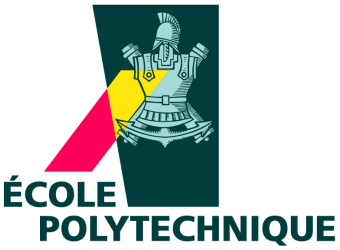
Prediction de structures secondaires d'ARN avec pseudo-noeuds

Michaël Bon



Thèse présentée pour obtenir le titre de

DOCTEUR DE L'ECOLE POLYTECHNIQUE

Spécialité : Physique théorique

par

Michaël Bon

Prédiction de structures secondaires d'ARN avec pseudo-nœuds

Soutenue le 21 septembre 2009 devant le jury composé de :

Peter CLOTE,	Boston College	Rapporteur
Sebastian DONIACH,	Stanford University	Rapporteur
Alain DENISE,	Université Paris-Sud	Examinateur
Henri ORLAND,	IPhT, CEA Saclay	Directeur de thèse
Thomas SIMONSON,	Ecole Polytechnique, Palaiseau	Examinateur

Remerciements

Je remercie Henri, mon directeur de thèse, pour ces trois années qui furent finalement assez agréables. Le souvenir de son attitude infailliblement encourageante, de sa bonne humeur, de son professionnalisme, de sa disponibilité, de sa tolérance, de sa subtilité et de sa compétence ne sera jamais éclipsé par celui de son humour douteux.

Je remercie mes rapporteurs pour avoir accepté d'effectuer cette tâche dans des délais très serrés et pour l'attention soutenue avec laquelle ils ont lu le présent manuscrit, allant jusqu'à me signaler une faute de frappe dans la bibliographie. Je suis honoré qu'ils aient également accepté de traverser l'Atlantique pour assister à la soutenance de ma thèse.

Je remercie les autres membres de mon jury pour le temps qu'ils ont consacré à mes travaux.

Je remercie les thésards du labo pour les deux heures quotidiennes de leur compagnie enjouée, intelligente et stimulante. J'en viendrais presque à regretter mes nombreuses victoires sans partage au jeu de société "Les colons de Catane". Je remercie Emeline particulièrement, pour tout.

Je remercie tous les membres du secrétariat de l'IPhT pour leur bienveillance et leur professionnalisme qui font de ce laboratoire un paradis administratif souriant.

Je remercie le personnel du restaurant 3 du CEA Saclay pour la qualité de sa programmation culinaire et de son service.

Je remercie mes parents et ma soeur pour d'innombrables raisons, la moindre d'entre elles étant d'avoir à trois reprises lu ce manuscrit à la recherche de fautes d'orthographe. Je regrette cependant de ne pas avoir été assez clair pour qu'ils sachent enfin replier de l'ARN.

Je remercie JB, l'ami.

Je remercie Erik Gerets d'avoir remis l'OM sur le chemin de la Ligue des Champions.

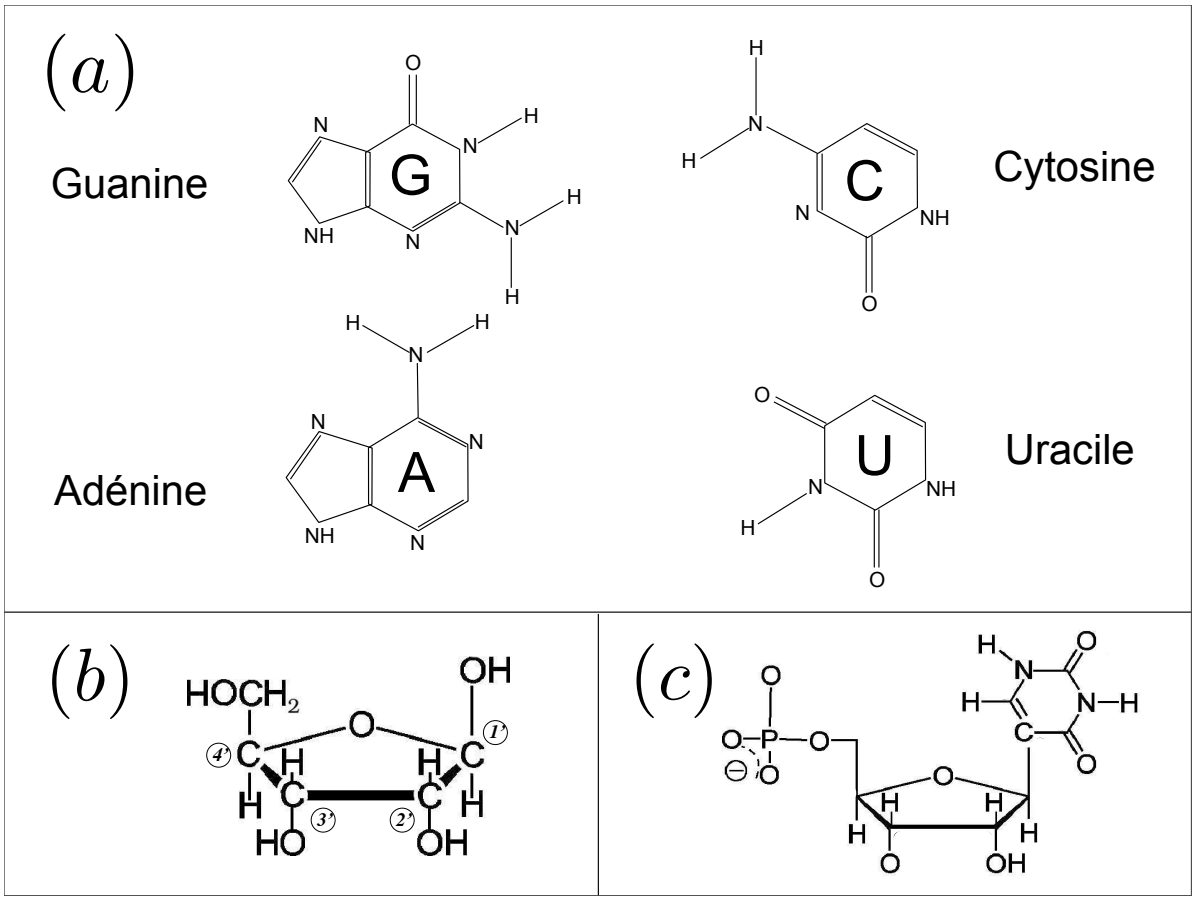
Table des matières

Chapitre 1

Introduction : structure et fonction de l'ARN

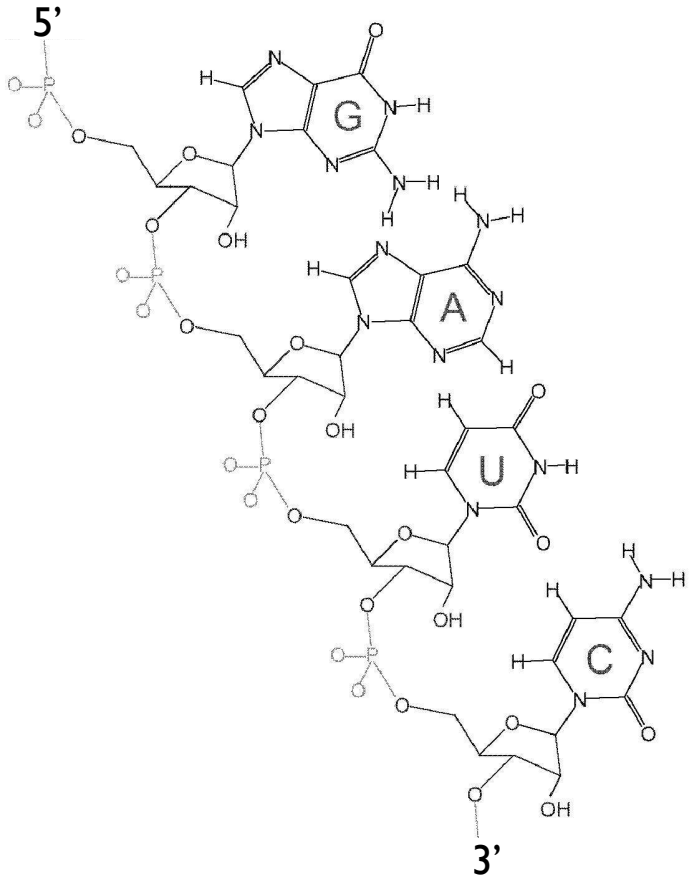
Structure de l'ARN

L'acide ribonucléique (ARN) constitue, avec l'ADN et les protéines, une des trois classes de biopolymères à séquence codante présentes dans toute cellule vivante. Chaque ARN est une chaîne pouvant être composée de 20 à 3000 nucléotides. Ces nucléotides sont chacun composés d'une base azotée, d'un sucre ribose et d'un groupe phosphate chargé négativement. Ils peuvent être de quatre types différents selon la nature de leur base azotée : adénine (A), guanine (G), cytosine (C), uracile (U). Les bases A et G sont appelées purines, C et U pyrimidines.



(a) les quatre bases azotées (b) le ribose (c) un nucléotide "U"

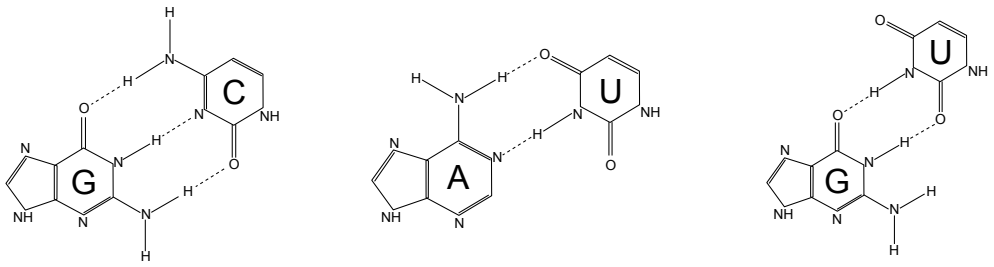
Le phosphate d'un nucléotide peut estérifier l'alcool situé en 3' d'un autre nucléotide. Ainsi, les nucléotides peuvent se polymériser et former une chaîne. En raison de l'asymétrie de cette liaison, l'ARN est une molécule orientée. Ses extrémités sont par convention notées 5' et 3'.



Un ARN de quatre bases de long. Sa séquence se note ${}^{5'}GAUC^{3'}$ *ou* $GAUC$

La séquence de nucléotides d'un ARN constitue sa *structure primaire*. Cette structure primaire s'écrit simplement avec un alphabet à quatre lettres : A, C, G et U.

Deux nucléotides peuvent effectuer des liaisons hydrogène entre eux. Pour ce faire, ils peuvent s'agencer selon plusieurs géométries répertoriées dans [1]. Cependant, trois de ces géométries sont particulièrement stables : ce sont les paires "Watson-Crick" $A = U$ et $G \equiv C$ et la paire "Wobble" $G = U$.



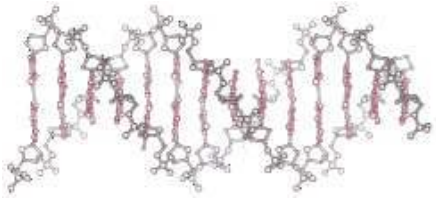
Géométrie des paires Watson-Crick et Wobble

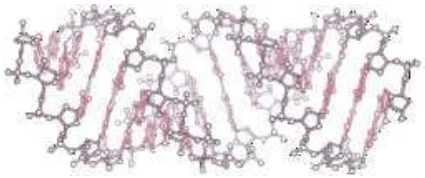
La paire $G \equiv C$ met en jeu trois liaisons hydrogène, les paires $A = U$ et $G = U$ en forment deux. Rappelons que la liaison hydrogène est une interaction d'origine purement quantique qui ne peut se produire qu'entre molécules polaires, par déplacement des nuages électroniques. Elle joue un rôle crucial en biologie. Sans entrer dans les détails chimiques, les particularités importantes de la liaison hydrogène sont sa courte portée et sa forte directionnalité qui lui confèrent un aspect saturant déterminant. Du fait de ces propriétés, on peut considérer que chaque nucléotide ne peut former de liaisons hydrogène qu'avec un et un seul autre nucléotide. En première approximation, il n'y a pas de triplets de nucléotides. Leurs bases ne peuvent que *s'apparier*.

La liste de ses bases appariées constitue la *structure secondaire* d'un ARN.

L'ARN ressemble beaucoup à l'ADN, sauf qu'il est en général simple-brin (alors que l'ADN est double-brin), que le sucre des bases est un ribose (et non un désoxyribose comme dans l'ADN) et que la base uracile remplace la thymine de l'ADN. Chez l'ADN double brin, la formation de paires stables entre chaque brin est à l'origine de sa structure bien connue en double hélice. Chez l'ARN simple brin, elles conduisent au repliement de l'ARN sur lui-même. En coudant la chaîne, l'ARN peut également former des motifs hélicoïdaux. En raison des particularités chimiques de l'ARN, cette hélice a une géométrie différente de celle de l'ADN.

(a) (b)





hélice de type "A" de l'ARN *hélice de type "B" de l'ADN*

1.1 Rôle de l'ARN dans la cellule

L'ARN a longtemps été pensé uniquement comme un acteur de la synthèse des protéines à laquelle il prend part à plusieurs niveaux. Dans cette vision, connue sous le nom de "dogme central de la biologie moléculaire", un ARN, qualifié de "messager" (ARNm), est créé à partir de l'ADN génomique au cours d'une opération appelée la *transcription*. Cet ARNm, vecteur inerte de l'information codée dans sa séquence, est ensuite livré au ribosome, complexe de protéines et d'ARN dits ribosomaux (ARNr), pour donner lieu à la synthèse de protéines par un mécanisme appelé *traduction*. Celle-ci requiert la lecture de la séquence de l'ARNm et l'ajout successif d'acides aminés à la protéine synthétisée : cette opération est assurée par l'ARN de transfert (ARNt).

Bien que correcte, on s'aperçoit depuis les années 90 que cette vision du rôle de l'ARN dans la cellule est très limitée. En effet, la liste de ses fonctions biologiques ne cesse de s'allonger. Tout d'abord, de nouvelles aptitudes des ARNm, ARNt et ARNr ont été découvertes.

- Alors qu'on pensait que l'ARNr ne servait que de plate-forme à diverses enzymes qui, elles, catalysaient l'élongation de la protéine, on a découvert que la réalité était en fait inverse : c'est l'ARNr qui, entre autres, catalyse la formation de liaisons peptidiques au sein des protéines nouvellement créées [2].

- Les ARNt ont une structure imitée par d'autres ARN, ce qui permet à ces derniers d'intervenir dans la réplication, l'épissage et la régulation de la traduction [3].

- L'ARNm n'est pas un vecteur passif de l'information génétique : sa propre structure

joue un rôle dans la régulation de sa traduction. Ceci est par exemple possible par l'intermédiaire de motifs codés dans la séquence de l'ARN comme les pseudo-nœuds (cf chapitre 3). L'article [4] présente un exemple particulièrement subtil de réarrangements conformationnels se déroulant à dessein lors de la traduction d'un ARNm de E. Coli. Les potentialités de la structure des ARNm éclairent d'une lumière nouvelle les avantages évolutifs de la dégénérescence du code génétique.

La grande révolution de l'ARN vient toutefois de la découverte des ARN non-codants (ARNnc). Ceux-ci représentent en fait la majorité des ARN transcrits par la cellule. Ces ARN ne sont pas traduits en protéines et interviennent directement dans plusieurs voies métaboliques.

- Les micro-ARN (μARN), une catégorie d'ARNnc, jouent un rôle crucial dans la régulation post-transcriptionnelle du génome. Les μARN sont de courts ARN de 22 bases de long qui régulent la traduction en s'appariant avec des ARNm, entraînant généralement leur inhibition ou leur dégradation. Des travaux récents ont montré que les μARN pouvait aussi promouvoir la synthèse de leur ARNm cible [5]. On pense actuellement que 30% de l'ADN codant est ainsi régulé par des μARN [6]. Leur importance s'est avérée dans de nombreux processus comme l'apoptose, la différenciation, le développement et la prolifération cellulaire [7].

- Les "ribocommutateurs" (*riboswitches*) constituent une autre catégorie d'ARNnc qui démontre les capacités catalytiques de l'ARN. Les ribocommutateurs sont des ARN capables de changer de conformation en réponse à certains stimuli de manière à exercer une fonction de régulation dans la cellule. Ces stimuli peuvent être par exemple la température [8] ou l'interaction avec un métabolite [9].

Les aptitudes des ribocommutateurs étayent l'hypothèse du "monde ARN" [10] qui place l'ARN comme forme primitive du génome dans le cadre de l'évolution. L'ARN est ainsi devenu un acteur polyvalent et incontournable du métabolisme cellulaire.

1.2 Le problème du repliement

La clé de la compréhension des mécanismes d'action de l'ARN réside dans la connaissance de sa structure tridimensionnelle. Ce point de vue est justifié de manière frappante par l'exemple des ARNt. En raison de leur rôle indispensable dans la synthèse des protéines, tous les organismes ont besoin de ce type d'ARN et des centaines d'études ont montré que la structure tertiaire des ARNt est universellement conservée alors que leurs séquences primaires peuvent différer considérablement d'une espèce à l'autre. Ainsi, c'est bien la structure tridimensionnelle, également appelée *structure tertiaire*, qui est nécessaire à la fonction et non la séquence de bases. Comment déterminer la structure tertiaire d'un ARN ?

Les méthodes expérimentales étant encore trop coûteuses en temps et en argent, il est justifié de développer des méthodes bio-informatiques de prédiction de structures. La prédiction directe de la structure tertiaire est cependant un problème encore trop complexe bien que quelques travaux s'y essayent [11]. Une étape intermédiaire prometteuse est celle de la prédiction de structures secondaires qui est l'objet de cette thèse. La connaissance de la structure secondaire apporte en effet une information très précieuse sur la structure tertiaire de l'ARN. Il a été montré expérimentalement que le repliement des ARN suit un processus hiérarchique : la structure primaire code la structure secondaire qui elle-même sert de support à la structure tertiaire [12]. De fait, la structure secondaire apporte la principale contribution énergétique qui explique la structure tertiaire. Ainsi la connaissance de la structure secondaire fournit un jeu de contraintes qui simplifie considérablement le problème de la détermination de la structure tertiaire. Prédire la structure secondaire est, en outre, un problème plus général. En effet, la structure tertiaire dépend fortement de facteurs externes comme les conditions salines. L'ARN est une molécule chargée négativement et les ions de la solution ont pour effet d'écranter l'interaction coulombienne en en diminuant grandement la portée, ce qui favorise les conformations tertiaires compactes. En particulier, la stabilisation de la structure tertiaire est très influencée par la présence d'ion magnésium divalents Mg^{++} : la cinétique et le taux de succès du repliement sont considérablement accrus lorsque la concentration $[Mg^{++}]$ croît [13]. La structure secondaire d'un ARN n'est au contraire définie que par la combinatoire des paires de bases de sa séquence.

Alors, comment déterminer la structure secondaire ?

Ce problème consiste à identifier quelles bases s'apparient. Sa difficulté est donc premièrement combinatoire car le nombre de possibilités croît exponentiellement avec la longueur de la séquence : c'est le problème informatique de l'*énumération*. Peut-on écrire un algorithme de prédiction capable de proposer efficacement la bonne structure secondaire quelle que soit la séquence ? La réponse est affirmative en se limitant aux structures ne contenant pas un certain motif structural, le pseudo-nœud. En incluant les pseudo-nœuds, le problème devient NP-complet [14] et des solutions efficaces sont toujours requises.

La deuxième difficulté réside dans la caractérisation de la structure secondaire réelle : comment distinguer la bonne structure secondaire de repliements concurrents ? Il est d'usage d'utiliser un critère inspiré de la mécanique statistique portant sur l'énergie libre de ces repliements. Ce problème de caractérisation s'identifie ainsi au problème de l'*estimation* de l'énergie libre d'une structure secondaire.

Dans ma thèse, je me suis intéressé à ces deux questions en explorant pour chacune des pistes nouvelles.

1.3 Plan de la thèse

Concernant le problème de l'estimation de l'énergie d'une structure secondaire, je procède dans le chapitre 1 à un examen critique des modèles d'énergie actuellement disponibles et propose une nouvelle démarche pour les reparamétrer. Les résultats, répartis entre les chapitres 1 et 3, sont très prometteurs. Le chapitre 4 est consacré à d'autres travaux de simulation par dynamique moléculaire ayant le même but mais dont les résultats n'ont pas été utilisés.

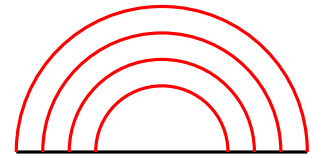
Pour le problème de l'énumération, je présente dans le chapitre 2 une nouvelle classification des pseudo-nœuds qui permet de simplifier le problème de leur prédiction. Je propose un nouvel algorithme de prédiction de pseudo-nœuds fondé sur cette classification dans le chapitre 3 après avoir rappelé le principe des algorithmes de prédiction sans pseudo-nœuds. La comparaison avec les meilleurs algorithmes actuels permet de conclure à une nette amélioration de la qualité de prédiction.

1.4 Notations

Représentation des structures secondaires

Les structures secondaires sont représentées par des diagrammes où la séquence est écrite sur une ligne et où les paires de bases sont matérialisées par des arcs tracés au-dessus. Considérons par exemple la séquence ${}^{5'}CCCCAAAAGGGG^{3'}$ où les quatre premières cytosines s'apparient avec les quatre dernière guanines.

Cette structure secondaire est représentée par le diagramme suivant :



Cette structure secondaire sera aussi parfois annotée à l'aide de parenthèses ouvrantes et fermantes, chacune représentant une des deux bases d'une même paire. Les bases libres sont représentées par des points.

```
CCCCAAAAGGGG
( ( ( ( . . . . ) ) ) )
```

Définitions

- la **sensibilité** d'une structure secondaire prédite est le rapport du nombre de paires de bases correctement prédites au nombre total de paires de bases de la structure secondaire réelle.
- la **valeur positivement prédite** (**VPP**) d'une structure secondaire prédite est le rapport du nombre de paires de bases correctement prédites au nombre total de paires de bases de la structure secondaire prédite.

Chapitre 2

Le modèle d'énergie libre

Le but de ce chapitre est d'expliquer les raisons qui m'ont amené à construire un nouveau modèle d'énergie libre. J'y expose les principaux paramétrages actuellement disponibles dans la littérature, en discute les hypothèses implicites et explicites et en étudie les limites. Constatant que ces différents modèles ne s'accordent pas sur leurs paramètres les plus essentiels, j'en conclus que cette question du paramétrage est encore ouverte et je propose une nouvelle démarche pour y répondre. Le plan de ce chapitre est le suivant :

I) Finalité du modèle d'énergie libre
Discussion sur le principe de la "minimisation d'énergie libre", principale méthode utilisée pour prédire des structures secondaires d'ARN

II) Structure du modèle d'énergie libre
Présentation des différents paramètres du modèle

III) Paramétrage à partir de données expérimentales
Présentation du "modèle de Turner" : description de la procédure expérimentale et discussion des hypothèses utilisées pour interpréter les résultats

IV) Paramétrage par analyse des bases de données de structures secondaires
1) CONTRAfold
Une estimation des paramètres fondée sur le formalisme des grammaires algébriques

2) CG
Une estimation des paramètres fondée sur une optimisation mélangeant mesures expérimentales et données structurales

V) Paramétrage par dynamique moléculaire

VI) Paramétrage par une nouvelle démarche : MC
Présentation d'une nouvelle manière d'estimer les paramètres par analyse des données structurales. Résultats et réflexions.

2.1 Finalité du modèle d'énergie

"*Comment caractériser la structure d'équilibre d'un ARN à température ambiante ?*"

Trouver la bonne manière dont une séquence d'ARN donnée se replie est désigné dans la littérature comme un problème de "minimisation de son énergie libre" ([15], [16], [17]). Une telle formule peut paraître surprenante aux yeux d'un physicien. L'énergie libre est en effet une grandeur caractérisant la variabilité d'un système : elle représente l'ensemble des configurations que peut adopter un système dans l'espace et ne saurait indiquer l'une d'entre elles en particulier. Un brin d'ARN immergé dans un solvant a une énergie libre donnée qui encode tous ses repliements possibles et celle-ci ne peut être l'objet d'une minimisation. Parler de minimisation de l'énergie libre est donc un abus de langage qui correspond à la manière de penser l'ARN en usage dans le domaine et j'explicite ci-dessous ce que cette appellation désigne réellement.

Soit $\mathcal{Z}$ la fonction de partition d'un système, par exemple une molécule d'ARN d'une certaine séquence dans un volume d'eau. Ce système peut adopter plusieurs configurations $\mathcal{C}$, ayant une énergie $E(\mathcal{C})$. $\mathcal{Z}$ est alors donnée par :

$$\mathcal{Z} = \sum_{configurations\ \mathcal{C}} e^{-\beta E(\mathcal{C})} \quad \text{où} \quad \beta = (k_B T)^{-1}$$

L'énergie libre F du système s'en déduit par :

$$e^{-\beta F} = \mathcal{Z}$$

En physique, on s'intéresse souvent à l'état fondamental du système : celui-ci correspond à la configuration dont l'énergie est minimale. En notant $\mathcal{C}_0$ cette configuration, on a :

$$\mathcal{Z} = e^{-\beta E(\mathcal{C}_0)}(1 + \sum_{\substack{configurations\ \mathcal{C} \\ \mathcal{C} \neq \mathcal{C}_0}} e^{-\beta[E(\mathcal{C}) - E(\mathcal{C}_0)]})$$

Les termes $[E(\mathcal{C}) - E(\mathcal{C}_0)]$ étant positifs, on voit que l'état fondamental est celui qui domine la fonction de partition lorsque $\beta \to \infty$, ce qui correspond à la limite de température nulle : la somme se réduit à un seul terme, le système se "gèle" dans un unique état qui est l'état fondamental. Ce n'est donc pas cette configuration définie sans ambiguïté qui intéresse le biologiste dont les systèmes d'intérêt ne le sont qu'à température ambiante. Alors comment définir "l'état natif" du système à température ambiante ? La réponse utilisée dans la littérature n'est pas un théorème de mécanique

statistique mais un choix de biophysicien reposant sur une certaine manière de penser l'ARN. La fonction de partition décrit une profusion de possibilités différemment pondérées de manière à décrire correctement les *fluctuations* du système. *A priori* cela s'accorde mal avec une hypothèse centrale de la biologie, étayée par d'abondantes mesures reproductibles, qui implique que la structure d'un biopolymère est rigoureusement organisée pour assurer sa fonction. Dans le cas de l'ARN, cette structure serait le résultat de trois niveaux d'organisation *hiérarchisés* : les structures primaire, secondaire et tertiaire [12]. La structure secondaire s'établit à partir de la structure primaire et la structure tertiaire se forme une fois la structure secondaire achevée. Un tel cheminement logique est étranger à la description du système par la mécanique statistique. Comment concilier ces deux descriptions ? La nécessité d'avoir une structure bien déterminée a amené à postuler que la fonction de partition d'un ARN naturel est telle qu'un ensemble de configurations proches les unes des autres la domine. Ce sont les nombreuses observations des expérimentateurs qui servent de base pour définir en quoi consiste cette proximité : sont considérées comme proches des configurations qui partagent la même structure secondaire, c'est-à-dire qui partagent les mêmes paires de bases Watson-Crick et Wobble.

Cette définition ne contraint pas la disposition des bases non-appariées, ce qui explique que plusieurs configurations correspondent à une même structure secondaire. Cette définition suggère de réécrire la fonction de partition avec la factorisation suivante :

$$\mathcal{Z} = \sum_{\substack{structures \\ secondaires\ S}} \left[\sum_{\substack{configurations\ \mathcal{C} \\ \mathcal{C}\ satisfait\ S}} e^{-\beta E(\mathcal{C})} \right]$$

et on définit "l'énergie libre d'une structure secondaire S" comme :

$$e^{-\beta F(S)} = \sum_{\substack{configurations\ \mathcal{C} \\ \mathcal{C}\ satisfait\ S}} e^{-\beta E(\mathcal{C})}$$

$$\text{de sorte que :} \qquad \mathcal{Z} = e^{-\beta F} = \sum_{\substack{structures \\ secondaires\ S}} e^{-\beta F(S)}$$

Ainsi, on définit l'état natif d'un ARN comme la structure secondaire S dont le poids $e^{-\beta F(S)}$ est maximal, c'est-à-dire dont l'énergie libre $F(S)$ est minimale. Ce repliement est le plus stable des repliements.

J'insiste sur l'importance cruciale du choix de cette définition : une autre définition entraînerait une autre factorisation pour laquelle il y aurait également un ou plusieurs poids maximal mais on suppose pour des raisons phénoménologiques que la factorisation la plus pertinente pour décrire l'état d'équilibre d'un ARN à température ambiante est celle associée à la donnée de l'ensemble de ses paires de bases. La pertinence de cette factorisation dépend en fait d'une autre hypothèse : elle suppose que les énergies d'activation nécessaires pour passer d'une structure secondaire à une autre sont fortes, tandis que les énergies d'activation pour passer d'une configuration à une autre au sein d'une même structure secondaire sont faibles. Cette hypothèse est cependant très raisonnable car, comme on le verra, les motifs principaux des structures secondaires ont des énergies de formation très supérieures à l'énergie thermique $k_B T$ à température ambiante.

J'insiste aussi sur le fait que cette méthode qui consiste à trier les configurations microscopiques d'un système n'est pas une méthode "canonique" de mécanique statistique où les propriétés d'équilibre sont calculées comme moyennes prises sur l'ensemble des configurations. Une description de "la" structure d'un ARN par ce moyen a été proposée dans l'article "*The equilibrium partition function and base pair probabilities for RNA secondary structure*" [18]. Son auteur y calcule les probabilités P_{ij} que la base i soit appariée avec la j, pour toutes les bases de la séquence de longueur L. Ainsi, la structure d'un ARN est représentée par une matrice de probabilités $L \times L$. Cette méthode ne donne cependant pas l'image d'une structure secondaire clairement déterminée et fixe, ni ne permet de répondre directement à la question posée par le biologiste qui est de savoir quelle est la meilleure des structures secondaires. En effet, les probabilités d'apparier i à j et k à l ne sont pas indépendantes et on ne peut donc pas librement combiner les probabilités lues dans cette matrice pour en effectuer l'analyse.

La méthode de "minimisation de l'énergie libre", dont le résultat est la prédiction d'une structure secondaire, a été ainsi construite après que les résultats expérimentaux ont établi comme réalité objective l'existence d'une structure secondaire déterminée et elle propose un critère pour la prédire. Cette méthode doit être jugée sur ses succès et échecs.

Unicité de la structure secondaire d'énergie libre minimale

A séquence donnée, la structure secondaire de poids maximal est-elle unique ?
Un argument biologique nous invite à le penser : le bon fonctionnement d'une molécule étant associé à son bon repliement, la nature a sélectionné les séquences pour lesquelles le poids de ce bon repliement se démarque clairement des autres. Cependant, de nouvelles possibilités ont été expérimentalement montrées dans "*One sequence, two rybozymes : implications for the emergence of new rybozyme folds*" [19]. Dans cet article, les auteurs ont sélectionné deux ARN de même longueur dont les structures secondaires respectives S1 et S2 sont connues et catalysent respectivement les réactions R1 et R2. Ils ont ensuite conçu une séquence pouvant alternativement se replier selon S1 et S2 et ont alors expérimentalement vérifié que cet ARN donnait bien lieu aux catalyses de R1 et R2. Une explication possible serait qu'une population de cet ARN existe sous les deux formes S1 et S2 car elles correspondraient à deux minima d'énergie libre, au sens ci-dessus. Cependant, on peut aussi objecter en considérant que ces repliements alternatifs sont induits par l'interaction avec les ligands propres à chacune des réactions. A ma connaissance, aucune expérience n'a jusqu'à présent démontré l'existence d'un ARN naturel sous une forme double, non induite par le milieu.

Interprétation de $F(S)$

L'enjeu du modèle d'énergie est ainsi de calculer le terme $F(S)$ à S donné.
Théoriquement, l'énergie d'une configuration inclut la contribution du solvant dans lequel baigne l'ARN. Pour éliminer la partie de cette contribution non spécifique au repliement, on travaille avec la quantité $e^{-\beta\Delta F(S)} = e^{-\beta(F(S)-F(\emptyset))}$, où $F(\emptyset)$ désigne l'énergie libre de l'état "déplié" où toutes les bases sont libres. Ce terme $\Delta F(S)$ comprend une partie enthalpique et une partie entropique.

- la partie enthalpique $\Delta H = H(S) - H(\emptyset)$ s'explique par une contribution interne à l'ARN, liée à la formation de paires de bases, et une différence d'enthalpie liée aux interactions avec l'eau. En considérant qu'une base non appariée dans S interagit de la même manière avec le solvant que dans $\emptyset$, la différence d'enthalpie ΔH est, au final, associée uniquement à la formation de paires de bases.
- la partie entropique est due à la variabilité conformationelle des bases libres et on peut par exemple l'estimer grâce à la théorie des polymères.

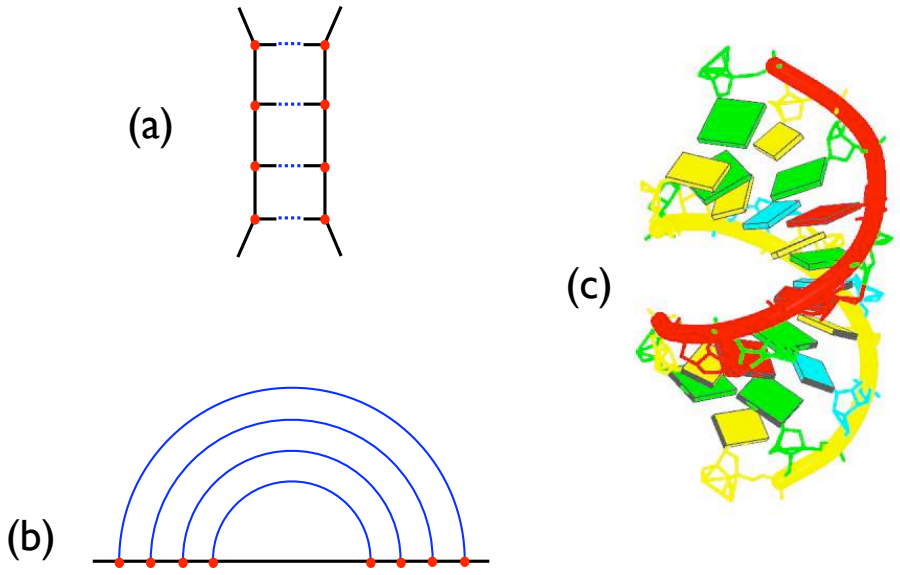
Ainsi est fait le modèle d'énergie développé depuis de nombreuses années : il calcule l'énergie libre d'un repliement comme la somme d'énergies d'appariement et de contributions entropiques pour les bases libres.

2.2 Structure du modèle d'énergie

Le principe qui domine le repliement de l'ARN est la formation de paires de bases successives, appelées hélices. La formation des ces hélices entraîne en retour une certaine organisation des bases non-appariées : celles-ci peuvent former des motifs différents et, sur la base d'observations expérimentales, on en distingue classiquement quatre types appelés renflements, têtes d'épingle, boucles internes et boucles multi-hélices. Je décris dans cette section ces différents motifs et le principe de leur paramétrage.

2.2.1 Base appariées : Hélices

Il est systématiquement observé dans la nature que les structures d'ARN s'organisent autour de successions de paires de bases qui engendrent une structure hélicoïdale régulière, à l'instar de l'ADN.

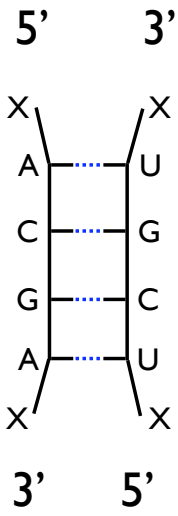


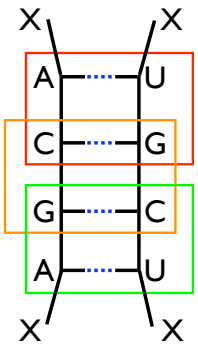
Hélices. (a) Représentation schématique (b) Représentation diagrammatique (c) Représentation 3D

Cette conformation hélicoïdale compacte est le résultat de contributions "verticales" et "horizontales". La contribution horizontale désigne la formation de liaisons hydrogène entre bases complémentaires de deux brins différents, tandis que la contribution verticale, dite d'empilement ("*stacking*"), désigne l'interaction entre cycles aromatiques de nucléotides voisins sur un même brin. Cette dernière est la somme de plusieurs effets : interactions électrostatiques, interactions π (de nature enthalpique) et interactions hydrophobes (de nature entropique). Ces interactions verticales tendent à superposer parallèlement et consécutivement les cycles aromatiques composant la séquence et sont assez fortes pour structurer un ARN simple brin en solution ([20], [21], [22]). Ainsi, dans l'optique de paramétrer la formation d'une liaison Watson-Crick ou Wobble au sein d'une hélice, il est insuffisant de considérer seulement la nature de celle-ci : sa contribution énergétique dépend aussi de ses interactions avec ses voisines et il est donc naturel de la paramétrer par un terme de la forme $\Delta F \begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \\ Z-\overline{Z} \end{pmatrix}$ où $Y-\overline{Y}$ est la paire d'intérêt et $X-\overline{X}$ et $Z-\overline{Z}$ sont ses voisines. En se limitant aux paires Watson-Crick et Wobble, le calcul d'énergie libre d'hélice requerrait ainsi $6^3 = 216$ paramètres. Pour des raisons d'économie du modèle, liées au peu de données disponibles lorsqu'il commençait à être développé, les chercheurs se sont limités à des termes de la forme $\Delta F \begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}$ qui ne sont plus qu'au nombre de 36 et même 21 si l'on tient compte des symétries (on doit avoir $\Delta F \begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix} = \Delta F \begin{pmatrix} \overline{Y}-Y \\ \overline{X}-X \end{pmatrix}$). Par la suite, j'appelerai le motif $\begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}$ une "dipaire" et les différences d'énergie libre associées seront notées ΔF_d.

Pour chaque paire en bout d'hélice, un terme ΔF_t dépendant de celle-ci et des bases libres voisines est ajouté, ce qui représente $6 \times 4^2 = 96$ nouveaux paramètres. Il n'y a pas de symétrie qui réduise ce nombre puisqu'il concerne quatre bases dont deux sont liées et deux sont libres.

Voici un schéma explicatif de la façon de calculer l'énergie libre d'une hélice avec ce modèle :





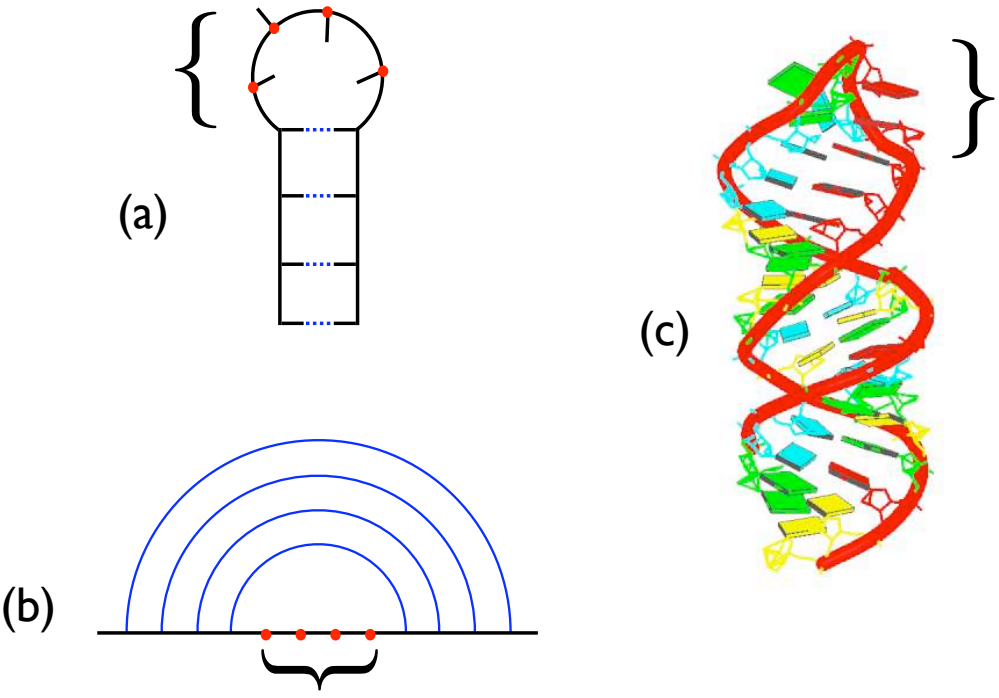
$$\Delta F(\text{hélice}) = \Delta F_t \begin{pmatrix} \text{X} & \text{X} \\ \text{A}-\text{U} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{A}-\text{U} \\ \text{C}-\text{G} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{C}-\text{G} \\ \text{G}-\text{C} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{G}-\text{C} \\ \text{A}-\text{U} \end{pmatrix} + \Delta F_t \begin{pmatrix} \text{A}-\text{U} \\ \text{X} & \text{X} \end{pmatrix}$$

2.2.2 Bases libres

Les bases libres forment des simple brins ou des boucles et on en distingue 4 types selon ce que ces boucles relient.

Têtes d'épingle (*hairpin loops*)

La tête d'épingle est le type de boucle qui relie les deux segments d'une hélice.

