$

on obtient

$$R_G^2 = \frac{(N-1)(N+1)}{12} b^2 \approx \frac{n_{stem}^2}{12} b^2 \tag{6.81}$$

avec $n_{stem} = N$ le nombre de segments dans la tige.

Nombre de chemins aléatoires idéaux entre deux points fixes dans un réseau CFC

On généralise l'approche développée par Dijkstra, Frenkel et Hansen dans [187] pour un réseau cubique simple à un réseau CFC. On cherche à calculer le nombre $\mathcal{N}_{RW}(P_1;P_2)$ de chemins aléatoires idéaux (sans volume, ni interdiction de retour en arrière) de taille N entre deux points P_1 et P_2 du réseau définis par leurs coordonnées (i_1, j_1, k_1) et (i_2, j_2, k_2) dans la base des vecteurs générateurs du réseau CFC (voir figure 3.7). Si on pose

$$(\Delta i, \Delta j, \Delta k) = (i_2 - i_1, j_2 - j_1, k_2 - k_1) \tag{6.82}$$

, calculer $\mathcal{N}_{RW}(P_1;P_2)$ est alors équivalent à calculer $\mathcal{N}_{RW}((0,0,0);(\Delta i, \Delta j, \Delta k))$. On note $\{\vec{v}_1, ..., \vec{v}_{12}\}$ les 12 directions possibles sur le réseau CFC que l'on définit dans l'ordre suivant : $\vec{v}_1 = (1,-1,0)$, $\vec{v}_2 = (-1,1,0)$, $\vec{v}_3 = (0,1,-1)$, $\vec{v}_4 = (0,-1,1)$, $\vec{v}_5 = (-1,0,1)$, $\vec{v}_6 = (1,0,-1)$, $\vec{v}_7 = (0,-1,0)$, $\vec{v}_8 = (0,0,-1)$, $\vec{v}_9 = (0,0,1)$, $\vec{v}_{10} = (0,1,0)$, $\vec{v}_{11} = (-1,0,0)$ et $\vec{v}_{12} = (1,0,0)$. Pour un nombre fixe de pas a_l dans chacune des directions l, le nombre de chemins aléatoires vaut

$$\mathcal{N}_{RW}(a_1, ..., a_{12}) = \frac{N!}{\prod_{l=1}^{12}(a_l!)} \tag{6.83}$$

Le 12-uplets $(a_1, ..., a_{12})$ $(a_l \in \mathbb{N})$ devant vérifier le système d'équations

$$N = \sum_{l=1}^{12} a_l \tag{6.84}$$

$$\Delta i = \sum_{l=1}^{12} a_l v_l^i \tag{6.85}$$

$$\Delta j = \sum_{l=1}^{12} a_l v_l^j \tag{6.86}$$

$$\Delta k = \sum_{l=1}^{12} a_l v_l^k \tag{6.87}$$

Les relations de fermeture précédentes nous permettent alors d'exprimer $a_9, ..., a_{12}$ en fonction de N, Δi, Δj, Δk et les $\{a_1, ..., a_8\}$:

$$a_9 = \Delta k + a_3 - a_4 - a_5 + a_6 + a_8 \tag{6.88}$$

$$a_{10} = \Delta j + a_1 - a_2 - a_3 + a_4 + a_7 \tag{6.89}$$

$$a_{11} = \frac{1}{2}\left[N - (\Delta i + \Delta j + \Delta k) - a_1 - a_2 - a_3 - a_4 - a_5 - a_6 - 2a_7 - 2a_8\right] \tag{6.90}$$

$$a_{12} = \frac{1}{2}\left[N + \Delta i - \Delta j - \Delta k - 3a_1 + a_2 - a_3 - a_4 + a_5 - 3a_6 - 2a_7 - 2a_8\right] \tag{6.91}$$

et donc

$$\mathcal{N}_{RW}((0,0,0);(\Delta i, \Delta j, \Delta k)) = \sum_{(a_1,...,a_8)/a_l \in \{0,...,N\}} \mathcal{N}_{RW}(a_1, ..., a_{12}) \tag{6.92}$$

Cette dernière formule permet alors de calculer $\mathcal{N}_{RW}(P_1;P_2)$ en $\mathcal{O}(N^8)$ opérations.

6.9 Statistique dans les simulations de Monte-Carlo

Erreurs statistiques dans l'évaluation des valeurs moyennes

On cherche à évaluer l'erreur statistique commise lors de l'évaluation de la valeur moyenne d'une observable A à l'aide de la formule 4.9. On utilise ici les notations définies dans la section 4.2. Pour

faciliter les calculs, on suppose que A et le poids W sont décorrélés entre eux et que $\langle A \rangle = 0$. Soit N_B le nombre de blocks indépendants dans la suite de micro-états $\{S_1, ..., S_N\}$, on pose $\widetilde{W}$ et $\widetilde{AW}$ la valeur moyenne de W et de AW dans chaque block. On peut alors réécrire l'équation 4.9

$$\bar{A} = \frac{\sum_{\text{block}} \widetilde{AW}}{\sum_{\text{block}} \widetilde{W}} = \frac{\sum_{\text{block}} \widetilde{AW}}{N_B \langle \widetilde{W} \rangle} \tag{6.93}$$

avec $\langle \widetilde{W} \rangle$ la valeur moyenne sur tous les blocks de $\widetilde{W}$. D'où

$$\sigma_{\bar{A}}^2 = \frac{N_B \sigma_{\widetilde{AW}}^2}{N_B^2 \langle \widetilde{W} \rangle^2} \tag{6.94}$$

Les hypothèses faites sur A et W impose

$$\langle \widetilde{AW} \rangle = \frac{N_B}{N} \sum_i \langle A_i W_i \rangle = \frac{N_B}{N} \sum_i \langle A_i \rangle \langle W_i \rangle = 0 \tag{6.95}$$

Ainsi

$$\begin{aligned} \sigma_{\widetilde{AW}}^2 = \langle \widetilde{AW}^2 \rangle &\leq \langle \widetilde{A}^2 \widetilde{W}^2 \rangle = \langle \widetilde{A}^2 \rangle \langle \widetilde{W}^2 \rangle \\ &\leq \langle A^2 \rangle \langle \widetilde{W}^2 \rangle = \sigma_A^2 \langle \widetilde{W}^2 \rangle \end{aligned} \tag{6.96}$$

Il y a égalité quand toutes les mesures sont indépendantes entre elles ($N_B = N$). En combinant l'équation 6.94 et l'inégalité 6.96, on trouve

$$\sigma_{\bar{A}}^2 \leq \frac{\langle \widetilde{W}^2 \rangle}{N_B \langle \widetilde{W} \rangle^2} \sigma_A^2 \tag{6.97}$$

Combinaison optimale de mesures indépendantes

Supposons que l'on ait plusieurs mesures indépendantes $\{A_i \pm \sigma_i\}$ d'une même observable A. On veut les combiner afin d'avoir une meilleure estimation pour A :

$$A^{est} = \sum_i w_i A_i \tag{6.98}$$

avec $\sum_i w_i = 1$ et $w_i \geq 0$. L'erreur statistique faite sur A^{est} vaut

$$\sigma^2 = \sum_i (\partial A^{est} / \partial A_i)^2 \sigma_i^2 = \sum_i w_i^2 \sigma_i^2 \tag{6.99}$$

avec σ_i l'erreur faite sur chaque mesure A_i. Pour trouver la valeur des poids $\{w_i\}$ qui minimise l'erreur σ, on effectue alors une minimisation sous contrainte de $I = \sigma^2 + \lambda \sum_i w_i$ par rapport aux w_i (avec λ le multiplicateur de Lagrange associé) et on obtient

$$\frac{\partial I}{\partial w_i} = 0 \quad \Longrightarrow \quad 2wi\sigma_i^2 + \lambda = 0 \tag{6.100}$$

soit

$$w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2} \tag{6.101}$$

Ainsi plus une mesure sera précise, plus son poids sera important.

Distribution statistique des entrées d'un histogramme

Soit un histogramme $\mathcal{N}(x)$ obtenu lors d'une simulation de Monte-Carlo où x est une mesure possible d'une observable X quelconque. Notons p_x la probabilité d'obtenir x à chaque fois qu'on mesure X. Après $\mathcal{N}^{tot}$ mesures, la probabilité d'avoir observé $\mathcal{N}(x)$ fois la valeur x est donnée par la loi binomiale

$$\mathcal{P}_{bin}(\mathcal{N}(x)) = \binom{\mathcal{N}^{tot}}{\mathcal{N}(x)} p_x^{\mathcal{N}(x)}(1-p_x)^{\mathcal{N}^{tot}-\mathcal{N}(x)} \tag{6.102}$$

Intéressons nous à la valeur moyenne et à l'écart-type de $\mathcal{N}(x)$. Par définition

$$\langle \mathcal{N}(x) \rangle = \sum_{\mathcal{N}(x)=0}^{\mathcal{N}^{tot}} \mathcal{N}(x)\mathcal{P}_{bin}(\mathcal{N}(x)) \tag{6.103}$$

$$\sigma^2_{\mathcal{N}(x)} = \left[\sum_{\mathcal{N}(x)=0}^{\mathcal{N}^{tot}} \mathcal{N}(x)^2\mathcal{P}_{bin}(\mathcal{N}(x))\right] - \langle \mathcal{N}(x) \rangle^2 \tag{6.104}$$

Or on sait que

$$\sum_{k=0}^{n} \binom{n}{k} u^k v^{n-k} = (u+v)^n \equiv f(u,v) \tag{6.105}$$

Ainsi

$$\langle \mathcal{N}(x) \rangle = \left(u\frac{\partial f(u,v)}{\partial u}\right)(p_x, 1-p_x) = \mathcal{N}^{tot}p_x \tag{6.106}$$

$$\sigma^2_{\mathcal{N}(x)} = \left\{\left[u\frac{\partial}{\partial u}\left(u\frac{\partial f}{\partial u}\right)\right](p_x, 1-p_x)\right\} - \langle \mathcal{N}(x) \rangle^2 = \mathcal{N}^{tot}p_x(1-p_x) \tag{6.107}$$

Soit, si $p_x << 1$, $\sigma^2_{\mathcal{N}(x)} = \langle \mathcal{N}(x) \rangle$.

6.10 Lien entre probabilités et poids de Rosenbluth dans la dimérisation avec biais

Pour $k \neq 0$, la probabilité d'avoir construit la conformation $\mathcal{C} = \mathcal{C}_A \cup \mathcal{C}_i$ vaut la probabilité d'avoir construit la demi-conformation $\mathcal{C}_A$ multipliée par la probabilité d'avoir générer les N_{trans} conformations à partir de $\mathcal{C}_B$ multipliée par la probabilité d'en choisir une $(\mathcal{C}_i)$ parmi les k_l conformations valides soit

$$\mathcal{P}(\mathcal{C}_A \cup \mathcal{C}_i) = \mathcal{P}(\mathcal{C}_A) \times [N_{trans}\mathcal{P}(\mathcal{C}_B)] \times \left(\frac{1}{k_l}\right) \tag{6.108}$$

Grâce à la définition du poids de Rosenbluth pour la dimérisation avec biais (équation 4.78) et à la relation 6.108, on montre facilement par récurrence sur k que pour une conformation $\mathcal{C}$ de taille 2^kN_0, la probabilité de construire $\mathcal{C}$ avec l'algorithme de la dimérisation avec biais vaut

$$\mathcal{P}(\mathcal{C}) = \frac{\mathcal{P}_0^{2^k}}{W_k(\mathcal{C})} \tag{6.109}$$

avec $\mathcal{P}_0 = \mathcal{N}_{SAW}(N_0)/(z(z-1)^{N_0-1})$ la probabilité de générer une conformations à l'étape $k = 0$. Ainsi, soit A_k le nombre d'appel à la fonction $dim(k)$ dans la dimérisation avec biais,

$$\sum_{i=1}^{A_k} W_k(\mathcal{C}_i) = A_k \sum_{\mathcal{C}} \mathcal{P}(\mathcal{C})W_k(\mathcal{C}) = A_k \sum_{\mathcal{C}} \mathcal{P}_0^{2^k} = A_k \times \mathcal{N}_{SAW}(2^kN_0) \times \mathcal{P}_0^{2^k} \tag{6.110}$$

De même, soit $\mathcal{C}_{i,A}$ et $\mathcal{C}_{i,B}$ les deux demi-conformations utilisées pour construire $\mathcal{C}_i$

$$\sum_{i=1}^{A_k} W_{k-1}(\mathcal{C}_{i,A})W_{k-1}(\mathcal{C}_{i,B}) = A_k \sum_{\mathcal{C}_A,\mathcal{C}_B} \mathcal{P}(\mathcal{C}_A,\mathcal{C}_B)W_{k-1}(\mathcal{C}_A)W_{k-1}(\mathcal{C}_B) \tag{6.111}$$

Or $\mathcal{C}_A$ et $\mathcal{C}_B$ sont construites indépendamment l'une de l'autre donc $\mathcal{P}(\mathcal{C}_A,\mathcal{C}_B) = \mathcal{P}(\mathcal{C}_A)\mathcal{P}(\mathcal{C}_B)$, d'où

$$\begin{aligned}
\sum_{i=1}^{A_k} W_{k-1}(\mathcal{C}_{i,A})W_{k-1}(\mathcal{C}_{i,B}) &= A_k \left[\sum_{\mathcal{C}_A} \mathcal{P}(\mathcal{C}_A)W_{k-1}(\mathcal{C}_A)\right] \times \left[\sum_{\mathcal{C}_B} \mathcal{P}(\mathcal{C}_B)W_{k-1}(\mathcal{C}_B)\right] \\
&= A_k \left[\sum_{\mathcal{C}_A} \mathcal{P}_0^{2^{k-1}}\right]^2 \\
&= A_k \times [\mathcal{N}_{SAW}(2^{k-1}N_0)]^2 \times \mathcal{P}_0^{2^k} \qquad (6.112)
\end{aligned}$$

Finalement, avec les équations 6.110 et 6.112, on obtient

$$\mathcal{P}_k \equiv \frac{\mathcal{N}_{SAW}(2^k N_0)}{[\mathcal{N}_{SAW}(2^{k-1}N_0)]^2} = \frac{\sum_{i=1}^{A_k} W_k(\mathcal{C}_i)}{\sum_{i=1}^{A_k} W_{k-1}(\mathcal{C}_{i,A})W_{k-1}(\mathcal{C}_{i,B})} \tag{6.113}$$

6.11 Cinétique du repliement de l'ARN

Etude avec une équation maîtresse

Dans cette annexe, on étudie quelques particularités de l'équation maîtresse 5.30 et les méthodes pour la résoudre. A l'équilibre, $\mathcal{P}_\mathcal{S}$ doit tendre vers la distribution de Boltzmann $\mathcal{P}_\mathcal{S}^{eq} = \exp[-\beta\Delta G(\mathcal{S})]/Z$. Ainsi on a la balance détaillée

$$k(\mathcal{S} \to \mathcal{S}') \times \mathcal{P}_\mathcal{S}^{eq} = k(\mathcal{S}' \to \mathcal{S}) \times \mathcal{P}_{\mathcal{S}'}^{eq} \tag{6.114}$$

$$\frac{k(\mathcal{S} \to \mathcal{S}')}{k(\mathcal{S}' \to \mathcal{S})} = \exp\left\{-\beta\left[\Delta G(\mathcal{S}') - \Delta G(\mathcal{S})\right]\right\} \tag{6.115}$$