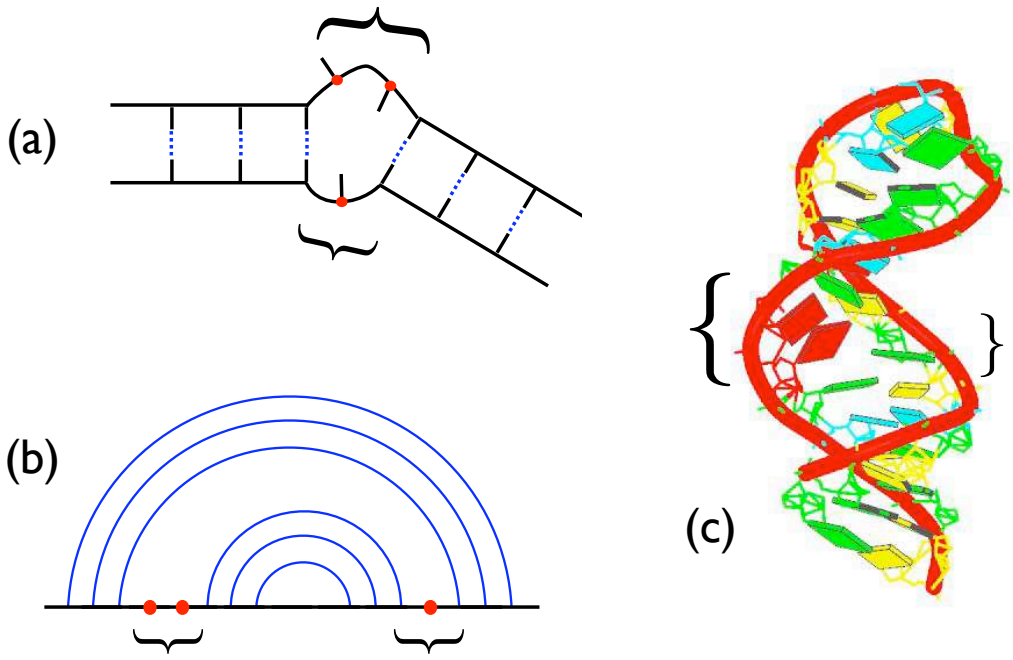


Têtes d'épingle. (a) Représentation schématique (b) Représentation diagrammatique (c) Représentation 3D

On lui attribue une pénalité d'initiation et une pénalité entropique dépendant de la longueur de la tête d'épingle.

Boucles internes (*internal loops*)

La boucle interne est le type de boucle qui relie deux hélices. Elle peut se voir comme une tête d'épingle interrompue par une hélice. L'ensemble des bases libres qui la constituent est disjoint.

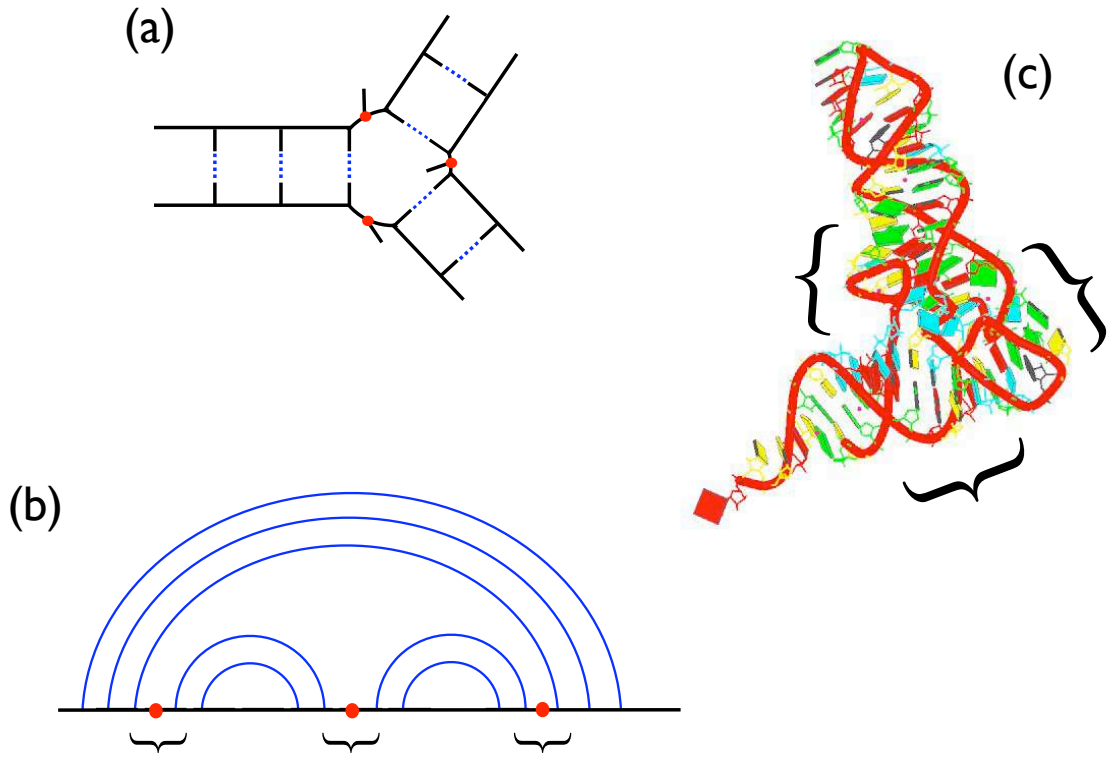


Boucles internes. (a) Représentation schématique (b) Représentation diagrammatique (c) Représentation 3D

On attribue à la boucle interne une pénalité d'initiation et une pénalité $\Delta F_{bi}(l_1, l_2)$, où l_1 et l_2 sont le nombre de bases libres sur chacun de ses segments. Cette pénalité contient un terme entropique dépendant de la longueur totale de la boucle $l_1 + l_2$ et une pénalité d'asymétrie dépendant de $|l_1 - l_2|$. Pour les boucle internes suffisamment courtes, des termes correctifs dépendant de la séquence peuvent être ajoutés.

Boucles multi-hélices (*multibranch loop*)

La boucle multi-hélices est une boucle reliant au moins 3 hélices. Elle peut se voir comme une tête d'épingle interrompue par au moins deux hélices.

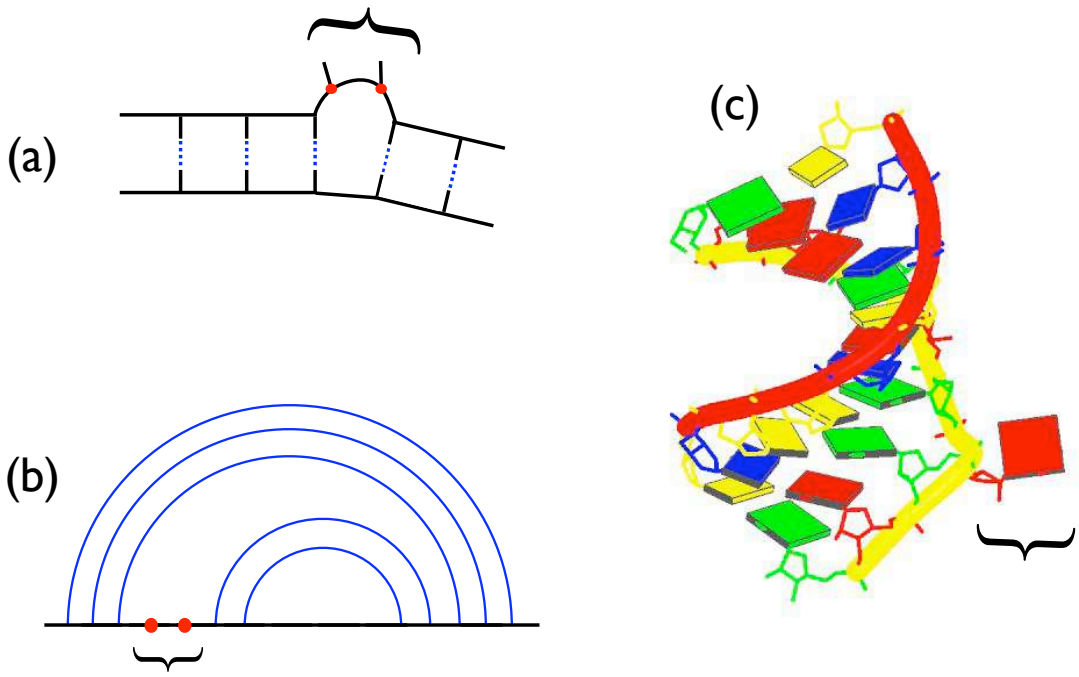


Boucles multi-hélices. (a) Représentation schématique (b) Représentation diagrammatique (c) Représentation 3D

On lui attribue une pénalité d'initiation, à laquelle s'ajoutent de nouvelles pénalités d'initiation pour chacune des k hélices issues de la boucle multi-hélices ainsi qu'une pénalité entropique liée au nombre de bases libres constituant la boucle. Une boucle multi-hélices dont sont issues k hélices sera par la suite appelée "boucle k-hélices".

Renflement (*bulge*)

Le renflement est une boucle nichée dans une hélice.



Renflements. (a) Représentation schématique (b) Représentation diagrammatique (c) Représentation 3D

On lui attribue une pénalité d'initiation et un terme entropique dépendant du nombre de bases libres le constituant. Comme pour les boucles internes, une dépendance à la séquence peut être ajoutée pour les renflements suffisamment courts.

Linéarité du modèle et notation

L'énergie libre d'une structure secondaire donnée est ainsi calculée : les différents motifs qui la composent sont énumérés et son énergie libre est la somme des énergies libres de ces motifs. Ainsi, on peut associer à une structure secondaire S le vecteur $\boldsymbol{n}(S)$, ayant autant de composantes que le modèle de paramètres, tel que $\boldsymbol{n}_i(S)$ soit le nombre d'occurrences du paramètre i dans S. En notant $\boldsymbol{p}$ le vecteur des paramètres du modèle, on a alors $\Delta F(S) = \boldsymbol{n}^T(S)\boldsymbol{p}$. Cette notation vectorielle sera conservée par la suite.

Comment estimer les valeurs numériques des paramètres du modèle $\boldsymbol{p}$?

Deux approches principales ont été développées dans ce sens :

- La première, l'approche expérimentale, est celle dont les résultats sont les plus communément utilisés : ce sont le plus souvent les paramètres par défaut des algorithmes de repliement actuellement disponibles. Cette méthode repose sur des expériences de mesure de l'absorption ultra-violette de courts ARN en fonction de la température. L'interprétation de ces expériences dépend d'une hypothèse forte dont je discuterai les conséquences.

- La deuxième approche consiste à déterminer à l'aide une certaine optimisation le paramétrage qui rende compte au mieux des bases de données de structures secondaires connues à ce jour. Après avoir illustré cette approche par deux exemples tirés de la littérature, CONTRAfold et CG, je présenterai une nouvelle démarche qui s'inscrit dans cette même catégorie.

2.3 Paramétrage du modèle à partir de données expérimentales

"*Comment déterminer expérimentalement les paramètres du modèle ?*"

La principale expérience, abondamment utilisée depuis plus de trois décennies (de [23] en 1974 à [24] en 2007) vise à établir les variations d'enthalpie et d'entropie ΔH et ΔS liés à un certain repliement grâce à la mesure par absorption ultra-violette (en anglais : "*optical melting*") de sa dénaturation sous l'effet de la température.

2.3.1 Principe de l'expérience

Plaçons-nous dans le cas le plus fréquent dans la littérature, utilisé pour étudier hélices, renflements et boucles internes : l'ARN à deux brins. On étudie la séparation, sous l'effet de la température, de deux brins A et B complémentaires (hélices) ou imparfaitement complémentaires (renflements et boucles internes). On suppose que :

- cette réaction a deux états : le premier, noté A+B, est celui où il n'y a pas d'appariement entre A et B et le second, noté AB, est celui où les deux brins sont complètement appariés.
- les réactifs A et B sont préparés dans les conditions stoechiométriques, c'est à dire que leurs concentrations initiales C_A^0 et C_B^0 sont égales. Notons $C_A^0 = C_B^0 = C_t$.

On s'intéresse à la constante d'équilibre de la réaction $A + B \rightarrow AB$

$$K(T) = e^{-\frac{\Delta G(T)}{RT}} = \frac{[AB]}{[A][B]} \tag{2.1}$$

En notant α l'avancement de la réaction, on a :

$$\begin{aligned} [A] = [B] &= (1-\alpha)C_t \\ [AB] &= \alpha C_t \\ K(T) &= \exp\left(-\frac{\Delta H}{RT} + \frac{\Delta S}{R}\right) = \frac{\alpha}{(1-\alpha)^2 C_t} \end{aligned} \tag{2.2}$$

Cette relation est utilisée de deux manières différentes, dites "analyses de Van't Hoff", afin de relier les quantités ΔH et ΔS aux résultats de la mesure par absorption à 280nm (notée $\epsilon(T)$) de l'hypochromicité de l'ARN.

Première analyse de Van'tHoff du profil $\epsilon(T)$: interpolation directe

Voici quelques profils typiques obtenus pour $\epsilon(T)$:

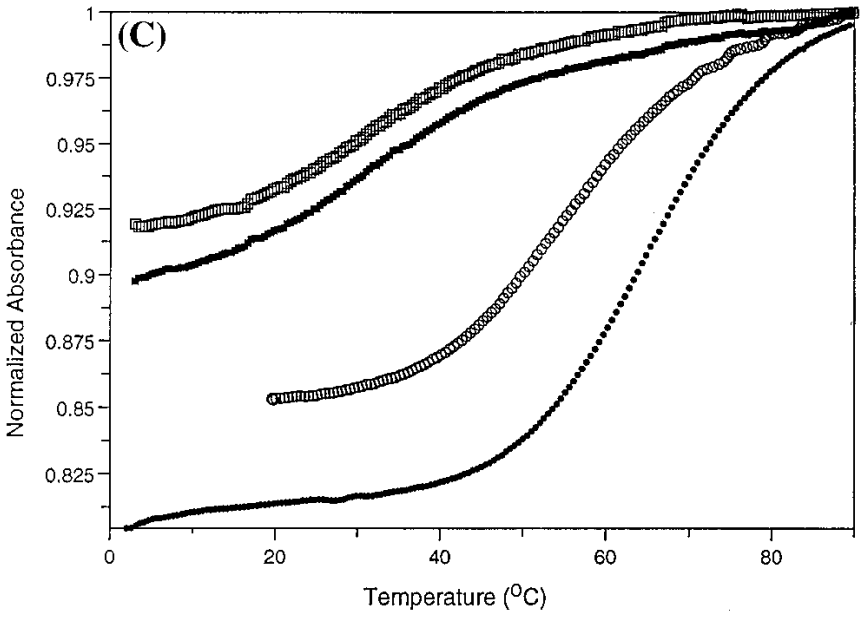


Figure extraite de [25] : absorption en fonction de la température pour quatre têtes d'épingle de séquences différentes

Ces profils sont classiquement interprétés en distinguant 3 régimes :

1. à basse température, le système est dans l'état AB et est caractérisé par une absorption $\epsilon_{AB}(T)$ approximativement linéaire :
$$\epsilon_{AB}(T) = a_{AB}T + b_{AB} \tag{2.3}$$
2. à haute température, le système est dans l'état A+B, caractérisé par une absorption $\epsilon_{A+B}(T)$ également linéaire :
$$\epsilon_{A+B}(T) = a_{A+B}T + b_{A+B} \tag{2.4}$$
3. un régime transitoire dans lequel les deux espèces sont présentes. En notant $\alpha(T)$ l'avancement de la réaction, l'absorption pendant ce régime transitoire est :
$$\epsilon(T) = \alpha(T)\epsilon_{A+B} + (1 - \alpha(T))\epsilon_{AB} \tag{2.5}$$

On a ainsi :

$$\alpha(T) = \frac{\epsilon(T) - \epsilon_{AB}(T)}{\epsilon_{A+B}(T) - \epsilon_{AB}(T)} = \frac{\epsilon(T) - (a_{AB}T + b_{AB})}{(a_{A+B} - a_{AB})T + (b_{A+B} - b_{AB})} \tag{2.6}$$

En combinant les relations (2.2) et (2.6), on obtient une équation qui relie l'observation $\epsilon(T)$ à six paramètres (ΔH, ΔS, a_{AB}, b_{AB}, a_{A+B}, a_{AB}). Des algorithmes d'interpolation non-linéaire sont alors utilisés pour ajuster ces paramètres, ce qui fournit en particulier une première estimation de ΔH et ΔS.

Seconde analyse de Van't Hoff du profil $\epsilon(T)$: dépendance à la concentration

On définit la "température de dénaturation" T_d comme la température à laquelle la quantité de matière totale est également répartie entre les états initiaux et finaux, c'est à dire à laquelle la moitié des brins A et B sont appariés. Ainsi définie, cette température se repère facilement sur les profils K(T). A T_d, on a :

$$[A] = [B] = [AB] = \frac{C_t}{2} \tag{2.7}$$

$$\text{d'où :} \qquad e^{-\frac{\Delta G}{RT_d}} = \frac{2}{C_t} \tag{2.8}$$

$$\text{qui se réécrit :} \qquad \frac{1}{T_d} = \frac{R}{\Delta H} \ln \frac{C_t}{2} + \frac{\Delta S}{\Delta H} \tag{2.9}$$

Lorsque le brin A est auto-complémentaire (c'est-à-dire que B≡A), la réaction s'écrit :

$$2A \rightarrow A_2 \tag{2.10}$$

C_t correspond maintenant à la concentration initiale de brins A. A T_d, on a $[A] = \frac{C_t}{2}$ et $[A_2] = \frac{C_t}{4}$ et on obtient :

$$\frac{1}{T_d} = \frac{R}{\Delta H} \ln C_t + \frac{\Delta S}{\Delta H} \tag{2.11}$$

Cette relation diffère de (2.9) par le coefficient de C_t.

T_d est mesurée pour plusieurs concentrations initiales C_t, ce qui permet de tracer les courbes $\left(\frac{1}{T_d}, \ln C_t\right)$ dont voici un exemple typique :

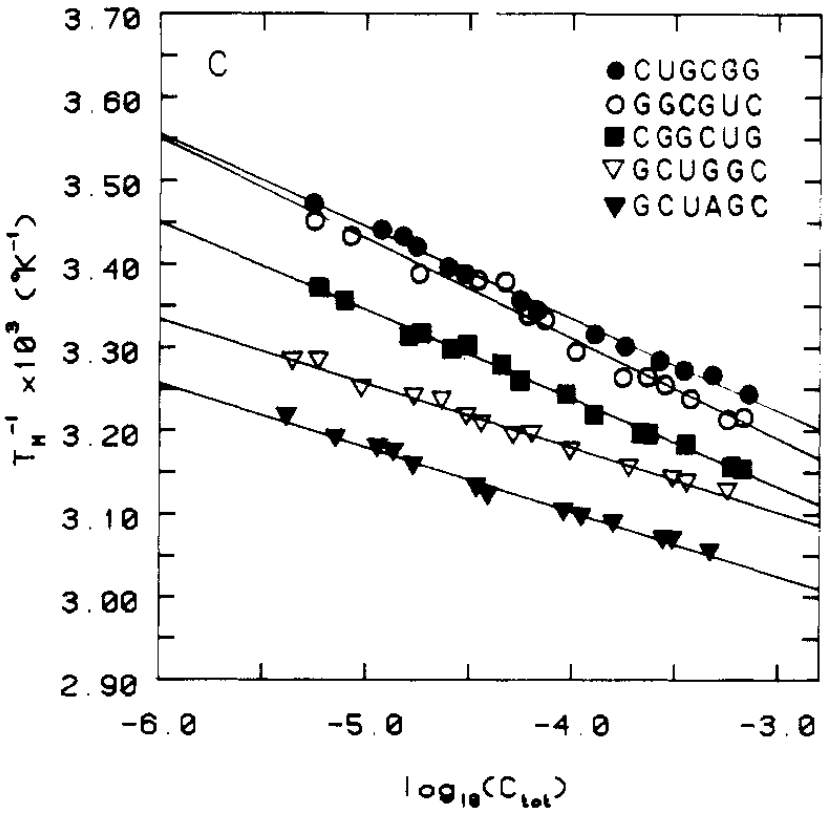


Figure extraite de [26] : courbes $\left(\frac{1}{T_d}, \ln C_t\right)$ *pour cinq séquences auto-complémentaires*

Ces courbes, en illustrant la relation linéaire entre $\frac{1}{T_d}$ et $\ln C_t$,

1. établissent que ΔH et ΔS ont des valeurs constantes sur les intervalles de température considérés.
2. permettent de calculer ces valeurs ΔH et ΔS avec une grande précision

Les valeurs de ΔH et ΔS calculées par ces deux méthodes sont jugées fiables si elles s'accordent à moins de 15%.

Cas particulier de la mesure d'énergie libre de formation de tête d'épingle

Les expériences étudiant la formation de têtes d'épingles mettent en jeu des séquences dont les extrémités sont complémentaires : la tête d'épingle est formée lorsque ces deux extrémités s'apparient. Cette réaction est du type :

$$A \rightarrow A' \tag{2.12}$$

En reprenant l'étude précédente il apparaît que dans ce cas la température de dénaturation T_d ne dépend plus de la concentration totale C_t. Ainsi, seule l'interpolation du profil $\epsilon(T)$ est utilisée pour estimer ΔH et ΔS.

Le modèle Turner99

Les données expérimentales s'accumulant petit à petit au cours des années, le modèle a été paramétré progressivement. Les valeurs les plus utilisées aujourd'hui ont été établies dans les articles "*Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs*" [27], "*Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*" [28] et elles ont été révisées dans "*Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*"[29]. J'appelle par la suite ce paramétrage "Turner99".

Dans le premier article, Xia *et al.* compilent les données de 90 mesures réalisées sur différentes séquences complémentaires et en déduisent un jeu de valeurs pour les 10 termes d'énergie libre de dipaires ΔF_d n'impliquant que les liaisons Watson-Crick. Dans le second, Mathews *et al.* complètent le modèle : ils estiment tous les autres paramètres (termes impliquant liaisons Wobble, têtes d'épingle, renflements...) en compilant les mesures idoines et en les interprétant à partir des valeurs du premier article. Ainsi donc, le modèle n'est pas le résultat d'une optimisation globale effectuée à partir de l'ensemble des mesures mais il s'est construit en deux temps, le premier conditionnant le deuxième.

Examinons attentivement ce qui a été fait dans le premier article.

2.3.2 Interprétation des expériences : estimation des énergies de dipaires Watson-Crick et discussion de l'hypothèse implicite de différence de capacité calorifique nulle

Dans Turner99, le modèle interpolé est précisément le suivant : l'énergie libre de formation d'une hélice est la somme des énergies libres de dipaires qui la constituent, d'un terme d'initiation d'hélice ΔG_{init} et d'une pénalité $\Delta G_{AU\ term}$ s'appliquant si les paires extrémales de l'hélice sont des paires A–U.

Comme mentionné plus haut, 90 mesures d'énergie libre de formation d'hélices (appariements de deux brins d'ARN complémentaires), de différentes séquences et de différentes longueurs, sont compilées. Chacune de ces expériences j donne lieu à une relation du type $\boldsymbol{n}^T(S_j)\boldsymbol{p} = f_j$, ce qui permet d'écrire l'ensemble de ces données sous forme matricielle $\boldsymbol{N}\boldsymbol{p} = \boldsymbol{f}$, où $\boldsymbol{N}$ est une matrice à 90 lignes. Le modèle d'énergie est alors paramétré en minimisant la quantité :

$$\chi(\boldsymbol{p}) = \|\boldsymbol{N}\boldsymbol{p} - \boldsymbol{f}\|^2 \tag{2.13}$$

Cette procédure repose sur plusieurs hypothèses : la plus importante est, à mon avis, celle portant sur la *différence de capacité calorifique* qui a été faite implicitement dans le précédent exposé des méthodes de Van't Hoff. En effet, on a supposé dans toutes les relations écrites dans cette partie que les différences ΔH et ΔS sont indépendantes de la température, c'est à dire que la *différence de capacité calorifique* Δc_p *entre les états A+B et AB est nulle*. Renoncer à cette hypothèse amène par rapport à un certain standard les corrections :

$$\Delta H(T) = \Delta H^0 + \Delta c_p(T - T^0) \tag{2.14}$$

$$\Delta S(T) = \Delta S^0 + \Delta c_p \ln \frac{T}{T^0} \tag{2.15}$$

$$\Delta G(T) = \Delta H^0 + \Delta c_p(T - T^0) - T\Delta S^0 - \Delta c_p T \ln \frac{T}{T^0} \tag{2.16}$$

où ΔH^0 et ΔS^0 désignent respectivement les différences d'enthalpie et d'entropie associées à la réaction pour une certaine température T^0 dite "standard" (en général la température ambiante).

Cette correction modifie les équations sur lesquelles reposent les analyses de Van't Hoff car K(T) devient :

$$K(T) \quad = \quad \exp\left(-\frac{\Delta H^0 - \Delta c_p T^0}{RT} + \frac{\Delta S^0 - \Delta c_p}{R} + \frac{\Delta c_p}{R}\ln\frac{T}{T^0}\right) \tag{2.17}$$

Pour pouvoir comparer entre elles les 90 mesures, il convient de les ramener à des conditions de concentration et de température identiques. Autrement dit, ce sont les quantités standard ΔH^0 et ΔS^0 qui nous intéressent. Est-il satisfaisant de simplement supposer que $\Delta c_p = 0$ et d'identifier ΔH et ΔS à ΔH^0 et ΔS^0 ? Un argument justifiant cette hypothèse est le fait qu'on constate expérimentalement une relation linéaire entre T_d et C_t. Cette linéarité est incompatible avec la présence du terme logarithmique apparaissant dans l'expression de K(T) si $\Delta c_p \neq 0$ (2.17). Cependant, cet argument n'est pas décisif. En effet, l'examen approfondi des données expérimentales montre qu'en pratique elles recouvrent un intervalle de température assez réduit, généralement inférieur à 15K, bien que la concentration varie sur plusieurs ordres de grandeur. On peut constater que c'est par exemple bien le cas dans l'exemple des courbes $\left(\frac{1}{T_d}, \ln C_t\right)$ rapportées plus haut. Cette linéarité apparente généralement observée ne pourrait être ainsi qu'une bonne approximation, sur cet intervalle de température, de la relation réelle entre $\frac{1}{T_d}$ et $\ln C_t$ mais qui ne nous informerait qu'imparfaitement sur ce qui se passe à des échelles plus grandes.

L'existence d'une différence de capacité calorifique non nulle a été démontrée de la manière la plus élégante dans l'article "*Heat capacity changes in RNA folding : application of perturbation theory to hammerhead ribozyme cold denaturation*" [30]. En effet, si $\Delta c_p \neq 0$, alors il apparaît avec (2.16) que l'équation $\Delta G(T) = 0$ a deux solutions, alors qu'elle n'en a qu'une seule si $\Delta c_p = 0$. Dans cet article, les auteurs montrent que l'ARN qu'ils étudient subit effectivement deux dénaturations à des températures séparées de 50K. Il est ainsi clair que Δc_p n'est pas nul et qu'il ne saurait être négligé dès lors qu'on travaille sur des intervalles de température de plusieurs dizaines de degrés.

Voici un extrait des mesures compilées dans Xia *et al.* (extrapolées pour $C_t = 1\text{M}$) :

	$-\Delta H(T_d)$ (kcal/mol)	$-\Delta S(T_d)$ (eu)	T_d (°C)
ACGCA/	45.40	130.4	29.4
AGCGA/	46.31	133.0	30.2
CACAG/	40.20	115.4	24.5
GCACG/	45.31	126.2	37.5
GCUCG/	43.38	120.1	37.2
GCGGCG/	58.50	155.0	61.8
GCGUCG/	52.38	140.6	53.7
GCUACG/	58.02	162.7	45.0
GCUAGC	59.13	165.1	49.3
GGAUCC	53.70	149.1	47.6
UUGGCCAA	63.68	169.8	65.3
UUGUACAA	49.45	137.8	43.6
CAAAAAAAG/	59.78	175.1	33.8

J'ai changé l'intitulé de la deuxième et troisième colonne, originellement $-\Delta H^\circ$ et $-\Delta S^\circ$, en $-\Delta H(T_d)$ et $-\Delta S(T_d)$. On voit que dans ces exemples les températures T_d s'étalent sur plus de 30K, ce qui suggère que l'inclusion d'une capacité calorifique non nulle peut significativement influer sur le paramétrage final du modèle.

La mesure précise de Δc_p pour des appariements entre brins d'ARN n'a pas bénéficié du même effort que pour les appariements ADN/ADN et ARN/ADN. Néanmoins, ces derniers nous donnent certainement une idée de ce qu'il en est pour les premiers. Dans l'article "*Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation*" [31], les auteurs effectuent les mesures des ΔH, ΔS et Δc_p de 41 hélices ARN/ADN longues de 5 à 9 bases. Voici les valeurs moyennes des capacités calorifiques obtenues, en fonction de la longueur de l'hélice :

$$\begin{aligned}
\Delta c_p(5) &= -0.38 \text{ kcal/mol/K} \\
\Delta c_p(6) &= -0.65 \text{ kcal/mol/K} \\
\Delta c_p(7) &= -0.76 \text{ kcal/mol/K} \\
\Delta c_p(8) &= -0.97 \text{ kcal/mol/K} \\
\Delta c_p(9) &= -1 \text{ kcal/mol/K}
\end{aligned} \tag{2.18}$$

Les auteurs ont utilisé leur mesures pour paramétrer un modèle d'énergie de dipaires ARN/ADN, avec ou sans l'hypothèse $\Delta c_p = 0$. Voici les résultats :

dipaire ARN/ADN	$-\Delta G(310K)$ $\Delta c_p = 0$	$-\Delta G(310K)$ $\Delta c_p \neq 0$
rAA/dTT	−1.0	−0.4
rAC/dTG	−2.1	−1.6
rAG/dTC	−1.8	−1.4
rAU/dTA	−0.9	−1.0
rCA/dGT	−0.9	−1.0
rCC/dGG	−2.1	−1.5
rCG/dGC	−1.7	−1.2
rCU/dGA	−0.9	−0.9
rGA/dCT	−1.3	−1.4
rGC/dCG	−2.7	−2.4
rGG/dCC	−2.9	−2.2
rGU/dCA	−1.1	−1.5
rUA/dAT	−0.6	−0.3
rUC/dAG	−1.5	−0.8
rUG/dAC	−1.6	−1.0
rUU/dAA	−0.2	−0.2

On constate des différences très significatives pour certaines valeurs de dipaires, comme rAA/dTT, rUA/dAT et rUC/dAG dont la valeur varie du simple au double. Le fait que d'autres dipaires ne soient par contre pas modifiés suggère que Δc_p dépend de la séquence. On peut ainsi s'attendre à des effets similaires pour les dipaires ARN/ARN et l'hypothèse $\Delta c_p = 0$ apparaît comme une approximation exagérée.

(2.18) montre que Δc_p dépend en moyenne linéairement de la longueur de l'hélice formée. Le Δc_p mis en jeu lors de la formation d'ARN double brin a été estimé par plusieurs travaux ([32], [33], [34]) à environ -0.1kcal/mol/K par paire de bases pour une concentration en NaCl de 1M.

Afin d'avoir une idée quantitative de l'impact d'un Δc_p non nul pour le modèle d'énergie ARN/ARN qui nous intéresse, j'ai réécrit le problème d'optimisation (2.13) comme la minimisation de la quantité :

$$\chi(\boldsymbol{p}, c_p) = \|\boldsymbol{N}\boldsymbol{p} - \boldsymbol{f^*}(\boldsymbol{c_p})\|^2 \tag{2.19}$$

où $\boldsymbol{f_j^*}(\boldsymbol{c_p})$ est maintenant la valeur de la mesure obtenue pour l'expérience j à 310K corrigée selon (2.16). J'ai supposé en bon accord avec les résultats (2.18) que Δc_p dépend linéairement de la longueur l de l'hélice à laquelle elle s'applique :

$$\Delta c_p(l) = a \times l + b \tag{2.20}$$

Ainsi, l'optimisation porte aussi sur ces deux coefficients a et b.

Les énergies libres obtenues sont, en kcal/mol :

ΔF_d	Turner99	Turner99 $\Delta c_p \neq 0$	Différence (%)
A–U A–U	-0.93	-0.83	11
A–U U–A	-1.10	-1.09	1
U–A A–U	-1.33	-1.14	14
A–U G–C	-2.08	-2.11	1
C–G A–U	-2.11	-2.06	2
A–U C–G	-2.24	-1.94	13
G–C A–U	-2.19	-1.95	11
C–G G–C	-2.36	-2.45	4
C–G C–G	-3.26	-2.95	11
G–C C–G	-3.42	-2.76	24
moyenne	-2.11	-1.93	10
écart-type	0.78	0.68	15
ΔG_{init}	+4.09	+3.23	21
$\Delta G_{AU\ terminal}$	+0.45	+0.43	4

Concernant Δc_p (2.20), cette optimisation a donné pour résultats les valeurs $a = -0.11\ kcal/mol/K$ et $b = 0.19\ kcal/mol/K$, valeurs tout à fait compatibles avec l'expérience [32] [34]. Tenir compte de cette capacité calorifique a pour effet d'uniformiser les énergies, ce qui a été également rapporté dans [33]. Au final, six paramètres sur 10 subissent une variation significative de plus de 10% et il est raisonnable de penser que cet effet serait accentué en prenant mieux en compte la dépendance de Δc_p à la séquence, comme il a été fait pour les interactions ARN/ADN. Il est donc vraisemblable qu'une mesure systématique de la capacité calorifique aurait entraîné des valeurs plus correctes

et significativement différentes de celles proposées dans Xia *et al.*, qui sont les valeurs sur lesquelles repose l'interprétation des mesures de tous les autres motifs.

2.3.3 Interprétation des expériences : interpolation complète du modèle

En prenant pour acquis les valeurs de dipaires Watson-Crick proposées dans Xia *et al.*, Mathews *et al.* [28] et [29] compilent et interprètent les mesures adéquates pour paramétrer le reste du modèle. L'hypothèse $\Delta c_p = 0$ est toujours faite. Cette interpolation a dû s'affranchir de nombreuses difficultés liées à l'insuffisance des données expérimentales pour certains paramètres et à l'incohérence apparente de certaines mesures qui a requis l'introduction de nouveaux bonus ou pénalités exceptionnels. Voici quelques exemples des difficultés rencontrées :

- Hélices
  1. Il n'y a qu'une seule mesure permettant d'estimer le terme $\Delta F_d \begin{pmatrix} \text{G–U} \\ \text{G–U} \end{pmatrix}$: les auteurs ont considéré *a posteriori* que la valeur qui s'en déduisait nuisait à la qualité de prédiction de structures secondaires et l'ont finalement ajustée par un processus d'essais et erreurs.
  2. Les énergies libres des hélices contenant le motif $\begin{matrix} \text{5'GGUC3'} \\ \text{3'CUGG 5} \end{matrix}$ se sont révélées difficiles à interpoler et les auteurs ont assigné une énergie spéciale à cette sous-séquence de 8 bases de long.

- Têtes d'épingles : le modèle de base pour les têtes d'épingle consiste en la somme d'un terme dépendant de la nature de la paire qui ferme la tête d'épingle et d'un terme entropique dépendant de la longueur de la tête d'épingle. D'autres termes ont dû être rajoutés pour rendre compte des mesures :

  1. une pénalité spéciale s'applique aux boucles constitutées uniquement de cytosines
  2. un bonus s'applique si la dernière paire de l'hélice formant la tête d'épingle est une paire Wobble 5' G–U 3' dont la guanine est précédée dans la séquence de deux autres guanines
  3. une autre pénalité s'applique si les bases libres jouxtant la dernière paire sont 5' U U 3' ou 5' G A 3' mais non 5' A G 3'

4. des tables spéciales sont introduites pour rectifier le calcul de l'énergie de têtes d'épingle de séquence particulière

- Boucles internes : les auteurs ont introduit des tables de valeurs pour les courtes boucles internes (1×1, 1×2, 2×2) tenant compte de leur séquence et de celle des paires extrêmales des hélices qui les forment. Ces tables représentent environ 2500 nouveaux paramètres – en tenant compte des symétries – pour lesquels bien moins de mesures sont disponibles. Des règles sont donc introduites pour estimer ces paramètres à l'aide d'un sous-ensemble d'une centaine.

Au final, quels résultats donne ce paramétrage ? La réponse varie selon les études : une sensibilité de 73% est annoncée par Mathews *et al.*, alors qu'elle est de 41% dans "*Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction*" [35]. La différence tient aux bases de données utilisées et à la façon de calculer la sensibilité. Dans ces deux articles, les bases de données utilisées contiennent beaucoup de pseudo-nœuds, motifs non paramétrés par le modèle ni pris en charge par les algorithmes de repliements utilisés. Cela pose la question du sens des mesures de sensibilité faites sur ces bases de données. Si ce modèle prédit la structure secondaire sans pseudo-nœud s'approchant le plus de la structure réelle avec pseudo-nœuds, rien ne garantit que ce modèle permettrait effectivement de trouver la structure réelle si on l'utilisait avec un algorithme permettant les pseudo-nœuds. En effet, cet algorithme pourrait alors trouver un nouveau minimum global de l'énergie libre qui corresponde à un repliement très différent. Les paires de bases constituant les pseudo-nœuds ont été comptabilisées dans le calcul de la sensibilité dans [35] et il n'est pas précisé si cela a été aussi le cas par Mathews *et al.*. Cela pourrait expliquer en partie la grande différence entre les sensibilités rapportées par ces deux articles.

Pour avoir une meilleure idée des performances de ce modèle, je cite ces mesures de prédiction réalisées dans [36] sur deux bases de données sans pseudo-nœud : 61% de sensibilité sur 375 séquences d'ARNr 5s et 69% sur 68 séquences d'ARN SRP.

Pour améliorer ces résultats, d'autres paramétrages ont été développés par une approche différente qui consiste à se donner un certain nombre de structures secondaires et de construire le meilleur modèle qui les explique. Je présente et discute ci-dessous deux de ces paramétrages, CONTRAfold et CG.

2.4 Paramétrage du modèle par l'analyse des bases de données

2.4.1 CONTRAfold

Les résultats et le détail de cette méthode sont publiés dans l'article *CONTRAfold : RNA secondary structure without physics-based models* [37].

Principe de l'optimisation

Dans cet article, les auteurs développent un équivalent du modèle d'énergie dans un autre cadre conceptuel : celui des grammaires algébriques [38] . Une structure secondaire d'ARN est construite récursivement selon les règles de transformations suivantes :

$$\begin{aligned} S &\rightarrow LS \mid \emptyset \\ L &\rightarrow a \mid c \mid g \mid u \mid aHu \mid uHa \mid gHc \mid cHg \mid gHu \mid uHg \\ H &\rightarrow aHu \mid uHa \mid gHc \mid cHg \mid gHu \mid uHg \mid LS \end{aligned} \tag{2.21}$$

S est le motif initial de la grammaire, H permet de créer des hélices, L permet de créer des base libres ou d'initier une hélice tandis que a, c, g, u représentent les nucléotides et sont les éléments terminaux de la grammaire. La règle $H \rightarrow aHu$ signifie qu'une paire A-U est ajoutée autour de la structure H.

Ainsi, la structure secondaire : 

est engendrée par la suite de transformations suivante :

$$\begin{aligned} S &\rightarrow LS \rightarrow LLS \rightarrow LLLS \rightarrow LLL \rightarrow aLa \rightarrow agHca \\ &\rightarrow aggHcca \rightarrow aggLScca \rightarrow aggLcca \rightarrow aggacca \end{aligned}$$

A chacune de ces transformations i est associée une probabilité p_i.

Supposons de plus que la grammaire ne soit pas ambiguë, c'est-à-dire qu'à chaque structure secondaire donnée correspond un unique ensemble de transformations à utiliser. Ainsi chaque structure secondaire S est caractérisée par un vecteur $\boldsymbol{n}(S) = \{n_1, ..., n_N\}$, où n_i est le nombre d'occurrences de la transformation i nécessaire à la création de S. On définit alors la probabilité de S comme :

$$p(S) = \prod_i p_i^{n_i} = \exp(\sum_i n_i \ln p_i) = \exp(\boldsymbol{p}^T \boldsymbol{n}(S)) \quad \text{avec } \boldsymbol{p}_i = \ln p_i$$

A séquence x et paramètres $\boldsymbol{p}$ donnés, le problème du repliement consiste à trouver la structure la plus probable, c'est-à-dire la structure dont la probabilité est la plus élevée *conditionnellement à la séquence x*. Ceci revient à rechercher la structure S dont le vecteur $\boldsymbol{n}$(S) maximise la quantité

$$P_{\boldsymbol{p}}(S|x) = \frac{\exp(\boldsymbol{p}^T \boldsymbol{n}(S))}{\sum_{\boldsymbol{m}} \exp(\boldsymbol{p}^T \boldsymbol{m})} \tag{2.22}$$

On peut maintenant définir le problème de l'estimation des paramètres $\boldsymbol{p}$:
soient $x_1, ..., x_K$ K séquences, chacune associée à sa structure secondaire expérimentale respective $S_{x_1}, ..., S_{x_K}$. Le modèle de probabilité optimal $\boldsymbol{p_{opt}}$ est défini comme celui qui maximise la quantité :

$$\prod_{i=1}^{K} P_{\boldsymbol{p}}(S_{x_i}|x_i) \tag{2.23}$$

et il existe des algorithmes pour effectuer cette optimisation.

Les auteurs ont procédé à une modification supplémentaire de manière à ce que la recherche de S_x tienne compte des objectifs de sensibilité et de VPP sur lesquels l'algorithme sera évalué : un poids γ est introduit dans le calcul de $P_{\boldsymbol{p}}(S|x)$ afin de favoriser l'un ou l'autre de ces aspects.

Ce formalisme se distingue *a priori* de celui de la minimisation d'énergie libre par deux aspects :

1. Les composantes de $\boldsymbol{p}$ sont négatives car ce sont des logarithmes de probabilité. Cependant, les auteurs ont levé cette restriction dans leurs algorithmes d'optimisation car celle-ci n'est pas nécessaire pour que la quantité (2.22) soit définie. Ainsi, cette quantité s'identifie au terme $e^{-\beta F(S)}/\mathcal{Z}$ défini en partie 2.1 et les résultats de l'optimisation peuvent être considérés comme un modèle d'énergie et non plus un modèle de probabilité.
2. Comme il a été démontré dans [39], l'emploi d'une grammaire ambiguë a des répercussions significatives et défavorables sur la qualité de prédiction. Les auteurs ne mentionnent pas si c'est le cas dans leur étude. (Les auteurs ont en effet approfondi la simple grammaire donnée en (2.21) afin de prendre en compte les paramètres *ad hoc* se rapportant aux différents motifs énoncés en partie 2, à quelques modifications près. Par exemple, un terme prenant en compte la longueur des hélices a été rajouté, aucune forme n'a été supposée pour la fonction d'asymétrie des boucles internes et les tables pour boucles internes $1 \times 1, 1 \times 2$ et 2×2 ont été abandonnées.)

Bases de données

L'optimisation du modèle par cette méthode s'est faite à partir de 151 structures obtenues par comparaison de séquences dans la base de données Rfam [40].

Rfam distingue plusieurs familles d'ARN au sein desquelles plusieurs alignements de séquences sont proposés. Les auteurs se sont restreints aux 151 familles dont la structure consensus a été expérimentalement vérifiée et, pour chaque famille, ont sélectionné la séquence qui, par rapport à sa structure consensus, contenait le moins de nucléotides manquants et le moins de paires qui ne soient pas Watson-Crick ou Wobble.