Plusieurs choix possibles existent pour les taux de transition vérifiant l'équation 6.114. Communément, on les définit via une équation d'Arrhenius 5.31. Là aussi, il existe plusieurs possibilités pour définir l'état de transition. Par exemple, Zhang et Chen [234, 235] choisissent $\Delta G^*(\mathcal{S},\mathcal{S}') = \Delta H(\mathcal{S}') - T\Delta S(\mathcal{S})$ si l'on passe de $\mathcal{S}$ à $\mathcal{S}'$ en ouvrant une ou plusieurs paires de bases. Cela revient à dire que lors de la création d'une paire de bases, l'énergie d'activation équivaut à la différence d'entropie entre les deux structures, alors que lors de l'ouverture d'une paire, c'est la différence d'enthalpie qui est prise en compte. On considère donc que le facteur limitant dans la formation d'une partie double-brin est imposé par le changement entropique totale (principalement dominé par la diminution de l'entropie de conformation).

Une fois choisis les taux de transition, reste à résoudre l'équation 5.30. Sa solution formelle est $\vec{\mathcal{P}}(t) = \exp[t\mathbf{K}]\vec{\mathcal{P}}(0)$ avec $\vec{\mathcal{P}} = \{\mathcal{P}_\mathcal{S}\}$ et $\mathbf{K} = \{k(\mathcal{S} \to \mathcal{S}')\}$. Si le nombre total de structures secondaires est faible, une possibilité est de diagonaliser $\mathbf{K}$ et ainsi d'avoir accès à $\vec{\mathcal{P}}(t)$ pour des temps arbitraires [235]. Cependant dès que la séquence devient un peu longue, la taille de l'espace des états explose et la diagonalisation de $\mathbf{K}$ devient numériquement délicate à effectuer. La solution couramment utilisée est d'appliquer l'algorithme de Gillespie [236] à l'équation 5.30 qui permet de simuler des systèmes d'équations stochastiques.

Approche avec un biais configurationnel

Afin de dériver l'expression 5.32 de l'acceptance $\mathrm{acc}(\mathcal{C}_o \to \mathcal{C}_n)$ pour l'algorithme cinétique avec biais configurationnel que nous avons décrit dans la section 5.5.2.2, reprenons le raisonnement vu pour l'acceptance dans les schémas dynamiques (section 4.3.1). La balance détaillée impose

$$\pi(\mathcal{C}_o)\mathcal{P}(\mathcal{C}_o \to \mathcal{C}_n) = \pi(\mathcal{C}_n)\mathcal{P}(\mathcal{C}_n \to \mathcal{C}_o) \tag{6.116}$$

avec $\pi(\mathcal{C}_o) = (1/Z)\exp[-\beta\Delta G_{latt}(\mathcal{C})]$ où $\Delta G_{latt}(\mathcal{C}) = \Delta G_{latt}(\mathcal{S}(\mathcal{C}))+\kappa^{tot}(\mathcal{C})$ (voir section 3.3.5), et avec $\mathcal{P}(\mathcal{C}_o \to \mathcal{C}_n)$ la probabilité de transition de $\mathcal{C}_o$ vers $\mathcal{C}_n$ qui se décompose en le produit de la probabilité α de construire un mouvement entre les deux conformations et la probabilité acc de l'accepter (voir équation 4.12). Contrairement au cas classique, α ici n'est pas symmétrique. En effet $\alpha(\mathcal{C}_o \to \mathcal{C}_n)$ est égale au produit de la probabilité de générer $\mathcal{S}_n$ à partir de $\mathcal{S}_o$ et de la probabilité de construire $\mathcal{C}_n$ à $\mathcal{S}_n$ fixée, soit

$$\alpha(\mathcal{C}_o \to \mathcal{C}_n) = \left(\frac{1}{\mathcal{N}_o}\right) \times \left(\frac{\exp[-\beta\kappa^{tot}(\mathcal{C}_n)]}{W_n}\right) \tag{6.117}$$

où $\mathcal{N}_o$ le nombre de structures secondaires voisines que l'on peut atteindre directement depuis $\mathcal{S}_o$ en ouvrant ou fermant une paire de bases, et $\exp[-\beta\kappa^{tot}(\mathcal{C}_n)]/W_n$ la probabilité de construire $\mathcal{C}_n$ (voir équation 4.67). En combinant les équations 6.116 et 6.117, on obtient

$$\frac{\mathrm{acc}(\mathcal{C}_o \to \mathcal{C}_n)}{\mathrm{acc}(\mathcal{C}_n \to \mathcal{C}_o)} = \exp\left(-\beta\left[\Delta G_{latt}(\mathcal{S}_n) - \Delta G_{latt}(\mathcal{S}_o)\right]\right) \times \left(\frac{W_n}{W_o}\right) \times \left(\frac{\mathcal{N}_o}{\mathcal{N}_n}\right) \tag{6.118}$$

Par défaut, on choisit la solution de Metropolis pour l'acceptance, aboutissant à l'équation 5.32.

Bibliographie

[1] J.D. Watson and F.H.C. Crick, "A structure for deoxyribose nucleic acid," Nature **171**, 737–738 (1953).

[2] Wikipedia, "http ://en.wikipedia.org," .

[3] C. Calladine, H. Drew, B. Luisi, and A. Travers, *Understanding DNA ; the molecule and how it works* (Elsevier Academic Press, 2004).

[4] R. Gesteland, T.R. Cech, and J.F. Atkins, *RNA World* (Cold Spring Harbor Laboratory Press, 2005).

[5] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello, "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans," Nature **391**, 806–811 (1998).

[6] J. Zeng, J.J. Birktoft, Y. Chen, T. Wang, R. Sha, P.E. Constantinou adn S.L. Ginell, C. Mao, and N.C. Seeman, "Folding DNA to create nanoscale shapes and patterns," Nature **461**, 74–77 (2009).

[7] P.W.K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," Nature **440**, 297–302 (2006).

[8] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, "An autonomous molecular computer for logical control of gene expression," Nature **429**, 423–429 (2004).

[9] N.C. Seeman, "DNA in a material world," Nature **421**, 427–431 (2003).

[10] E. Winfree, "Algorithmic self-assembly of DNA : theoretical motivations and 2D assembly experiments," J. Biomol. Struct. Dyn. **11**, 263–270 (2000).

[11] D.D. Shoemaker, E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engele, and al, "Experimental annotation of the human genome using microarray technology," Nature **409**, 922–927 (2001).

[12] D. Nykypanchuk, M.M. Maye, D. van der Lelie, and O. Gang, "DNA-guided crystallization of colloidal nanoparticles," Nature **451**, 549–552 (2008).

[13] S.Y. Park, A.K.R. Lytton-Jean, B. Lee, S. Weigand, and G.C. Schatz, "DNA-programmable nanoparticle crystallization," Nature **451**, 553–556 (2008).

[14] H.H. El-Hajj, S.A.E. Marras, S. Tyagi, E. Shashkina, M. Kamboj, T.E. Kiehn, M.S. Glickman, F.R. Kramer, and D. Alland, "Use of sloppy molecular beacon probes for identification of mycobacterial species," J. Clin. Microbiol. **47**, 1190–1198 (2009).

[15] Protein Data Bank, "http ://www.rcsb.org/pdb/," .

[16] R. Thomas, "The denaturation of deoxyribonucleic acids," Trans. Faraday Soc. **50**, 304 (1954).

[17] R.M. Wartell and A.S. Benight, "Thermal denaturation of DNA molecules : a comparison of theory with experiment," Phys. Rep. **126**, 67–107 (1985).

[18] R.W. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswick, and A. Zamir, "Structure of a ribonucleic acid," Science **147**, 1462–1465 (1965).

[19] J.D. Robertus, J.E. Ladner, J.T. Finch, D. Rhodes, R.S. Brown, B.F.C. Clark, and A. Klug, "Structure of yeast phenylalanine transfer RNA at 3 å resolution," Nature **250**, 546–551 (1974).

[20] P. Schuster, "Prediction of RNA secondary structures : from theory to models and real molecules," Rep. Prog. Phys. **69**, 1419–1477 (2006).

[21] David Ussery Homepage, "http ://www.cbs.dtu.dk/staff/dave/roanoke/," .

[22] Felix Ritort group, "http ://www.ffn.ub.es/ritort/," .

[23] V.A. Bloomfield, D.M. Crothers, and I. Tinoco, *Nucleic acids, structures, properties and functions* (University Science Books, 2000).

[24] V.V. Filimonov, *Thermodynamic data for biochemistry and biotechnology* (Hinz, 1986).

[25] J. Liphardt, B. Onoa, S.B. Smith, I. Jr. Tinoco, and C. Bustamante, "Reversible unfolding of single RNA molecules by mechanical force," Science **292**, 733–737 (2001).

[26] D. Poland and H.A. Scheraga, "Phase transition in one dimension and the helix-coil transition in polyamino acids," J. Chem. Phys. **45**, 1456–1464 (1966).

[27] B.H. Zimm and J.K. Bragg, "Theory of the phase transition between helix and random coil in polypeptide chains," J. Chem. Phys. **31**, 526–535 (1959).

[28] H. DeVoe and I. Tinoco, "The stability of helical polynucleotides : base contributions," J. Mol. Biol. **4**, 500–517 (1962).

[29] D.M. Crothers and B.H. Zimm, "Theory of the melting transition of synthetic polynucleotides : evaluation of the stacking free energy," J. Mol. Biol. **9**, 1–9 (1964).

[30] J. Jr. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," Proc. Natl. Acad. Sci. USA **95**, 1460–1465 (1998).

[31] D. Poland and H.A. Scheraga, *Theory of helix-coil transition in biopolymers* (Academic Press, New York, 1970).

[32] D. Poland, "Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations," Biopolymers **13**, 1859–1871 (1974).

[33] R. Everaers, S. Kumar, and C. Simm, "Unified description of poly- and oligonucleotide DNA melting : Nearest-neighbor, poland-scheraga, and lattice models," Phys. Rev. E **75**, 041918 (2007).

[34] T.A. Knotts, N. Rathore, C. Schwartz, and J.J de Pablo, "A coarse grain model for DNA," J. Chem Phys. **126**, 084901 (2007).

[35] T.E. Ouldridge, I.G. Johnston, A.A. Louis, and J.P.K Doye, "The self-assembly of DNA Holliday junctions studied with a minimal model," J. Chem. Phys. **130**, 065101 (2009).

[36] M. Peyrard and A.R. Bishop, "Statistical mechanics of a nonlinear model for DNA denaturation," Phys. Rev. Lett. **62**, 2755–2758 (1989).

[37] T. Dauxois, M. Peyrard, and A.R. Bishop, "Dynamics and thermodynamics of a nonlinear model for DNA denaturation," Phys. Rev. E **47**, 684–695 (1993).

[38] M. Feig and B.M. Pettitt, "Structural equilibrium of DNA represented with different force fields," Biophys. J. **75**, 134–149 (1998).

[39] F. Merzel, F. Fontaine-Vive, M.R. Johnson, and G.J. Kearley, "Atomistic model of DNA : phonons and base-pair opening," Phys. Rev. E **76**, 031917 (2007).

[40] O. Gotoh and Y. Tagashira, "Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles," Biopolymers **20**, 1033–1042 (1981).

[41] M.J. Doktycz, R.F. Goldstein, T.M. Paner, F.J. Gallo, and A.S. Benight, "Studies of DNA dumbbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T_4 endloops : evaluation of the nearest-neighbor stacking interactions in DNA," Biopolymers **32**, 849–864 (1992).

[42] J. Jr. SantaLucia, H.T. Allawi, and P.A. Seneviratne, "Improved nearest-neighbor parameters for predicting DNA duplex stability," Biochemistry **35**, 3555–3562 (1996).

[43] H.T. Allawi and J. Jr. SantaLucia, "Thermodynamics and NMR of internal G·T mismatches in DNA," Biochemistry **36**, 10581–10594 (1997).

[44] A.L. Oliver, R.M. Wartell, and R.L. Ratliff, "Helix coil transitions of $d(A)_n \cdot d(T)_n$, $d(A-T)_n \cdot d(A-T)_n$, and $d(A-A-T)_n \cdot d(A-T-T)_n$; evaluation of parameters governing DNA stability," Biopolymers **16**, 1115–1137 (1977).

[45] B.R. Amirikyan, A.V. Vologodskii, and Y.L. Lyubchenko, "Determination of DNA cooperativity factor," Nucleic Acids Res. **9**, 5469–5482 (1981).

[46] R.D. Blake and S.G. Delcourt, "Thermal stability of DNA," Nucleic Acids Res. **15**, 3323–3332 (1998).

[47] Y. Zeng, A. Montrichok, and G. Zocchi, "Bubble nucleation and cooperativity in DNA melting," J. Mol. Biol. **339**, 67–75 (2004).

[48] Y. Zeng and G. Zocchi, "Mismatches and bubbles in DNA," Biophys. J. **90**, 4522–4529 (2006).

[49] E. Yeramian, "Genes and physics of the DNA double-helix," Gene **255**, 139–150 (2000).

[50] E. Yeramian, "The physics of DNA and the annotation of the Plasmodium falciparum genome," Gene **255**, 151–168 (2000).

[51] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal, "RNA codewords and protein synthesis, VII. On the general nature of the RNA code," Proc. Natl. Acad. Sci. USA **53**, 1161–1168 (1965).

[52] J.C. Venter et al, "The sequence of the human genome," Science **16**, 1304–1351 (2001).

[53] The International Human Genome Mapping Consortium, "Initial sequencing and analysis of the human genome," Nature **409**, 860–921 (2001).

[54] The International Human Genome Mapping Consortium, "A physical map of the human genome," Nature **409**, 934–941 (2001).

[55] M.S. Gelfand, A.A. Mironov, and P.A. Pevzner, "Gene recognition via spliced sequence alignment," Proc. Natl. Acad. Sci. USA **93**, 9061–9066 (1996).

[56] M.Q. Zhang, "identification of protein coding regions in the human genome by quadratic discrimant analysis," Proc. Natl. Acad. Sci. USA **94**, 565–568 (1997).

[57] M.R. Brent and R. Guigo, "Recent advances in gene structure prediction," Curr. Opin. Struct. Biol. **14**, 264–272 (2004).

[58] C.B. Burge and S. Karlin, "Finding the genes in genomic DNA," Curr. Opin. Struct. Biol. **8**, 346–354 (1998).

[59] J. Besemer and M. Borodovsky, "Heuristic approach to deriving models for gene finding," Nucleic Acids Res. **27**, 3911–3920 (1999).