Résultats

Au final, ce nouveau modèle CONTRAfold donne des prédictions significativement meilleures que Turner99 sur cette base de données de 151 séquences, la qualité de prédiction étant estimée suivant les critères habituels de sensibilité et de VPP (74% et 67% pour CONTRAfold contre 69% et 60% pour Turner99). Ce paramétrage a été testé sur d'autres séquences dans [36] et on observe qu'en moyenne il fait de moins bonnes performances que Turner99. Une nouvelle version de ces paramètres, optimisée sur de plus grandes bases de données, a été rendue disponible récemment mais aucun test de celle-ci n'a encore été publié.

2.4.2 CG

Les travaux relatés dans cette section ont été publiés dans "*Efficient parameter estimation for RNA secondary structure prediction* [36]. La méthode CG effectue un paramétrage du modèle à l'aide d'optimisations conjointes sur les bases de données de structures et sur les mesures expérimentales d'énergie libre.

Principe de l'optimisation

Selon un modèle d'énergie idéal, la structure secondaire native S_x^r d'une séquence x a une énergie libre inférieure à toute autre structure secondaire S_x :

$$\begin{array}{rrcll} & \Delta F(S_x^r, x) & < & \Delta F(S_x, x) & \forall S_x \neq S_x^r \\ \text{soit :} & \left[\boldsymbol{n}(S_x^r) - \boldsymbol{n}(S_x)\right]^T \boldsymbol{p} & < & 0 & \forall S_x \neq S_x^r \end{array} \tag{2.24}$$

En considérant plusieurs séquences x_i et en créant à l'aide d'un programme comme MFold plusieurs structures candidates S_{x_i}, la relation (2.24) permet d'écrire sous forme matricielle un jeu de contraintes que doit satisfaire le modèle d'énergie idéal :

$$(N^r - N)\boldsymbol{p} < \mathbf{0} \tag{2.25}$$

où les lignes N_i et N_i^r des matrices N et N^r sont les vecteurs $\boldsymbol{n}^T(S_{x_i})$ et $\boldsymbol{n}^T(S_{x_i}^r)$. Anticipant qu'un modèle satisfaisant toutes ces contraintes n'existe pas, les auteurs les assouplissent en permettant une erreur $\boldsymbol{\delta}$ qui doit être aussi petite que possible. Ainsi, le problème d'optimisation consiste à minimiser $||\boldsymbol{\delta}||$ sous la contrainte $(N^r - N)\boldsymbol{p} - \boldsymbol{\delta} < \mathbf{0}$.

Les mesures expérimentales d'énergie libre sont également utilisées d'une manière similaire. En notant f_i l'énergie libre mesurée lors de l'expérience i et $\boldsymbol{n}_i$ le vecteur du nombre d'occurrences des paramètres du modèle dans cette mesure, le modèle d'énergie idéal doit satisfaire :

$$\boldsymbol{n}_i^T \boldsymbol{p} = f_i \quad \forall i \tag{2.26}$$

qui se réécrit également de manière matricielle :

$$N^{exp}\boldsymbol{p} - \boldsymbol{f} = 0 \tag{2.27}$$

où N^{exp} est la matrice telle que la ligne N_i^{exp} corresponde au vecteur $\boldsymbol{n}_i$.
Là encore, une erreur $\boldsymbol{\zeta}$ est autorisée et on cherche ainsi à minimiser $||\boldsymbol{\zeta}||$ sous la contrainte $N^{exp}\boldsymbol{p} - \boldsymbol{f} - \boldsymbol{\zeta} = 0$.

Le problème global d'optimisation traité dans l'article est finalement le suivant :

$$\begin{aligned} &\text{trouver } \boldsymbol{p} \text{ qui minimise} && c(\lambda) = (1-\lambda)||\boldsymbol{\delta}|| + \lambda||\boldsymbol{\zeta}|| && (2.28)\\ &\text{sous les contraintes :} && (N^r - N)\boldsymbol{p} - \boldsymbol{\delta} < \mathbf{0} && (2.29)\\ &\text{et} && N^{exp}\boldsymbol{p} - \boldsymbol{f} - \boldsymbol{\zeta} = 0 && (2.30) \end{aligned}$$

où λ est un paramètre ajustable par l'utilisateur.

Bases de données

La base de données utilisée pour le calcul des paramètres et leur test comporte 1660 séquences issues de différentes compilations de structures : ARNt, RNase P, ARNr 5s, ARNr 16s, ARNr 23s, ARN SRP et ribozymes.

Résultats

Les auteurs comparent la qualité des prédictions de leur modèle avec CONTRAFold et Turner99 et concluent à une significative amélioration de 7%. Ces résultats sont obtenus avec un choix de $\lambda = 0.995$. Pour cette valeur très proche de 1, comme on peut le voir avec (2.28), le problème se ramène à la construction d'un modèle optimisé globalement sur l'ensemble des données expérimentales. Le rôle des données structurales se cantonne ainsi à fournir des contraintes sur les paramètres apparaissant rarement dans les données expérimentales. Les auteurs justifient ce choix de λ en signalant que la moitié des paramètres du modèle ne sont représentés qu'au plus une fois dans les données structurales. Si le choix de λ donnait trop d'importance aux données structurales, cela entraînerait de grandes incertitudes dans la détermination de ces paramètres rares.

2.4.3 Comparaison des énergies de dipaires Watson-Crick obtenues avec chacune de ces méthodes et discussion

ΔF_d	Turner99	Turner99 $c_p \neq 0$	CONTRAFold v1	CONTRAFold v2	CG
A–U A–U	-0.93	-0.83	-1.66	-0.99	-0.5
A–U U–A	-1.10	-1.09	-1.65	-1.27	-0.73
U–A A–U	-1.33	-1.14	-1.60	-1.53	-0.69
A–U G–C	-2.08	-2.11	-2.18	-2.47	-1.42
C–G A–U	-2.11	-2.06	-2.09	-2.12	-1.26
A–U C–G	-2.24	-1.94	-2.16	-2.0	-1.32
G–C A–U	-2.19	-1.95	-2.29	-2.78	-1.48
C–G G–C	-2.36	-2.45	-2.35	-2.64	-1.37
C–G C–G	-3.26	-2.95	-2.69	-3.10	-2.25
G–C C–G	-3.42	-2.76	-2.39	-2.78	-2.18
moyenne	-2.11	-1.93	-2.11	-2.11	-1.32
écart-type	0.78	0.68	0.34	0.68	0.55

Ce tableau dresse un bilan des paramètres de dipaires Watson-Crick que calculent les méthodes précédentes. Les paramètres de CONTRAfold étant donnés à une constante près, j'ai ajusté celle-ci pour faire correspondre leurs valeurs moyennes à Turner99. En comparant deux à deux les distributions de ces paramètres et leurs hiérarchies, on s'aperçoit que chaque méthode se démarque significativement des autres. Les paramètres de dipaires ont une importance capitale dans tout modèle d'énergie car ce sont les principaux paramètres associés à une différence d'énergie libre négative, traduisant le fait qu'ils dirigent le repliement. A ce titre, les disparités constatées dans ce tableau sont étonnantes. On pourrait s'attendre à ce que les valeurs d'énergie libre de dipaires, du fait de leur petit nombre et de leur importance unique, soient suffisamment robustes

pour que toute méthode y converge. Les résultats ci-dessus infirment cette intuition et illustrent la grande sensibilité de chacune de ces méthodes à leurs hypothèses :

- aucune méthode ne fournit la même hiérarchie de valeurs qu'une autre
- l'inclusion d'une capacité calorifique dans *l'interprétation* des mesures expérimentales induit des différences de plus de l'ordre de 10% sur plusieurs valeurs. En particulier, l'énergie libre de $\begin{pmatrix} \text{G}-\text{C} \\ \text{C}-\text{G} \end{pmatrix}$ est diminuée de 25%.
- les distributions des valeurs obtenues par la méthode CONTRAfold diffèrent grandement selon *l'ensemble sur lequel l'algorithme a été entraîné.*
- les paramètres de Turner99 et CG ont des moyennes significativement différentes alors que toutes deux s'appuient sur les mesures expérimentales sans inclure de capacité calorifique. L'"énergie manquante" de CG est en fait compensée par un paramétrage différent des paires extrémales d'une hélice et ceci illustre la grande *interdépendance des paramètres du modèle.*

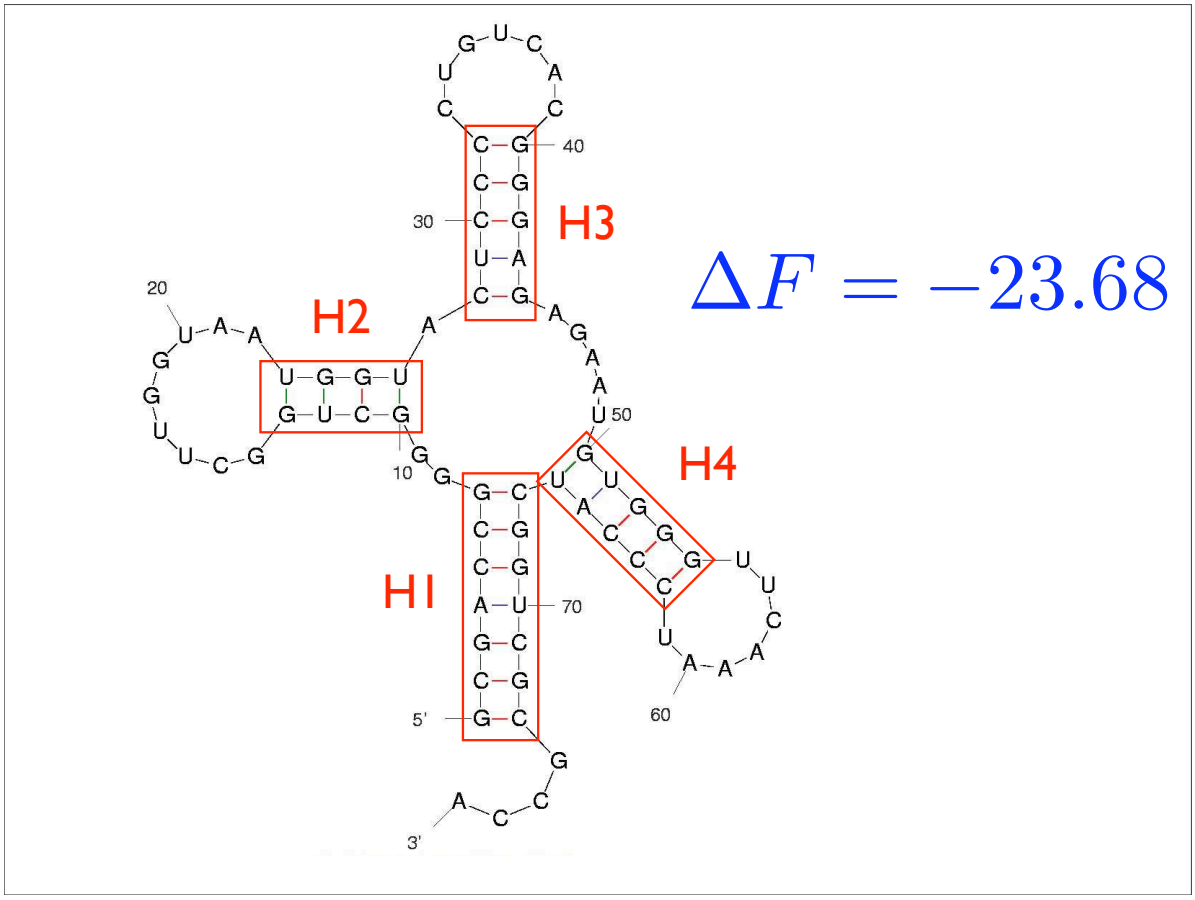
Il m'apparaît donc que les fondements du modèle d'énergie nécessaire à la prédiction du repliement de l'ARN ne sont pas encore bien établis dans la mesure où il n'y a pas de vision unifiée des paramètres centraux du modèle. Cette perception est personnelle et, à ma connaissance, il n'y a pas de discussion de cet aspect dans la littérature. Je cite en exemple la conclusion de [35] : après avoir calculé que la meilleure structure proposée par l'algorithme Mfold v3.1 utilisé en conjonction avec Turner99 ne partage en moyenne que 41% de paires de bases avec la structure réelle, les auteurs proposent deux pistes pour expliquer ce faible pourcentage :

- la défaillance du paramétrage des boucles mutli-hélices, notamment l'évaluation des pénalités d'initiation de boucle a et d'hélices b.
- la possibilité que les structures réelles d'ARN ne soient pas à l'équilibre thermodynamique mais piégées dans des configurations métastables du fait de la cinétique particulière de leur repliement cotranscriptionnel.

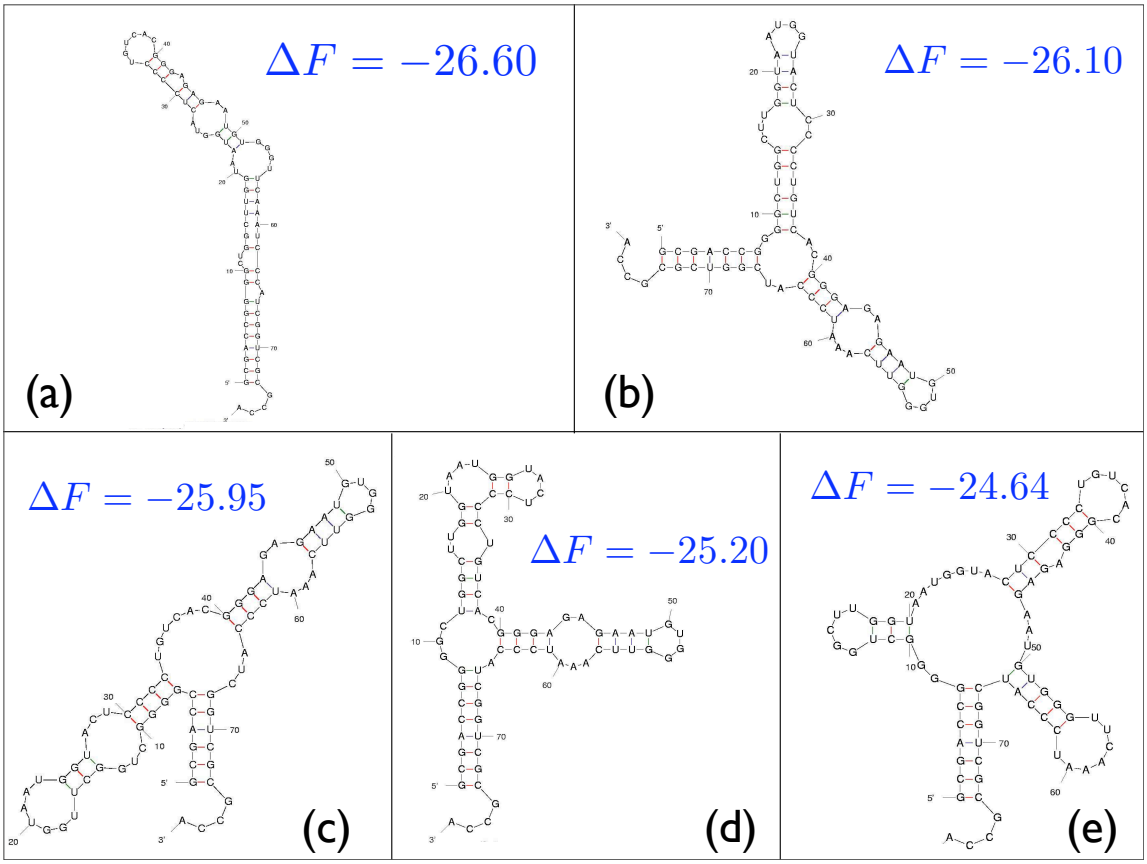
Un examen attentif des erreurs de MFold montre cependant que d'autres explications sont nécessaires.

Considérons l'exemple de la prédiction de l'ARNt-Asp du phage T5 (dont la séquence est répertoriée dans la *Sprinzl database* [41] avec la référence RD0260).

Voici la structure secondaire réelle de cet ARNt et les cinq meilleures propositions de Mfold :



Structure réelle de l'ARNt-Asp du phage T5. Elle s'organise autour de quatre hélices notées H1, H2, H3 et H4.



Les cinq meilleures structures proposées par Mfold

Pourquoi la structure secondaire réelle n'est-elle pas la structure d'énergie minimum proposée par Mfold ?

1. En comparant la structure secondaire réelle aux structures (b), (c), (d) et (e), l'hypothèse de la défaillance des paramètres de boucles multi-hélices est peu vraisemblable. Toutes ces structures contenant des boucles mutli-hélices sensiblement de même taille, on s'attend à ce que les corrections soient très similaires, ce qui ne permettra pas de perturber cette hiérarchie. Plus précisément, la structure (e) et la structure secondaire réelle ont toutes deux une boucle 4-hélices, donc leur différence d'énergie ne peut s'expliquer par le paramétrage des termes d'initiation de ce motif. De plus, la structure (b) comportant seulement une boucle 3-hélices, il faudrait que le terme d'initiation d'hélice diminue de 2.5 kcal/mol pour favoriser la structure secondaire réelle, ce qui est impossible car celui-ci est actuellement fixé à +0.4kcal/mol.

2. L'hypothèse que l'état natif fonctionnel de l'ARN puisse être en fait un état métastable doit être discutée. En effet, c'est une hypothèse très forte qui remet en cause le principe de la minimisation de l'énergie libre.

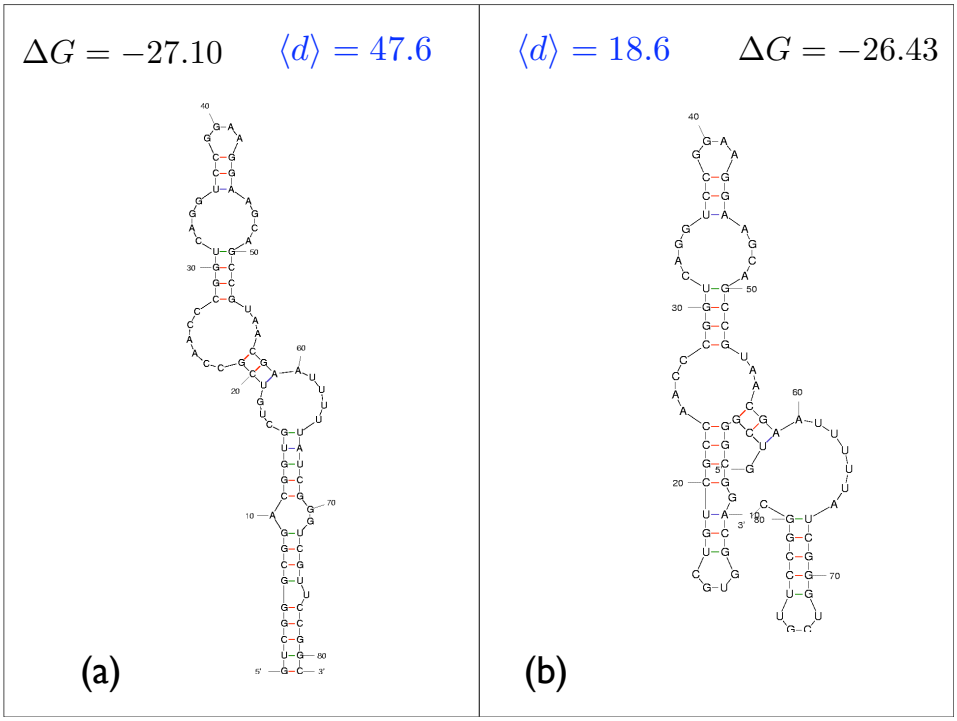
   Le repliement cotranscriptionnel serait le mécanisme biologique qui induit cette alternative. Le repliement cotranscriptionnel est un repliement progressif au fur et à mesure que la séquence est allongée de 5' vers 3' par la polymérase à ARN. Dans cette vision, les hélices formées seraient les premières disponibles au cours de l'allongement dans le temps de la séquence et elles concerneraient donc des sous-séquences complémentaires proches les unes des autres. Le biais que le repliement cotranscriptionnel pourrait introduire dans notre problème serait alors de diminuer la distance moyenne dans la séquence entre bases appariées. Dans le cas donné ici en exemple, on s'aperçoit que la meilleure structure (a) a la topologie d'une longue tige, topologie qui maximise cette distance moyenne, tandis que la structure secondaire réelle comporte trois hélices dont les deux brins se succèdent dans la séquence. On pourrait effectivement attribuer cette erreur de prédiction au repliement cotranscriptionnel.

   Tout d'abord, précisons que l'éventualité d'un repliement métastable peut être expérimentalement détectée en effectuant des cycles de recuit (*annealing*), consistant à dénaturer et renaturer l'ARN en jouant sur la température. Observer des structures différentes au cours de ces cycles démontre l'existence de structures métastables. Cette procédure a été suivie dans [42]. Pan *et al.* montrent qu'au sein d'une population de ribozymes de "*Tetrahymena* " les structures natives cohabitent effectivement avec d'autres structures non fonctionnelles. La question est de savoir si les structures natives correspondent au minimum d'énergie libre ou si elles sont en fait des structures métastables. On sait que l'exécution de cycles de recuits fait converger une population vers la structure d'énergie minimum. Les auteurs démontrent que l'exécution de cycles de recuits fait converger la population vers la structure native : il y a donc dans ce cas-là identité entre structure native et structure d'énergie minimum. Ainsi, l'hypothèse que l'état natif soit un état métastable peut être testée expérimentalement. A ma connaissance, il n'a pas encore été trouvé d'exemple où cette hypothèse est vérifiée.

   Dans le cas des ARNt, il y a déjà des arguments biologiques qui permettent de relativiser l'importance du repliement cotranscriptionnel. Certains nucléotides d'ARNt sont en effet modifiés chimiquement *après* leur transcription et des travaux

expérimentaux ont montré que ces modifications étaient *nécessaires* au bon repliement de ces ARNt [43]. Ces études démontrent, au moins sur les exemples étudiés, que l'ARNt n'est pas bloqué dans l'hypothétique structure métastable qu'il aurait à la fin de sa transcription mais qu'il adopte sa configuration fonctionnelle après avoir été intégralement transcrit. De plus, il existe beaucoup d'exemples dans les bases de données où la structure secondaire réelle a la topologie d'une longue tige (ARN SRP), alors qu'existent des structures concurrentes très proches en énergie pour laquelle la distance moyenne $\langle d \rangle$ entre bases appariées est bien plus faible. Ces exemples montrent que dans certains cas le biais induit par le repliement cotranscriptionnel ne joue pas.

Voici un exemple parmi de nombreux autres :

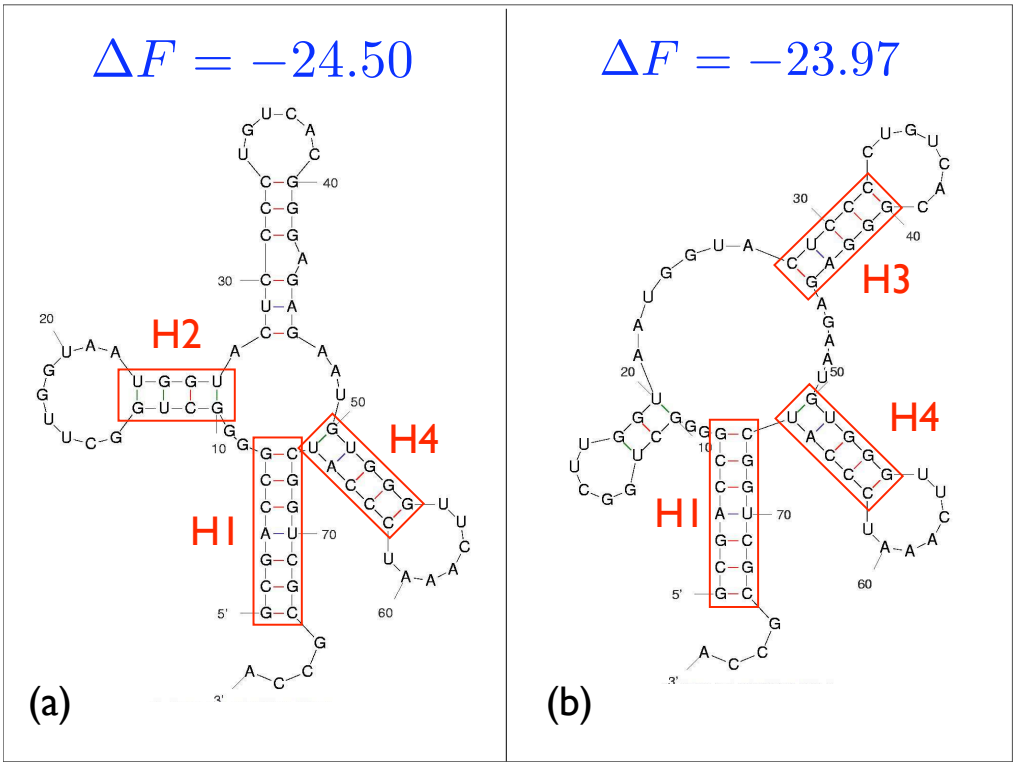


(a) Structure secondaire réelle d'un ARN SRP et meilleure structure prédite par Mfold (b) Seconde meilleure structure prédite par Mfold.
La prédiction de Mfold est correcte et tenir compte du repliement cotranscriptionnel pourrait favoriser de manière erronée la structure (b) au détriment de (a)

Les ARNt et les ARN SRP sont des séquences relativement courtes. Il a été discuté dans [44], article antérieur à [42], que les aspects cinétiques du repliement, cotranscriptionnel ou non, peuvent avoir un impact pour les séquences plus longues. Les auteurs y démontrent en effet que pour de longues séquences Mfold prédit mieux les parties d'une séquence appelés "domaines" que la totalité de la séquence qui n'est pourtant que la juxtaposition de ces domaines. De plus, la distance moyenne entre paires de bases de la structure prédite est systématiquement plus grande que celle de la structure réelle. Cependant, les auteurs se gardent de conclure définitivement car toute leur étude dépend du modèle d'énergie Turner99 (construit à partir de mesures expérimentales de courts motifs) et les pseudo-nœuds sont exclus de l'analyse.

Dans la mesure où il n'y a pas encore de support expérimental à l'hypothèse d'un repliement fonctionnel métastable de l'ARN et que les modèles d'énergie disponibles peuvent être pris en défaut sur de courtes séquences, je pense qu'il est encore trop tôt pour explorer cette voie.

3. Dans le cas de l'ARNt-Asp du phage T5, la source de l'erreur peut être facilement mise en lumière. En relançant Mfold en imposant comme contrainte l'existence des hélices 1, 3 et 4, il apparaît que Mfold n'arrive toujours pas à prédire l'hélice 2. De même, en imposant l'existence des hélices 1, 2 et 4, Mfold ne complète pas la structure par l'hélice 3. La plus simple explication des échecs de Mfold est donc que les paramètres utilisés pour les dipaires impliquant des paires Wobble (hélice 2) et les renflements (hélice 3) ne sont pas satisfaisants.



Structures prédites par Mfold (a) en imposant la présence des hélices H1, H2 et H4 (b) en imposant la présence des hélices H1, H3 et H4

La comparaison des modèles d'énergie existants et l'examen de nombreux autres exemples similaires à celui ci-dessus m'ont finalement convaincu de déterminer un nouveau paramétrage. Pour ce faire, j'ai tout d'abord imaginé employer des méthodes de calculs d'énergie libre par dynamique moléculaire. Celles-ci s'étant révélées infructueuses, une nouvelle optimisation du modèle à partir de données structurales est proposée.

2.5 Paramétrage du modèle par dynamique moléculaire

Du fait de sa taille, cette section fait l'objet d'un chapitre auquel le lecteur peut se référer immédiatement.

2.6 Paramétrage par une nouvelle méthode d'optimisation à partir de bases de données structurales : MC

Les études précédentes ont montré toute l'importance de la base de données utilisée. Ce choix doit donc faire l'objet d'une réflexion particulière.

2.6.1 Base de données : considérations préliminaires

Bases de données et modèle d'énergie

Une information importante tirée des études précédentes est que la moitié des paramètres du modèle est représentée au plus une fois dans les bases de données actuelles. Ceux-ci sont principalement des termes de pénalité entropique. Le modèle prévoit en effet d'affecter aux bases libres des pénalités d'entropie de têtes d'épingle, renflements, boucles internes et boucles multi-hélices dépendant de la longueur de ces boucles mais beaucoup de ces longueurs sont rarement observées dans les bases de données. De même, beaucoup de termes dépendant de la séquence pour l'estimation d'énergie libre des courtes boucles internes ne sont pas représentés. Cet état de fait implique de faire un choix : bien qu'il soit concevable que "dans l'absolu" ces termes aient des valeurs significativement différentes dont il faille tenir compte, ne vaut-il pas mieux s'en passer en pratique ? Du fait de leur rareté, leur estimation sera de toute manière incertaine et pourrait significativement perturber les autres paramètres.

Ainsi, le modèle actuel, bien que cohérent formellement, est à mon avis trop riche. Les données actuelles, structurales et expérimentales, ne permettent d'en estimer de manière pertinente qu'une partie qu'il convient de définir.

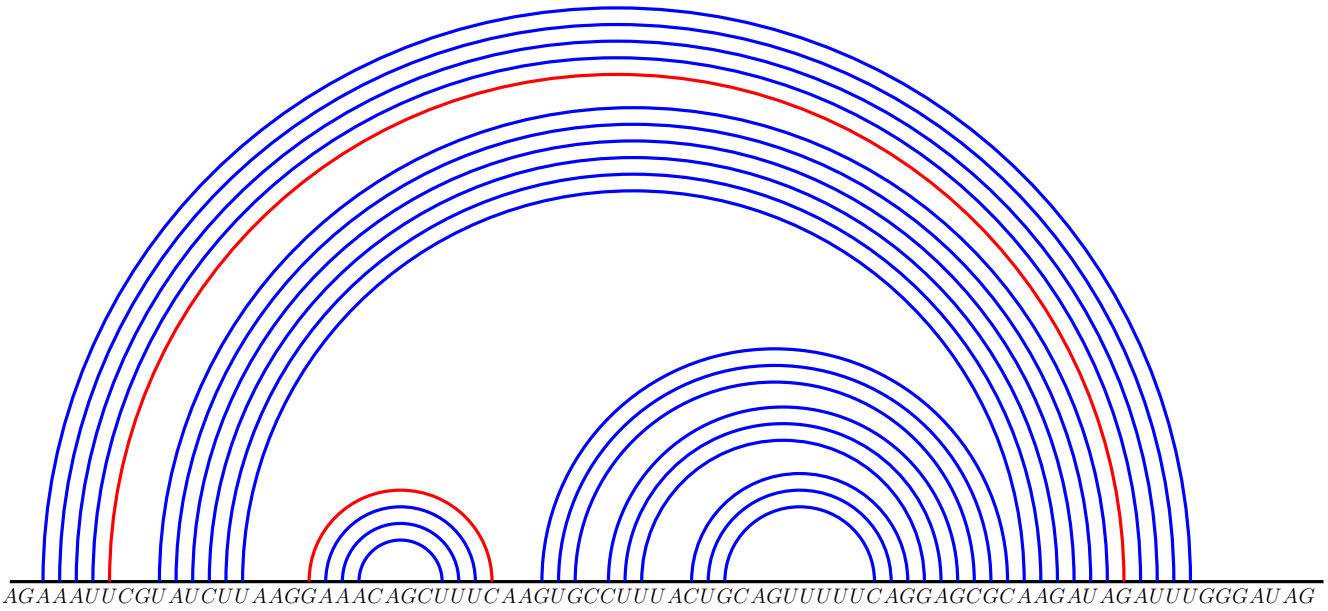
Fiabilité des bases de données

Les erreurs d'annotation dans les bases de données sont problématiques. On peut raisonnablement penser que le risque d'erreur d'annotations augmente avec la taille des séquences et le fait que celles-ci sont déterminées par *comparaison de séquences* plutôt que par l'expérience. En effet, le repliement par comparaison de séquences consiste à reporter un repliement connu pour une séquence sur une autre légèrement différente : cette méthode peut oublier la formation de paires qui seraient propres à cette nouvelle séquence.

Considérons par exemple la séquence 5' AGGUXXXXACCU 3', se repliant selon le schéma ((((....)))). L'énergie de cette hélice, consitutée de 4 paires successives, est alors donnée par la somme de trois termes d'énergie de dipaires p_1, p_2 et p_3 (négligeons les énergies extrémales dans cet exemple qualitatif). Supposons que cette séquence soit mal annotée dans une base de donnée, de sorte qu'on y lise le schéma .(((....))). . Les algorithmes d'optimisation vont alors mettre cette dernière hélice, d'énergie $p_2 + p_3$, en concurrence avec l'hélice (((.....))) d'énergie $p_1 + p_2$. Considérer que la première structure est la bonne résulte en la contrainte "$p_3 \leq p_1$". En supposant que de telles erreurs d'annotation soient uniformément distribuées dans la base, on obtiendra ainsi toute une série de nouvelles contraintes contradictoires $p_i \leq p_j$ et $p_j \leq p_i$ avec i et j quelconques parmi les indices possibles. La seule manière de satisfaire ces nouvelles contraintes étant d'avoir $p_i = p_j \ \forall i, j$, l'ajout de ces contraintes va donc intuitivement avoir tendance à uniformiser les paramètres autour de leur valeur moyenne, c'est-à-dire de diminuer la variance de leur distribution.

Sans prétendre que cette raison soit la seule, on observe de manière frappante ce phénomène d'uniformisation sur les valeurs de CONTRAfold v1 dont le paramétrage a été construit uniquement à partir de séquences annotées par comparaison.

Voici par exemple l'annotation de la séquence "RF00109_A.bpseq" dans la base de données que les auteurs ont utilisée :



Les arcs bleus représentent les paires annotées dans le fichier, les deux arcs rouges représentent deux autres possibles paires supplémentaires. La paire G–C a sans doute été oubliée et cet oubli engendrera des contraintes qui n'ont pas lieu d'être ainsi que j'en ai discuté préalablement.
Les paramètres de CONTRAfold v2, entraîné sur une base de données contenant plus de structures expérimentales, a un écart-type deux fois plus grand que ceux de CONTRAfold v1.

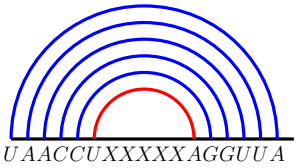
Homologie de séquences au sein d'une base de données

Une autre source de biais à prendre en compte est l'homologie entre séquences au sein d'une base de données. Si deux séquences identiques ou très proches sont incluses dans la base de données, aucune information nouvelle n'est apportée mais un poids supplémentaire leur est donné, poids dont l'impact dépend du critère d'optimisation. Un tel poids peut être souhaitable. Le fait que plusieurs séquences proches se replient de la même manière est un gage de confiance en ce repliement qui peut être signifié à l'algorithme en augmentant le poids de la séquence consensus.

2.6.2 Base de données : choix et construction

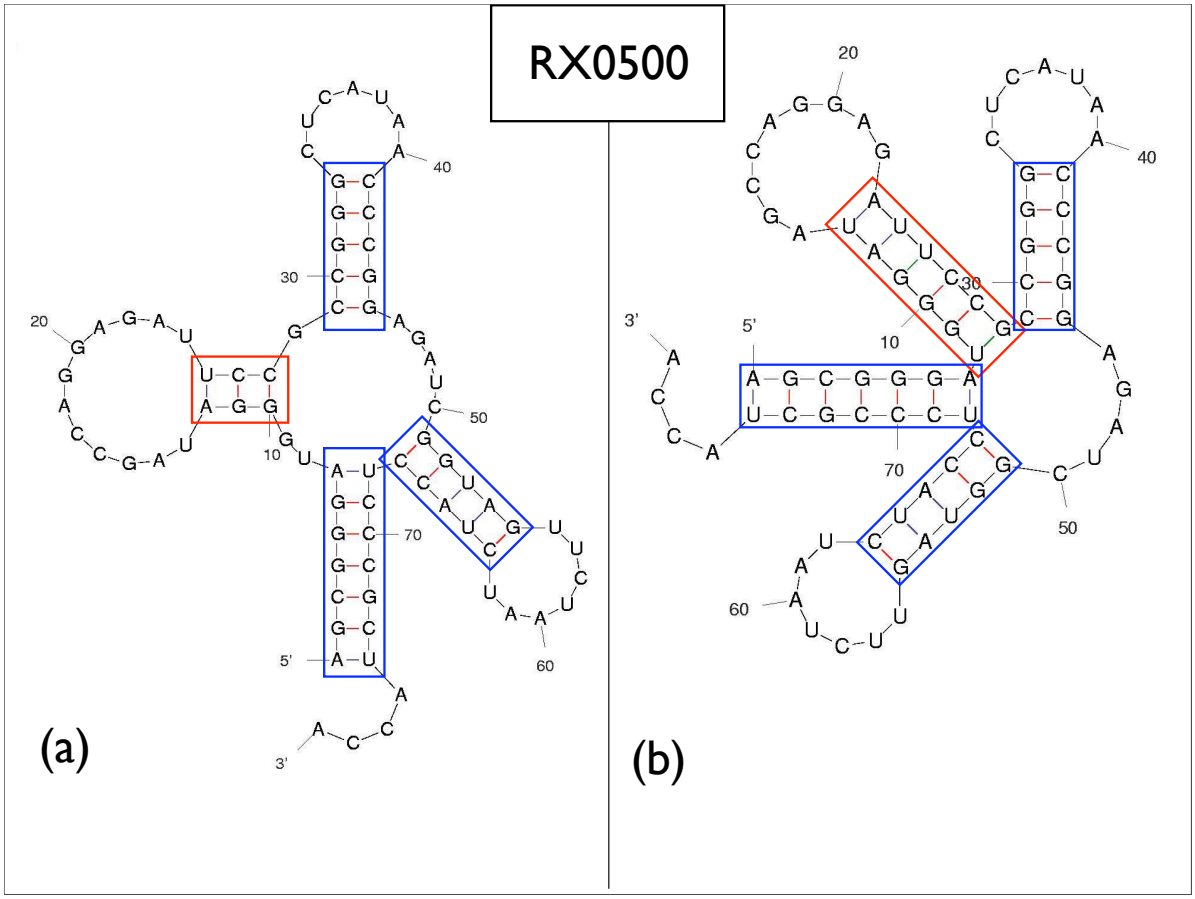
Comme discuté plus haut, le choix de la base de données est critique. Celle-ci doit être la plus fiable possible et comporter un grand nombre de paramètres du modèle afin de pouvoir les estimer de la manière la plus robuste. Le type d'ARN le plus étudié, dont la structure a été le mieux établie, est sans conteste l'**ARNt**.
La remarquable conservation de la structure *tertiaire* de l'ARNt parmi le vivant offre une garantie supplémentaire de la fiabilité des analyses. La "*Sprinzl database*" (http ://www. staff.uni-bayreuth.de/ btc914/search/index.html) répertorie 3768 séquences connues, parmi lesquelles 561 dont le repliement a été vérifié expérimentalement.
La structure consensus de l'ARNt a déjà été illustrée plus haut : elle est constituée de quatre hélices le plus souvent parfaites (c'est-à-dire sans renflements ni courtes boucles internes), d'une boucle multi-hélices et de trois têtes d'épingle. 25 structures contiennent en plus une boucle interne ou un renflement.

J'ai examiné une par une les annotations de ces 561 séquences afin de vérifier que toutes les hélices répertoriées sont bien maximales. Ainsi, on peut constater par exemple que la troisième hélice de la séquence RD5280 ne l'est pas :



La paire signalée en rouge n'est pas annotée dans la *Sprinzl database*. Dans ces cas-là, la structure déclarée comme correcte est celle où cette hélice est complétée par la paire manquante. Le fait que cette paire n'ait pas été annotée ne signifie pas nécessairement un oubli lors de la saisie de la structure ni une erreur expérimentale. Il est possible que ces deux bases ne soient réellement pas appariées mais les raisons d'un tel état de fait sont inexplicables par le modèle d'énergie : les termes de dipaires y sont toujours favorables. Ce modèle ne peut pas déclarer comme optimale une structure non saturée. Toutes les hélices non maximales ont été ainsi prolongées pour que l'algorithme d'optimisation ne soit pas soumis à des contraintes qu'il ne puisse gérer.

Les premières tentatives de repliement m'ont amené à retirer d'autres séquences de la base de données, pour des raisons illustrées par la figure suivante :



(a) La structure secondaire réelle de la séquence RX0500 (b) Une structure secondaire concurrente

Les structures secondaires données en (a) et (b) diffèrent seulement par une hélice encadrée en rouge. Dans la structure secondaire réelle, cette hélice est composée de trois paires alors qu'elle en comporte six dans la structure concurrente. Le modèle d'énergie centré sur les hélices est donc inapte à déclarer le repliement (a) comme meilleur que celui de (b) et cette séquence a été supprimée de la base de données. La structure (b) est sans doute moins favorable que la (a) car elle requiert la formation d'une boucle 4-hélices fortement asymétrique et très contrainte mais le modèle simplifié n'est pas assez fin pour rendre compte de tels effets.