[60] S.L. Salzberg, M. Pertea, A.L. Delcher, M.J. Gardner, and H. Tettelin, "Interpolated markov models for eukaryotic gene finding," Genomics **59**, 24–31 (1999).

[61] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structure to guide sequence motif finding towards single-stranded regions," Nucleic Acids Res. **34**, e117 (2006).

[62] S.E. Cawley, A.I. Wirth, and T.P. Speed, "Phat-a gene finding program for Plasmodium falciparum," Mol. Biochem. Parasitol. **118**, 167–174 (2001).

[63] S.L. Salzberg and A.L. Delcher, *Microbial Genome* (Humana Press, Totowa,NJ, 2004).

[64] A.T. Sumner, J. de la Torre, and L. Stuppia, "The distribution of genes on chromosomes : a cytological approach," J. Mol. Evol. **37**, 117–122 (1993).

[65] J.L. Oliver and A. Marin, "A relationship between GC content and coding-sequence length," J. Mol. Evol. **43**, 216–223 (1996).

[66] G. Bernardi, "Isochores and the evolutionary genomics of vertebrates," Gene **241**, 3–17 (2000).

[67] J.L. Oliver, P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan, "Isochore chromosome maps of eukaryotic genomes," Gene **276**, 47–56 (2001).

[68] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," Proc. Natl. Acad. Sci. USA **95**, 11163–11168 (1998).

[69] M. Frank-Kamenetskii, "Simplification of the empirical relationship between melting temperature of DNA, its GC content and concentration of sodium ions in solution," Biopolymers **10**, 2623–2624 (1971).

[70] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes, "Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences," J. Biol. Phys. **30**, 33–81 (2004).

[71] E. Carlon and R. Blossey, "Exons, introns and DNA thermodynamics," Phys. Rev. Lett. **94**, 178101 (2005).

[72] G. Kalosakas, K.O. Rasmussen, A.R. Bishop, C.H. Choi, and A. Usheva, "Sequence-specific thermal fluctuations identify start sites for DNA transcription," Europhys. Lett. **68**, 127–133 (2004).

[73] T.S. van Erp, S. Cuesta-Lopez, J.G. Hagmann, and M. Peyrard, "Can one predict DNA transcription start sites by studying bubbles ?." Phys. Rev. Lett. **95**, 218104 (2005).

[74] C.J. Benham, "Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatiry loci," Proc. Natl. Acad. Sci. USA **90**, 2999–3003 (1993).

[75] F. Liu, E. Tostesen, J.K. Sundet, T.-K. Jenssen, C. Bock, G.I. Jerstad, W.G. Thilly, and E. Hovig, "The human genomic melting map," PLoS Comput. Biol. **3**, e93 (2007).

[76] M. Rubinstein and R.H. Colby, *Polymer Physics* (Oxford University Press, 2003).

[77] M.E. Fisher, "Effects of excluded volume on phase transitions in biopolymers," J. Chem. Phys. **45**, 1469–1473 (1966).

[78] Y. Kafri, D. Mukamel, and L. Peliti, "Melting and unzipping of DNA," Eur. Phys. J. B **27**, 135–146 (2002).

[79] Y. Kafri, D. Mukamel, and L. Peliti, "Kafri, Mukamel, and Peliti reply," Phys. Rev. Lett. **90**, 159802 (2003).

[80] R. Blossey and E. Carlon, "Reparametrizing the loop entropy weights : effect on DNA melting curves," Phys. Rev. E **68**, 061911 (2003).

[81] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca NY, 1979).

[82] C. Vanderzande, *Lattice Models of Polymers* (Cambridge University Press, 1998).

[83] R.A. Dimitrov and M. Zuker, "Prediction of hybridization and melting for double-stranded nucleic acids," Biophys. J. **87**, 215–226 (2004).

[84] T. Garel and H. Orland, "Generalized poland-scheraga model for DNA hybridization," Biopolymers **75**, 453–467 (2004).

[85] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," J. Mol. Biol. **288**, 911–940 (1999).

[86] I.L. Hofacker, "Vienna RNA secondary structure server," Nucleic Acids Res. **31**, 3429–3431 (2003).

[87] M. Zuker, “Mfold web server for nucleic acid folding and hybridization prediction,” Nucleic Acids Res. **31**, 3406–3415 (2003).

[88] N.R. Markham and M. Zuker, “DINAMelt web server for nucleic acid melting prediction,” Nucleic Acids Res. **33**, W577–W581 (2005).

[89] M. Fixman and J.J. Freire, “Theory of DNA melting curves,” Biopolymers **16**, 2693–2704 (1977).

[90] D.M. Gray, “Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. Simple sets of independent sequences and the influence of absent nearest neighbors,” Biopolymers **42**, 783–793 (1997).

[91] R.F. Goldstein and A.S. Benight, “How many numbers are required to specify sequence-dependent properties of polynucleotides ?.” Biopolymers **32**, 1679–1693 (1992).

[92] D.M. Gray, “Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA/RNA hybrids and DNA duplexes,” Biopolymers **42**, 795–810 (1997).

[93] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in fortran 77 : the Art of Scientific Computing* (Cambridge University Press, Cambridge,UK, 1996).

[94] R. Owczarzy, Y. You, B.G. Moreira, J.A. Manthey, L. Huang, M.A. Behlke, and J.A. Walder, “Effects of sodium ions on DNA duplex oligomers : improved predictions of melting temperatures,” Biochemistry **43**, 3537–3554 (2004).

[95] N. Le Novere, “MELTING, computing the melting temperature of nucleic acid duplex,” Bioinformatics **17**, 1226–1227 (2001).

[96] D. Erie, N. Sinha, W. Olson, R. Jones, and K. Breslauer, “A dumbbell-shaped, double-hairpin structure of DNA : a thermodynamic investigation,” Biochemistry **26**, 7150–7159 (1987).

[97] M.T. Record and T.M. Lohman, “A semiempirical extension of polyelectrolyte theory to the treatment of oligoelectrolytes : application to oligonucleotide helix-coil transitions,” Biopolymers **17**, 159–166 (1978).

[98] G.S. Manning, “Limiting laws and counterion condensation in polyelectrolyte solutions,” J. Chem. Phys. **51**, 924–933 (1969).

[99] G.S. Manning, “On the application of polyelectrolyte "limiting laws" to the helix-coil transition of DNA. I. Excess univalent cations,” Biopolymers **11**, 937–949 (1972).

[100] R. Owczarzy, B.G. Moreira, Y. You, M.A. Behlke, and J.A. Walder, “Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations,” Biochemistry **47**, 5336–5353 (2008).

[101] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, London, 1993).

[102] A. Spassky and D. Angelov, “Temperature-dependence of UV laser one-electron oxidative guanine modifications as a probe of local stacking fluctuations and conformational transitions,” J. Mol. Biol. **323**, 9–15 (2002).

[103] M. Peyrard, S. Cuesta-Lopez, and D. Angelov, “Experimental and theoretical studies of sequence effects on the fluctuation and melting of short DNA molecules,” J. Phys. : Condens. Matter **21**, 034103 (2009).

[104] A. Vologodskii, B. Amirikyan, Y. Lyuchenko, and M. Krank-Kamenetskii, “Allowance for heterogeneous stacking in the DNA helix-coil transition theory,” J. Biomol. Struct. Dyn. **2**, 131–148 (1984).

[105] R.D. Blake, J.W. Bizzaro, J.D. Blake, G.R. Day, S.G. Delcourt, J. Knowles, K.A. Marx, and J. Jr. SantaLucia, “Statistical mechanical simulation of polymeric DNA melting with MELTSIM,” Bioinformatics **15**, 370–375 (1999).