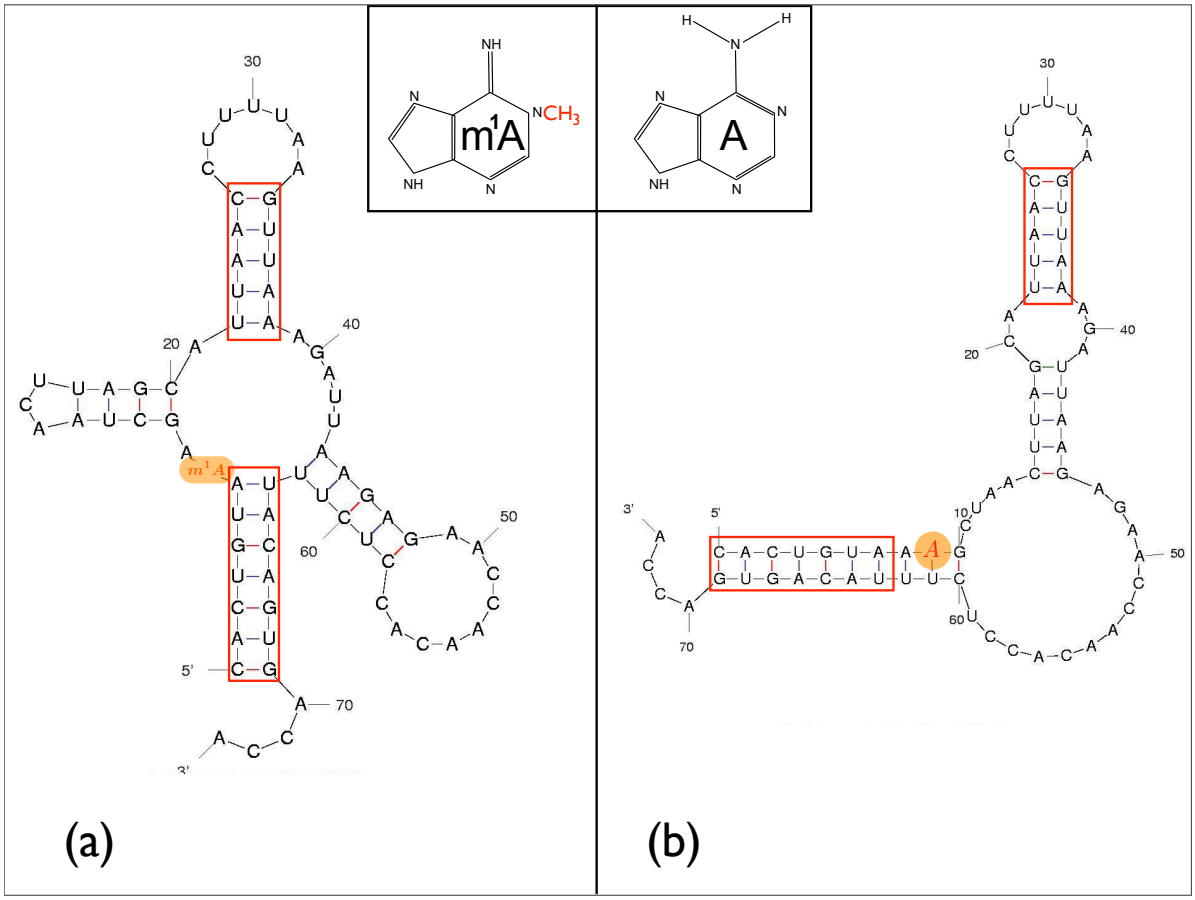
Les autres séquences que j'ai jugées inaccessibles au modèle, à tort ou à raison, sont : RC0500, RD0500, RE0500, RE0501, RE4800, RH0500, RK0501, RK6230, RL0260, RL0501, RL0502, RL0503, RL0504, RL2120, RL2840, RL9280, RM4400, RP0180, RP0500, RP0501, RQ0500, RR0380, RR0501, RR6230, RS0500, RS7661, RV0381, RV0382, RV0500, RV0501, RX0500, RY1140, RY2120.

Les structures d'ARNt étant fiables, les séquences présentant une homologie supérieure à 90% ont été supprimées, en ne gardant à chaque fois que l'une d'entre elles.

Au final, la base de données d'ARNt utilisée contient 413 séquences.

Les bases modifiées

Travailler avec des séquences d'ARNt comporte une difficulté spécifique qu'il ne faut pas occulter : la présence de nucléotides chimiquement modifiés. En effet, les ARNt subissent différentes altérations post-transcriptionnelles dont certaines sont remarquablement conservées, comme T54 et Ψ55 sur la boucle T et les dihydrouridines D16, D17, D20a et D20b sur la boucle D. A ce jour, une centaine de ces modifications ont été répertoriées. Elles jouent des rôles multiples et il en reste sûrement d'autres à découvrir. Les bases modifiées permettent ainsi de déterminer subtilement la géométrie de la boucle anti-codon de l'ARNt ([45], [46]) et par là l'efficacité de la traduction et notamment de la répression des glissements de cadres de lecture. Elles améliorent l'efficacité de l'interaction avec le ribosome [47]. Une autre aspect fonction des bases modifiées est de stabiliser la structure des ARNt. Elles expliquent, par exemple, le fait que les températures de dénaturation d'ARNt de certaines archaebactéries soient particulièrement élevées [46]. La structure tertiaire de l'ARNt est également stabilisée par les bases modifiées T54 et Ψ55 qui forment des liaisons avec G18 et G19 de la boucle D. L'importance des bases modifiées dans le repliement de l'ARNt a été cependant montré de la manière la plus spectaculaire par la série de travaux relatés dans "*The presence of modified nucleotides is required for cloverleaf folding of a mitochondrial tRNA*" [43] et "*A Watson-Crick base-pair-disrupting Methyl Group (m^1A9) is sufficient for cloverfolding folding of human mitochondrial $tRNA^{Lys}$*" [48] : les auteurs y démontrent sur un exemple que le remplacement d'une adénine méthylée en position 9, naturellement présente, par une adénine non-modifiée entraîne le repliement de l'ARNt en une structure secondaire non native, en forme de longue tige et non fonctionnelle.



(a) La structure secondaire réelle de l'ARNt étudié dans [48] (b) Structure secondaire de la même séquence où m^1A a été remplacée par une adénine

La méthylation de l'adénine en position 1 empêche celle-ci de former une liaison hydrogène avec l'uracile : elle ne peut donc pas former de liaison Watson-Crick. En la remplaçant par une adénine, on permet la formation d'une liaison Watson-Crick et il existe une meilleure structure où cette base est effectivement appariée.

Comment tenir compte de ces bases modifiées dans le modèle ? Il est évidemment hors de question de chercher les paramètres thermodynamiques qui leur sont propres, ce qui rajouterait au modèle des milliers de paramètres faiblement contraints. J'ai décidé de plutôt les traiter comme il a été fait dans [28] : les bases modifiées sont assimilées aux bases canoniques dont elles dérivent (par exemple Ψ est traitée comme U) à l'exception de certaines d'entre elles, comme m^1A, pour lesquelles la formation d'une paire Watson-Crick ou Wobble est interdite . La liste des bases ne pouvant former de paires Waston-Crick ou Wobble a été déterminée en fonction de leur composition chimique et de la littérature :

- guanines modifiées : m_2^2Gm, $galQ$, m^1G, Gm, $manQ$, yW, $o2yW$
- cytosines modifiées : k^2C, m^5Cm, m^3C
- uraciles modifiées : D, I, m^1I, acp^3U
- adénines modifiées : io^6A, m^2A, m^1A, $Ar(p)$, ms^2t^6A

Ainsi donc, dans le modèle, la contrainte $\Delta F_d \begin{pmatrix} X & -\overline{X} \\ Y & -\overline{Y} \end{pmatrix} = +\infty$ est rajoutée lorsqu'au moins une des 4 bases impliquées est une base modifiée présente dans la liste ci-dessus.

Il a été montré de plus que la dihydrouridine D ne se superposait pas à ses voisines [49] : cela se traduit par l'ajout de la contrainte $\Delta F_t \begin{pmatrix} W & Z \\ X & -\overline{X} \end{pmatrix} = +\infty$ si W ou Z est une dihydrouridine. Dans tous les autres cas, les bases modifiées se comportent comme les bases non-modifiées dont elles dérivent.

2.6.3 Une forme simplifiée du modèle d'énergie

La base de données choisie contient beaucoup d'hélices de compositions différentes et de têtes d'épingle sensiblement de même taille mais peu de boucles internes et renflements. Il n'y a pas assez de données pour paramétrer les boucles internes et les renflements dans toute leur généralité. Le modèle d'énergie entraîné sur cette base doit donc tenir compte de ces limites : l'accent a été mis sur les paramètres se rapportant aux hélices en limitant la quantité de termes autres. A séquence donnée, on s'attend à ce que les meilleures structures concurrentes soient toutes constituées approximativement du même nombre de paires et donc du même nombre de bases libres. Il est donc vraisemblable que l'entropie des bases libres soit un facteur peu discriminant, d'autant plus que la théorie des polymères prévoit que l'entropie des boucles ne dépend "que" logarithmiquement de leur longueur. Autrement dit, tenir compte de l'entropie des bases libres ne perturbe pratiquement pas la hiérarchie établie par la restriction du modèle aux termes de dipaires. L'entropie des bases libres est par contre nécessaire pour rendre compte de la dénaturation de l'ARN à des températures élevées mais l'étude de cet aspect n'est pas un objectif de cette thèse. Ainsi, il est convenable, dans un premier temps, de réduire le paramétrage de tous les types de boucles à seulement un terme d'initiation. Tous les bonus et pénalités exceptionnels sont supprimés comme, par exemple, le fait qu'une tête d'épingle ne soit constituée que de cytosines. Les tables pour évaluer les boucles internes 1×1, 1×2, 2×2 en fonction de leur séquence sont également mises de côté. Un terme d'initiation indépendant de la séquence est néanmoins utilisé pour chacune de ces boucles.

Voici une énumération complète des paramètres de ce modèle simplifié :

- $6 \times 6 = 36$ termes d'énergie libre de dipaires $\Delta F_d \begin{pmatrix} X - \overline{X} \\ Y - \overline{Y} \end{pmatrix}$, se réduisant à 21 par symétrie.
- $6 \times 4 \times 4 = 96$ termes d'énergie libre de paire terminale $\Delta F_t \begin{pmatrix} X - \overline{X} \\ Y \quad Z \end{pmatrix}$
- 1 terme d'initiation pour les têtes d'épingle
- 5 termes pour les boucles internes :
  - 3 termes d'initiation pour les boucles internes 1×1, 1×2, 2×2
  - 2 termes pour les autres boucles internes : un terme d'inititation a_{bi} et un terme b_{asym} intervenant dans la pénalité d'asymétrie choisie de la forme :

$$F_{asym}(l_1, l_2) = b_{asym} \times \frac{|l_1 - l_2|}{l_1 + l_2} \tag{2.31}$$

- 1 terme d'initiation pour les boucles multi-hélices

- 5 termes pour les renflements :
  - 4 termes dépendant de la séquence pour les renflements de taille 1
  - 1 terme d'initiation pour les renflements de longueur plus grande

Ce modèle compte en tout 129 paramètres, dont 117 s'appliquent aux hélices.
Une simplification supplémentaire a été utilisée : les termes d'initiation de têtes d'épingle et de boucles internes ont été supposés égaux.

2.6.4 Critère d'optimisation

Si le modèle d'énergie idéal est encore hors de notre portée, alors le critère d'optimisation détermine de quelle manière on s'en écarte.

Considérations préliminaires

Le critère utilisé par CONTRAfold est équivalent à la maximisation des produits des poids de Boltzmann des structures réelles. Le critère utilisé par CG consiste à minimiser la différence entre l'énergie libre de la structure réelle et celle de ses meilleures concurrentes. Ces deux critères peuvent produire des effets différents selon les caractéristiques de la base de données. Initialisons le modèle d'énergie d'une certaine manière et supposons que la base de données contienne des séquences "faciles" et "difficiles" à prédire. "Facile" signifie que la structure réelle a une énergie libre assez inférieure à ses concurrentes pour pouvoir rester la structure d'énergie minimale lorsque le modèle d'énergie fluctue significativement. "Difficile" signifie l'inverse.
Les critères d'optimisation de CONTRAfold et CG peuvent alors intuitivement donner lieu à des stratégies différentes :

- CG ne s'occupe que de l'erreur, de sorte que les structures faciles à trouver ne pèsent pas dans l'optimisation. CG s'occupe alors seulement de minimiser l'erreur commise sur les structures difficiles.

- *A contrario*, toutes les séquences contribuent au critère d'optimisation de CONTRAfold. Une stratégie possible d'optimisation serait ainsi de maximiser les poids de Boltzmann des structures faciles à trouver, sachant que celui des difficiles variera peu.

Une mesure permettant qualitativement d'apprécier ces stratégies d'optimisation est la *variance de la sensibilité* : on s'attend à ce que celle des prédictions de CG soit plus faible que pour CONTRAfold car CG s'attache plus à prédire des structures proches de la structure réelle que CONTRAfold. Malheureusement, seules la valeurs moyenne de la sensibilité et la VPP sont utilisées dans la littérature et il n'est pas possible de caractériser clairement l'impact du choix du critère d'optimisation.

La minimisation du classement (MC)

Le critère d'optimisation que j'ai choisi est celui de la minimisation du classement de la structure réelle dans la liste des structures sous-optimales à séquence donnée. Le modèle qui s'en déduit sera appelé "MC". Formellement, ce critère s'écrit comme suit. Soient $S^r_{x_1},\ ..., S^r_{x_N}$ les structures réelles des N séquences de la base de données. Pour chaque séquence x_i, des structures candidates $S^1_{x_i},\ ..., S^{K_i}_{x_i}$ sont proposées par un algorithme de repliement. Le modèle d'énergie optimal $\boldsymbol{p}_{opt}$ est défini comme celui qui minimise la quantité :

$$\chi(\boldsymbol{p}) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} \Theta(\Delta F_{\boldsymbol{p}}(S^r_{x_n}) - \Delta F_{\boldsymbol{p}}(S^k_{x_n})) \tag{2.32}$$

$$\text{où :} \qquad \Theta(t) = \begin{cases} 0 & \text{si } t < 0 \\ 1 & \text{sinon} \end{cases} \tag{2.33}$$

Ainsi, si pour la séquence j il y a trois séquences candidates qui ont une énergie libre inférieure à la structure réelle, on a $\sum_{k=1}^{K_j} \Theta(F_{\boldsymbol{p}}(S^r_{x_j}) - F_{\boldsymbol{p}}(S^k_{x_j})) = 3$. La minimisation de χ revient donc bien à minimiser le *classement moyen* de la structure réelle. Ce critère est clairement exigeant et traite à égalité toutes les séquences de la base de données. Son principal inconvénient est qu'il n'y a pas de méthode de calcul du minimum global de $\chi(\boldsymbol{p})$.

Un optimum est déterminé par une suite de minimisations locales opérées par l'algorithme suivant :

1. initialiser $\boldsymbol{p}$
2. engendrer K structures candidates pour chaque séquence à l'aide d'un algorithme de repliement utilisant $\boldsymbol{p}$.
3. ajuster $\boldsymbol{p}$ en le faisant évoluer selon l'équation :

$$\frac{d\boldsymbol{p}}{dt} = -\frac{d\chi}{d\boldsymbol{p}}(\boldsymbol{p}) \tag{2.34}$$

   qui converge vers un minimum local de $\chi(\boldsymbol{p})$. (Note : pour réaliser cette opération, il faut remplacer Θ par une approximation continue)
4. reprendre l'opération (2) jusqu'à convergence

En initialisant le modèle avec les paramètres de Turner corrigés par une capacité calorifique non nulle et en utilisant $K = 15$ structures candidates pour chaque séquence, cet algorithme converge au bout de trois itérations en utilisant 5000 pas de temps pour l'étape (3).

2.6.5 Résultats

En retenant la meilleure des deux premières structures, la sensibilité des prédictions pour la base sur laquelle le modèle a été entraîné est de 98,5%.

390 structures sur 413 sont retouvées parmi les deux premières propositions (364 en se limitant à la première). La valeur de 98,5% correspond au nombre total de paires correctement prédites (à savoir 8494) rapporté au nombre total de paires de bases présentes dans la base de données (à savoir 8620). La VPP correspondante est de 98,2%. L'examen individuel des erreurs montre qu'en général elles consistent en la mauvaise prédiction d'une ou deux hélice(s) de l'ARNt au profit d'une ou deux autre(s) hélice(s) plus longue(s).

Voici les 23 structures non prédites, avec indiqués entre parenthèses le nombre de paires non prédites dans la meilleure des deux premières structures proposées et le classement de la structure correcte :

RD0260 (9,10^e)	RD1580 (9,5^e)	RD9290 (1,3^e)	RE2640 (5,3^e)
RF7780 (4,4^e)	RG9230 (6,5^e)	RL1540(8, $> 15^e$)	RL7631(7,4^e)
RL7650 (9,4^e)	RL7661 (4,12^e)	RL9200 (6,12^e)	RP3280 (1,3^e)
RQ0260 (1,4^e)	RS8560 (4,10^e)	RV0380 (3,4^e)	RX0380 (8,10^e)
RX0540 (3,5^e)	RX0900 (10,4^e)	RX3280 (9,7^e)	RX4400 (10,12^e)
RF7590 (2,5^e)	RG6240 (6,$> 15^e$)	RV5360 (1,4^e)	

Voici les paramètres d'énergie libre de dipaires Watson-Crick obtenus par cette méthode, ainsi qu'un extrait des paramètres d'énergie libre de paire extrémale :

ΔF_d	Turner99 $\Delta c_p \neq 0$	MC		ΔF_t	MC	Turner99
A–U A–U	-0.83	-1.21		A A A–U	1.34	-0.3
A–U U–A	-1.09	-1.30		A C A–U	0.86	-0.5
U–A A–U	-1.14	-0.82		A G A–U	1.99	-0.3
A–U C–G	-1.94	-2.10		C A A–U	1.17	-0.1
G–C A–U	-1.95	-1.81		C C A–U	1.37	-0.2
C–G A–U	-2.06	-2.07		C U A–U	1.04	-0.2
A–U G–C	-2.11	-2.12		G A A–U	2.03	-1.1
C–G G–C	-2.45	-2.65		G G A–U	1.96	-0.2
G–C C–G	-2.76	-2.65		U C A–U	0.96	-0.3
C–G C–G	-2.95	-2.62		U U A–U	0.68	-1.1
moyenne	-1.93	-1.93		moyenne	1.47	-0.43
écart-type	0.68	0.61		écart-type	0.47	0.34

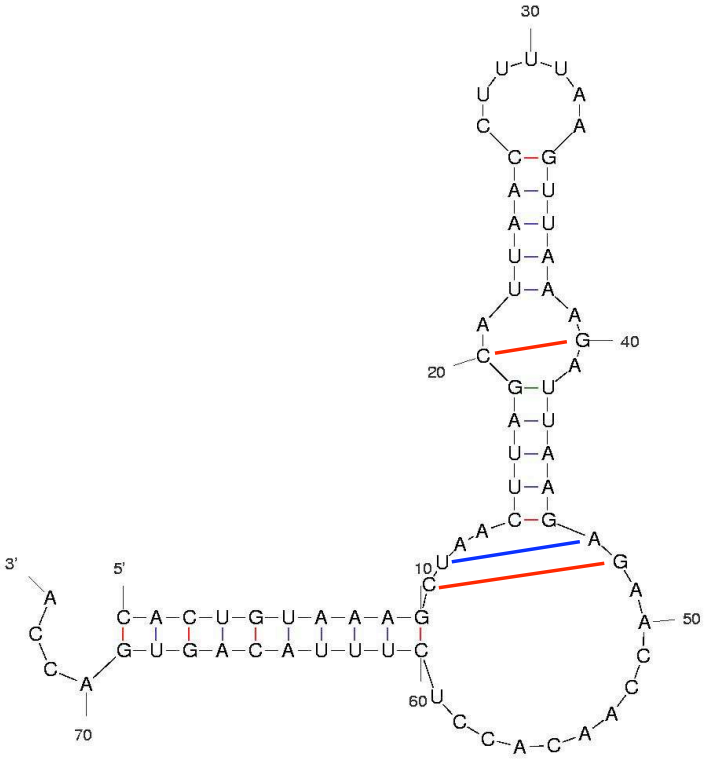
Là encore, des modifications significatives ont été effectuées par rapport au modèle avec lequel l'optimisation a été initialisée. Ce nouveau modèle présente des spécificités : il est par exemple le seul à déclarer la dipaire $\begin{pmatrix} \text{A}-\text{U} \\ \text{U}-\text{A} \end{pmatrix}$ comme le moins favorable des dipaires et il attribue des valeurs très similaires aux trois termes de dipaires uniquement constitués de paires G–C. On constate également une grande variabilité des termes de paire extrémale. Dans l'exemple ci-dessus, la différence entre le terme le moins favorable et le plus favorable est de 1.35 kcal/mol, soit une valeur comparable à l'énergie de dipaires A–U. Une telle variabilité est observée sur l'ensemble des termes de paire extrémale : ce modèle attribue une grande importance à celles-ci. Il n'y a pas d'équivalent direct à ces paramètres de paire extrémale dans le modèle Turner99 : ceux-ci dépendent de la nature de la boucle connectée à cette paire extrémale (tête d'épingle, boucle interne ou boucle multi-hélices) et doivent être complétés de la pénalité entropique attachée à cette boucle. Les paramètres rapportés dans le tableau sont ceux afférents aux têtes d'épingles. Dans le modèle d'énergie présenté dans la partie (1.6.3), il n'y a pas de pénalité entropique de boucle. Lors de l'optimisation, la pénalité entropique "réelle", si elle existe, a été nivelée et absorbée dans le terme de paire extrémale. Ceci explique la différence de signe observée pour ΔF_t entre MC et Turner99. Dans Turner99, les paramètres de paires extrémales sont très proches les uns des autres, sauf lorsque les premières bases libres voisines sont 5' G A 3' ou 5' U U 3', auquel cas un bonus a été ajouté. MC calcule que 5' U U 3' sont les paires voisines les plus favorables mais à l'inverse déclare 5' G A 3' et 5' A G 3' comme les moins favorables.

Il est intéressant de constater que ce simple modèle, qui paramétre les hélices finement et le reste grossièrement, est capable d'avoir une sensibilité de 98,5% sur une base de données d'ARNt : leur structure secondaire peut être simplement expliquée comme le résultat d'une compétition entre hélices, sans avoir à faire intervenir de liaisons tertiaires.

Deux succès intéressants

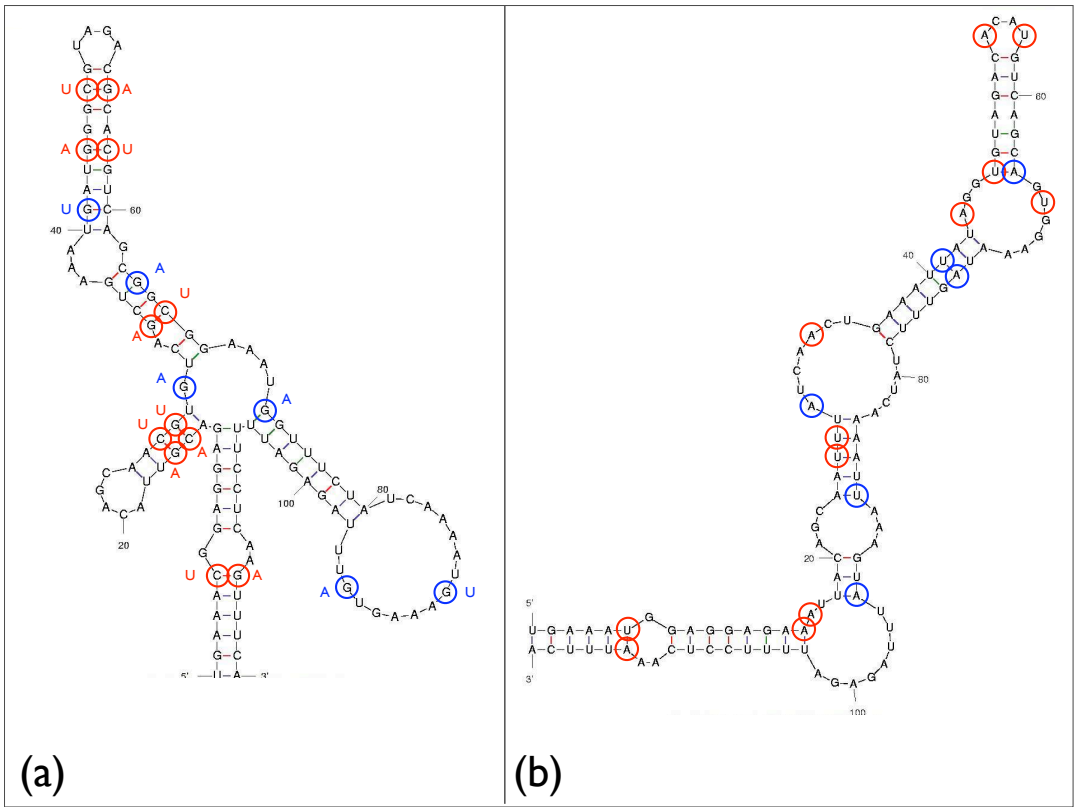
Ce nouveau modèle d'énergie prédit avec succès deux structures remarquables, que le modèle Turner99 ne reproduit pas.

(i) La première est l'exemple de l'ARNt se repliant différemment en fonction de la présence d'un résidu m^1A9 [48], illustré plus haut : en présence de ce résidu, le repliement prédit est bien la structure secondaire en "feuille de trèfle". En son absence, on obtient la structure en tige, saturée à l'aide de trois paires supplémentaires :



Ainsi, ce modèle est assez *sensible* pour rendre compte de l'effet d'une unique modification chimique.

(ii) La seconde est rapportée dans *"Distinctive structures between chimpanzee and human in a brain noncoding RNA"* [50]. Les auteurs y ont élucidé la structure de l'ARN codé par la région HAR1 du génome de l'homme et du chimpanzé. Cette région fait 118 bases de long et, en dépit d'une homologie de 85%, les ARN correspondants se replient distinctivement chez l'homme et chez le chimpanzé.



(a) Structure secondaire de l'ARN codé par HAR1 chez l'homme
(b) Idem chez le chimpanzé

Les bases différentes entre l'homme et le chimpanzé sont cerclées de rouge si elles participent à des paires chez l'homme, en bleu sinon. On voit que 12 des 18 mutations ont pour effet de remplacer des paires Watson-Crick G–C chez l'homme par des paires Watson-Crick A–U. Ces mutations déstabilisent légèrement les hélices, ce qui, avec la possibilité d'en former de nouvelles via les 6 autres mutations, entraîne un repliement différent. Les structures prédites par le modèle MC sont correctes dans les deux cas.

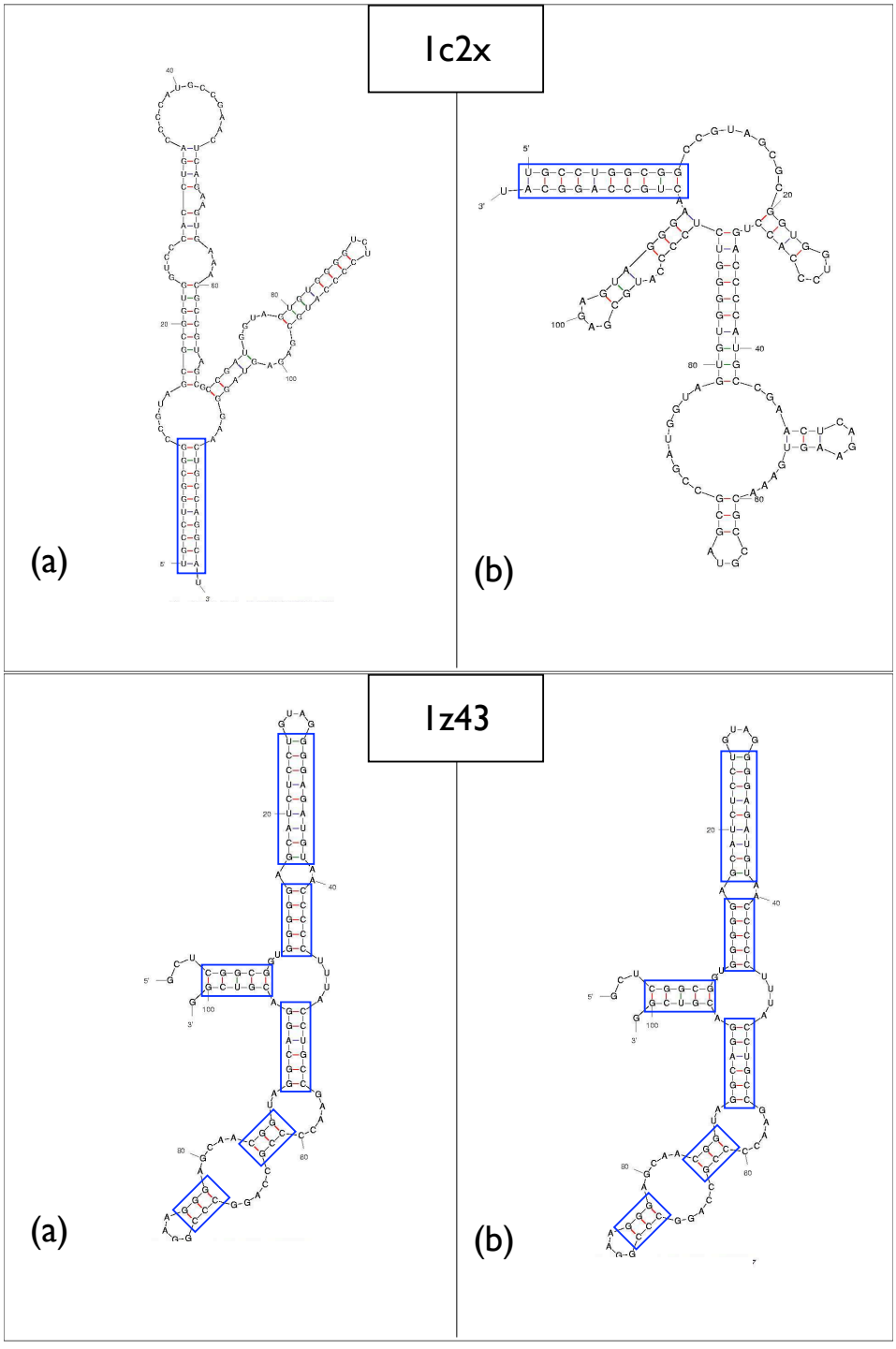
Cet exemple et le précédent sont de fortes illustrations des limites de la prédiction de structure par comparaison de séquences.

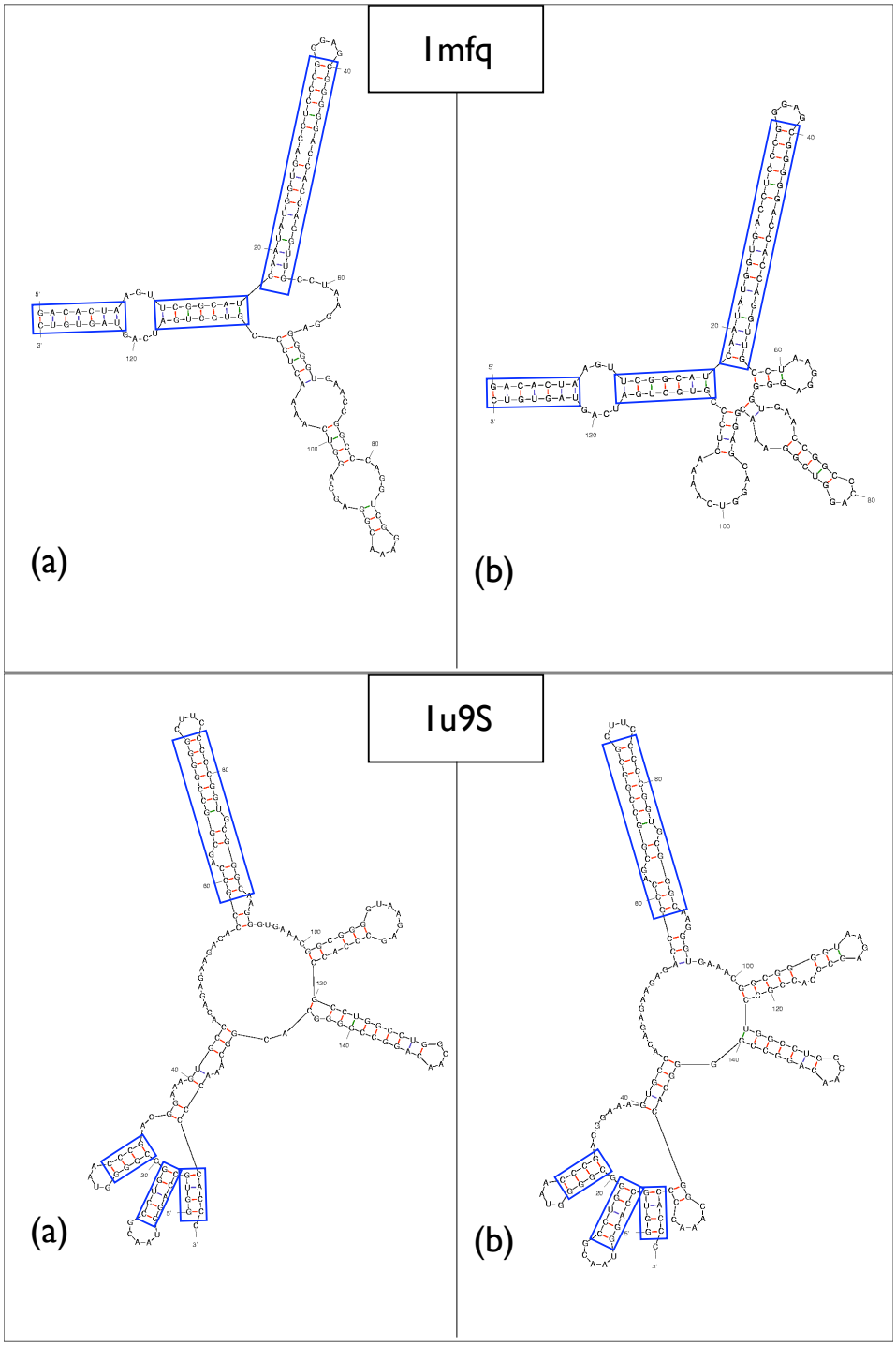
Dans [48] , la modification d'une unique base perturbe globalement la structure secondaire. Dans [50], les mutations transforment des paires Watson-Crick en d'autres paires Watson-Crick mais la structure secondaire n'est pas maintenue. Ces résultats entachent d'un doute la plupart des bases de données actuelles, comme par exemple la *Rfam database* [40] et la *SRPDB* [51], massivement fondées sur la comparaison de séquences.

Ces deux exemples montrent également qu'*a contrario* l'approche thermodynamique peut donner les bons résultats. C'est pourquoi je ne teste pas ce nouveau modèle sur des structures obtenues par comparaison de séquences : il n'est pas aisé de démêler le vrai du faux. Toutes les évaluations d'algorithmes de repliement publiées dans la littérature doivent être nuancées du fait de l'incertitude pesant sur les bases de données, incertitude dont on ne connaît l'ampleur. L'objectif d'une sensibilité de prédiction de 100% sur ces bases de données n'est sans doute pas souhaitable !

Quelques tests sur des séquences plus longues

Voici quelques tests supplémentaires effectués sur des structures de plus de 100 bases répertoriées dans la PDB, c'est-à-dire dont la structure tertiaire a été résolue. Leur structure secondaire en a été extraite avec le logiciel RNAview [52]. Les structures (a) sont les structures réelles, les structures (b) sont les meilleures prédictions. Les hélices correctement prédites sont encadrées en bleu.


1c2x
(a)
(b)
1z43
(a)
(b)


1mfq
(a)
(b)
1u9S
(a)
(b)

Sur ces relativement longues séquences, les résultats sont plus mitigés, puisque seulement la séquence 1z43 (ARN SRP) est prédite correctement. Un examen attentif de ces résultats montre qu'ils sont dus à la faiblesse du paramétrage des renflements et boucles internes. En effet, certains de ces motifs présents dans les structures secondaires réelles, comme la boucle interne 1×1 $\begin{matrix} 5'\text{CCGAU}3' \\ 3'\text{GGAUG}5' \end{matrix}$ de 1c2x, ne sont pas déclarés comme favorables thermodynamiquement par le modèle d'énergie. De manière générale, on s'aperçoit que le modèle prédit moins de renflements et de boucles internes et plus de boucle multi-hélices. On s'attend donc à ce que la qualité de prédiction diminue à mesure que la structure réelle contienne beaucoup de renflements et boucles internes, ce qui statistiquement est de plus en plus fréquent avec des séquences de plus en plus longues.

Il est donc apparent, comme on pouvait s'y attendre, que le modèle est incomplet. Il a été construit pour paramétrer correctement les hélices et à ce titre permet de rendre compte de la structure des ARNt de manière très satisfaisante. Ce modèle sera testé plus encore dans la partie "Prédiction de pseudo-nœuds". Il y sera montré que ce modèle continue à être des plus efficaces pour prédire des structures composées d'hélices parfaites, même avec pseudo-nœuds : ceci démontrera clairement que l'efficacité du modèle ne se limite pas à la base de données sur laquelle il a été entraîné mais, au contraire, s'élargit au type de structures secondaires pour lesquelles il a été conçu, c'est-à-dire toutes celles constituées d'hélices parfaites. Ainsi, il est raisonnable de penser que la base de ce modèle est très bien établie et il sera développé dans de futurs travaux par la même démarche rationnelle pour mieux rendre compte des renflements et boucle internes.

Conclusion

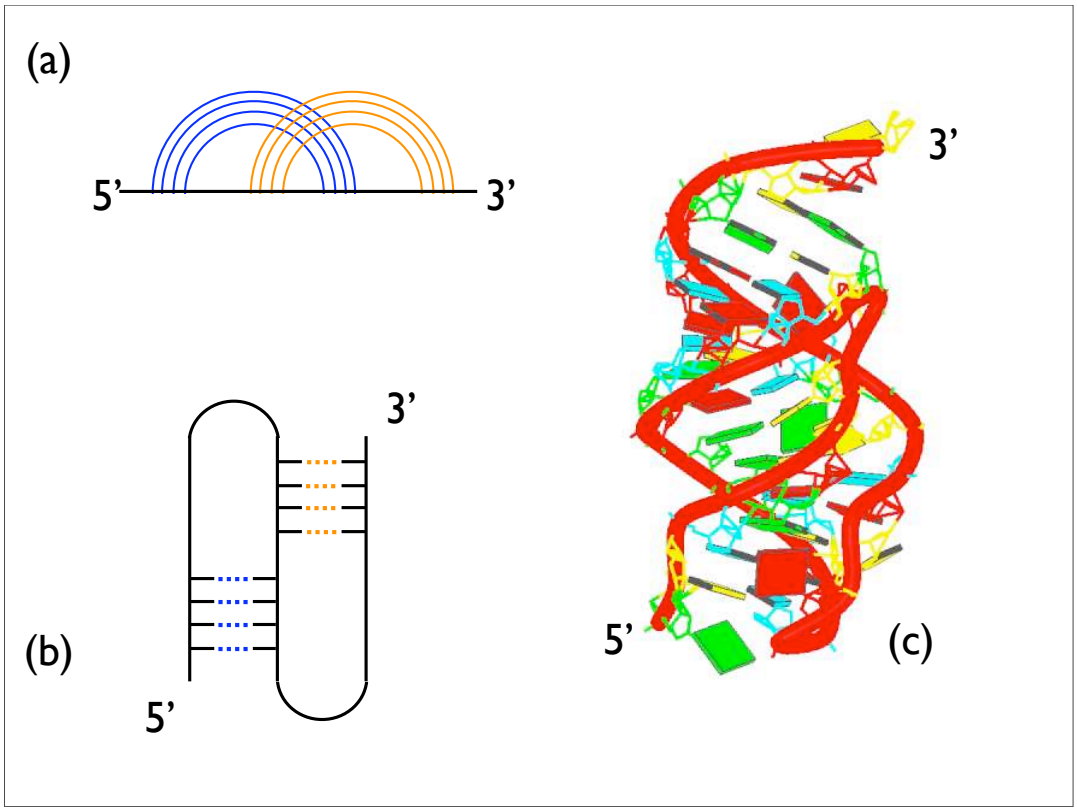
Ainsi donc, j'ai effectué une revue critique des différents paramétrages actuellement disponibles et ai tenu compte de leurs faiblesses pour proposer une nouvelle méthodologie qui donne des résultats encourageants. Ces résultats ne permettent évidemment pas de dire que ce nouveau modèle s'approche plus des "valeurs réelles" que les autres, cette prérogative étant réservée à l'expérimentateur. Certains de mes choix ne sont pas exempts eux-mêmes de critique, comme le fait d'enlever de la base de données des structures estimées inexplicables par le modèle ou le choix de n'avoir que des hélices maximales. Cependant, j'espère avoir démontré que cette question du paramétrage n'était pas encore résolue et que de grandes incertitudes pesaient encore sur les énergies libres de dipaires qui sont essentielles au repliement. Ce point de vue va à l'encontre de récents travaux expérimentaux qui visent à compléter le modèle en paramétrant l'effet subtil dit d'"empilement coaxial" (c'est-à-dire l'estimation d'un bonus attaché à l'alignement des axes de deux hélices issues d'une même boucle multi-hélices) en maintenant l'hypothèse $\Delta c_p = 0$ [53]. Je pense que l'effort expérimental requis vise plutôt à l'estimation des Δc_p associées aux mesures déjà disponibles.

Chapitre 3

Le genre : un critère de classification des structures secondaires d'ARN

3.1 Introduction : les pseudo-nœuds

Dans le chapitre précédent, “Modèle d’énergie libre”, différents motifs constitutifs des structures secondaires, comme les hélices ou les têtes d’épingle, ont été caractérisés et paramétrés. Un d’entre eux a cependant été laissé de côté : le pseudo-nœud. Celui-ci est un motif de repliement comprenant des bases libres et des paires de bases. Il met en jeu au moins deux hélices et se caractérise par l’apparition d’un croisement dans la représentation diagrammatique des structures secondaires. Le pseudo-nœud le plus simple est le pseudo-nœud “H” :



Le pseudo-nœud “H” (a) Représentation diagrammatique (b) Représentation schématique (c) Représentation 3D

Les pseudo-nœuds peuvent avoir une importance fonctionnelle comme il a été montré à diverses reprises.

- La présence de pseudo-nœuds permet ainsi à certains ARN, notamment viraux, de s'auto-cliver sans l'intervention de protéines. Cette propriété a été bien étudiée dans le cas du virus satellite de l'hépatite B, le HDV (*hepatitis delta virus*). Le HDV se présente sous la forme d'un ARN simple brin *circulaire* de 1700 bases de long. La réplication de ce génome circulaire par la polymérase II à ARN de l'hôte *eucaryote* va conduire à la formation d'un long génome linéaire contenant plusieurs unités du génome initial. Un pseudo-nœud codé dans le génome initial va permettre à ce génome *multimérique* de se cliver efficacement en ses unités adéquates [54] [55]. La structure tridimensionnelle de ce pseudo-nœud complexe a été élucidée dans [56].

  Chez les eucaryotes, certains introns ont également la capacité de s'auto-épisser et sont désignés dans la littérature comme le "*group I self-splicing introns*" [57]. La structure de ces introns s'organise autour d'un pseudo-nœud [58].

- Des pseudo-nœuds sont observés en 3' d'ARN viraux, comme pour le TYMV (*turnip yeow mosaic virus*) [59] ou le TMV (*tobacco mosaic virus*) [60]. Un des rôles de ces pseudo-nœuds, démontré expérimentalement, est de favoriser la traduction des ARN viraux en compensant l'absence de la queue poly-(A) normalement présente dans les ARN messagers de l'hôte eucaryote du virus [61]. Cependant, le mécanisme sous-jacent n'est pas encore compris.

- Certains pseudo-nœuds d'ARN viraux, comme le HIV, ont la capacité d'induire un changement du cadre de lecture lors de leur traduction [62]. Ce changement provient de l'action combinée d'une séquence glissante composée de 7 bases et d'un pseudo-nœud situé en aval. Alors que la séquence glissante est traduite, le pseudo-nœud enraye le coulissement de l'ARN dans le ribosome de sorte que le ribosome marque une pause sur la séquence glissante, ce qui statistiquement entraîne un déplacement du cadre de lecture d'une base à rebours (*-1 frameshifting*) [63].

Ainsi donc, de nombreux exemples montrent l'importance fonctionnelle des pseudo-nœuds. Cette importance est d'autant plus grande qu'ils sont les éléments centraux de stratégies de prolifération virale, ce qui en fait une cible privilégiée pour de futures thérapies.

Dans le problème de la prédiction des structures secondaires, les pseudo-nœuds sont distingués des autres motifs car ils présentent une difficulté propre. Alors qu'existent des relations de récurrence permettant d'étudier efficacement les structures secondaires sans pseudo-nœud (cf chapitre suivant), la prédiction de pseudo-nœuds est un problème NP-complet [64]. Le nombre de structures comportant un pseudo-nœud à séquence donnée est exponentiellement grand et il n'y a pas d'algorithme permettant de les énumérer efficacement. De plus, il n'y a pas de données expérimentales permettant de caractériser les propriétés thermodynamiques d'un pseudo-nœud.

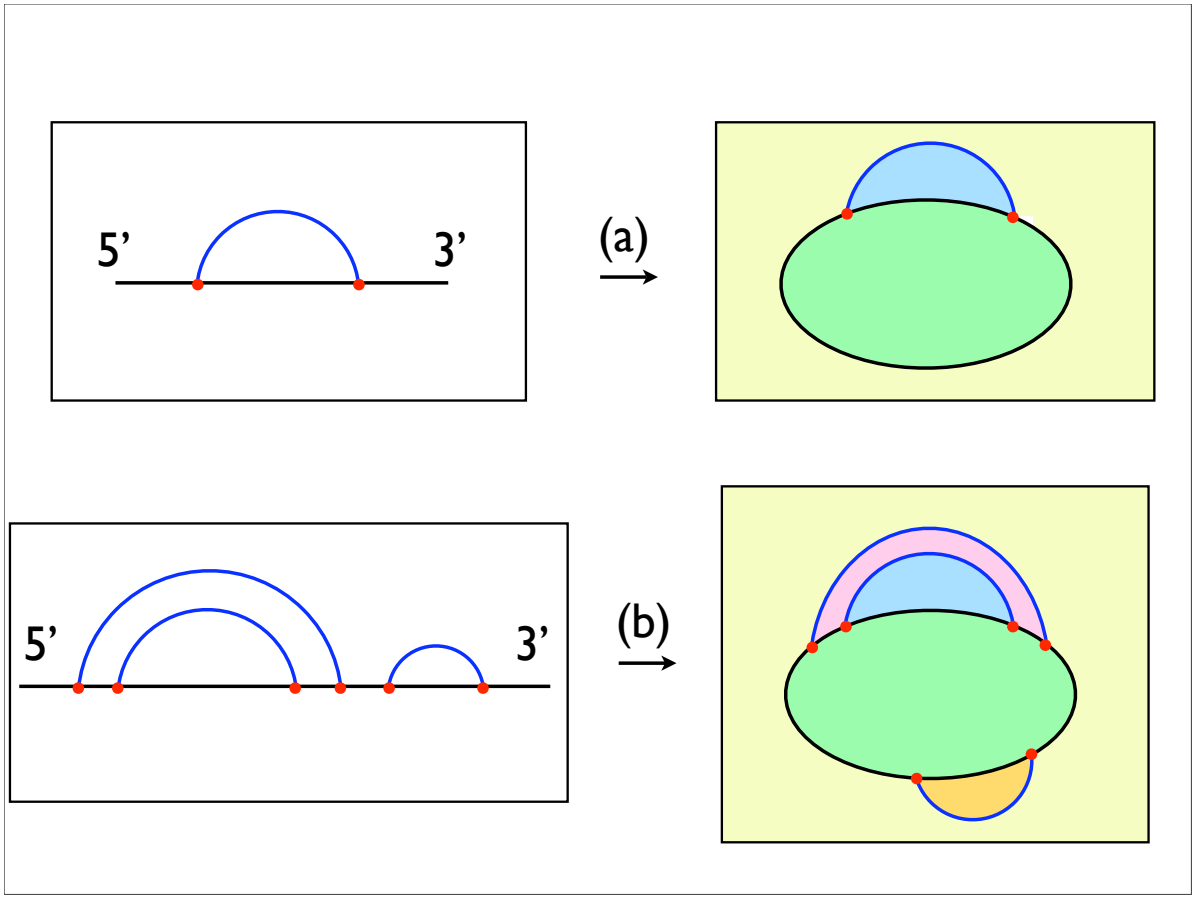
Le but de ce chapitre est de proposer une méthode permettant de pallier à ces problèmes d'énumération et de paramétrage. Le "genre", un indice topologique, est introduit et permet d'effectuer une classification des pseudo-nœuds. L'étude des bases de données montrera que les topologies des structures réelles n'ont pas la diversité prodigieuse qu'elles pourraient théoriquement avoir. Au contraire, les topologies réelles se distribuent de manière très particulière parmi les différentes classes induites par le genre. Le problème de la prédiction s'en retrouvera ainsi considérablement simplifié car nous disposerons à la fin de ce chapitre d'un critère contraignant permettant de caractériser les pseudo-nœuds réels. Ce critère permettra à tout algorithme de prédiction de pseudo-nœuds d'éliminer simplement et efficacement beaucoup de pseudo-nœuds invraisemblables de sa recherche. Le paramétrage des pseudo-nœuds proposé sera de nature topologique et non thermodynamique.

Le plan suivi dans ce chapitre est le suivant :

I) Une nouvelle représentation des diagrammes de structures secondaires
II) Définition du genre
III) Calcul du genre
IV) Propriétés du genre
V) Bilan et intérêt du genre dans la prédiction de pseudo-nœuds

3.2 Une nouvelle représentation des diagrammes de structures secondaires

En connectant les extrémités 5' et 3' des diagrammes de structures secondaires et en dessinant les arcs à l'extérieur de la courbe fermée ainsi formée, on obtient un objet pouvant être vu comme un polyèdre. Ses sommets sont les bases appariées et les arcs représentant les paires de bases en deviennent des arêtes.



La caractéristique d'Euler χ d'un polyèdre est définie par :

$$\chi = S - A + F \tag{3.1}$$

où S est le nombre de sommets du polyèdre, A est le nombre d'arêtes et F le nombre de faces. χ est un nombre entier ne dépendant que du type de la surface –plus exactement la *variété* (*manifold*)– sur laquelle est dessiné le polyèdre.

Lorsqu'un polyèdre peut-être dessiné dans un plan, sa caractéristique d'Euler est égale à 2, comme on peut le vérifier sur les deux exemples précédents :

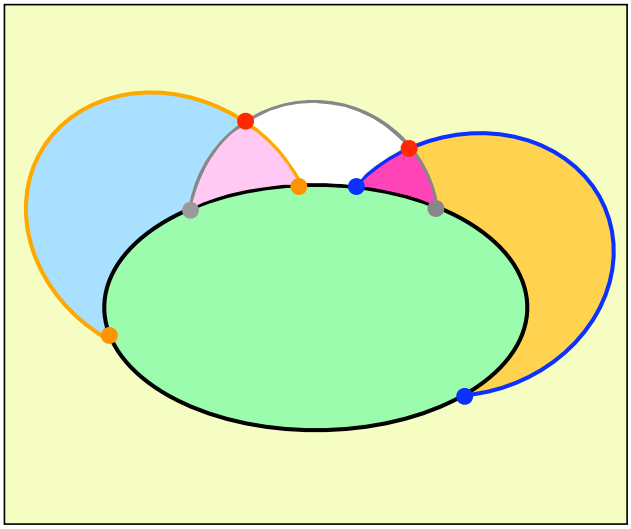
a) $S = 2$, $A = 3$ (2 arêtes noires et 1 arête bleue) et $F = 3$ (signalées par trois couleurs différentes) $\rightarrow \chi = 2 - 3 + 3 = 2$
b) $S = 6$, $A = 9$, $F = 5 \rightarrow \chi = 6 - 9 + 5 = 2$

Il est clair que tout diagramme dans lequel les arcs ne se croisent pas peut être dessiné dans un plan : c'est pourquoi de telles structures secondaires peuvent être également appelées "structures planaires". Les structures planaires ont toutes la même caractéristique d'Euler, qui vaut 2.

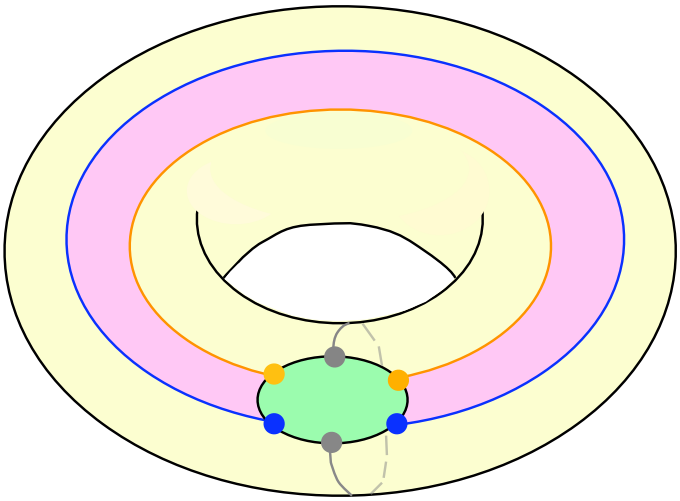
Qu'en est-il des diagrammes dans lesquels des arcs s'intersectent, c'est-à-dire des structures secondaires comportant un pseudo-nœud ?

Considérons l'exemple du diagramme : 

Ce diagramme **n'est pas** représenté par le polyèdre suivant :

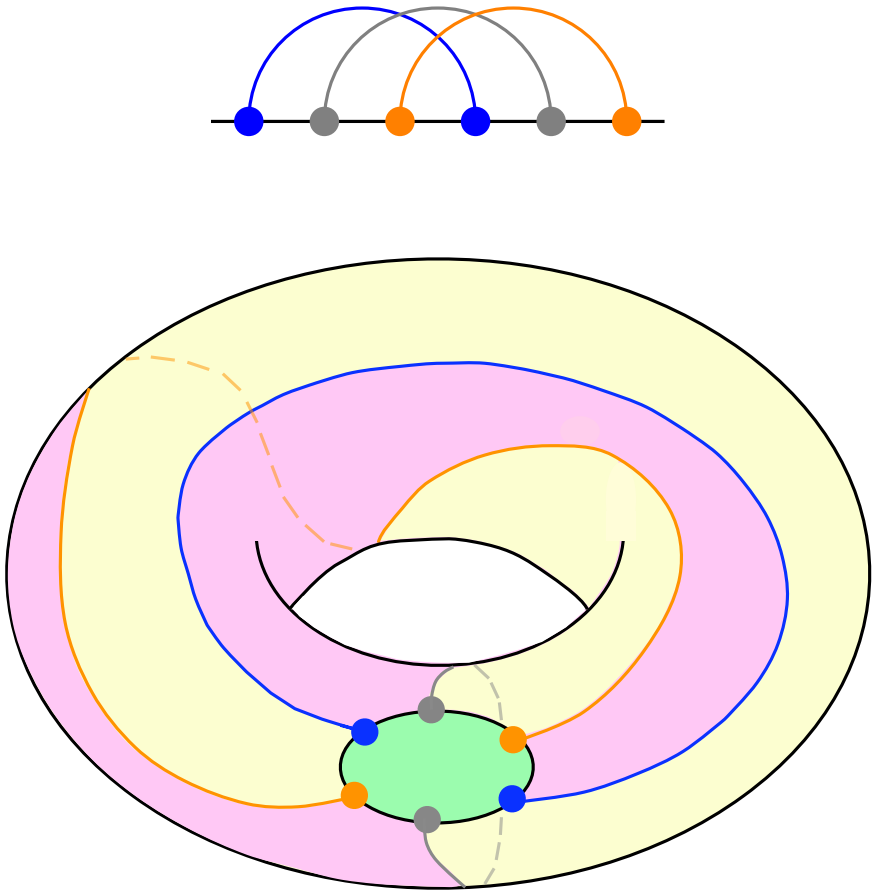


En effet, en dessinant dans le plan, il faut rajouter un sommet à chaque intersection (indiqués en rouge) pour que les faces soient bien définies. Ce polyèdre a ainsi une caractéristique d'Euler de 2 ($S = 8$, $E = 13$, $F = 7$) mais n'a pas la topologie demandée. Les intersections du diagramme peuvent être levées sans rajouter de sommet en dessinant le diagramme non plus dans un plan mais sur un tore qui est une variété comprenant une *poignée* et caractérisée par $\chi = 0$:



Sur le tore, ce polyèdre délimite 3 faces (en jaune, violet et vert), de sorte qu'on a bien $\chi = 6 - 9 + 3 = 0$.

Voici un autre exemple de diagramme se dessinant sans intersection sur un tore :



Ce polyèdre a également 3 faces et on retrouve $\chi = 0$.

3.2.1 Définition du genre

Le genre g d'un diagramme est le nombre minimum de poignées que doit contenir une variété pour que le diagramme puisse y être dessiné sans intersection.

Le genre se relie à la caractéristique d'Euler par :

$$\chi = 2 - 2g \tag{3.2}$$

Le tore est une variété à 1 poignée : les diagrammes qui s'y dessinent ont un genre de 1 et une caractéristique d'Euler de 0. Les structures planaires ont un genre nul et une caractéristique d'Euler de 2.

Ainsi, le genre est un nombre entier qui caractérise la topologie d'un diagramme. Intuitivement, le genre est une mesure de la *complexité* d'un repliement.

3.2.2 Calcul du genre

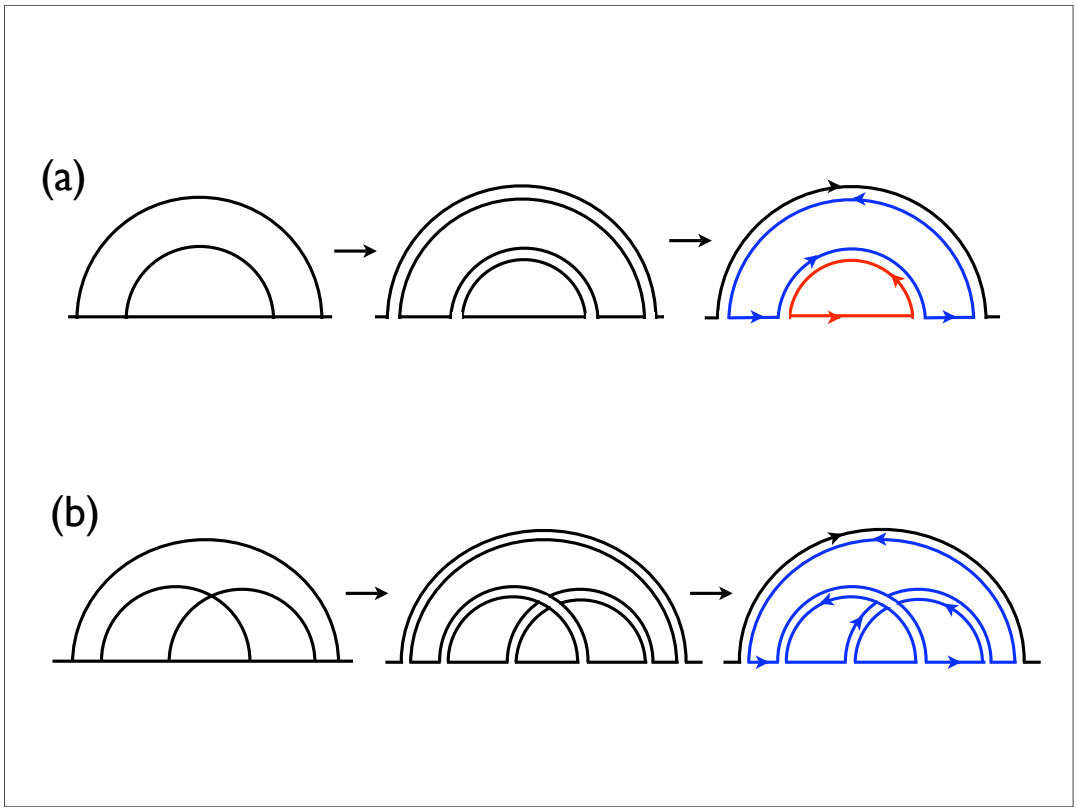
Comment calculer simplement le genre d'un diagramme ?

La combinaison de (3.1) et (3.2) donne :

$$g = \frac{A - S - (F - 2)}{2} \tag{3.3}$$

Il est assez simple de relier le nombre d'arcs P d'un diagramme (c'est-à-dire le nombre de paires de bases de la structure secondaire de l'ARN représenté) au nombre de sommets et d'arêtes du polyèdre correspondant : on a $S = 2P$ et $A = 3P$. Le nombre de faces est par contre moins évident à lire directement sur le diagramme. Plus précisément, c'est la quantité $(F - 2)$ qui est recherchée : si on considère que 2 faces sont créées "par défaut" lorsqu'on relie les extrémités 5' et 3' (l'intérieure et l'extérieure, représentées en vert et en jaune dans tous les exemples donnés précédemment), alors $(F - 2)$ représente le nombre de faces supplémentaires créées par les arcs du diagramme.

Il est possible de calculer $(F-2)$ en représentant les arcs du diagramme par une double ligne et non plus par une simple ligne [65] [66] :



Cette opération fait apparaître des boucles, représentées en bleu et en rouge dans les exemples précédents : *ces boucles sont les bordures des faces créées par le diagramme lorsque dessiné sur la variété adéquate*. Ainsi la quantité $(F-2)$ est égale au nombre B de boucles apparaissant dans cette nouvelle représentation et elle peut se calculer en $O(P)$ opérations. Au final, le genre se calcule simplement par :

$$g = \frac{P-B}{2} \tag{3.4}$$

Dans les deux exemples précédents, on a ainsi :

a) $P=2$ et $B=2$ (deux boucles, rouge et bleue) $\rightarrow g=0$, comme attendu pour une structure planaire

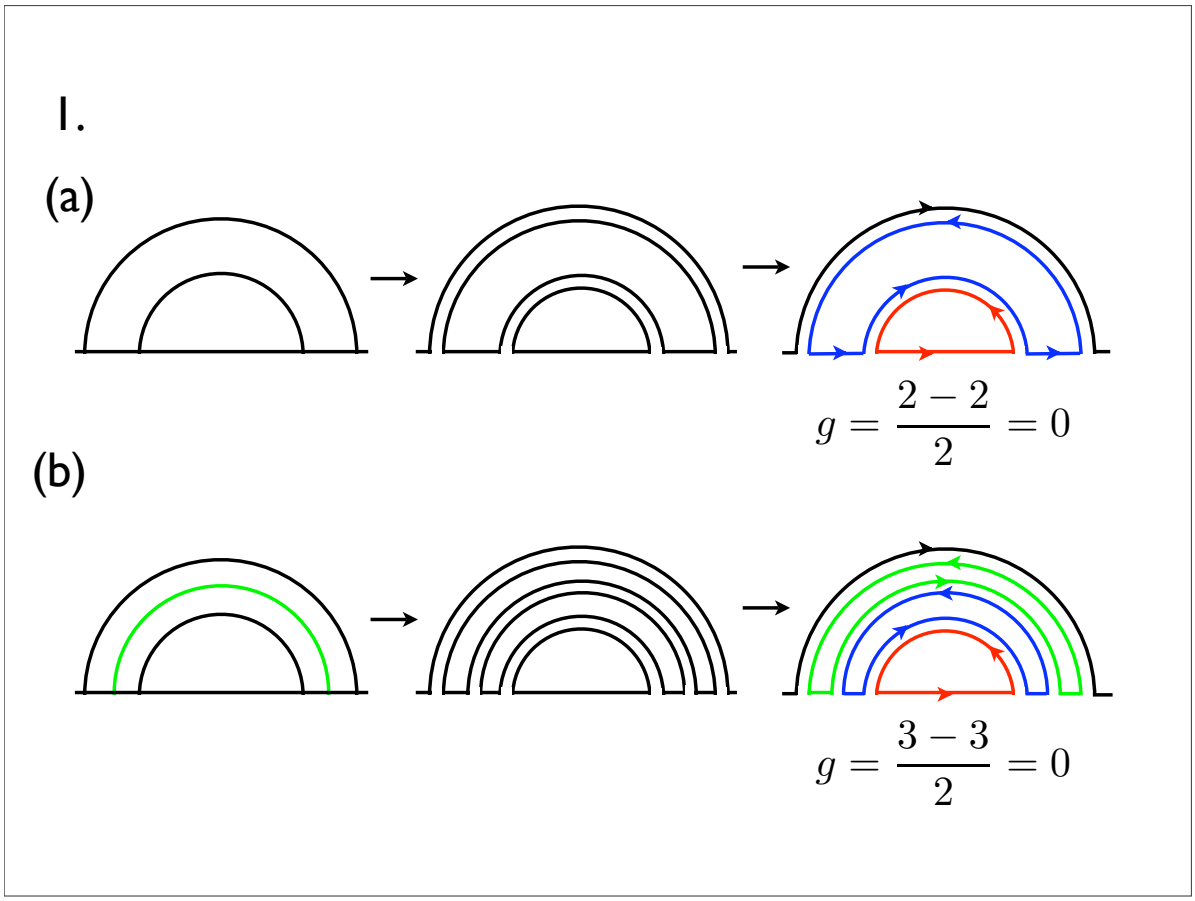
b) $P=3$ et $B=1 \rightarrow g=1$: ce diagramme se dessine donc sans croisement sur un tore, comme on peut le vérifier.

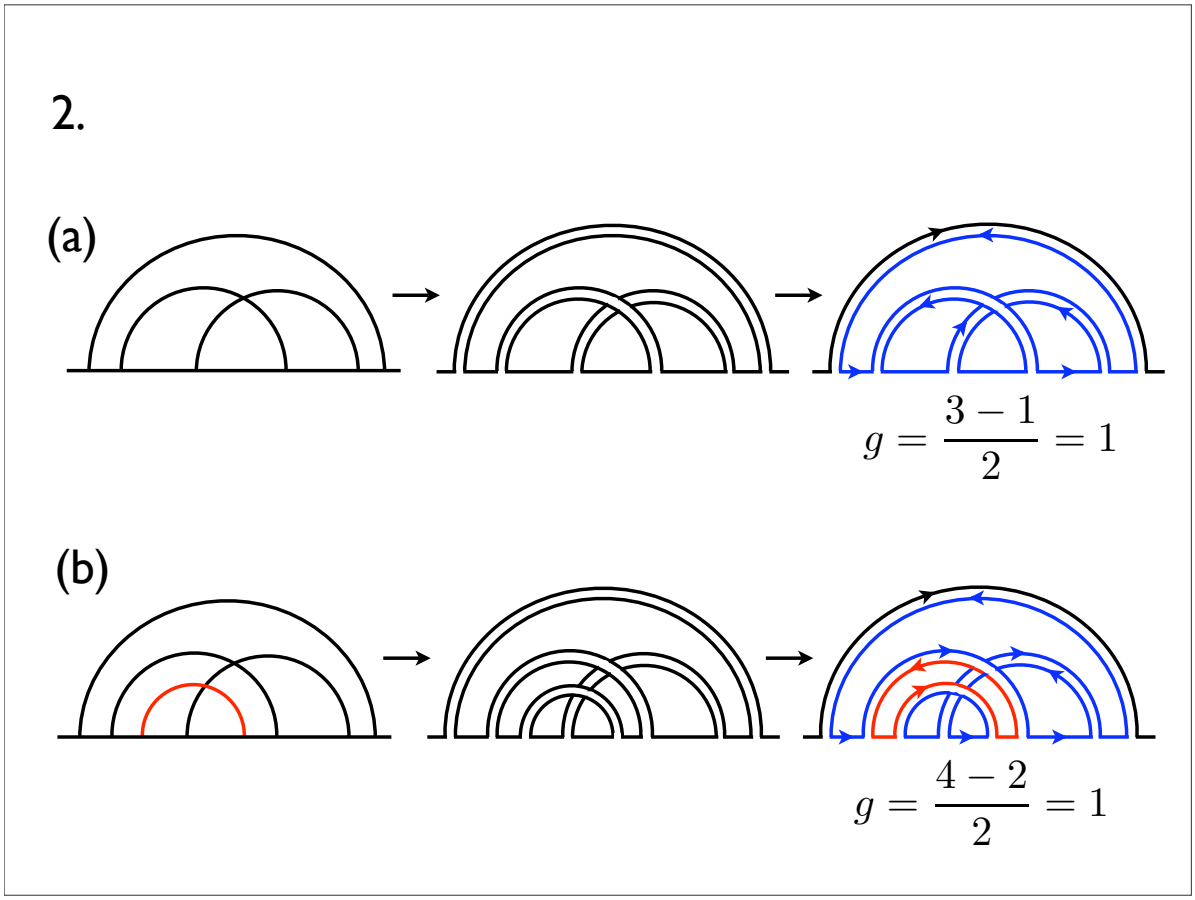
3.3 Propriétés du genre

Afin d'étudier l'intérêt que présente la connaissance du genre d'un diagramme pour la prédiction de pseudo-nœuds, il convient d'énoncer quelques-unes de ses propriétés.

3.3.1 Invariance du genre par ajout d'arcs parallèles les uns aux autres

L'ajout d'un arc A2 parallèlement à un autre arc A1 sur le diagramme ne modifie pas le genre. En effet, comme on peut le constater par exemple avec la représentation en double ligne, la création d'une telle paire s'accompagne toujours de la création d'une boucle impliquant A1 et A2, de sorte qu'avec (3.4) le genre ne varie pas. En voici deux illustrations :

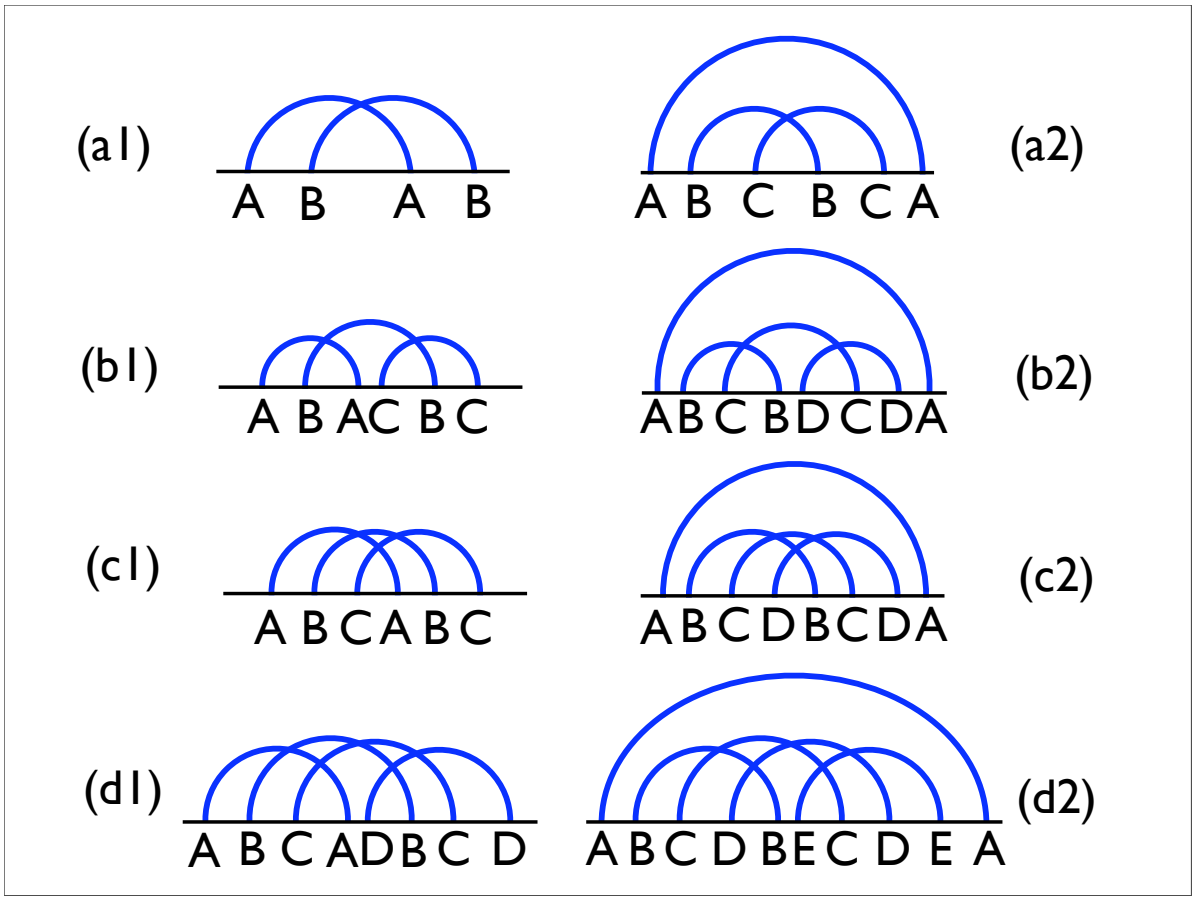




Ramenée à l'ARN, une succession d'arcs parallèles les uns aux autres représente une hélice. Ainsi, cette propriété du genre signifie qu'il *n'est pas modifié par la longueur des hélices* présentes dans un repliement. Topologiquement, les hélices s'identifient à une simple paire. En vertu de cette propriété, les hélices ne seront par la suite plus dessinées intégralement mais représentées par un unique arc :

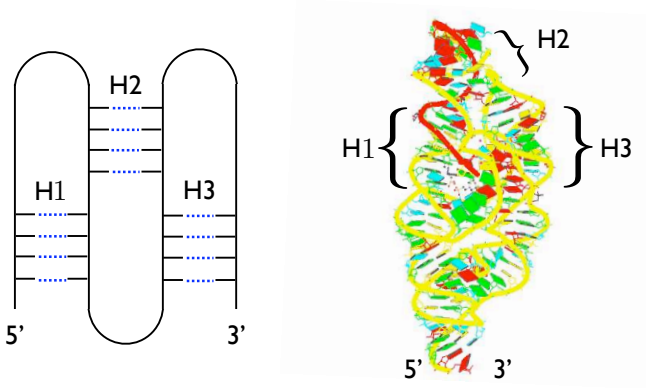
3.3.2 Topologies de genre 1

Si le genre mesure la complexité d'un repliement, alors les diagrammes de genre 1 sont les plus simples des diagrammes contenant un pseudo-nœud. Pillsbury *et al.* [67] ont montré qu'il n'existe que 8 topologies de diagrammes de genre 1, qui sont :

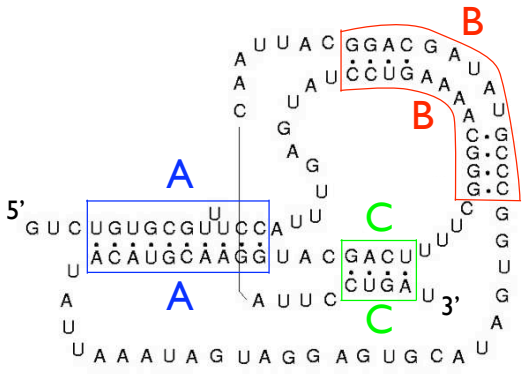


(a1) et (a2) représentent le type de pseudo-nœud le plus fréquemment observé dans la nature : le pseudo-nœud “H” (PNH), déjà illustré en introduction de ce chapitre.

(b1) et (b2) représentent un autre type de pseudo-nœud commun dans la nature, le “*kissing-hairpins*” (KH, par exemple [68]) :



(c1) et (c2) représentent un type de pseudo-nœud qui n’a pour l’instant été rapporté qu’une seule fois [69] :



(d1) et (d2) n’ont jamais été observés jusqu’à présent.

Ainsi, les pseudo-nœuds de genre 1 correspondent aux pseudo-nœuds les plus observés dans la nature. Pour étayer quantitativement cette affirmation, la PseudoBase [70], base de données de pseudo-nœuds, a été analysée. Les pseudo-nœuds répertoriés se répartissent ainsi :

- 271 sont des pseudo-nœuds H (ABAB ou ABCBCA)
- 6 sont des KH (ABACBC ou ABCBDCDA)
- 1 est de type ABCABC
- 1 est de genre 2, de type ABCDCADB

Ainsi, dans la PseudoBase, tous les pseudo-nœuds sauf un sont de genre 1.
La caractérisation par le genre est donc très efficace puisqu'elle permet d'énoncer un point commun entre la grande majorité des pseudo-nœuds répertoriés. Cette information sera très utile à un algorithme cherchant à les prédire.

3.3.3 Propriétés statistiques du genre

Vernizzi *et al.* ont étudié la distribution du genre pour des modèles d'homopolymères, c'est à dire pour un ARN composé d'un seul type de base pouvant former des paires [71]. A longueur L fixée, le nombre de diagrammes de genre g est asymptotiquement :

$$n_{L,g} \approx \kappa_g 3^L L^{3g-\frac{3}{2}} \text{ avec } \kappa_g = \frac{1}{3^{4g-\frac{3}{2}} 2^{2g+1} g! \sqrt{\pi}} \tag{3.5}$$

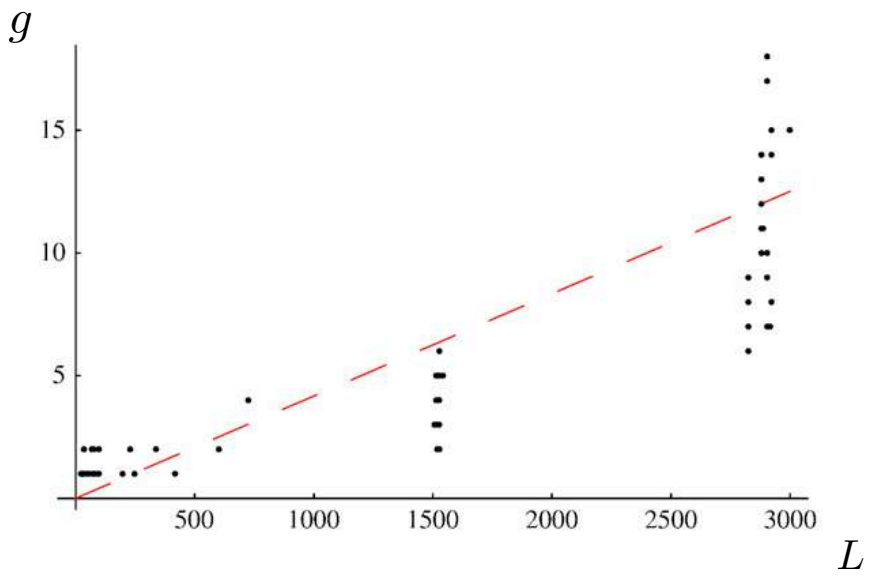
Ceci permet de calculer le genre moyen d'un homopolymère de longueur L :

$$\langle g \rangle = 0.24L \tag{3.6}$$

Ceci signifie que le genre d'une structure aléatoire typique est proche de $L/4$, sa valeur maximale (vérifiable avec (3.4), P étant majoré par $L/2$).

L'analyse de la PseudoBase a montré que les genres observés dans la nature sont bien éloignés de cette asymptote, puisque le genre maximal est de 2, alors qu'il correspond à une séquence de 87 bases de long. Il y a cependant un biais : la PseudoBase ne contient que des pseudo-nœuds étudiés expérimentalement (ou déduits des premiers par comparaison de séquences) car ils ont un intérêt fonctionnel. D'autres pseudo-nœuds qui n'ont pas d'intérêt fonctionnel particulier, qui ne sont "que" des éléments de structure au même titre que les boucles internes et têtes d'épingle, ont pu ne pas être rapportés. Les structures de la PDB (*Protein Data Bank*) contenant un pseudo-nœud ont donc également été analysées : leur structure secondaire a été extraite avec le logiciel RNAview [52] et leur genre a été calculé en ne conservant que les paires Watson-Crick et Wobble.

Voici la distribution des genres obtenus, en fonction de la longueur des séquences :



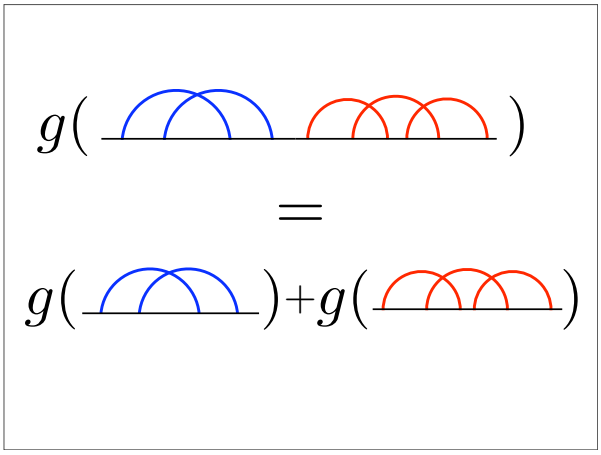
Ainsi donc, le genre le plus élevé observé dans la PDB vaut 18. Les structures de moins de 500 bases de long ont un genre inférieur ou égal à 2, ce qui confirme bien la tendance constatée dans la PseudoBase : les genres des structures réelles sont bien moindres que les genres de diagrammes aléatoires. On a approximativement cette relation linéaire :

$$g(L) \approx \frac{L}{250} \tag{3.7}$$

Cette étude a été réalisée en automatisant l'analyse des fichiers .pdb disponibles sur la PDB. La structure secondaire en a été extraite avec le logiciel RNAview et le genre a été calculé en ne gardant que les liaisons Watson-Crick et Wobble. Dans cette procédure, toutes les paires G–C, A–U et G–U annotées par RNAview ont été conservées. Il n'est pas rare que beaucoup de pseudo-nœuds pris en compte dans cette étude ne soient en fait créés que par une seule paire, ce qui n'est pas pertinent d'un point de vue topologique. En effet, ces paires ne dirigent pas le repliement et ne sont formées qu'"opportunément" pour stabiliser la structure. En éliminant ces paires de l'analyse, il est clair que les genres des structures étudiées seront encore plus faibles que ceux rapportés ici.

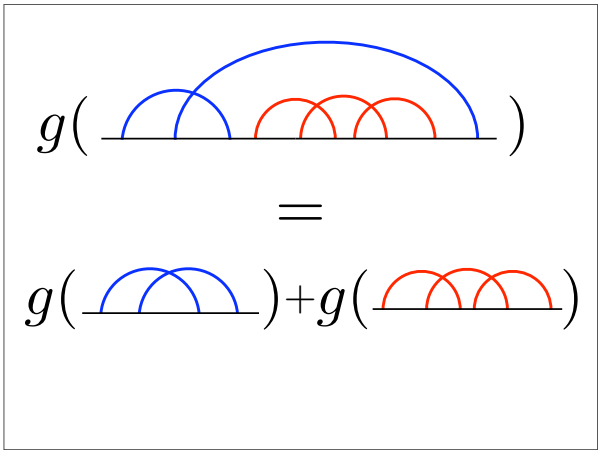
3.3.4 Propriétés d'additivité du genre

Le genre a des propriétés d'additivité permettant de mieux caractériser la complexité intrinsèque d'un pseudo-nœud. Le genre d'un diagramme composé de deux pseudo-nœuds successifs et disjoints est la somme des genres de ces deux pseudo-nœuds.

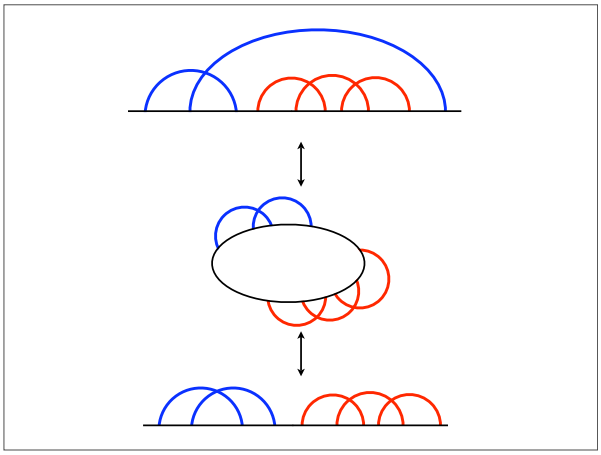


Cette propriété se démontre simplement en constatant que les nombres de paires et de boucles ont bien cette même propriété d'additivité et en concluant avec (3.4).
Des pseudo-nœuds agencés consécutivement sur un même diagramme sont dits "*en série*". Un diagramme ne comportant pas de pseudo-nœuds en série est dit "*irréductible*".

Cette propriété d'additivité s'étend aux diagrammes constitués de pseudo-nœuds disjoints imbriqués les uns dans les autres, comme illustré dans l'exemple suivant :

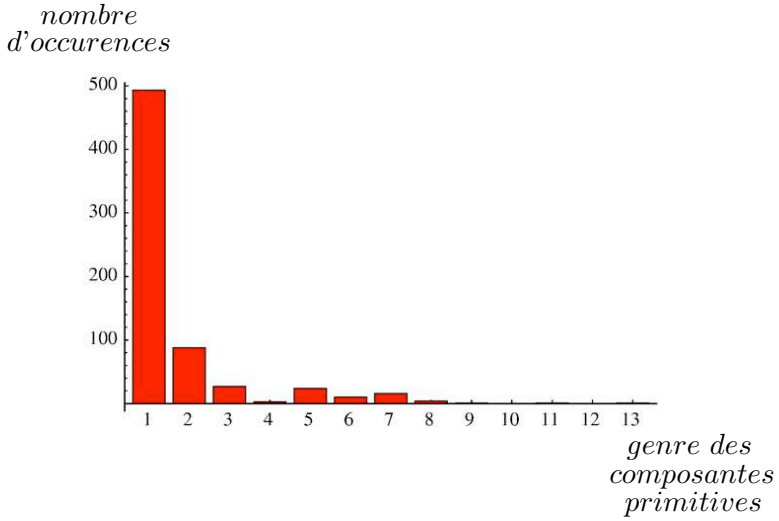


En effet, cela se démontre en constatant que ces diagrammes ont exactement la même topologie que les diagrammes "en série" dans la représentation circulaire sur laquelle s'appuie le genre :



Des pseudo-nœuds disjoints imbriqués les uns dans les autres sont dits "*en parallèle*". Un diagramme ne comportant ni pseudo-nœuds en série ni pseudo-nœuds en parallèle est dit "*primitif*".
Les pseudo-nœuds primitifs sont ainsi les briques élémentaires ("*building-blocks*") à partir desquelles peut être engendré l'ensemble des pseudo-nœuds par concaténation ou imbrication.