[106] E. Yeramian and L. Jones, "GeneFizz : a web tool to compare genetic (coding/non-coding) and physical (helix/coil) segmentations of DNA sequences. gene discovery and evolutionary perspectives," Nucleic Acids Res. **31**, 3843–3849 (2003).

[107] D. Jost and R. Everaers, "A unified poland-scheraga model of oligo- and polynucleotide DNA melting : salt effects and predictive power," Biophys. J. **96**, 1056–1067 (2009).

[108] P.R. Bevington and D.K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1992).

[109] R.M. Wartell and E.W. Montroll, "Equilibrium denaturation of natural and of periodic synthetic DNA molecules," Adv. Chem. Phys. **22**, 129–203 (1972).

[110] T. Garel and C. Monthus, "Numerical study of the disordered Poland-Scheraga model of DNA denaturation," J. Stat. Mech. : Theor. Exp., P06004(2005).

[111] E.W. Montroll, "Statistical mechanics of nearest neighbor systems," J. Chem. Phys. **9**, 706–721 (1941).

[112] T.S. van Erp, S. Cuesta-Lopez, and M. Peyrard, "Bubbles and denaturation in DNA," Eur. Phys. J. E **20**, 421–434 (2006).

[113] E. Yeramian, S. Bonnefoy, and G. Langsley, "Physics-based gene identification : proof of concept for Plasmodium falciparum," Bioinformatics **18**, 1–4 (2002).

[114] B.Y. Tong and S.J. Battersby, "Comparison of theoretical denaturation maps of ϕX174 and SV40 with their gene maps," Nucleic Acids Res. **6**, 1073–1079 (1979).

[115] O. Gotoh, "Prediction of melting profiles and local helix stability for sequenced DNA," Adv. Biophys. **16**, 1–52 (1983).

[116] A. Suyama and A. Wada, "Correlation between thermal stability maps and genetic maps of double-stranded DNAs," J. Theor. Biol. **105**, 133–145 (1983).

[117] E. Carlon A. Dkhissi, M.L. Malki, and R. Blossey, "Stability domains of actin genes and genomic evolution," Phys. Rev. E **76**, 051916 (2007).

[118] D. Jost and R. Everaers, "Genome wide application of DNA melting analysis," J. Phys. : Condens. Matter **21**, 034108 (2009).

[119] T.B. Liverpool, S.A. Harris, and C.A. Laughton, "Supercoiling and denaturation of DNA loops," Phys. Rev. Lett. **100**, 238103 (2008).

[120] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson, *Molecular biology of the cell* (Garland Science, 2002).

[121] M. Gellert, R. Menzel, K. Mizuuchi, and M. O'Dea an D. Friedman, "Regulation of DNA supercoiling in Escherichia coli," Cold Spring Harbor Symp. Quant. Biol. **47**, 763–767 (1983).

[122] C.J. Benham, "Torsional stess and local denaturation in supercoiled DNA," Proc. Natl. Acad. Sci. USA **76**, 3870–3874 (1979).

[123] C.J. Benham, "Energetics of the strand separation transition in superhelical DNA," J. Mol. Biol. **225**, 835–847 (1992).

[124] R.M. Fye and C.J. Benham, "Exact method for numerically analyzing a model of local denaturation in superhelically stressed DNA," Phys. Rev. E **59**, 3408–3426 (1999).

[125] P. Ak and C.J. Benham, "Susceptibility to superhelically driven DNA duplex destabilization : a highly conserved property of yeast replication origins," PLos Comp. Biol. **1**, e7 (2005).

[126] H. Wang and C.J. Benham, "Superhelical destabilization in regulatory regions of stress response genes," PLos Comp. Biol. **4**, e17 (2008).

[127] C.J. Benham and C. Bi, "The analysis of stress-induced duplex destabilization in long genomic DNA sequences," J. Comp. Biol. **11**, 519–543 (2004).

[128] C. Bi and C.J. Benham, "WebSIDD ; server for predicting stress-induced duplex destabilized (SIDD) sites in superhelical DNA," Bioinformatics **20**, 1477–1479 (2004).

[129] B.A. Shapiro, Y.G. Yingling, W. Kasprzak, and E. Bindewald, "Bridging the gap in RNA structure prediction," Curr. Op. in Struc. Biol. **17**, 157–165 (2007).

[130] E. Capriotti and M.A. Marti-Renom, "Computational RNA structure prediction," Curr. Bioinformatics **3**, 32–45 (2008).

[131] E. Rivas and S.R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," J. Mol. Biol. **285**, 2053–2068 (1999).

[132] S. Cao and S.J. Chen, "Predicting RNA folding thermodynamics for chain molecules with a reduced chain representation model," RNA **11**, 1884–1897 (2005).

[133] T.R. Einert, P. Nager, H. Orland, and R.R. Netz, "Impact of loop statistics on the thermodynamics of RNA folding," Phys. Rev. Lett. **101**, 048103 (2008).

[134] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner, "Incorporating chemical modification constraints into dynamic programming algorithm for prediction of RNA secondary structure," Proc. Natl. Acad. Sci. USA **101**, 7287–7292 (2004).

[135] K.E. Deigan, T.W. Li, D.H. Mathews, and K.M. Weeks, "Accurate SHAPE-directed RNA structure determination," Proc. Natl. Acad. Sci. USA **106**, 97–102 (2009).

[136] A.P. Gultyaev, F.H. van Batenburg, and C.W. Pleij, "The computer simulation of RNA folding pathways using a genetic algorithm," J. Mol. Biol. **250**, 37–51 (1995).

[137] B.A. Shapiro, W. Kasprzak, C. Grunewald, and J. Aman, "Exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm," J. Mol. Graph. Model **25**, 514–531 (2006).

[138] H. Isambert and E.D. Siggia, "Modeling RNA folding paths with pseudoknots : application to hepatitis delta virus ribozyme," Proc. Natl. Acad. Sci. USA **97**, 6515–6520 (2000).

[139] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, "Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations," Proc. Natl. Acad. Sci. USA **100**, 15310–15315 (2003).

[140] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, "Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots," Nucleic Acids Res. **33**, 605–610 (2005).

[141] Y.G. Yingling and B.A. Shapiro, "The prediction of the wild-type telomerase RNA pseudoknot structure and the pivotal role of the bulge in its formation," J. Mol. Graph. Model **25**, 261–274 (2006).

[142] F. Ding, S. Sharma, P. Chalasani, V.V. Demidov, N.E. Broude, and N.V. Dokholyan, "Ab initio RNA folding by discrete molecular dynamics : from structure prediction to folding mechanisms," RNA **14**, 1164–1173 (2008).

[143] M. Parisien and F. Major, "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," Nature **452**, 51–45 (2008).

[144] J. Burks, C. Zwieb, F. Muler, I. Wower, and J. Wower, "Comparative 3-D modeling of tmRNA," BMC Mol. Biol. **6**, 14–31 (2005).

[145] C. Hyeon and D. Thirumalai, "Mechanical unfolding of RNA : from hairpins to structures with internal multiloops," Biophys. J. **92**, 731–743 (2007).

[146] R. Das and D. Baker, "Automated de novo prediction of native-like RNA tertiary structures," Proc. Natl. Acad. Sci. USA **104**, 14664–14669 (2007).

[147] R. Das, M. Kudaravalli, M. Jonikas, A. Laederach, R. Fong, J.P. Schwans, D. Baker, J.A. Piccirilli, R.B. Altman, and D. Herschlag, "Structural inference of native and partially folded RNA by high-throughout contact mapping," Proc. Natl. Acad. Sci. USA **105**, 4144–4149 (2008).

[148] I. Jr. Tinoco and C. Bustamante, "How RNA folds," J. Mol. Biol. **293**, 271–281 (1999).