Afin d'estimer la complexité intrinsèque des pseudo-nœuds naturels, l'analyse des pseudo-nœuds de la PseudoBase et de la PDB a été reprise en décomposant les pseudo-nœuds en leurs composantes primitives.



Le pseudo-nœud primitif le plus compliqué observé n'a plus qu'un genre de 13. La plupart des pseudo-nœuds de genre supérieur à 2 sont en fait des composés de pseudo-nœuds primitifs plus simples. Une analyse plus poussée montre que de tels pseudo-nœuds sont organisés simplement. Des pseudo-nœuds primitifs de genre faible sont imbriqués dans un unique pseudo-nœud primitif plus compliqué. Il n'y pas plusieurs niveaux d'imbrication où un pseudo-nœud A serait imbriqué dans un pseudo-nœud B, lui-même imbriqué dans un pseudo-nœud C.

3.4 Bilan et intérêt du genre pour la prédiction de pseudo-nœuds dans les structures secondaires d'ARN

Le genre est un nombre entier qui peut être interprété comme une mesure de la complexité d'un repliement. Il est de nature topologique, ce qui signifie en particulier qu'il ne dépend pas de la longueur des hélices mises en jeu dans la formation du pseudo-nœud. Ainsi, différents pseudo-nœuds peuvent être comparés entre eux et il apparaît alors que leurs topologies présentent des similarités frappantes bien caractérisées par le genre. Les pseudo-nœuds les plus courants tombent dans une même catégorie qui est celle des pseudo-nœuds de genre 1. Les pseudo-nœuds de genre supérieur peuvent en général se construire à partir de topologies primitives plus simples, par concaténation et imbrication. En particulier, le pseudo-nœud primitif le plus compliqué observé parmi les séquences de moins de 500 bases de long a un genre valant 2.

Le genre apporte donc une information extrêmement utile à la prédiction de pseudo-nœuds. En effet, la difficulté principale de la prédiction de pseudo-nœuds vient du fait que leur nombre est beaucoup trop élevé même pour des séquences courtes. Savoir que le genre des structures réelles est faible permet donc de considérablement réduire l'espace de recherche. En pratique, le genre peut être intégré au modèle d'énergie comme une *pénalité topologique*. L'énergie libre d'une structure S se calculerait ainsi comme :

$$\Delta F'(S) = \Delta F(S) + f(g(S)) \tag{3.8}$$

où $\Delta F(S)$ est l'énergie libre calculée selon un des modèles présentés dans le chapitre précédent et où f est une fonction croissante. Le rôle de cette pénalité f est de reproduire la distribution du genre observée dans les bases de données, par exemple en pénalisant lourdement les topologies de genre élevé qui ne sont jamais observées. *A priori*, ce terme de pénalité topologique n'a pas vocation à caractériser une contribution énergétique réelle et mesurable à diagramme donné auquel cas on pourrait se demander pourquoi les pseudo-nœuds de type ABAB (PNH), ABACBC (KH) et ABCABC ont la même pénalité alors que leurs topologies sont manifestement différentes. f se contente de filtrer les structures vraisemblables et c'est le reste du modèle d'énergie qui permettra de déterminer laquelle est la meilleure.

Dans le chapitre suivant,"Prédiction de pseudo-nœuds", la forme utilisée pour f sera simplement :

$$f(g) = \mu g \quad \text{avec} \quad \mu > 0 \tag{3.9}$$

Le coefficient μ est un "potentiel chimique topologique". Dans l'article [72], il a été montré qu'une augmentation de la concentration en ions magnésium a pour effet de défaire les liaisons formant la structure tertiaire parmi lesquelles sont comptabilisées les hélices des pseudo-nœuds. Le genre d'une structure diminue quand $[Mg^{2+}]$ croît. Le paramètre μ peut ainsi être physiquement interprété comme une fonction de $[Mg^{2+}]$ permettant d'en imiter l'influence dans le calcul de l'énergie libre d'un repliement donné. Une forte concentration correspond à une grande valeur de μ, rendant plus difficile la formation de pseudo-nœuds.

L'étude du genre développée dans ce chapitre a été également présentée dans l'article "*Topological classification of RNA structures*" [73].

Chapitre 4

Algorithmes de repliement

Dans ce chapitre, je présente le principe d'algorithmes de prédiction de structures secondaires avec et sans pseudo-nœud.

Les idées et la structure générale des algorithmes de prédiction sans pseudo-nœud seront exposées. Le principe général en est connu depuis [74] et [75] et a fait l'objet de raffinements successifs pour pouvoir prendre en charge un modèle d'énergie de plus en plus riche. Plusieurs implémentations en sont actuellement disponibles ([76], [77], [78]). La principale difficulté rencontrée par ces algorithmes est la prise en charge des composantes non-linéaires du modèle d'énergie comme l'entropie des boucles multi-hélices (logarithmique) ou l'asymétrie des boucles internes pour laquelle un exemple d'approximation sera exposé.

Après un bref exposé de l'état de l'art, un nouvel algorithme de prédiction de structures secondaires avec pseudo-nœuds utilisant le paramétrage topologique présenté dans le chapitre précédent sera détaillé. Ses résultats seront évalués et comparés à ceux d'autres algorithmes, concluant à une significative amélioration de la qualité de prédiction.

Le plan suivi dans ce chapitre est :

I) Algorithme de prédiction de structures secondaires sans pseudo-nœud
   1. Principe général : la récurrence
   2. Un exemple d'implémentation avec un modèle d'énergie libre détaillé
   3. Une approximation pour traiter l'asymétrie des boucle internes

II) Algorithme de prédiction de structures secondaires avec pseudo-nœuds
   1. Etat de l'art
   2. Réflexions préliminaires
   3. Un nouvel algorithme : TT2NE
   4. Résultats
   5. Discussion

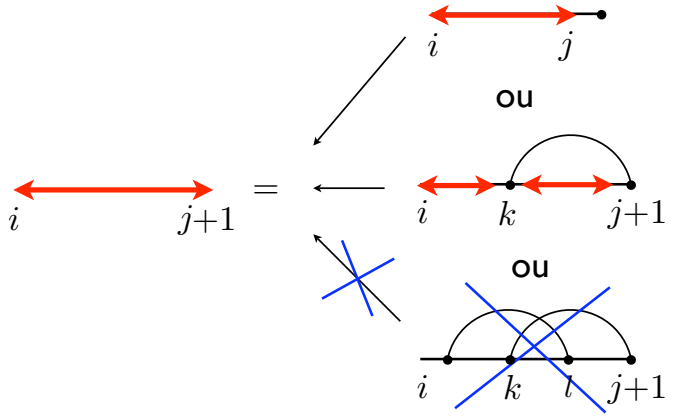
4.1 Algorithmes de repliement sans pseudo-nœud

4.1.1 Principe général : la récurrence

Le nombre de structures secondaires sans pseudo-nœud constructibles à partir d'une séquence de longueur L varie comme $3^L L^{-\frac{3}{2}}$ [71]. Il est ainsi hors de propos de rechercher la meilleure structure secondaire par l'énumération complète des possibilités.
Les algorithmes de repliement tirent parti de la linéarité du modèle d'énergie libre qui permet d'exprimer l'énergie libre d'un repliement comme la somme d'énergies libres de ses sous-structures. La fonction de partition peut ainsi être efficacement calculée à l'aide de relations de récurrence.

Soit une séquence x de longueur L dont on numérote les bases de 1 à L. Soit $\mathcal{Z} = \mathcal{Z}(1, L)$ sa fonction de partition. Soit $\mathcal{Z}(i, j)$ la fonction de partition associée au segment [i,j] de la séquence *ie* $\mathcal{Z}(i, j)$ est la fonction de partition de la séquence de longueur j-i+1 correspondant à la sous-séquence [i,j] de x. Supposons pour l'instant un modèle d'énergie minimal où l'énergie d'une structure secondaire serait simplement la somme d'énergies ϵ_{kl} portant sur les seules paires de bases {(k,l)} la constituant.
On peut alors écrire la relation suivante :

$$\mathcal{Z}(i, j+1) = \mathcal{Z}(i, j) + \sum_{k=i}^{j} \mathcal{Z}(i, k-1) e^{-\beta \epsilon_{k,j+1}} \mathcal{Z}(k+1, j)$$



Cette relation de récurrence permet de sommer toutes les structures secondaires sans pseudo-nœud avec leur bon poids de Boltzmann. Chaque $\mathcal{Z}(i,j)$ se calculant en $O(L)$ opérations, la fonction de partition totale requiert, au final, $O(L^3)$ opérations. Cette complexité est très satisfaisante et permet l'étude des séquences d'ARN les plus longues connues à ce jour. Cette relation de récurrence peut être adaptée pour rechercher la structure d'énergie libre minimale plutôt que pour calculer la fonction de partition. En appelant $\Delta F[i,j]$ l'énergie libre du repliement d'énergie libre minimale sur [i,j], on a :

$$\Delta F[i,j+1] = min\left\{\Delta F[i,j]\ ,\ \min_k\left(\Delta F[i,k-1] + \epsilon_{k,j+1} + \Delta F[k+1,j]\right)\right\} \qquad (4.1)$$

et $\Delta F[1,L]$ se calcule également en $O(L^3)$.

Cette idée peut se développer pour prendre en charge le modèle d'énergie linéaire désiré. Parmi les variantes disponibles, je présente une structure générale du programme s'inspirant de "*An improved algorithm for RNA secondary structure prediction*" [64].

4.1.2 Un exemple d'implémentation avec un modèle d'énergie détaillé

5 tableaux sont introduits, notés V, VBI, VM, WM et W :

- V est un tableau $L \times L$ tel que $V(i,j)$ soit l'énergie du meilleur repliement sur la sous-séquence [i,j] où i et j sont appariés.

- VBI est un tableau $L \times L \times L$ tel que $VBI(i,j,l)$ soit l'énergie du meilleur repliement sur la sous-séquence [i,j] où i et j sont appariés et forment une boucle interne ou un renflement de longueur l.

- VM est un tableau $L \times L$ tel que $VM(i,j)$ soit l'énergie du meilleur repliement sur la sous-séquence [i,j] où i et j sont appariés et forment une boucle multi-hélices.

- WM est un tableau $L \times L$ utilisé comme intermédiaire dans les relations de récurrence portant sur les boucles multi-hélices. Il représente la meilleure juxtaposition d'hélices entre i et j avant que la boucle mutli-hélices ne soit créée par la formation d'une paire de bases (i',j') avec $i' < i < j < j'$.

- W est un tableau de taille L utilisé pour la récurrence finale.

Ces tableaux sont remplis à l'aide de différentes relations de récurrence qui les relient les uns aux autres. La valeur d'un de ces tableaux en (i, j) s'exprime en fonction des valeurs en (i', j') des autres tableaux, avec $i \leq i' < j' \leq j$. Ainsi la récurrence porte sur la longueur $|j - i|$. Les tableaux sont initialisés en $(i, i + 1)$ pour tout i. A partir de ces conditions initiales, les valeurs en $(i, i + 2)$ sont calculées pour tout i, puis $(i, i + 3)$ et ainsi de suite. La récurrence prend fin quand la valeur de tous les tableaux en $(1, L)$ a été calculée.

V

La formation d'une paire (i, j) peut donner lieu à la création d'une tête d'épingle, d'un dipaire dans le cas où la paire $(i+1, j-1)$ serait déjà présente, d'une boucle interne ou enfin d'une boucle multi-hélices. Ceci est exprimé ainsi :

$$V(i,j) = \min \begin{cases} \Delta F_{t.e.}(i,j) \\ V(i+1,j-1) + \Delta F_d \begin{pmatrix} i+1 & - & j-1 \\ i & - & j \end{pmatrix} \\ \min\limits_{l} \{VBI(i,j,l)\} \\ VM(i,j) \end{cases} \tag{4.2}$$

où $\Delta F_{t.e.}(i, j)$ désigne le coût de création de la tête d'épingle fermée par la paire (i, j). Si $|i - j| < 4$, $\Delta F_{t.e}(i, j)$ vaut $+\infty$ car, du fait de la rigidité de l'ARN, une tête d'épingle doit comporter au moins 4 bases.
Le calcul de V(i,j) s'effectue en $O(L)$ (à cause de la troisième ligne) : le tableau V se remplit donc en $O(L^3)$.

W

Une fois calculés les $V(i, j)$, $W(k)$ construit la meilleure juxtaposition de ces sous-structures $V(i, j)$ pour $i < j \leq k$, avec la relation :

$$W(k) = \min\{W(k-1), \min_{1<i\leq k}\{W(i-1) + V(i,k)\}\} \tag{4.3}$$

Ainsi, l'énergie de la meilleure structure est donnée par $W(L)$ qui s'obtient en $O(L^2)$ opérations.

VM et WM

Les relations engendrant les boucles multi-hélices dépendent de la forme de la fonction d'énergie libre associée. En notant l la longueur d'une boucle k-hélices, une façon de faire a été trouvée dans [18] en supposant une pénalité linéaire, de la forme :

$$\Delta F_{bmh}(k,l) = a + b \times k + c \times l \tag{4.4}$$

Les relations utilisées sont les suivantes :

$$WM(i,j) = \min \begin{cases} V(i,j) + b \\ WM(i,j-1) + c \\ WM(i+1,j) + c \\ \min\limits_{l}\{WM(i,k-1) + WM(k,j)\} \end{cases} \tag{4.5}$$

et :

$$VM(i,j) = \Delta F(i,j) + WM(i+1,j-1) + a \tag{4.6}$$

où $\Delta F(i,j)$ désigne le coût de formation de la paire isolée (i,j).

WM construit des hélices affectées d'une pénalité b (ligne 1) et les juxtapose (ligne 4) en affectant une pénalité c à toutes les bases libres entre ces hélices (ligne 2 et 3). VM(i,j) construit la boucle mutli-hélices en formant une paire (i,j) autour de la structure représentée par $WM(i+1,j-1)$et en ajoutant la pénalité a de création de boucle multi-hélices. $WM(i,j)$ se calcule en $O(L)$ et $VM(i,j)$ en $O(1)$. (Note : le tableau VM est en fait facultatif car l'équation (4.6) peut directement être écrite dans (4.2).)

VBI

Les boucles internes d'énergie minimale sont calculées comme suit :

$$VBI(i,j,l) = \min \begin{cases} VBI(i+1,j-1,l-2) + \Delta F(i,j) - \Delta F(i+1,j-1) + \delta \\ V(i+1,j-l-1) + \Delta F_r\{(i,j);(i+1,j-l-1)\} \\ V(i+l+1,j-1) + \Delta F_r\{(i,j);(i+l+1,j-1)\} \end{cases} \tag{4.7}$$

Les deuxième et troisième lignes construisent les deux renflements de taille l fermés par la paire (i,j), délimités respectivement par les paires $(i+1,j-l-1)$ et $(i+l+1,j-1)$ et dont l'énergie libre est notée ΔF_r. La première ligne stipule que la meilleure boucle interne de longueur l se déduit de la meilleure boucle interne $VBI(i+1,j-1,l-2)$ en ôtant la paire $(i+1,j-1)$ et en ajoutant la paire (i,j) ainsi qu'une correction δ liée à la forme utilisée pour l'entropie des boucles internes.

Cette relation est rigoureusement exacte avec le modèle utilisé dans l'article [64], pour lequel une boucle interne délimitée par les paires (i,j) et (i',j') avec $i < i' < j' < j$ a une énergie de la forme :

$$\begin{aligned}\Delta F_{bi}\{(i,j);(i',j')\} = \Delta F(i,j) + \Delta F(i',j') \\ + a_{bi} + b_{bi} \times [(i'-i)+(j-j')] + c_{bi} \times |(i'-i)-(j-j')| \quad (4.8)\end{aligned}$$

a_{bi} est un terme d'initiation, b_{bi} est une pénalité sanctionnant la longueur $[(i'-i)+(j-j')]$ de la boucle interne et c_{bi} une pénalité sanctionnant son asymétrie $|(i'-i)-(j-j')|$. Avec (4.8), la validité de la première ligne de (4.7) se démontre par :

$$\begin{aligned}VBI(i,j,l) &= \min_k \{\Delta F_{bi}\{(i,j);(i+l-k,j-k)\}\} \\ &= \min_k\{\Delta F(i,j) + \Delta F(i+l-k,j-k) \\ &\qquad +a_{bi} + b_{bi} \times l + c_{bi} \times |l-2k|\} \quad (4.9)\end{aligned}$$

$$\begin{aligned}&= \Delta F(i,j) - \Delta F(i+1,j-1) + 2b_{bi} \\ &\quad + \min_k\{\Delta F(i+1,j-1) + \Delta F(i+l-k,j+k) \\ &\qquad +a_{bi} + b_{bi} \times (l-2) + c_{bi} \times |l-2k|\}\end{aligned}$$

$$\begin{aligned}&= \Delta F(i,j) - \Delta F(i+1,j-1) + 2b_{bi} \\ &\qquad +VBI(i+1,j-1,l-2) \quad (4.10)\end{aligned}$$

La forme requise dans (4.7) est obtenue avec $\delta = 2b_{bi}$. Un ressort essentiel de cette démonstration est la linéarité de la pénalité d'asymétrie qui assure que la boucle interne délimitée par $\{(i,j),(i',j')\}$ a la même pénalité d'asymétrie que $\{(i+1,j-1),(i',j')\}$.

Ainsi $VBI(i,j,l)$ peut se calculer en $O(1)$ avec (4.7) et le tableau VBI se remplit en $O(L^3)$. L'article [64] présente également une astuce élégante permettant d'économiser les L^3 unités de mémoire nécessaires au stockage de VBI.

Complexité de l'algorithme

Tous les tableaux se remplissent par au plus $O(L^3)$ opérations : cet algorithme a donc une complexité de $O(L^3)$. L'énergie libre minimale est donnée par $W(L)$.

4.1.3 Une approximation pour une autre fonction d'asymétrie des boucles internes

La fonction d'asymétrie linéaire ci-dessus n'est sans doute pas physiquement réaliste et pose un problème de cohérence interne au modèle. Par exemple, dans Turner99, l'entropie des renflements dépend logarithmiquement de leur longueur et non linéairement. Ainsi, l'énergie d'une boucle interne $1 \times L$ et celle d'un renflement de taille L divergent lorsque L croît. Or on s'attend plutôt à une convergence étant donné que le renflement n'est qu'une boucle interne particulière, $0 \times L$. Il convient donc de chercher de nouvelles manières de traiter l'asymétrie des boucle internes. Supposons par exemple cette forme :

$$\Delta F_{bi}\{(i,j);(i',j')\} = \Delta F(i,j) + \Delta F(i',j') + a_{bi} + c_{bi} \times \frac{|(i'-i)-(j-j')|}{(i'-i)+(j-j')} \quad (4.11)$$

Avec cette forme, la pénalité d'asymétrie est bornée par c_{bi} et il est alors possible de mettre a_{bi} et c_{bi} en relation avec les coefficients paramétrant les renflements pour assurer plus de cohérence interne au modèle. Reprenons la démonstration (4.10) :

$$\begin{aligned}
VBI(i,j,l) &= \min_k\{\Delta F(i,j) + \Delta F(i+l-k,j+k) + a_{bi} + c_{bi}\frac{|l-2k|}{l}\} \\
&= \Delta F(i,j) - \Delta F(i+1,j-1) \\
&\quad + \min_k\{\Delta F(i+1,j-1) + \Delta F(i+l-k,j+k) \\
&\quad\quad + a_{bi} + c_{bi}\frac{|l-2k|}{l-2} + \delta(k)\} \\
&= \Delta F(i,j) - \Delta F(i+1,j-1) \\
&\quad + \min_k\{\Delta F_{bi}\left(\{i+1,j-1\};\{i+l-k,j+k\}\right) + \delta(k)\} \qquad (4.12)
\end{aligned}$$

avec :

$$\begin{aligned}
\delta(k) &= c_{bi}\frac{|l-2k|}{l} - c_{bi}\frac{|l-2k|}{l-2} \\
&= -2c_{bi}\frac{|l-2k|}{l(l-2)}
\end{aligned}$$

Ainsi, le résidu δ dépend de k et la relation (4.7) ne peut plus être utilisée pour déduire $VBI(i,j,l)$ de $VBI(i+1,j-1,l-2)$ en temps constant. En utilisant (4.9), le calcul de $VBI(i,j,l)$ requiert $O(l)$ opérations de sorte que le tableau VBI se remplit en $O(L^4)$, rendant l'algorithme trop long à utiliser sur des séquences de quelques centaines de bases. Cependant, en remarquant que $\delta(k)$ varie en l^{-1} et devient négligeable lorsque l est grand, il est justifié de calculer $VBI(i,j,l)$ en distinguant ces deux cas :

1. Concernant les boucles internes de longueur au plus l_{max}, $VBI(i,j,l)$ se détermine en examinant exhaustivement les l boucles internes en compétition.
2. Pour les boucles internes de taille supérieure, on peut utiliser l'approximation suivante : garder en mémoire les l_{max} meilleures boucles internes fermées par (i,j) de taille $l-2$ et déterminer $VBI(i,j,l)$ selon (4.12) à partir de ces l_{max} candidates.

Le calcul de $VBI(i,j,l)$ se fait ainsi en au plus l_{max} opérations dans chacun de ces cas. Avec $l_{max} = 10$, cette approximation est très bonne. Le coût en mémoire de $VBI(i,j,l)$ est par contre augmenté et devient $O(l_{max}L^3)$. J'ai utilisé cette implémentation du programme pour les travaux relatés dans le chapitre "Modèle d'énergie libre" nécessitant l'emploi d'un algorithme de repliement sans pseudo-nœud.

Dans cette dernière section, il a été montré comment une pénalité d'asymétrie non linéaire peut-être gérée par un algorithme de repliement sans pseudo-nœud tout en gardant une complexité de $O(L^3)$. L'entropie des boucles multi-hélices est une autre composante non linéaire du modèle d'énergie mais à l'heure actuelle aucune solution n'a été trouvée pour la prendre en charge en $O(L^3)$.
Einert *et al.* [79] proposent une nouvelle relation de récurrence en $O(L^4)$ supportant n'importe quelle forme pour l'entropie des boucles et montrent que de significatives différences sont observées selon que la modélisation de l'entropie des boucles multi-hélices soit linéaire ou logarithmique. La théorie des polymères prédit une dépendance logarithmique de la forme $\Delta S = cRT \ln l$ avec $c = 1.75$ pour des polymères auto-évitants. Les auteurs montrent en particulier que cette valeur théorique de c ne peut s'appliquer directement à l'ARN et que certaines données expérimentales sont mieux reproduites avec $c = 2.3$. Ces travaux récents doivent être testés plus amplement pour savoir si le gain qu'amène une meilleure modélisation de l'entropie des boucles compense la perte d'un ordre de grandeur dans la complexité de l'algorithme.

4.2 Un algorithme de repliement avec pseudo-nœuds

La prédiction de structures secondaires d'énergie libre minimale pouvant comporter des pseudo-nœuds est un problème NP-complet [14]. Ceci signifie que sa résolution requiert inévitablement des calculs dont la complexité augmente exponentiellement avec la longueur de la séquence étudiée. Ainsi, tout algorithme s'y essayant aura à choisir entre :

- ◇ effectuer des calculs exacts, ce qui entraînera forcément une *limite à la longueur des séquences* que l'algorithme peut prendre en charge en un temps raisonnable
- ◇ effectuer des approximations qui repoussent la limite de longueur mais *perdent la garantie* que la structure prédite corresponde effectivement au minimum d'énergie libre
- ◇ effectuer des calculs exacts en se *restreignant à certaines classes de pseudo-nœuds* pour lesquelles le problème n'est plus NP-complet

4.2.1 Bref état de l'art

Plusieurs algorithmes se répartissant dans ces différentes catégories sont disponibles actuellement. Voici une brève description des principaux :

- Gultyaev *et al.* [80] (1995, *STAR*) : les auteurs utilisent un algorithme génétique qui ajoute et retire des paires de bases avec une certaine probabilité fonction de leur énergie et qui intègre un mécanisme de "fusion" de structures candidates.

- Rivas & Eddy [81] (1999) : cet article présente le premier algorithme capable de prédire des pseudo-nœuds par minimisation de l'énergie libre en un temps polynomial. Les auteurs présentent des relations de récurrence contenant un mécanisme de génération de pseudo-nœuds. Ce mécanisme engendre une certaine classe de pseudo-nœuds difficile à caractériser *a priori* mais qui contient au moins les trois topologies de genre 1 observées dans la nature. Cet algorithme a une complexité polynomiale en $O(L^6)$ ce qui, en pratique, rend son usage difficile : il faut compter plusieurs heures pour replier une séquence de 100 bases de long.

- Dirks & Pierce [82] (2003, *NUPACK*) : les auteurs se restreignent à une classe de pseudo-nœuds uniquement composés de pseudo-nœuds H agencés en série et en parallèle et sont capables d'écrire des relations de récurrence l'engendrant en $O(L^5)$.

- Reeder & Giegerich [83] (2004, *PKnots-RG*) : les auteurs présentent un algorithme en $O(L^4)$ décrivant la même classe de pseudo-nœuds que l'article précédent mais où des contraintes supplémentaires pèsent sur les hélices formant des pseudo-nœuds.

- Isambert & Siggia [84] (2000) : les auteurs utilisent une simulation Monte-Carlo 3D incorporant des considérations cinétiques et thermodynamiques. Cette simulation peut prédire des états métastables et converge vers la distribution boltzmannienne mais il n'a pas été montré à quelle vitesse.

- Ruan & *et al.* [85] (2004, *ILM*) : Les auteurs présentent un algorithme glouton qui construit itérativement une structure secondaire S ainsi :
  1. identifier la meilleure hélice H de la séquence et l'ajouter à S.
  2. supprimer de la séquence toutes les bases de cette hélice et reprendre en (1) jusqu'à ce que S soit saturée.

- Ren & *et al.* [86] (2005, *HotKnots*) : les auteurs présentent un algorithme heuristique qui construit des structures secondaires par ajouts d'hélices successifs. L'algorithme organise sa recherche sous forme d'arbre dont les nœuds représentent des structures secondaires "en construction", non saturées. L'algorithme choisit les nœuds "fils" à explorer selon un critère fondé sur l'estimation de l'énergie libre gagnée en saturant avec un algorithme comme Mfold (sans pseudo-nœud) la structure secondaire du nœud "père".

- Pillsbury *et al.* [67] (2005) : les auteurs décrivent la manière de construire des relations de récurrence calculant la fonction de partition d'une séquence incluant des repliements de genre maximal donné. Quand ce genre maximal est fixé à 1, les relations de récurrence obtenues se calculent en $O(L^6)$. Cette méthode n'a jamais été testée en pratique.

- Metzler & Nebel [87] (2007, *McQfold*) : les auteurs utilisent une modélisation des pseudo-nœuds par grammaire algébrique stochastique. Pour ce faire, ils enrichissent d'un nouveau symbole terminal q la grammaire habituellement utilisée pour les structures secondaires sans pseudo-nœud. Leur algorithme engendre une structure sans pseudo-nœud dans laquelle des sous-séquences non appariées sont marquées par un symbole q. Une chaîne de Markov est ensuite utilisée pour apparier ces différentes sous-séquences de manière à former des pseudo-nœuds.

4.2.2 Quelques réflexions préliminaires

Ces algorithmes sont construits en regard d'un certain modèle d'énergie : il s'agit bien sûr d'inclure un paramétrage des dipaires, renflements, etc... mais surtout un paramétrage des pseudo-nœuds. Pour les termes habituels du modèle, la plupart des algorithmes suivent Turner99 mais peuvent différer en ce qui concerne les pseudo-nœuds. En effet, chacun a son propre mécanisme de génération de pseudo-nœuds auquel se rattache naturellement une manière de les paramétrer. On peut cependant observer une démarche commune : prenant appui sur une représentation diagrammatique, ces paramétrages consistent à pénaliser le nombre d'intersections des éléments de structures que manipulent les algorithmes. Ces éléments de structure peuvent être des hélices (HotKnots) ou des matrices *gap* (Rivas & Eddy). Ainsi donc, ces paramétrages ne sont pas de nature topologique. Par exemple, les pseudo-nœuds ABAB et ABACBC auront des pénalités respectives de p et $2p$ en notant p la pénalité unitaire de formation de pseudo-nœud. Avec le paramétrage topologique présenté dans le chapitre précédent, ces deux pseudo-nœuds auraient la même pénalité, étant tous deux de même genre. En l'absence de données expérimentales, décider si le pseudo-nœud ABACBC doit avoir une pénalité différente du ABAB est une *affaire de goût*. Cependant, je donnerai à la fin de ce chapitre un exemple pratique qui illustre une différence plus fondamentale entre ces deux paramétrages, touchant à la définition de ce qu'est un pseudo-nœud. On pourra déjà discuter de cette différence sans données expérimentales.

Cahier des charges pour un nouvel algorithme de prédiction de pseudo-nœuds

Dans les algorithmes précités et notamment ceux reposant sur des relations de récurrence, le mécanisme de création de pseudo-nœuds induit leur paramétrage. Dans mon cas, le problème se pose différemment : disposant en premier lieu d'un nouveau paramétrage topologique, je dois imaginer un algorithme qui puisse engendrer des pseudo-nœuds tout en calculant leur genre. De plus, afin de pouvoir clairement analyser les résultats et en particulier les effets de ce nouveau paramétrage, il est préférable que l'algorithme effectue des calculs exacts et garantisse de trouver le minimum d'énergie libre. Comme mentionné en préambule de cette section, l'exactitude des calculs dans un problème NP-complet s'accompagne nécessairement d'une limite à la taille des séquences que l'algorithme peut traiter. Comme nous le verrons plus loin, les temps d'exécution de l'algorithme que je proposerai deviennent effectivement trop longs pour des séquences de plus de 150 bases.

Une telle limite peut néanmoins être nuancée :

- il est à mon avis encore utopique de vouloir prédire les pseudo-nœuds de structures de grande taille alors qu'on peut constater que les modèles d'énergie échouent le plus souvent à prédire de grandes structures secondaires pourtant sans pseudo-nœud.
- la limite de taille est repoussée par la puissance de calcul de machines de plus en plus accessibles.

Ce choix de l'exactitude n'a été fait pour aucun des algorithmes pré-cités. A modèle d'énergie donné, le principe de chacun exclut d'emblée la garantie de trouver le minimum d'énergie libre, même pour de courtes séquences. Ceci pose un problème pour l'évaluation de ces algorithmes. Sans cette garantie, il est après tout théoriquement possible qu'une prédiction correcte soit en fait une "heureuse erreur", c'est-à-dire que l'algorithme ait fait une bonne prédiction parce qu'il n'a pas trouvé le minimum d'énergie qui correspondait pourtant à une structure autre. Ayant un modèle d'énergie intégralement reparamétré à tester, je me suis donc appliqué à chercher des méthodes exactes pour pouvoir faire la part des choses dans l'analyse de mes résultats.

4.2.3 Un nouvel algorithme : TT2NE

Le nouvel algorithme présenté dans cette section, TT2NE, est le seul algorithme de prédiction de pseudo-nœud pouvant prédire n'importe quel type de pseudo-nœud en garantissant de trouver le minimum d'énergie libre. La structure de cet algorithme s'inspire de [88] : à séquence donnée, la meilleure structure secondaire est construite à partir de la liste des hélices possibles. Cette liste d'hélices est organisée en graphe et ce graphe est parcouru de manière adéquate pour trouver la structure d'énergie libre minimale. Quelques modifications sont apportées afin que ce parcours s'effectue aussi rapidement que possible. TT2NE est testé sur 35 séquences de moins de 130 bases de long et est comparé à HotKnots et QmacFold, démontrant une nette amélioration de la qualité de prédiction.

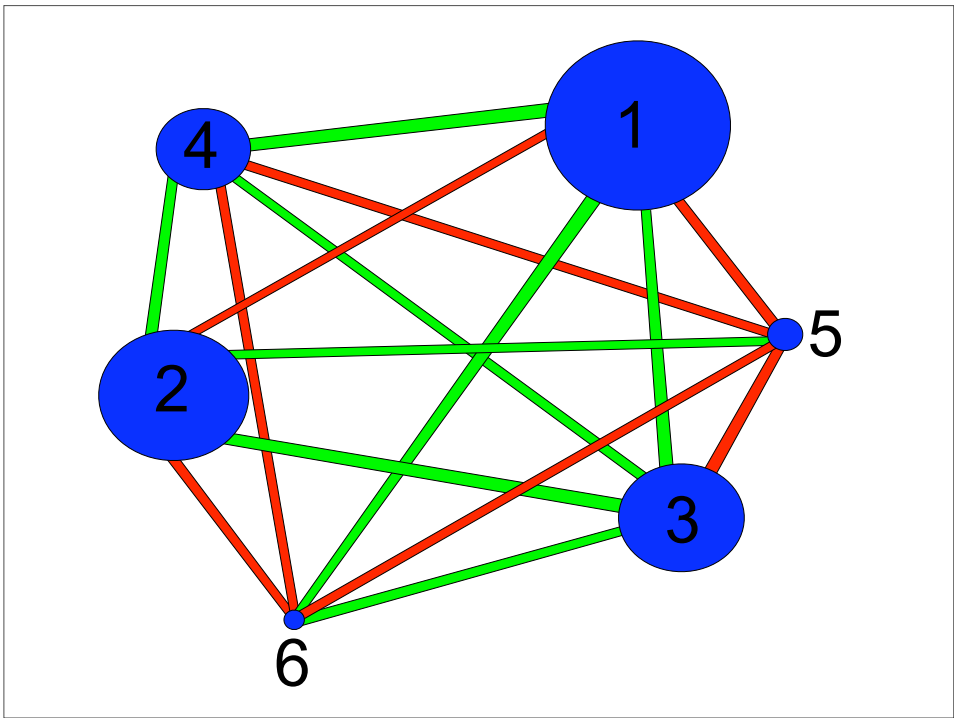
Représentation des structures secondaires

Pour pouvoir calculer la structure d'énergie libre minimale à séquence donnée, il faut que l'algorithme puisse avoir accès à tous ses repliements possibles. Leur nombre est exponentiel mais il est possible de tous les coder simultanément en mémoire grâce à un objet combinatoire : le graphe (au sens informatique du terme).

Soit x une séquence d'ARN et soit $\{H_i\}_{i \leq N_x}$ l'ensemble des hélices pouvant être formées à partir de x. Ces hélices sont ordonnées par énergie libre croissante : H_1 est l'hélice la plus stable et H_{N_x} la moins stable. La construction de cet ensemble consiste à identifier toutes les sous-séquences de x complémentaires. Elle s'exécute en $O(l_{max}L^2)$ où l_{max} est la longueur maximale des hélices de la séquence. Ces hélices constituent les nœuds d'un graphe complet et tout couple de nœuds est relié par un arc qui peut être de deux types, "vert" ou "rouge" selon que les hélices associées sont compatibles ou non. Des hélices sont déclarées incompatibles lorsqu'elles ne peuvent coexister dans une même structure, par exemple si elles ont une base en commun. En effet, une base ne peut former deux paires, de sorte que seule l'une ou l'autre des hélices peut l'accaparer. Afin d'éviter de futures redondances dans l'algorithme, sont aussi déclarées incompatibles des hélices dont la concaténation crée une hélice déjà présente dans le graphe.

Une structure secondaire étant un ensemble d'hélices toutes compatibles entre elles, les structures secondaires possibles sont alors codées dans le graphe $\mathcal{G}$ comme *l'ensemble des sous-graphes constitués de nœuds tous reliés deux à deux par des arcs verts*. Un tel sous-graphe est dit *admissible*.

Les hélices trop courtes peuvent être supprimées de $\mathcal{G}$. En effet, elles ne peuvent pas faire partie d'une structure secondaire d'énergie minimale si leur énergie libre de formation n'est pas favorable, c'est-à-dire si elle est positive. Les hélices de longueur 1 (paires isolées), la plupart des hélices de longueur 2 et beaucoup d'hélices de longueur 3 peuvent être ainsi négligées sans perte d'exactitude. Cette opération est évidemment conditionnée par le modèle d'énergie : un certain modèle peut déclarer comme favorable une courte hélice alors qu'un autre la négligera, de sorte que le nombre N_x de nœuds de $\mathcal{G}$ dépend du modèle utilisé.



Un exemple de graphe obtenu pour une séquence contenant six hélices possibles. L'hélice 1 est la plus favorable et 6 la moins favorable. L'analyse de ce graphe montre que le meilleur des sous-graphes admissibles est celui constitué des hélices 1, 3 et 4.

Un algorithme exhaustif, exact et inefficace pour parcourir les sous-graphes admissibles de $\mathcal{G}$

Les sous-graphes admissibles de $\mathcal{G}$ sont parcourus exhaustivement une et une seule fois grâce à l'algorithme récursif suivant qui fait appel à deux procédures ,"*Trouve_Minimum*" et "*Explore*", manipulant les variables globales S, S_{min}, ΔF et ΔF_{min} :

Initialisation des variables globales :
Structure secondaire courante $S = \emptyset$
Structure secondaire d'énergie libre minimale $S_{min} = \emptyset$
Energie libre courante $\Delta F = 0$
Energie libre minimale $\Delta F_{min} = 0$

Procédure Trouve_Minimum
for($i = 1..N_x$)
Explore(H_i, S)
end for

Procédure Explore(hélice H_i , structure secondaire S)
(1) *Test de compatibilité entre S et H_i*
If(H_i n'est pas compatible avec S) exit

(2)*Ajout de H_i à S et mise à jour de son énergie libre*
$S = S \cup H_i$
$\Delta F = \Delta F(S)$

(3)*Test pour savoir si la structure obtenue est la meilleure trouvée jusqu'à présent*
If($\Delta F < \Delta F_{min}$)
$\Delta F_{min} = \Delta F$
$S_{min} = S$

(4)*Appel récursif pour continuer à ajouter des hélices à S, qui contient maintenant H_i*
for ($j = i + 1...N_x$)
Explore(H_j, S)
end for

(5)*Retrait de H_i de S : à la fin de la fonction Explore, S est la même qu'au début. Ceci est nécessaire pour que l'opération (4) soit valide telle qu'elle est écrite. Il est en effet supposé que S est la même à chaque appel de la boucle "for".*
$S = S - H_i$
$\Delta F = \Delta F(S)$

(6) *fin*
exit

A la fin de l'exécution de la procédure *Trouve_Minimum*, S_{min} contient la structure d'énergie libre minimale et ΔF_{min} son énergie libre. La fonction *Explore* peut être facilement modifiée pour trouver en plus les n meilleures structures sous-optimales avec $n > 1$ et les détails techniques de cette modification ne seront pas exposés. Cet algorithme résout exactement le problème puisqu'il explore exhaustivement tous les sous-graphes admissibles de $\mathcal{G}$. Cependant, pour cette même raison, il est inefficace et devient en

pratique trop long à utiliser avec des séquences comportant plus de 400 hélices possibles.

Ainsi formulé, l'algorithme est indépendant du modèle d'énergie utilisé. Il permet en particulier d'incorporer le calcul du genre, qui doit s'effectuer à chaque mise à jour de l'énergie ΔF dans l'étape (2) de la fonction *Explore*.

Une méthode de séparation et évaluation progressives, adaptée au modèle d'énergie simplifié construit dans le chapitre "Modèle d'énergie libre"

Pour accélérer l'algorithme, une méthode de séparation et évaluation progressives ("*Branch-and-bound*" [89]) a été utilisée. Dans le cas présent, le principe de cette méthode se traduit ainsi : une fois qu'une bonne structure secondaire candidate a été trouvée, alors beaucoup de sous-graphes admissibles, constitués d'hélices faibles, ne peuvent plus concourir et n'ont plus besoin d'être explorés de manière approfondie. Une façon d'accélérer l'algorithme est donc de savoir estimer le plut tôt possible quels sous-graphes n'ont pas ou plus besoin d'être explorés : il faudra pour cela trouver une borne inférieure au gain d'énergie libre attendu en explorant un sous-graphe.

Notations

Les détails de cette opération nécessitent l'introduction de nouvelles notations :

- Soit M_0 le modèle d'énergie construit dans le chapitre "Modèle d'énergie libre", auquel s'ajoute la pénalité topologique liée au genre. Il s'agit du modèle selon lequel on désire trouver le minimum d'énergie libre. Je rappelle que dans M_0, l'entropie des différents types de boucles a été négligée et les termes d'initiation de têtes d'épingle et de boucles internes sont absorbés dans les énergies de paire extrémale d'hélice. Selon M_0, l'énergie libre d'une structure secondaire S est :

$$\Delta F^{M_0}(S) = \sum_{helices\ H_i} F(H_i) + \sum_{boucles\ internes\ BI_i} f_{asym}(BI_i) + n_{bmh}(S) \times a_{bmh} + \mu g(S) \quad (4.13)$$

  où a_{bmh} est la pénalité d'initation de boucles multi-hélices, $n_{bmh}(S)$ est le nombre de ces boucles dans S et f_{asym} est la fonction pénalisant l'asymétrie des boucles internes.

- Soit M_1 le modèle d'énergie obtenu à partir de M_0 en supprimant les pénalités de boucles internes, de sorte que

$$\Delta F^{M_1}(S) = \sum_{helices\ H_i} F(H_i) + n_{bml}(S) \times a_{bml} + \mu g(S) \tag{4.14}$$

- Soit M_2 le modèle d'énergie ne comportant que les énergies libres de formation d'hélices :

$$\Delta F^{M_2}(S) = \sum_{helices\ H_i} F(H_i) \tag{4.15}$$

- Soit $S^{M_i}_{min}(j)$ la structure secondaire d'énergie libre minimale obtenue avec le modèle M_i et construite uniquement avec des hélices d'indices supérieurs à j, c'est à dire avec les $N_x - j + 1$ hélices les moins favorables. Je note $\Delta F^{M_i}_{min}(j)$ l'énergie libre de cette structure. Avec cette notation, la structure d'énergie libre minimale recherchée est $S^{M_0}_{min}(1)$.

Minoration de $\Delta F(S)$

Soit $S = H_{i_1} \cup H_{i_2} \cup ... \cup H_{i_n}$ une structure composée de n hélices, où les indices sont ordonnés : $1 \leq i_1 < i_2 < ... < i_n \leq N$. Soit $S_k = H_{i_1} \cup H_{i_2} \cup ... \cup H_{i_k}$ la restriction de cette structure à ses k meilleures hélices. Symétriquement, notons $\overline{S_k} = H_{i_{k+1}} \cup H_{i_{k+2}} \cup ... \cup H_{i_n}$, de sorte que $S = S_k \cup \overline{S_k}\ \forall k, 1 \leq k \leq n$. On a alors les inégalités suivantes :

$$\forall k < n \ : \ \Delta F^{M_1}(S) \ \geq \ \Delta F^{M_1}(S_k) + \Delta F^{M_2}_{min}(i_k + 1) \tag{4.16}$$

$$\forall k < n \ : \ \Delta F^{M_2}(S) \ \geq \ \Delta F^{M_2}(S_k) + \Delta F^{M_2}_{min}(i_k + 1) \tag{4.17}$$

La preuve est immédiate :

$$\begin{aligned}
\Delta F^{M_1}(S) &= \sum_{p=1}^{n} \Delta F(H_{i_p}) + n_{bmh}(S) \times a_{bmh} + \mu g(S) \\
&= \left[\left(\sum_{p=1}^{k} \Delta F(H_{i_p}) \right) + n_{bmh}(S) \times a_{bmh} + \mu g(S) \right] + \sum_{p=k+1}^{n} \Delta F(H_{i_p})
\end{aligned}$$

or :

$$\left.\begin{array}{l} n_{bmh}(S) \geq n_{bmh}(S_k) \\ g(S) \geq g(S_k) \end{array}\right\} \text{ car } S_k \subset S$$

$$\sum_{p=k+1}^{n} F(H_{i_p}) \geq \Delta F^{M_2}_{min}(i_{k+1}) \text{ par définition de } \Delta F^{M_2}_{min}(j)$$

$$\Delta F^{M_2}_{min}(i_{k+1}) \geq \Delta F^{M_2}_{min}(i_k + 1) \text{ car } i_{k+1} \geq i_k + 1$$

donc :

$$\begin{aligned}
\Delta F^{M_1}(S) &\geq \left[\left(\sum_{p=1}^{k} \Delta F(H_{i_p}) \right) + n_{bmh}(S_k) \times a_{bmh} + \mu g(S_k) \right] + \Delta F^{M_2}_{min}(i_k + 1) \\
&\geq \Delta F^{M_1}(S_k) + \Delta F^{M_2}_{min}(i_k + 1)
\end{aligned}$$

Note : cette démonstration est correcte uniquement si a_{bmh} et μ sont *positifs*. La preuve de (4.17) n'est qu'un cas particulier de la démonstration ci-dessus.

Une nouvelle version de *Explore* plus rapide

Supposons que les $\Delta F_{min}^{M_2}(i)$ aient été calculés $\forall i \leq N$. La relation (4.16) peut être utilisée pour accélérer la recherche des $\Delta F_{min}^{M_1}(i)$ en introduisant ce nouveau test entre les étapes (2) et (3) de la procédure *Explore*, renommée *Explore2* :

Procédure Explore2(hélice H_i , structure secondaire S)
(1) *Test de compatibilité entre S et H_i*
If(H_i n'est pas compatible avec au moins une hélice de S) exit

(2)*Ajout de H_i à S et mise à jour de son énergie libre*
$S = S \cup H_i$
$\Delta F = \Delta F^{M_k}(S)$ *avec $M_k = M_1$ ou M_2*
selon qu'on veuille calculer $\Delta F_{min}^{M_1}$ ou $\Delta F_{min}^{M_2}$

(2b) *Ce nouveau test détecte s'il est impossible d'obtenir une meilleure structure que S_{min} en continuant à rajouter des hélices à la structure courante S*
If($\Delta F + \Delta F_{min}^{M_2}(i+1) > \Delta F_{min}$) aller à étape (5)

(3)*Ajout de H_i à S_0 et mise à jour de son énergie libre*
...

(5)*Retrait de H_i de S*

(6) *fin*
exit

Si, à un moment donné, la structure courante S constituée d'hélices dont l'indice le plus bas est i vérifie $\Delta F^{M_1}(S) + \Delta F_{min}^{M_2}(i+1) > \Delta F_{min}$, alors il est inutile d'essayer de chercher le minimum d'énergie libre en rajoutant à S des hélices d'indices inférieurs à i comme le fait l'étape (3). En effet, toutes les structures S' qu'on pourrait obtenir de la sorte vérifieraient alors $S \subset S'$ et donc d'après (4.16) :

$$\Delta F^{M_1}(S') \geq \Delta F^{M_1}(S) + \Delta F_{min}^{M_2}(i+1) > \Delta F_{min} \tag{4.18}$$

$\Delta F_{min}^{M_2}(i{+}1)$ représente l'énergie qu'on peut *au plus* gagner en continuant la recherche. Si celle-ci est insuffisante pour obtenir une énergie inférieure à ΔF_{min}, il faut arrêter d'explorer cette voie.

Ainsi, la connaissance des $\Delta F_{min}^{M_2}(i)$ permet d'économiser de nombreux appels récursifs inutiles lors de l'étape (3). Le test (2b) sera d'autant plus efficace que la valeur ΔF_{min}, variable mise à jour en cours d'exécution, sera proche du minimum.

Il n'est cependant pas question de calculer l'ensemble des $\Delta F_{min}^{M_2}(i)$ car ce serait uniquement reporter le problème. Un *juste milieu* consiste à calculer uniquement ces termes pour les M dernières valeurs de i. Il y a une manière efficace de procéder. Une mauvaise manière de calculer les quantités $\Delta F_{min}^{M_2}(N_x - M)$, $\Delta F_{min}^{M_2}(N_x - M + 1)$ ... , $\Delta F_{min}^{M_2}(N_x)$ serait la suivante :

Procédure Calcule_$\Delta F_{min}^{M_2}$
  for($i = N_x - M ... N_x$)
    Explore(H_i, S)
    $\Delta F_{min}^{M_2}(i) = \Delta F_{min}$
    $\Delta F_{min} = 0$
  end for

En effet, celle-ci est bien meilleure :

Procédure Calcule_$\Delta F_{min}^{M_2}$
  for($i = N_x ... N_x - M$)
    Explore2(H_i, S)
    $\Delta F_{min}^{M_2}(i) = \Delta F_{min}$
    $\Delta F_{min} = 0$
  end for

L'intérêt de calculer $\Delta F_{min}^{M_2}(i)$ avec i décroissant est le suivant : lorsque la procédure *Calcule_*$\Delta F_{min}^{M_2}$ appelle la fonction *Explore*$(H_j,\ S)$ pour une certaine valeur de j, les quantités $\Delta F_{min}^{M_2}(i)$ avec $i > j$ ont déjà été calculées. On peut donc d'ores et déjà avoir recours au test (2b) de la fonction *Explore2*. Ainsi, le calcul de $\Delta F_{min}^{M_2}(i)$ avec i décroissant est coopératif et incomparablement plus rapide qu'avec i croissant. Il est possible en général de calculer ces termes en un temps satisfaisant, de l'ordre de la minute sur un Mac PowerBook G4, pour $M = 400$.

Cette accélération permet au final de trouver le minimum d'énergie libre dans un graphe contenant environ 750 nœuds. Des graphes de cette taille correspondent généralement à des séquences de 100 bases de long.

Revenons maintenant sur la nécessité de travailler avec les modèles M_1 et M_2 au lieu de M_0.

Comme mentionné précédemment, l'idée de la méthode est d'estimer à tout moment l'énergie libre qu'une structure secondaire S en construction peut gagner *au mieux* en explorant telle ou telle branche du graphe. Pour que cette méthode permette les meilleures économies de temps de calcul, il convient de surestimer le moins possible cette quantité afin que le test (2b) soit le plus exigeant. Comme il a été démontré, il est exact d'employer le terme $\Delta F_{min}^{M_2}(i+1)$ où i est l'indice de la plus faible hélice constituant S. Peut-on faire mieux ? La quantité $\Delta F_{min}^{M_1}(i+1)$, qui inclut en plus les pénalités de boucle multi-hélices et de genre, est un candidat naturel. Mais a-t-on l'équivalent de (4.16), à savoir :

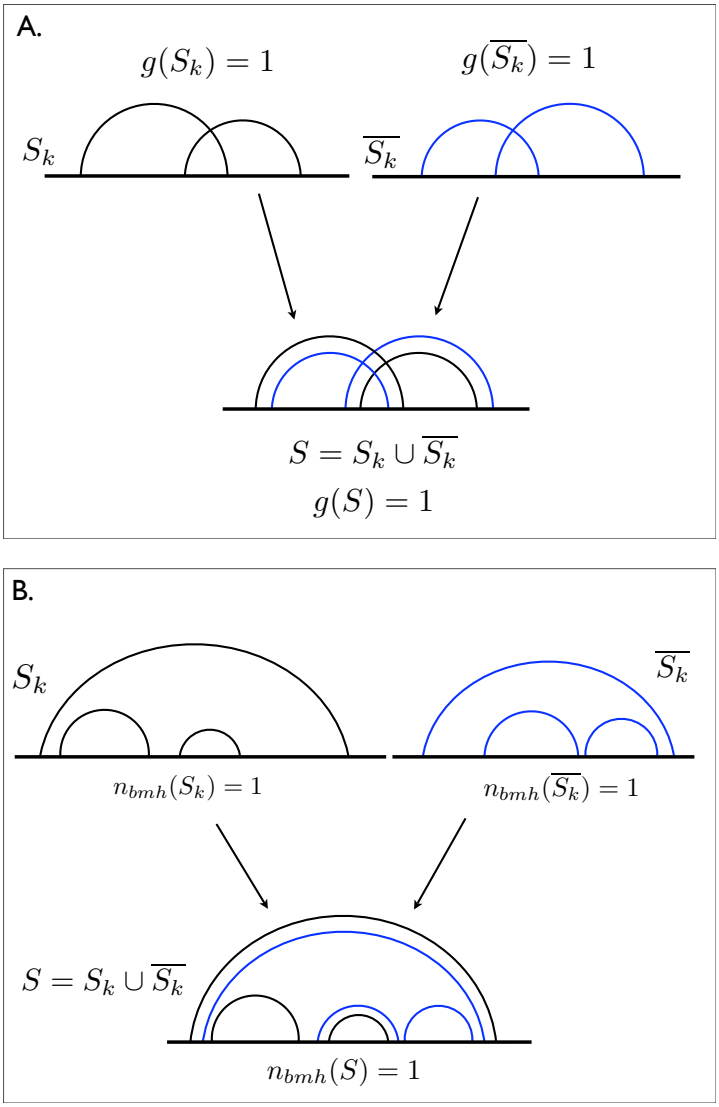
$$\forall k < n \ : \ \Delta F^{M_1}(S) \geq \Delta F^{M_1}(S_k) + \Delta F_{min}^{M_1}(i_k + 1) \tag{4.19}$$

En reprenant les notations et la démonstration de (4.16), on s'aperçoit que (4.19) n'est vraie que si :

$$n_{bmh}(S) \ \geq \ n_{bmh}(S_k) + n_{bmh}(\overline{S_k}) \tag{4.20}$$

$$g(S) \ \geq \ g(S_k) + g(\overline{S_k}) \tag{4.21}$$

Maheureusement, ces deux relations ne sont pas toujours satisfaites. Voici un exemple pour (4.21) :



Dans l'exemple (A), S_k, $\overline{S_k}$ et S ont toutes trois la topologie d'un pseudo-nœud H et sont de genre 1, contredisant la relation (4.21). L'exemple (B) contredit aussi (4.20). Des contre-exemples similaires portant sur les boucles internes montrent qu'on ne peut non plus remplacer M_1 par M_0 dans (4.16). La relation (4.16) est au final la meilleure que j'aie pu trouver.

Il y a encore une manière d'économiser du temps de calcul sans perte d'exactitude mais cette fois pour des raisons biologiques et non théoriques. Le parcours est initié par la procédure *Trouve_Minimum* dont le corps est :

```
S est initialisé à ∅
 Procédure Trouve_Minimum
    for(i = 1...N_x)
        Explore2( H_i, S )
    end for
```

Cette procédure construit tour à tour la structure d'énergie libre minimale dont la meilleure hélice est H_i, avec i croissant. En pratique, comme je l'ai constaté systématiquement pour toutes les bases de données que j'ai étudiées, les structures secondaires des ARN naturels contiennent toujours au moins une hélice forte qui se classe parmi les 10% les plus stables. Il est donc pleinement satisfaisant de limiter la boucle *for* de la procédure à $N_x/10$:

```
Procédure Trouve_Minimum
    for(i = 1...N_x/10)
        Explore2( H_i, S )
    end for
```

L'algorithme décrit jusqu'ici permet donc de trouver le minimum d'énergie libre d'une séquence d'environ 100 bases de long avec pseudo-nœuds en un temps raisonnable selon le modèle d'énergie M_1. Or, la question de départ concernait le modèle M_0, c'est à dire un modèle prenant en compte en plus l'asymétrie des boucles internes. Ces corrections étant faibles, je propose de :

1. utiliser TT2NE pour trouver les K meilleures structures selon M_1 (exemple : $K = 25$).
2. recalculer les énergies de ces K structures selon M_0 et les reclasser si besoin est.

Cette idée est la même que pour les algorithmes Mfold et *efn2* [28] : *efn2* recalcule l'énergie des meilleures prédictions de Mfold en utilisant un meilleur modèle pour les boucles multi-hélices et les reclasse. Quelques résultats seront présentés après la section suivante.

Une heuristique raisonnable pour aller plus loin

Afin de pouvoir traiter des séquences encore plus longues, j'introduis une heuristique reposant sur l'idée que les ARN naturels se structurent préférentiellement autour des hélices fortes disponibles dans leur séquence. Cette idée peut être vérifiée quantitativement sur les bases de données, bien que je ne l'aie pas fait de manière systématique. Etant donné un graphe $\mathcal{G}$ contenant beaucoup de nœuds, l'heuristique consiste à résoudre exactement le problème pour un sous-graphe $\mathcal{G}'$ de $\mathcal{G}$ constitué des N_h meilleures hélices $\{H_i\}_{1\leq i\leq N_h}$ où N_h est choisi tel que $\mathcal{G}'$ puisse être étudié suffisamment rapidement avec les techniques précédentes. Le résultat de cette étude est la donnée des K_h meilleures structures de $\mathcal{G}'$ avec, par exemple, $K_h = 500$. Ces structures sont ensuite utilisées comme *amorces* : elles sont saturées de la meilleure manière possible par les hélices restantes $\{H_i\}_{N_h+1\leq i\leq N}$ en utilisant encore la procédure *Explore2*. La structure d'énergie minimale trouvée par ce processus est la prédiction finale de l'algorithme. Avec cette heuristique, des séquences contenant jusqu'à 1200 hélices peuvent être traitées de manière satisfaisante : ceci correspond à des séquences d'environ 130 bases. L'algorithme devient trop long au-delà de 150 bases.

4.2.4 Résultats

J'ai comparé, sur une sélection de séquences tirées de la Pseudobase [70] et de la PDB, la sensibilité du programme HotKnots, de TT2NE utilisé avec le modèle d'énergie INN-HB et de TT2NE utilisé avec M_1. Je n'ai pas comparé TT2NE à d'autres algorithmes car il a été montré que les performances de HotKnots sont meilleures que celles des algorithmes l'ayant précédé. McQfold a été également testé, étant postérieur à HotKnots. Voici des précisions techniques que le lecteur doit avoir à l'esprit pour apprécier les résultats :

- Pour chaque séquence, les hélices possibles et leur énergie ont été déterminées en permettant des renflements de taille 1 et des boucles internes 1×1. Les renflements et les boucles internes plus grands restent réalisables en concaténant les hélices adéquates mais sans contribuer à l'énergie.
- la pénalité de formation de pseudo-nœud μ a été fixée à 0.9u.
- j'ai rajouté dans TT2NE l'*interdiction* de prédire des pseudo-nœuds de type ABCABC. La raison sera précisée ultérieurement.
- les résultats, obtenus pour le modèle d'énergie M_1, sont présentés sans avoir été recalculés selon M_0.

sequence	longueur	HotKnots	McQfold	TT2NE+INN-HB	TT2NE + M_0
1u8d	68	0.69	0.69	0.88	0.88
AMV3	113	0.84	0.76	0.98	0.94
BBMV	116	0.81	0.86	1*	1*
BVDV	74	0.56*	0.72	0.96	0.85
Bp_PK2	90	0.81	0.84	1	1
BWYV	50	1*	1*	1*	0.88
Bt-PrP	45	0.41	0.41	1	0.41
CGMMV	85	0.64	0.32	0.54	0.64
CcTMV	73	0.42*	0.53	0.53	0.42
CoxB3	73	0.68	0.92	0.92	0.96
EC_PK4	52	1*	0.63	1	1
Ec-RpmI	72	0.58*	0.48	0.55	0.48
Ec_PK1	30	1	1	1	1
Ec_S15	67	1	0.94*	1	1
Ec_alpha	108	0.45	0.61	0.62*	0.62
GLRaV-3	75	0.65	1	1	1
HAV	55	0.58	0.58 *	1*	1
HCV_229E	74	0.79	1	1	0.95
HDV	87	1*	0.75	1	1
HDV_anti	91	1*	1	0.4	1*
Hs_PrP	45	0	0.81*	0	0.45
IBV	56	1	0.94	1	1
Lp_PK1	30	1*	0.40	1*	1
MMTV	34	1	1	1	1
MMTV-vpk	34	1	1	1	1
Mengo-PKC	26	0.37*	0.37	0.25*	1*
RSV	128	0.97	1	0.97	0.97*
SRV-1	38	1	1	1	1
T2_gene32	33	1	1*	1	1
T4_gene32	28	1*	0.63	1	1
TMV	74	0.52	0.48	0.48*	0.8
TYMV	74	0.7	0.72	0.72	1
Tt-LSU	65	0.95	0.85*	0.95	1*
minimalIBV	45	1	1	1	1
pKA-A	36	1	1	1	1
satRPV	73	0.59	0.81	0.81	1
moyenne		0.76	0.78	0.84	0.91

Pour chaque séquence et chaque algorithme, les sensibilités des deux meilleures prédictions ont été mesurées et c'est la meilleure de ces deux sensibilités qui est rapportée dans le tableau. Lorsque celle-ci est en fait la sensibilité de la seconde meilleure prédiction, cela est signalé par une astérisque (*).

- Cette sélection de séquences inclut la plupart des 33 séquences contenant un pseudo-nœud sur lesquelles HotKnots a été montré comme plus performant que ses prédécesseurs. La comparaison avec HotKnots sur cette sélection permet donc "par transitivité" de situer TT2NE par rapport à l'état de l'art.
- TT2NE-INNHB a une sensibilité bien meilleure que Hotknots sur cette liste de séquences. Le modèle INN-HB est un sous-ensemble de Turner99, modèle d'énergie utilisé par HotKnots. Ceci montre que les meilleurs résultats de TT2NE-INNHB sont dus à l'algorithme et au nouveau paramétrage des pseudo-nœuds.
- Le modèle M_1 donne lui-même de meilleurs résultats que INN-HB. Ceci prouve que M_1 permet une nette amélioration moyenne de la sensibilité de prédiction sur une base de données autre que celle d'ARNt sur laquelle il a été entraîné. La plupart de ces séquences étant composées d'hélices parfaites, ces résultats étayent l'intuition que M_1, et par là M_0, estiment bien cette classe de structures secondaires.

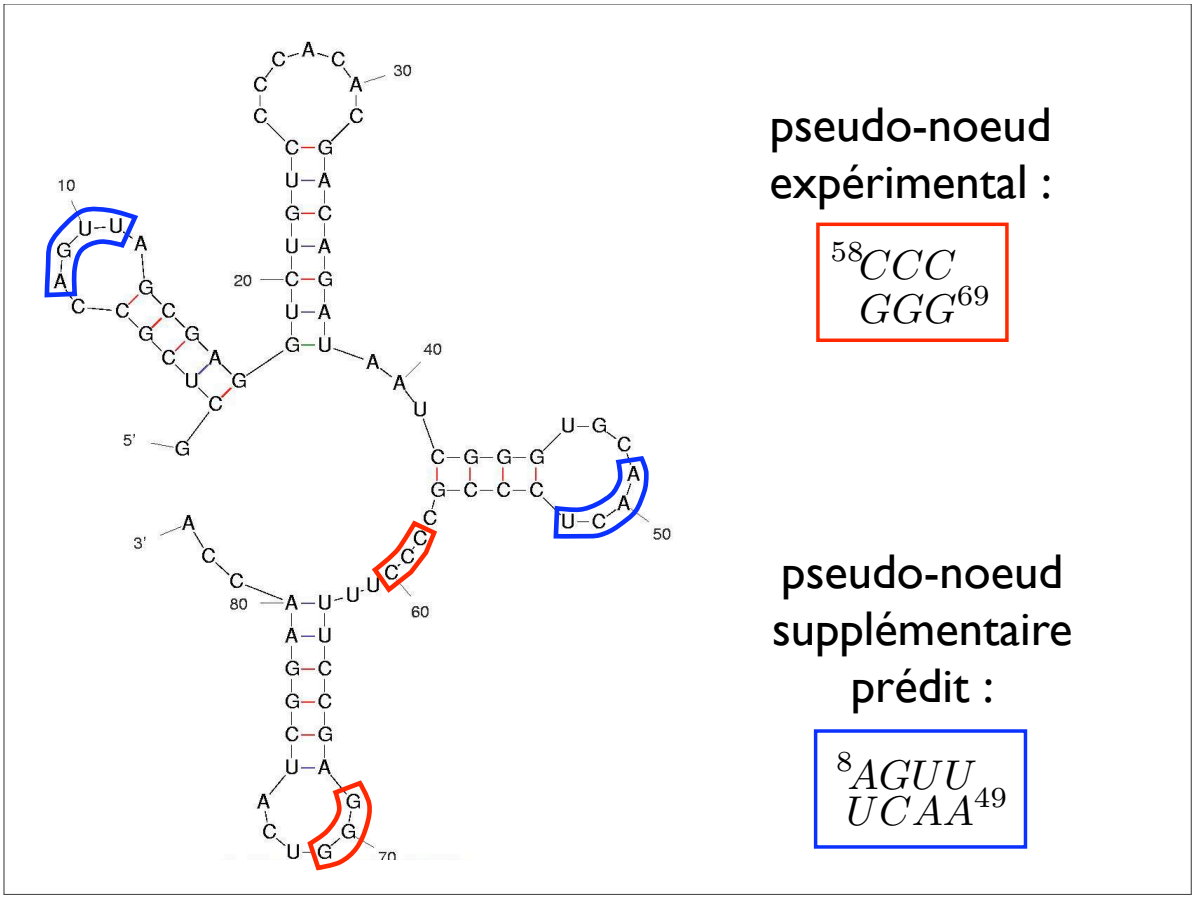
Bien que la sensibilité soit le critère habituel d'estimation de la performance d'un algorithme de prédiction, cette seule mesure est trop limitée pour se faire une idée précise de ses capacités et de ses limites. Je vais donc analyser plusieurs des échecs de TT2NE et de HotKnots pour comprendre leurs causes et pouvoir dire comment améliorer encore ces résultats.

Analyse approfondie des résultats de TT2NE utilisé avec M_1

1. 22 séquences ont été correctement prédites
2. 7 le sont avec une sensibilité supérieure à 85%
3. 1 avec une sensibilité de 80%
4. 6 autres le sont avec une sensibilité inférieure à 65%

Structures correctement prédites

Parmi ces 22 structures, la prédiction de TYMV ("*turnip yellow mosaic virus*") est particulièrement intéressante car elle annonce un pseudo-nœud additionnel par rapport à la structure expérimentalement déterminée dans [90] :



Comme mentionné dans [90], la structure de TYMV a une forte ressemblance avec celle des ARNt et cela a un intérêt fonctionnel. Le pseudo-nœud supplémentaire prédit par TT2NE est très pertinent car il accentue encore plus cette ressemblance en imitant l'interaction entre les boucles D et T de l'ARNt. Ce pseudo-nœud doit être vérifié expérimentalement. Cette prédiction est ainsi de genre 2. La meilleure structure de genre 1, obtenue en jouant sur la pénalité μ de formation de pseudo-nœud, est bien la structure expérimentale.

Structures prédites avec une sensibilité supérieure à 85%

Etant donné la relative petite taille des séquences concernées, une sensibilité supérieure à 85% signifie en fait que moins de trois paires de bases ont été incorrectement prédites par séquence. L'examen de ces structures montre que les paires incorrectement prédites sont dues à des imperfections du modèle d'énergie. Ces erreurs ne concernent jamais le pseudo-nœud mais le plus souvent des paires extrémales mal décidées selon ce principe :

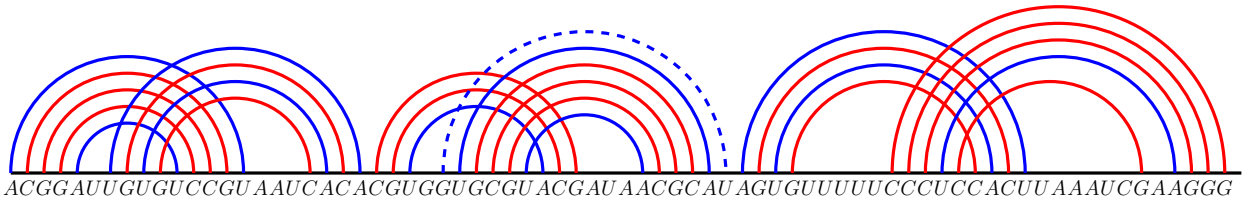
Chacune des deux hélices rouges peut être prolongée d'une dipaire supplémentaire en formant une paire A–U. Ces deux hélices sont donc en compétition pour la base U et le modèle d'énergie doit décider à qui l'attribuer, ce qu'il fait parfois de manière erronée. Dans ces cas-là, la structure secondaire réelle est toujours très proche de la meilleure structure prédite dans la hiérarchie établie par TT2NE.

Dans les diagrammes dessinés par la suite, la convention suivante sera toujours appliquée :

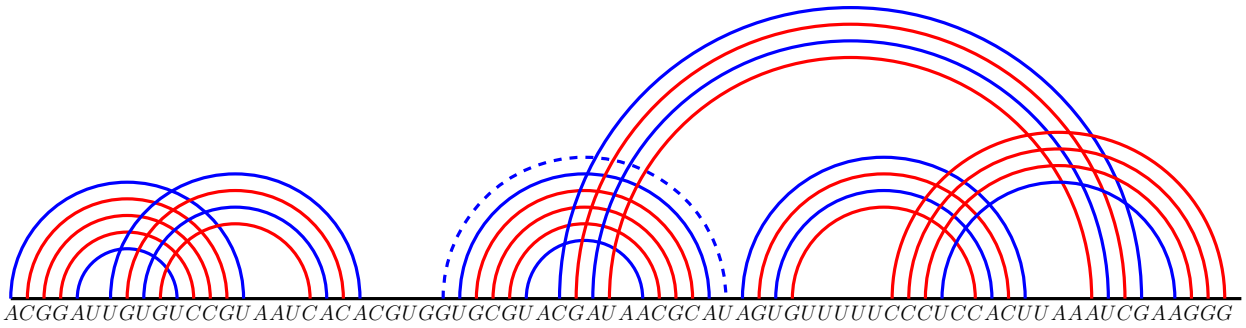
- les arcs rouges désignent une paire G–C ou C–G.
- les arcs bleus désignent une paire A–U ou U–A.
- les arcs pointillés bleus désignent une paire G–U ou U–G.

Structure prédite avec une sensibilité de 80% : TMV

Cette structure est constituée de trois pseudo-nœuds H en série. La structure secondaire réelle est [91] :



tandis que la structure prédite est :



Ainsi, une hélice de quatre paires de bases est prédite au détriment d'une hélice comparativement moins stable dans la structure réelle. Les structures réelle et prédite sont toutes deux de genre 3.

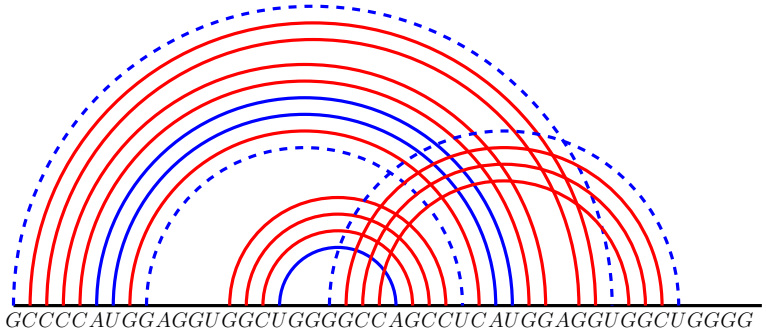
Structures prédites avec une sensibilité inférieure à 65%

Je présente les détails de toutes les structures prédites avec une sensibilité de moins de 65% afin de les discuter ultérieurement.

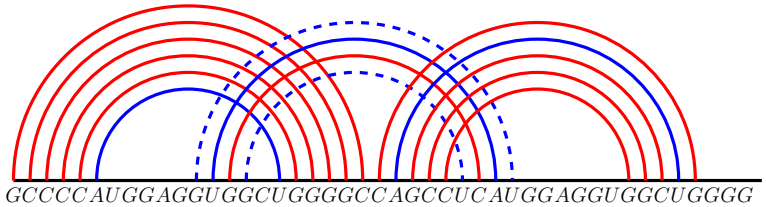
- *Bt_PrP* [92]

  La structure secondaire réelle est en fait la troisième meilleure prédite par TT2NE. Les deux premières contiennent des pseudo-nœuds différents, tous deux de genre 1.

  La meilleure structure prédite a la topologie d'un pseudo-nœud H :



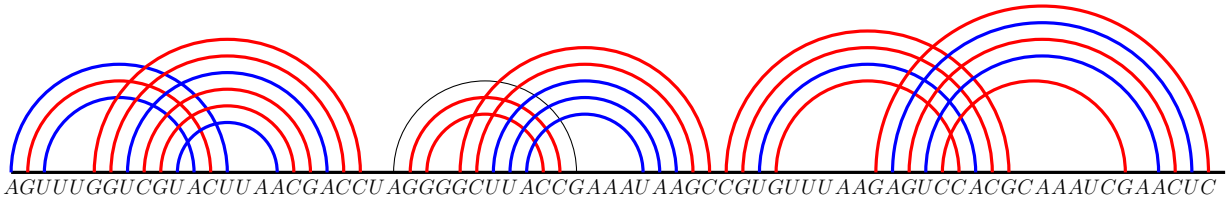
  La seconde meilleure structure prédite a la topologie d'un KH :



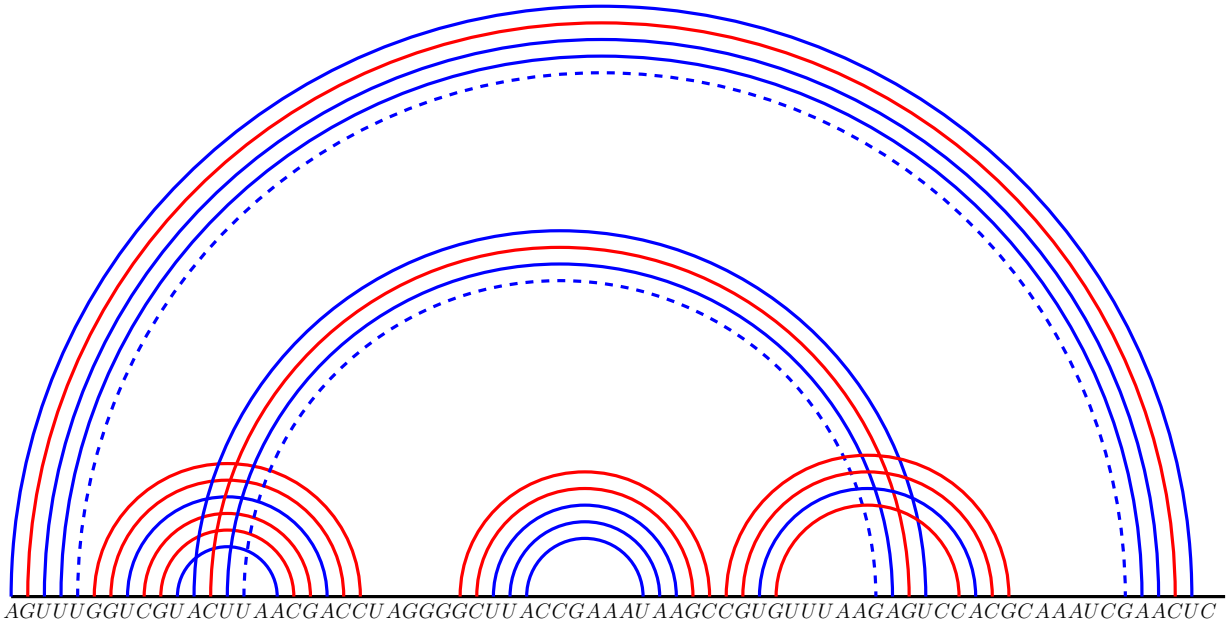
- *CGMMV et CcTMV* [91]

  Ces deux structures ont été déterminées par comparaison de séquences avec TMV qui, elle, l'a été expérimentalement.

  Voici la structure théorique de CcTMV :



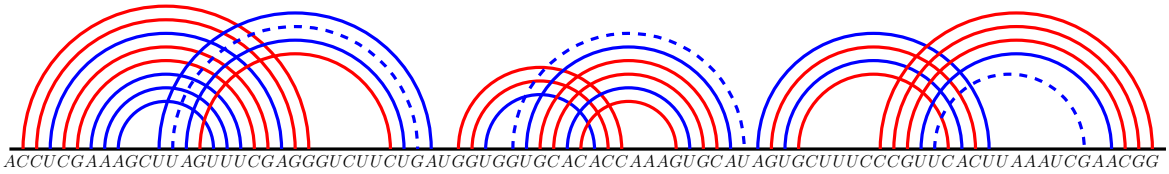
  et voici la meilleure structure prédite par TT2NE :



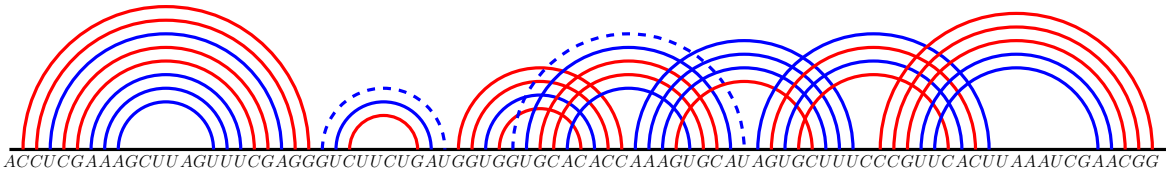
  En comparant la structure obtenue par comparaison de séquences et la structure prédite par TT2NE, cette dernière apparaît plus vraisemblable. La structure obtenue par comparaison de séquences comprend une hélice de trois bases de long contenant une paire G–A. Si on considère que cette hélice qui, de plus, forme un pseudo-nœud

est trop peu stable pour apparaître dans la structure réelle, alors on s'aperçoit que les structures théorique et prédite sont toutes deux constituées de 5 hélices, dont 3 sont en commun. Les deux autres hélices sont constituées de 7 paires de bases pour la structure théorique et de 9 pour la structure prédite. La structure prédite fait de plus l'économie de la création d'un pseudo-nœud : elle est seulement de genre 1. Ainsi, il y a de bonnes raisons de penser que la structure prédite est plus vraisemblable que celle théorique et cela demande une confirmation expérimentale.

Le structure théorique de CGMMV est :

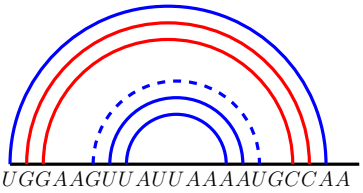


tandis que la structure prédite est :



- *Ec_RpmI* [93]

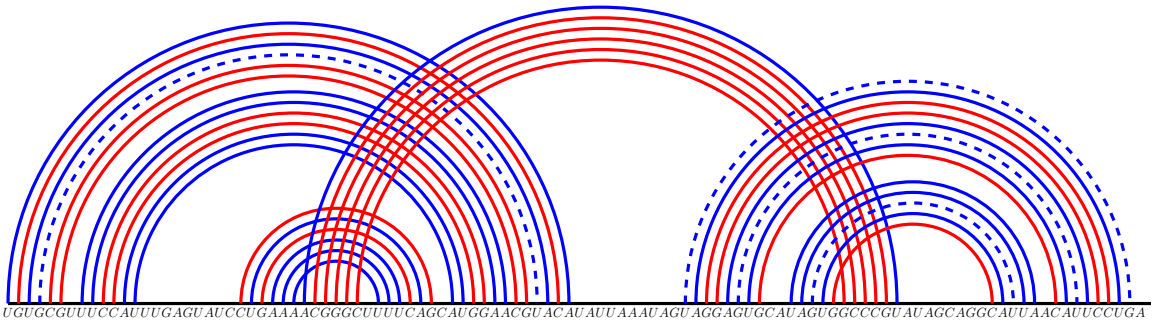
  La structure secondaire réelle contient une boucle interne 2×1 que ni M_1 ni M_0 ne déclarent favorable :



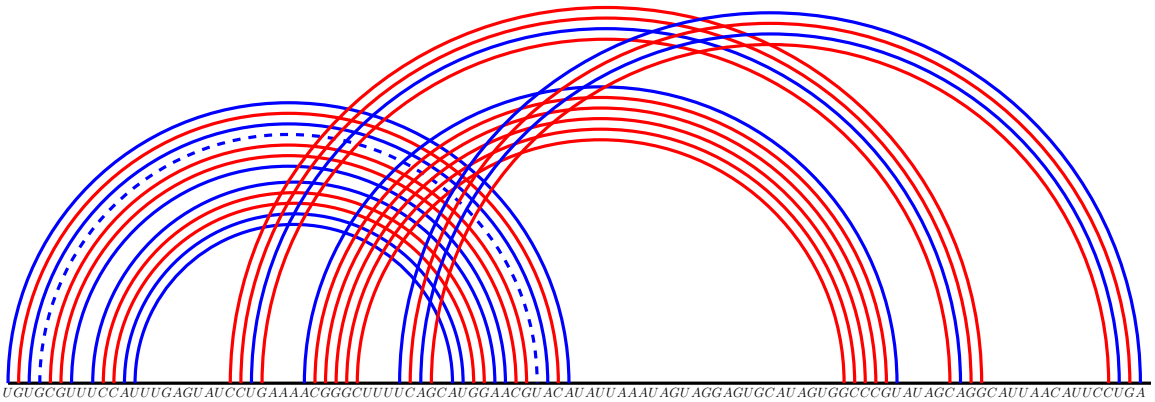
  La structure secondaire réelle privée de l'hélice intérieure de cette boucle interne est prédite en septième position.

- *Ec_alpha* [69]

  Ec_alpha est, pour l'instant, la seule structure connue dont le pseudo-nœud a la topologie ABCABC. Comme mentionné plus haut, prédire cette topologie a été interdite à TT2NE. En l'autorisant uniquement pour cette séquence, la structure prédite n'est toujours pas la bonbe. Elle a la topologie d'un KH (ABACBC) et l'énergie qui lui est associée est considérablement inférieure à celle de la topologie ABCABC attendue : la première est de -19.9u tandis que l'autre vaut -15.6u. La structure prédite (ci-contre) contient une boucle interne 9×1 fortement asymétrique et une autre 2×1 mais la prise en compte des pénalités qui leur sont associées ne change pas la hiérarchie.

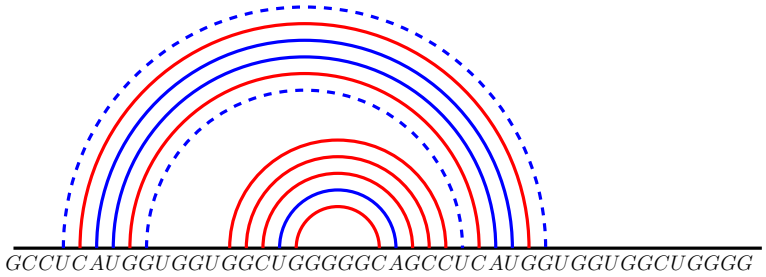


La structure réelle est :

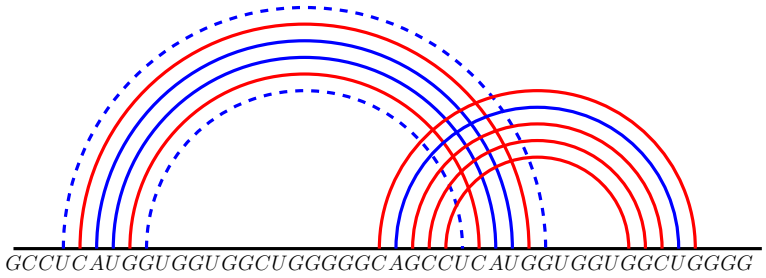


- *Hs_PrP* [94]

  La meilleure structure proposée ne contient pas de pseudo-nœud :



  alors que la structure réelle est :



  En comparant ces deux structures, on s'aperçoit qu'elles sont constituées d'hélices en tout point semblables : composition et bases libres voisines. La seule différence que fait le modèle d'énergie entre ces structures vient de :

  1. la présence d'un renflement de taille 4 dans la structure secondaire prédite.
  2. la présence d'un pseudo-nœud de genre 1 dans la structure réelle.

  Je rappelle que ces résultats ont été obtenus en négligeant les contributions des boucles internes asymétriques et des longs renflements et que l'énergie des structures obtenues n'a pas été recalculée en les incluant. Ici, cet exemple fournit une contrainte portant sur la pénalité des renflements de taille 4 $\Delta F_r(4)$ et la pénalité de formation de pseudo-nœuds μ. Pour pouvoir prédire la structure correcte, il faut que μ soit inférieure à $\Delta F_r(4)$. C'est bien le cas dans M_0 avec le choix $\mu = 0.9u$ et, en reclassant les structures, la structure réelle devient la structure d'énergie minimale. Cette erreur de prédiction n'est donc due qu'à l'approximation du modèle d'énergie utilisée par l'algorithme.