[149] H. Jacobson and W.H. Stockmayer, "Intramolecular reaction in polycondensations. I. The theory of linear systems," J. Chem. Phys. **18**, 1600–1606 (1950).

[150] E. Carlon, E. Orlandini, and A.L. Stella, "Roles of stiffness and excluded volume in DNA denaturation," Phys. Rev. Lett. **88**, 198101 (2002).

[151] S. Cao and S.J. Chen, "Predicting RNA pseudoknot folding thermodynamics," Nucleic Acids Res. **34**, 2634–2652 (2006).

[152] S. Cao and S.J. Chen, "Predicting structures and stabilities for H-type pseudoknots with interhelix loops," RNA **15**, 696–706 (2009).

[153] G. Vernizzi, H. Orland, and A. Zee, "Prediction of RNA pseudoknots by monte-carlo simulations," arXiv **q-bio**, 0405014 (2004).

[154] M. Bon, G. Vernizzi, H. Orland, and A. Zee, "Topological classification of RNA structures," J. Mol. Biol. **379**, 900–911 (2008).

[155] A. Lucas and K.A. Dill, "Statistical mechanics of pseudoknot polymers," J. Chem. Phys. **119**, 2414–2421 (2003).

[156] Z. Kopeikin and S.J. Chen, "Statistical thermodynamics for chain molecules with simple RNA tertiary contacts," J. Chem. Phys. **122**, 094909 (2005).

[157] Z. Kopeikin and S.J. Chen, "Folding thermodynamics of pseudoknotted chain conformations," J. Chem. Phys. **124**, 154903 (2006).

[158] M.S. Causo, B. Coluzzi, and P. Grassberger, "Simple model for the DNA denaturation transition," Phys. Rev. E **62**, 3958 (2000).

[159] P. Leoni and C. Vanderzande, "Statistical mechanics of RNA folding : a lattice approach," Phys. Rev. E **68**, 051904 (2003).

[160] A. Kabakcioglu and A.L. Stella, "Pseudoknots in a homopolymer," Phys. Rev. E **70**, 011802 (2004).

[161] M. Sales-Pardo, R. Guimera, A.A. Moreira, J. Widom, and L.A.N Amaral, "Mesoscopic modeling for nucleic acid chain dynamics," Phys. Rev. E **71**, 051902 (2005).

[162] N. Xia, J. Jr. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner, "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base-pairs," Biochemistry **37**, 14719–14735 (1998).

[163] Y. Byun and K. Han, "PseudoViewer : web application and web service for visualizing RNA pseudoknots and secondary structures," Nucleic Acids Res. **34**, W416–W422 (2006).

[164] W. Humphrey, A. Dalke, and K. Schulten, "VMD-Visual Molecular Dynamics," J. Molec. Graphics **14**, 33–38 (1996).

[165] A. Dkhissi, G. Renvez, and R. Blossey, "Y-DNA melting : a short tale of three scales," J. Phys. : Condens. Matter **21**, 034115 (2009).

[166] J.L. Martin, M.F. Sykes, and F.T. Hioe, "Probability of initial ring closure for self-avoiding walks on the face-centered cubic and triangular lattices," J. Chem. Phys. **46**, 3478–3481 (1967).

[167] M.J. Serra and D.H. Turner, "Predicting thermodynamic properties of RNA," Methods Enzym. **259**, 242–261 (1995).

[168] T. Ohmichi, S.-I. Nakano, D. Miyoshi, and N. Sugimoto, "Long RNA dangling end has large energetic contribution to duplex stability," J. Am. Chem. Soc. **124**, 10367–10372 (2002).

[169] A.E. Walter, D.H. Turner, J. Kim, M.H. Lyttle, P. Muller, D.H. Mathews, and M. Zuker, "Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding," Proc. Natl. Acad. Sci. USA **91**, 9218–9222 (1994).

[170] A.E. Walter and D.H. Turner, "Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces," Biochemistry **33**, 12715–12719 (1994).

[171] J. Kim, A.E. Walter, and D.H. Turner, "Thermodynamics of coaxially stacked helices with GA and CC mismatches," Biochemistry **35**, 13753–13761 (1996).

[172] B.M. Znosko, M.E. Burkard, S.J. Schroeder, T.R. Krugh, and D.H. Turner, "Sheared A_{anti}·A_{anti} base-pairs in a destabilizing 2×2 internal loop : the NMR structure of 5'$(rGGCAAGCCU)_2$," Biochemistry **41**, 14969–14977 (2001).

[173] S.J. Johnson and L.S. Beese, "Structures of mismatch replication errors observed in a DNA polymerase," Cell **116**, 803–816 (2004).

[174] A.M. Yoffe, P. Prinsen, A. Gopal, C.M. Knobler, W.M. Gelbart, and A. Ben-Shaul, "Predicting the sizes of large RNA molecules," Proc. Natl. Acad. Sci. USA **105**, 16153–16158 (2008).

[175] D.H. Mathews, "Using an RNA secondary structure partition function to determine confidence in base-pairs predicted by free energy minimization," RNA **10**, 1178–1190 (2004).

[176] M. Doi and S.F. Edwards, *The theory of polymer dynamics* (Oxford University Press, New York, 1986).

[177] C. Bustamante, J.F. Marko, E.D. Siggia, and S. Smith, "Entropic elasticity of λ-phage DNA," Science **265**, 1599–1600 (1994).

[178] D. Frenkel and B. Smit, *Understanding molecular simulation : from algorithms to applications* (Academic press, Oxford, UK, 2002).

[179] N. Madras and A.D. Sokal, "The pivot algorithm : a highly efficient Monte Carlo method for the self-avoiding walk," J. Stat. Phys. **50**, 109–186 (1988).

[180] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087–1092 (1953).

[181] N. Madras, A. Orlitsky, and L.A. Shepp, "Monte-Carlo generation of self-avoiding walks with fixed endpoints and fixed length," J. Stat. Phys. **58**, 159–183 (1990).

[182] A.D. Sokal, "Monte-carlo methods for the self-avoiding walk," ArXiv **hep-lat**, 9509032 (1995).

[183] E.J. Janse van Rensburg, S.G. Whittington, and N. Madras, "The pivot algorithm and polygons : results on the FCC lattice," J. Phys. A : Math. Gen. **23**, 1589–1612 (1990).

[184] A.M. Ferrenberg and R.H. Swendsen, "Optimized Monte-Carlo data analysis," Phys. Rev. Lett. **63**, 1195–1198 (1989).

[185] M.N. Rosenbluth and A.W. Rosenbluth, "Monte-Carlo calculation of the average extension of molecular chains," J. Chem. Phys. **23**, 356–359 (1955).

[186] J. Bartoulis and K. Kremer, "Statistical properties of biased sampling methods for long polymer chains," J. Phys. A : Math. Gen. **21**, 127–146 (1988).

[187] M. Dijkstra, D. Frenkel, and J.-P. Hansen, "Phase separation in binary hard-core mixtures," J. Chem. Phys. **101**, 3179–3189 (1994).

[188] P. Grassberger, "Pruned-enriched rosenbluth method : simulations of θ polymers of chain length up 1000000," Phys. Rev. E **56**, 3682–3693 (1997).

[189] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, "New Monte Carlo algorithm for protein folding," Phys. Rev. Lett. **80**, 3149–3152 (1998).

[190] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, "Growth algorithms for lattice heteropolymers at low tempratures," J. Chem. Phys. **118**, 444–451 (2003).

[191] Z. Alexandrowicz, "Monte Carlo of chains with excluded volume : a way to evade sample attrition," J. Chem. Phys. **51**, 561–565 (1969).

[192] E. Dijkstra, "A note on two problems in connexion with graphs," Numerische Math. **1**, 269–271 (1959).

[193] G. Varani, "Exceptionally stable nucleic acid hairpins," Annu. Rev. Biophys. Biomol. Struct. **24**, 379–404 (1995).

[194] M.A. Glucksmann-Kuis, X. Dai, P. Markiewicz, and L.B. Rothman-Denes, "E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition," Cell **84**, 147–154 (1996).

[195] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J.E. Darnell, *Molecular Cell Biology* (W.H. Freeman, New York, 2000).

[196] O.C. Uhlenbeck, "Tetraloops and RNA folding," Nature **346**, 613–614 (1990).

[197] C.W. Hilbers, C.A.G. Haasnoot, S.H. de Bruin, J.J.M. Joordens, G.A. Van Der Marel, and J.H. Van Boom, "Hairpin formation in synthetic oligonucleotides," Biochimie **67**, 685–695 (1985).

[198] M.J. Serra, T.W. Barnes, K. Betschart, M.J. Gutierrez, K.J. Sprouse, C.K. Riley, L. Stewart, and R.E. Temel, "Improved parameters for the prediction of RNA hairpin stability," Biochemistry **36**, 4844–4851 (1997).

[199] D. Collin, F. Ritort, C. Jarzynski, S.B. Smith, I. Tinoco Jr., and C. Bustamante, "Verification of the crooks fluctuation theorem and recovery of RNA folding free energy," Nature **437**, 231–234 (2005).

[200] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," BMC Bioinformatics **5**, 104 (2004).

[201] R.M. Dirks and N.A. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," J. Comput. Chem. **24**, 1664–1677 (2003).

[202] A. Ke, K. Zhou, F. Ding, J.H. Cate, and J.A. Doudna, "A conformational swith controls hepatitis delta virus ribozyme catalysis," Nature **429**, 201–205 (2004).

[203] P.L. Adams, M.R. Stahley, A.B. Kosek, J. Wang, and S.A. Strobel, "Crystal structure of a self-splicing group I intron with both exons," Nature **430**, 45–50 (2004).

[204] C.A. Theimer, C.A. Blois, and J. Feigon, "Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function," Mol. Cell **17**, 671–682 (2005).

[205] L.X. Shen and I. Tinoco Jr., "The structure of an RNA pseudoknot that causes efficient frameshift in mouse mammary tumor virus," J. Mol. Biol. **247**, 963–978 (1995).

[206] D.W. Staple and S.E. Butcher, "Pseudoknots : RNA structures with diverse functions," PLos Biology **3**, e213 (2005).

[207] A.P. Gultyaev, F.H. van Batenburg, and C.W. Pleij, "An approximation of loop free energy values of RNA H-pseudoknots," RNA **5**, 609–617 (1999).

[208] C.A. Theimer and D.P. Giedroc, "Equilibrium unfolding pathway of an H-type RNA pseudoknot which promotes programmed-1 ribosomal frameshifting," J. Mol. Biol. **289**, 1283–1299 (1999).

[209] S. Cao and S.-J. Chen, "Predicting ribosomal frameshifting efficiency," Phys. Biol. **5**, 16002 (2008).

[210] F.H. van Batenburg, A.P. Gultyaev, and C.W. Pleij, "Pseudobase : a database with RNA pseudoknots," Nucleic Acids Res. **28**, 201–204 (2000).

[211] V.K. Misra, R. Shiman, and D.E. Draper, "A thermodynamic framework for the magnesium-dependent folding of RNA," Biopolymers **69**, 118–136 (2003).

[212] D. Thirumalai and C. Hyeon, "RNA and protein folding : common themes and variations," Biochemistry **44**, 4957–4970 (2005).

[213] N.B. Leontis and E. Westhof, "Geometric nomenclature and classification of RNA base pairs," RNA **7**, 499–512 (2001).

[214] D.E. Draper, "A guide to ions and RNA structure," RNA **10**, 335–343 (2004).

[215] K. Takamoto, R. Das, Q. He, S. Doniach, M. Brenowitz, D. Herschlag, and M.R. Chance, "Principles of RNA compaction : insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations," J. Mol. Biol. **343**, 1195–1206 (2004).

[216] J.H. Cate, A.R. Gooding, E. Podell, K. Zhou, B.L. Golden, C.E. Kundrot, and et al., "Crystal structure of a group I ribozyme domain : principles of RNA packing," Science **273**, 1678–1685 (1996).

[217] F.L. Murphy and T.R. Cech, "An independently folding domain of RNA tertiary structure within *tetrahymena* ribozyme," Biochemistry **32**, 5291–5300 (1993).

[218] R. Zandi and P. van der Schoot, "Size regulation of ssRNA viruses," Biophys. J. **96**, 9–20 (2009).

[219] Y. Hu, R. Zandi, A. Anavitarte, C.M. Knobler, and W.M. Gelbart, "Packaging of a polymer by a viral capsid : the interplay between polymer length and capsid size," Biophys. J. **94**, 1428–1436 (2008).

[220] P. van der Schoot and R. Bruinsma, "Electrostatics and the assembly of an RNA virus," Phys. Rev. E **71**, 061928 (2005).

[221] D. Jost and R. Everaers, "Prediction of RNA multi-loop and pseudoknot conformations from a lattice-based, coarse-grain tertiary structure model," J. Chem. Phys. **132**, 095101 (2010).

[222] A.M. Soto, V. Misra, and D.E. Draper, "Tertiary structure of an RNA pseudoknot is stabilized by "diffuse" Mg ions," Biochemistry **46**, 2973–2983 (2007).

[223] B.A. Berg and T. Neuhaus, "Multicanonical ensemble : a new approach to simulate first-order phase," Phys. Rev. Lett. **68**, 9–12 (1992).

[224] M. Bachmann and W. Janke, "Multicanonical chain-growth algorithm," Phys. Rev. Lett. **91**, 208105 (2003).

[225] M. Bachmann and W. Janke, "Thermodynamics of lattice heteropolymers," J. Chem. Phys. **120**, 6779–6791 (2004).

[226] A. Xayaphoummine, V. Viasnoff, S. Harlepp, and H. Isambert, "Encoding folding paths of RNA switches," Nucleic Acids Res. **35**, 614–622 (2007).

[227] C. Flamm and I.L. Hofacker, "Beyond energy minimization : approaches to the kinetic folding of RNA," Monatsh. Chem. **139**, 447–457 (2008).

[228] F. Wang and D.P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," Phys. Rev. Lett. **86**, 2050–2053 (2001).

[229] J.I. Siepmann and D. Frenkel, "Configurational-bias Monte-Carlo : a new sampling scheme for flexible chains," Mol. Phys. **75**, 59–70 (1992).

[230] C.R. Cantor and P.R. Schimmel, *Biophysical Chemistry* (Freeman, New York, 1980).

[231] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," Monat. Chem. **125**, 167–188 (1994).

[232] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," Nucleic Acids Res. **9**, 133–148 (1981).

[233] J.S. McCaskill, "The equilibrium partition function and base-pair binding probabilities for RNA secondary structure," Biopolymers **29**, 1105–1119 (1990).

[234] W. Zhang and S.-J. Chen, "RNA hairpin-folding kinetics," Proc. Natl. Acad. Sci. USA **99**, 1931–1936 (2002).

[235] W. Zhang and S.-J. Chen, "Exploring the complex folding kinetics of RNA hairpins : I. General folding kinetics analysis," Biophys. J. **90**, 765–777 (2006).

[236] D.T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem. **81**, 2340–2361 (1977).