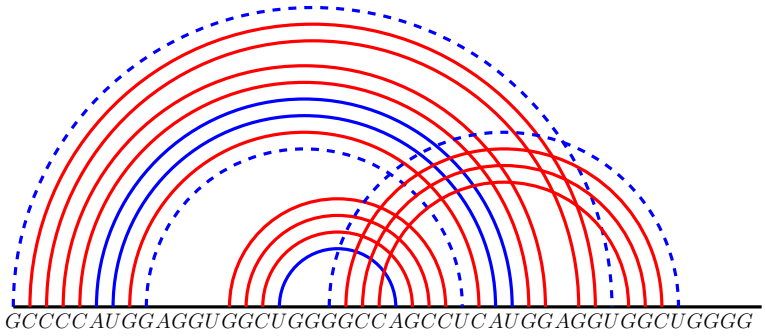
4.2.5 Discussion : une explication commune aux erreurs de prédiction

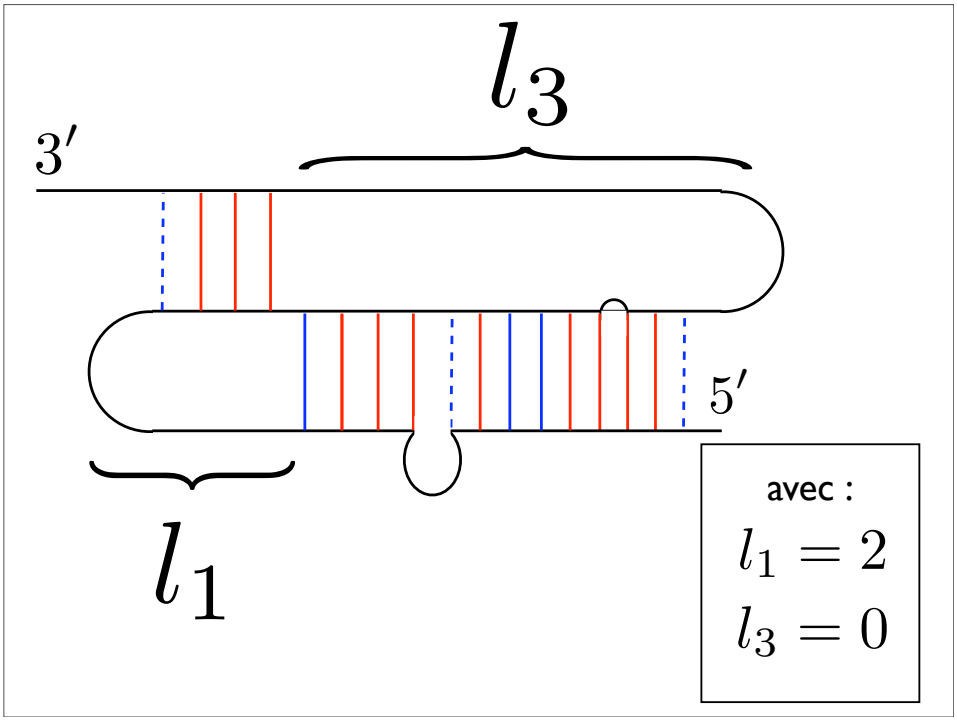
22 séquences sur 36 sont correctement prédites. Parmi les 14 structures imparfaitement prédites, 9 le sont à cause de légères imperfections du modèle d'énergie et sont proches de la structure réelle. Une prédiction diffère de sa structure théorique obtenue par comparaison de séquences mais il est permis de douter de la validité de cette dernière. Enfin 4 prédictions (Ec_alpha, Bt_Prp, TMV, CGMMV) présentent des différences significatives avec les structures expérimentales.

Comment expliquer ces mauvaise prédictions ? L'algorithme de prédiction étant exact, celles-ci sont bien les structures d'énergie libre minimale. Le grand nombre de structures par ailleurs correctement prédites laisse penser que le modèle utilisé M_1 n'est pas un coupable si facile pour expliquer ces échecs. Ceci est d'autant plus vrai que les structures incorrectement prédites font peu appel aux paramètres que ce modèle estime mal.
Alors, y-a-t-il une autre raison ?

Je propose une explication remettant en cause une "hypothèse cachée" dans toute la théorie du repliement de l'ARN, stipulant qu'une structure secondaire d'ARN réelle est représentée par un diagramme, de sorte que la prédiction d'une structure particulière revient à prédire son diagramme. Cette affirmation n'est absolument correcte que si la réciproque est vérifiée, à savoir : "*à tout diagramme correspond une structure secondaire réelle*". L'examen des fausses prédictions de TT2NE pour les séquences TMV, Bt_PrP, CGMMV et Ec_alpha amène à penser que ce n'est pas le cas : la représentation en diagramme néglige les *contraintes stériques* auxquelles est soumise une structure secondaire réelle. Un certain repliement dessiné sur un diagramme peut en fait être irréalisable par un ARN dans l'espace à trois dimensions. Contrairement à ce que j'ai fait jusqu'ici, il faut distinguer les notions de "meilleure structure" et "meilleur diagramme".

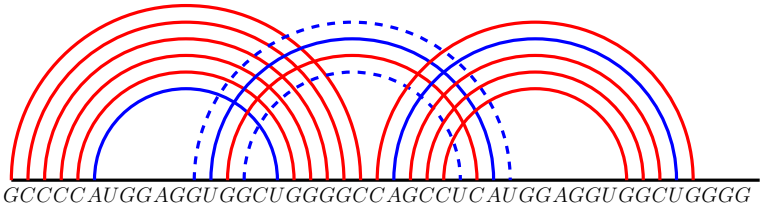
Essayons de dessiner le pseudo-nœud représenté par “le meilleur diagramme” calculé pour la séquence Bt_PrP :

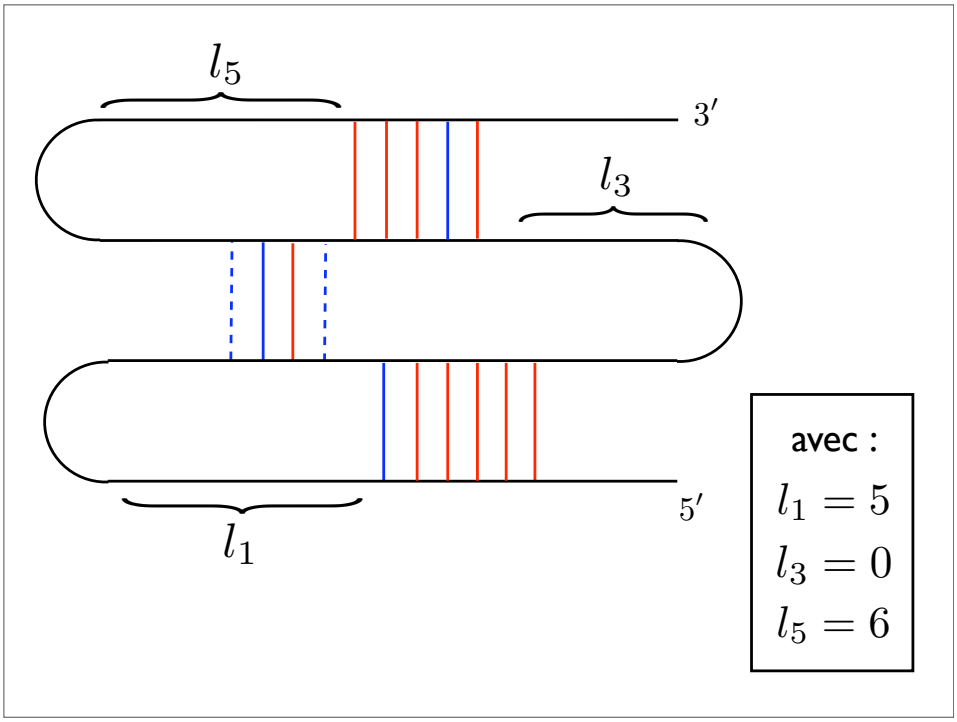




Le pseudo-nœud H dessiné sur le diagramme est clairement irréalisable en pratique. La longueur de ses hélices est trop grande par rapport aux longueurs l_1 et l_3 des segments de bases libres qui les rejoignent. l_1 ne contient que deux bases libres et l_3 n'en contient pas.

Examinons maintenant le second meilleur diagramme :



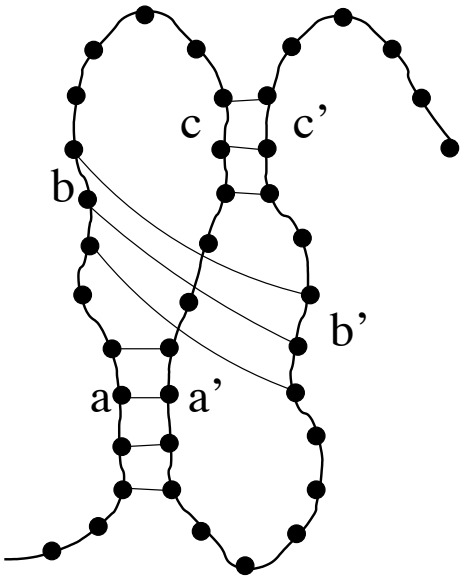


Là encore il est vraisemblable que ce pseudo-nœud ne soit pas réalisable à cause de la contrainte $l_3 = 0$. Ainsi donc, dans le cas de Bt_PrP, les deux meilleurs diagrammes proposés par l'algorithme ne correspondent pas à des structures secondaires réelles. La prédiction à retenir est donc le troisième meilleur diagramme qui correspond bien à la structure secondaire réelle. Pour Bt_PrP, la meilleure structure prédite est correcte mais elle ne correspond qu'au troisième meilleur diagramme.

En les dessinant, il est très clair que les pseudo-nœuds prédits pour TMV et CGMMV ne sont pas réalisables et cela est également très vraisemblable pour le KH prédit pour Ec_alpha. Dans tous les cas, réaliser ces topologies implique de diminuer la taille des hélices pour avoir plus de bases libres dans les segments qui les rejoignent, ce qui signifie aussi augmenter leur énergie. On peut raisonnablement espérer qu'à ce jeu la structure réelle devienne à chaque fois la structure d'énergie minimale.

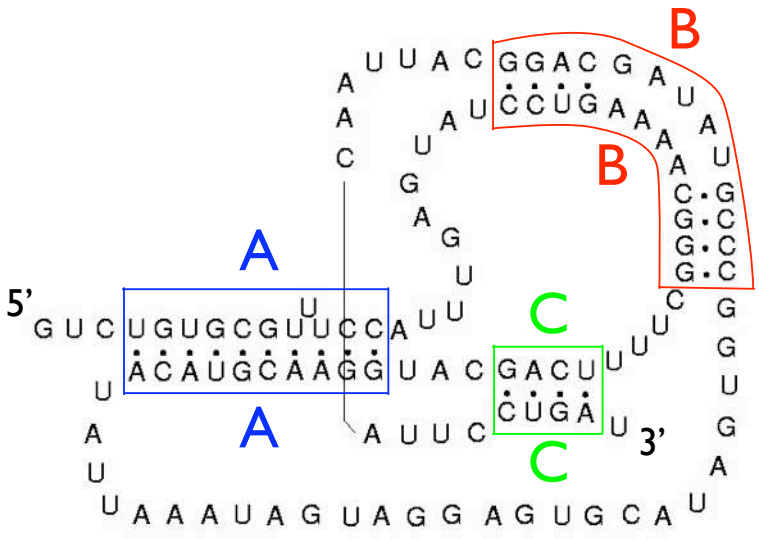
La présence de contraintes stériques est la raison pour laquelle j'ai interdit à TT2NE la formation de pseudo-nœuds ABCABC, comme précisé en début de section. TT2NE a en effet tendance à prédire beaucoup de ces pseudo-nœuds alors que leur topologie particulière est très exigeante en bases libres.

Il n'est pas aisé de se représenter mentalement ce pseudo-nœud. Dans l'article [83], il est représenté ainsi :



Malheureusement, ce schéma est faux puisque les appariements entre b et b' sont parallèles, alors qu'ils devraient être *anti-parallèles*. Pour réaliser ce pseudo-nœud, l'ARN doit effectuer plus de "contorsions".

Dans l'article où il a été originellement présenté [69], ce pseudo-nœud ABCABC est représenté ainsi :



où on observe que seulement 41% de bases sont appariées. Il y a en particulier un long segment de bases libres, en comprenant 26, qui relie les hélices A et B.

J'ai observé une nette amélioration de la qualité de prédiction en interdisant la formation de ce pseudo-nœud et ceci doit être vu comme une grossière mais efficace prise en compte des contraintes stériques.

L'absence de contraintes stériques est donc la principale source d'erreurs de TT2NE et cette question doit faire l'objet de futures recherches. Celles-ci peuvent se développer selon deux axes :

1. *L'étude exhaustive des pseudo-nœuds primitifs de genre 1.*
   En étudiant les bases de données et en utilisant des logiciels de modélisation 3D ou de dynamique moléculaire, il est possible de déterminer exactement sous quelles conditions stériques peuvent se former les pseudo-nœuds les plus simples. Ces conditions peuvent être alors écrites dans des tables auxquelles se réfère l'algorithme de repliement. Concernant les pseudo-nœuds H qui ne mettent en jeu que deux hélices, ces conditions peuvent être directement codées dans le graphe

en reliant par un arc rouge deux hélices formant un pseudo-nœud H stériquement impossible.

2. *Le paramétrage des pseudo-nœuds primitifs de genre supérieur.*
   La partie précédente a montré que les pseudo-nœuds primitifs de genre strictement supérieur à 1 sont beaucoup plus rares et apparaissent généralement à partir de séquences longues de plus de 200 bases, bien qu'il y ait des contre-exemples comme le HDV ("*hepatitis delta virus*"). Cette observation peut-être reproduite en affectant des pénalités particulières selon que des pseudo-nœuds de genre strictement supérieur à 1 sont primitifs ou non. Deux pseudo-nœuds de genre 1 agencés en série seraient moins pénalisés qu'un pseudo-nœud primitif de genre 2.

La prise en charge de ces futures contraintes se rajoute facilement à TT2NE, sous la forme d'un test adéquat entre les étapes (2b) et (3) de la procédure *Explore2*. Ce nouveau test, en plus de rendre les résultats meilleurs, permettra aussi d'accélérer l'algorithme puisqu'il évitera le parcours de nombreux sous-graphes.

Analyse des résultats de HotKnots

L'examen des erreurs de HotKnots montre clairement la source de ses insuffisances : 16/20 des structures incorrectement prédites par HotKnots ne contiennent en fait pas de pseudo-nœud. Par exemple, les structures prédites pour AMV3, BBMV, CoxB3 sont les structures secondaires réelles privées d'une hélice constituant le pseudo-nœud. La structure prédite pour TMV ne contient aucun pseudo-nœud alors qu'elle est composée de trois pseudo-nœuds H en série. TT2NE prédit pourtant correctement beaucoup de ces structures lorsqu'utilisé avec le modèle INN-HB qui n'est qu'un sous-ensemble de Turner99 utilisé par HotKnots. Cette différence vient du paramétrage des pseudo-nœuds : tandis que TT2NE utilise une pénalité $\mu = 0.9u$ pour pénaliser le genre, HotKnots affecte une pénalité $p = 3u$ à chaque croisement d'hélices observé sur le diagramme. Lorsque le pseudo-nœud considéré est un simple pseudo-nœud H, μ et p jouent le même rôle. Pourquoi les auteurs ont-ils décidé de choisir une pénalité si élevée alors qu'une plus faible aurait ainsi donné de meilleurs résultats sur plusieurs cas ?

Les auteurs suivent le paramétrage adopté dans [82]. Dans cet article, Dirks et Pierce utilisent Turner99 comme modèle d'énergie et établissent la valeur de p comme le meilleur compromis entre deux impératifs :

1. Prédire correctement des pseudo-nœuds.
2. Ne pas induire de pseudo-nœud dans des structures n'en comprenant pas.

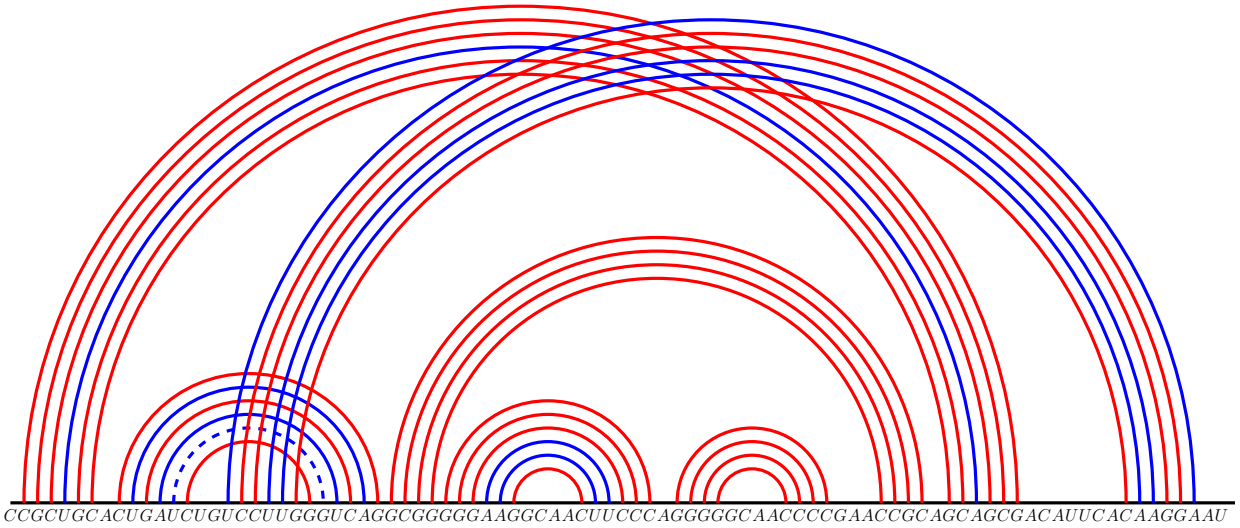
Le premier point utilise la Pseudobase et contraint p positivement. Le second s'appuie sur une sélection *aléatoire* de 200 séquences d'ARNt et contraint p négativement. Les auteurs ont considéré p comme trop petit si les structures prédites pour ces ARNt contiennent un pseudo-nœud.

Cette méthodologie est incorrecte car les auteurs ne se sont pas assurés que Mfold peut prédire la structure secondaire des ARNt sélectionnés. Si Mfold le peut, alors la prédiction erronée d'un pseudo-nœud ne serait effectivement due qu'au paramètre p. Si Mfold ne peut pas prédire la meilleure structure secondaire sans pseudo-nœud, alors la prédiction d'un pseudo-nœud n'est pas forcément le fait d'un mauvais paramétrage de p. Elle peut simplement être une nouvelle façon de se tromper qui reflète une défaillance générale du modèle d'énergie. Il n'y a aucune raison de préférer une prédiction fausse sans pseudo-nœud à une prédiction fausse avec pseudo-nœud. Ainsi, en attribuant à p la responsabilité de toute prédiction erronée de pseudo-nœud, les auteurs induisent une contrainte négative excessive qui tend à surestimer cette pénalité.

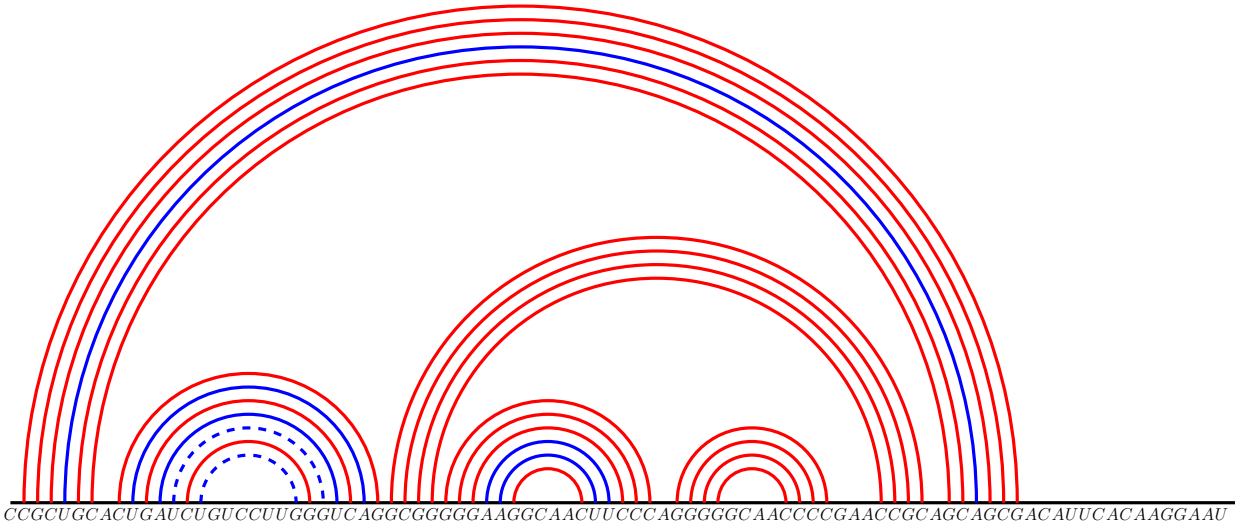
J'ai vérifié que le choix de $\mu = 0.9u$ n'induisait pas de pseudo-nœud dans une dizaine de séquences d'ARNt correctement prédites par un algorithme de repliement sans pseudo-nœud.

Un avantage d'utiliser une pénalité topologique

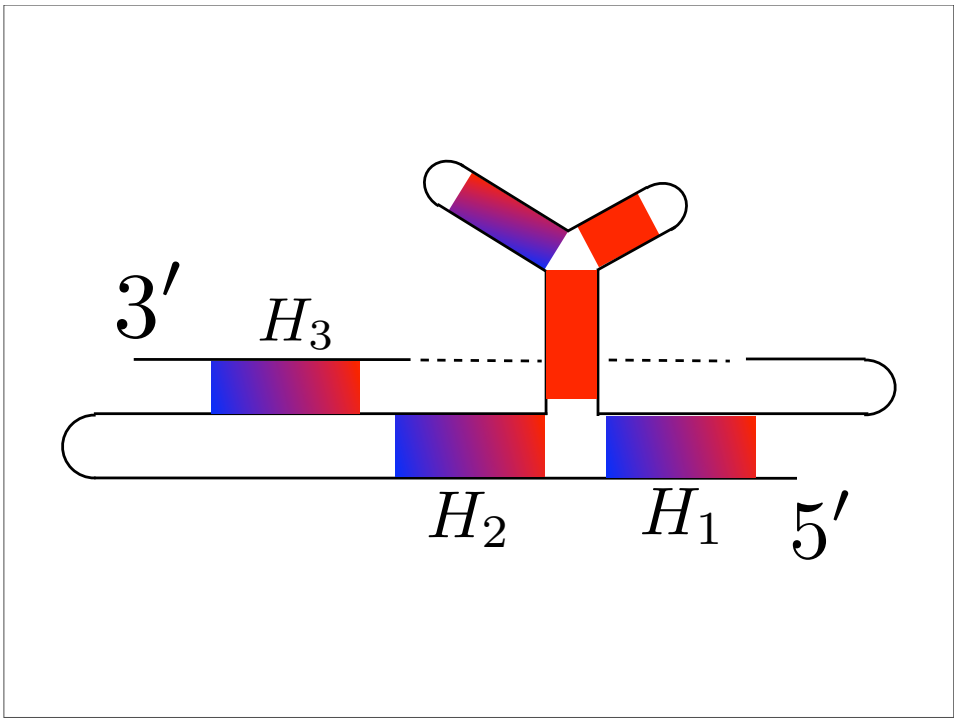
En comparant les prédictions de HotKnots et de TT2NE avec INN-HB, l'exemple de Bp_PK2 [95] a un intérêt particulier. Son diagramme, correctement prédit par TT2NE, est :



tandis que la structure prédite par Hotknots est :



HotKnots échoue à prédire une hélice formant un pseudo-nœud. Cependant, ce n'est, dans ce cas, pas à cause d'une pénalité p excessive car p est fixée à 3u tandis que l'hélice manquante a une énergie de -5u. Il serait donc favorable de la créer. La source de l'erreur de HotKnots est que la pénalité appliquée n'est pas p mais $2p$ car ajouter l'hélice manquante entraîne deux intersections sur le diagramme. Le genre, quant à lui, vaut 1 ce qui signifie que ce diagramme a en fait une des quatre topologies de genre 1 et que la pénalité μ ne s'applique qu'une fois. Voici un schéma donnant une idée de la structure 3D de Bp_PK2 :



Avec ce schéma, il est apparent que ce pseudo-nœud a la topologie ABAB. Les hélices H_1 et H_2 peuvent être vues comme une seule hélice imparfaite, interrompue par une boucle multi-hélices. Une fois qu'une de ces deux hélices est formée, former la deuxième ne revient pas à créer un pseudo-nœud supplémentaire et il est logique de n'appliquer qu'une seule fois μ ou p. La complexité d'un pseudo-nœud ne se relie pas systématiquement au nombre de croisements d'hélices sur un diagramme : la description topologique a ainsi l'avantage de détecter les hélices *équivalentes*.

Analyse des résultats de McQfold

Le principe de McQfold repose sur une description des repliements d'ARN en termes de grammaire. Dans cette approche, l'équivalent du modèle d'énergie est la donnée de probabilités associées aux différentes règles de la grammaire. De ce point de vue, le modèle utilisé par McQfold est beaucoup plus simple que M_1 puisqu'il ne fait intervenir que 15 paramètres. Il n'y a par exemple pas de terme de dipaires et seules les simples paires de bases sont prises en compte. La principale source d'erreur de McQfold réside ainsi dans ce modèle simplifié. Beaucoup d'hélices contenant des renflements ou des boucles internes ne sont pas prédites. On observe également que la contrainte stérique interdisant aux têtes d'épingle de contenir moins de trois bases n'a pas été implémentée.

Il est cependant remarquable que McQfold donne en moyenne de meilleurs résultats que HotKnots sur la base de données testée. Cette performance s'explique encore par la grande difficulté de HotKnots à prédire des pseudo-nœuds car la perte de sensibilité qui s'ensuit est plus dommageable sur cette base de données que la difficulté de McQfold à prédire des boucles internes et renflements. Il est clair que McQfold sera bien plus performant lorsqu'il pourra prendre en charge un modèle d'énergie complet.

Conclusion

Ainsi donc, TT2NE améliore clairement l'état de l'art. Il permet de prédire n'importe quelle topologie de pseudo-nœud tout en garantissant de trouver le minimum d'énergie libre. Le nouveau paramétrage des pseudo-nœuds de nature topologique s'avère pertinent. Au-delà des succès de TT2NE, on s'aperçoit que les prédictions erronées ont le même genre que les structures réelles ce qui permet de dire que le genre permet à TT2NE de cibler la bonne catégorie de structures secondaires. Toutes les erreurs ne découlant pas des faiblesses du modèle d'énergie s'expliquent par l'absence de contraintes stériques dans la représentation diagrammatique des structures secondaires. Le principe de la minimisation d'énergie libre n'est en particulier remis en cause par aucun des tests effectués. Les résultats de TT2NE suggèrent plutôt d'approfondir la prise en charge des contraintes stériques.

TT2NE souffre principalement de deux limites qui doivent également faire l'objet de travaux futurs :

- TT2NE ne peut pas calculer de fonction de partition

- La taille des séquences pouvant être traitées par TT2NE est limitée. En effet, l'exécution de l'algorithme prend plusieurs heures pour des séquences de 150 bases de long sur un processeur simple. Cette limite de taille peut être repoussée en parallélisant le programme car, la routine de TT2NE consistant en l'exploration de sous-graphes d'un graphe construit une fois pour toute en début de programme, rien ne s'oppose à ce que cette recherche soit partagée entre plusieurs processeurs. La qualité de prédiction de TT2NE sur des séquences plus longues sera néanmoins toujours tributaire de la qualité du modèle d'énergie utilisé.

Chapitre 5

Dynamique moléculaire

Comme mentionné dans le chapitre “Modèle d'énergie libre”, j'ai entrepris d'effectuer des simulations de dynamique moléculaire pour calculer les différents paramètres du modèle d'énergie libre avec, en premier lieu, les énergies libres de dipaires. Afin de tester la validité des résultats, j'ai mis en place un test de cohérence interne qui a révélé que les nombreuses problématiques de la dynamique moléculaire n'ont pas toutes été résolues dans mes travaux.

Le but de ce chapitre est de rapporter l'idée que j'ai utilisée pour calculer les paramètres du modèle et de présenter le plus clairement possible la difficulté à maîtriser cet outil de dynamique moléculaire. Au-delà de la difficulté purement technique liée à la manipulation d'un grand nombre d'algorithmes compliqués, concevoir une simulation soulève également des difficultés conceptuelles qui seront exposées.

Le plan suivi dans ce chapitre est :

I) Introduction à la dynamique moléculaire
Cette introduction présente le principe général de la dynamique moléculaire et certaines de ses nécessaires approximations

II) Le calcul d'énergie libre en dynamique moléculaire
Cette partie présente différentes méthodes de calcul d'énergie libre. Une d'entre elles, l'intégration thermodynamique, sera retenue pour le calcul des paramètres du modèle d'énergie

III) Application de l'intégration thermodynamique au calcul de l'énergie libre d'appariement de deux brins d'ARN
Cette partie détaille le principe de la simulation que j'ai conçue pour calculer l'énergie libre d'appariement de deux brins d'ARN

IV) Détails techniques de la simulation
Cette partie traite des nombreux réglages requis par la dynamique moléculaire

V) Résultats
Cette partie expose brièvement quelques résultats

5.1 Introduction à la dynamique moléculaire

La dynamique moléculaire est un outil puissant en constant développement dont le but est la simulation du comportement de molécules représentées au niveau atomique selon les équations de la mécanique classique. Elle permet l'étude de propriétés d'équilibre, comme la valeur moyenne de l'énergie potentielle d'un système ou la distance moyenne entre deux atomes d'intérêt mais aussi d'aspects de dynamique hors d'équilibre comme la viscosité d'un fluide ou la cinétique de la fixation d'un ligand sur une protéine. Elle permet d'avoir accès à des détails atomiques encore trop fins pour être mesurés par l'expérience et permet à ce titre de donner des éclairages pionniers sur le fonctionnement des biopolymères du vivant.

5.1.1 Approximations de la dynamique moléculaire

L'emploi de la mécanique classique nécessite quelques approximations pour représenter des phénomènes d'origine quantique cruciaux, parmi lesquels on peut citer la liaison hydrogène et la liaison covalente. Ces phénomènes jouant un rôle primordial dans la structuration des biopolymères, il faut donc les formuler de manière à ce que la dynamique moléculaire puisse rendre compte de leurs effets. Les solutions communément adoptées distinguent "interactions liées" (comme la liaison covalente) de "interactions non-liées" (comme l'interaction coulombienne).

Interactions liées

Les liaisons covalentes ne pouvant émerger naturellement d'une simulation en mécanique classique, celles-ci doivent être spécifiées en amont par l'utilisateur. Il incombe à l'utilisateur de fournir la liste des paires d'atomes liés de manière covalente et cette liste ne sera pas modifiée lors de l'exécution du programme. Autrement dit, l'utilisateur fixe une fois pour toutes la topologie des molécules de son système avant de lancer la simulation. Cela signifie en particulier qu'on ne peut pas simuler de réaction chimique par dynamique moléculaire car décrire par exemple la simple réaction $H^+ + OH^- \rightarrow H_2O$ demanderait la création d'une nouvelle liaison covalente. Une fois définie la topologie d'une molécule, sa géométrie est contrainte par l'introduction de potentiels portant sur les liaisons et les angles destinés à imiter le plus fidèlement les comportements observés expérimentalement.

Interactions non liées

A l'impossibilité de traiter adéquatement des interactions de nature quantique s'ajoute une contrainte informatique : les seules interactions permises sont celles mettant en jeu deux et seulement deux atomes car autoriser les interactions à plus de 3 atomes seraitaccompagné d'un coût supplémentaire de temps de calcul dissuasif. Cette condition exclut entre autres les interactions entre dipôles.

Les effets des interactions non liées sont capturés en jouant sur la calibration des paramètres des atomes en interaction, comme par exemple leur charge électrique.
Citons-en deux conséquences qui peuvent paraître surprenantes au premier abord :

- Dans le modèle d'eau populaire TIP3P, la charge de l'atome d'oxygène est fixée à -0.834e et celle des atomes d'hydrogène à +0.417e, au lieu de -2e et +1e.
- Dans les modèles d'ARN, au sein d'un même nucléotide, des atomes d'oxygène peuvent ne pas avoir les mêmes paramètres, selon qu'ils participent ou non à la liaison hydrogène que doit pouvoir former le nucléotide avec son complémentaire Watson-Crick.

5.1.2 Aspects techniques généraux

Une configuration du système étudié est la donnée de l'ensemble des positions et des impulsions $\{\boldsymbol{r_i}, \boldsymbol{p_i}\}$ des atomes i qui le constituent. Le but de la dynamique moléculaire est de faire évoluer ces positions et impulsions selon les équations du mouvement afin d'engendrer une suite de configurations dans le temps représentant fidèlement les états possibles du système. Cette suite de configurations dans le temps est appelée "trajectoire".

Les ingrédients de la dynamique moléculaire sont :

- un hamiltonien définissant les interactions entre les atomes d'un système.
- une collection d'algorithmes permettant de faire évoluer le système selon ce hamiltonien et d'analyser cette évolution.

Collection d'algorithmes

Mon choix s'est porté sur la collection GROMACS (GROningen Machine of Chemical Simulation, www.gromacs.org, [96]), du fait de sa rapidité, de son ergonomie, de sa gratuité et du dynamisme de la communauté des utilisateurs de GROMACS qui répondent rapidement aux questions posées dans les forums. J'ai utilisé la dernière version disponible au moment où j'ai effectué ces travaux, à savoir GROMACS 3.3.

Description du hamiltonien

Le hamiltonien utilisé par GROMACS a une forme standard. Ses variables sont les vecteurs position $\boldsymbol{r_i}$ et impulsion $\boldsymbol{p_i}$. Chaque atome est représenté par 4 paramètres : sa masse m_i, sa charge électrique q_i et deux paramètres $C_i^{(6)}$ et $C_i^{(12)}$ intervenant dans le potentiel Lennard-Jones décrit plus loin.

Notons $\boldsymbol{r_{ij}} = \boldsymbol{r_j} - \boldsymbol{r_i}$ et $r_{ij} = |\boldsymbol{r_{ij}}|$. On a :

$$
\begin{aligned}
\mathcal{H}(\{\boldsymbol{r_i}, \boldsymbol{p_i}\}) = & \sum_i E_c(\boldsymbol{p_i}) \\
& \left.\begin{aligned} &+ \sum_{i,j} V_{LJ}(\boldsymbol{r_{ij}}) \\ &+ \sum_{i,j} V_{Coulomb}(\boldsymbol{r_{ij}}) \end{aligned}\right\} \text{interactions non liées} \\
& \left.\begin{aligned} &+ \sum_{i,j} V_{cov}(\boldsymbol{r_{ij}}) \\ &+ \sum_{i,j,k} V_{angle}(\boldsymbol{r_{ij}}, \boldsymbol{r_{jk}}) \\ &+ \sum_{i,j,k,l} V_{torsion}(\boldsymbol{r_{ij}}, \boldsymbol{r_{jk}}, \boldsymbol{r_{kl}}) \\ &+ \sum_{i,j,k,l} V_{dihedre}(\boldsymbol{r_{ij}}, \boldsymbol{r_{ik}}, \boldsymbol{r_{il}}) \end{aligned}\right\} \text{interactions liées}
\end{aligned}
$$

- $E_c(\boldsymbol{p_i})$ est l'énergie cinétique de l'atome i :

$$E_c(\boldsymbol{p_i}) = \frac{p_i^2}{2m_i}$$

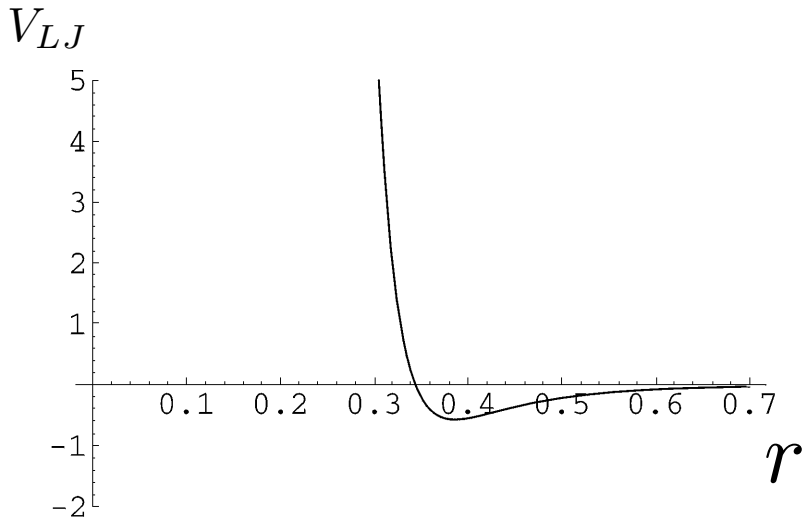
- $V_{Coulomb}$ est le potentiel de l'interaction coulombienne :

$$V_{Coulomb}(\boldsymbol{r_{ij}}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

- V_{LJ} est un potentiel de Lennard-Jones qui modélise le principe de répulsion de Pauli à courte distance par un terme en r_{ij}^{-12} et l'interaction attractive de Van der Waals à grande distance par un terme en r_{ij}^{-6} :

$$V_{LJ}(\boldsymbol{r_{ij}}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \quad \text{où} \quad C_{ij}^{(k)} = \sqrt{C_i^{(k)} C_j^{(k)}}$$

Ce potentiel a un minimum en $r_{min} = (\frac{2C_{ij}^{(12)}}{C_{ij}^{(6)}})^{\frac{1}{6}}$



- V_{cov} est le potentiel reproduisant la liaison covalente. La liaison covalente est de nature quantique et ne peut être idéalement prise en charge en dynamique moléculaire. Elle est imitée par un potentiel harmonique qui la maintient à sa longueur convenable, notée b^{cov}. Pour des atomes i et j, on a :

$$V_{cov}(\boldsymbol{r_{ij}}) = \frac{1}{2} k_{ij}^{cov} (r_{ij} - b_{ij}^{cov})^2$$

- V_{angle} est le potentiel contraignant l'angle entre deux liaisons successives ij et jk. Là encore, un potentiel harmonique est couramment utilisé :

$$V_{angle}(\boldsymbol{r_{ij}}, \boldsymbol{r_{jk}}) = \frac{1}{2} k_{ijk}^{a} (\theta_{ijk} - \theta_{ijk}^{0})^2 \quad \text{où} \quad \theta_{ijk} = \arccos\left(\frac{\boldsymbol{r_{ij}} . \boldsymbol{r_{jk}}}{r_{ij} r_{jk}}\right)$$

- $V_{torsion}$ contraint l'angle de torsion entre trois liaisons successives ij, jk et kl.

$$V_{torsion}(\boldsymbol{r_{ij}}, \boldsymbol{r_{jk}}, \boldsymbol{r_{kl}}) = k_t^{ijkl}(\ 1 + cos(n\phi_{ijkl} - \phi_0)\ )$$
où ϕ_{ijkl} est l'angle entre les plans ijk et jkl

- $V_{dihedre}$ est un potentiel harmonique utilisé pour contraindre la géométrie de trois liaisons ij, jk et jl ayant en commun l'atome j :

$$V_{dihedre}(\boldsymbol{r_{ij}}, \boldsymbol{r_{jk}}, \boldsymbol{r_{jl}}) = \frac{1}{2} k_d (\phi_{ijkl} - \phi_0)^2$$
où ϕ_{ijkl} est l'angle entre les plans ijk et jkl

Principe général de l'algorithme de dynamique moléculaire

A partir d'un jeu de conditions initiales, l'algorithme de dynamique moléculaire effectue des mises à jours successives des positions et des impulsions du système. Ces mises à jour sont appelées "pas" et sont associées à l'intervalle de temps τ, généralement de l'ordre de la femtoseconde, utilisé pour discrétiser les équations du mouvement. Un pas comporte deux étapes :

1. Calcul des forces $\boldsymbol{F_i}$ exercées sur l'atome i :

$$\boldsymbol{F_i} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{r_i}}$$

2. Résolution numérique des équations du mouvement :

$$\frac{d\boldsymbol{p_i}}{dt} = \boldsymbol{F_i} \quad \text{et} \quad \frac{d\boldsymbol{r_i}}{dt} = \frac{\boldsymbol{p_i}}{m_i}$$

en utilisant le schéma de discrétisation dit du "saut de grenouille", où les positions sont estimées aux temps discrets de la forme $t_0 + n\tau$ avec n entier, tandis que les impulsions le sont aux temps de la forme $t_0 + (n + \frac{1}{2})\tau$:

$$\begin{cases} \boldsymbol{p_i}(t + \frac{\tau}{2}) = \boldsymbol{p_i}(t - \frac{\tau}{2}) + \tau \boldsymbol{F_i}(t) \\ \boldsymbol{r_i}(t + \tau) = \boldsymbol{r_i}(t) + \tau \boldsymbol{p_i}(t + \frac{\tau}{2}) \end{cases}$$

L'article "*Molecular dynamics simulation of nucleic acids : successes, limitations and promise*" [97] récapitule quelques-uns des succès de la dynamique moléculaire comme l'étude de la courbure de l'ADN ou des interactions de l'ADN avec les ions Na^+ et Cl^-.

5.2 Le calcul d'énergie libre en dynamique moléculaire

Le calcul de différences d'énergie libre est ce qu'on peut faire de plus compliqué à l'heure actuelle par la dynamique moléculaire. Cela implique de trouver des stratégies permettant d'estimer correctement l'entropie, étant donné qu'une énumération exhaustive des configurations possibles du système est hors de portée de la simulation. Je présente dans cette section quatre techniques de calcul d'énergie libre, en expliquant pourquoi les trois premières sont inadéquates et comment la dernière s'applique au problème de mesure des énergies de dipaires. L'idée générale de chacune de ces méthodes est de relier l'énergie libre F à des valeurs moyennes d'observables effectivement calculables par dynamique moléculaire.

Il convient tout d'abord de préciser certaines notations et effectuer quelques rappels de mécanique statistique.

5.2.1 Notations et rappels de mécanique statistique

Fonction de partition

On note $\mathcal{Z}$ la fonction de partition du système :

$$\mathcal{Z} = A \int_{\mathbb{R}^{6N}} (\prod_i d\boldsymbol{r_i} d\boldsymbol{p_i}) e^{-\beta \mathcal{H}(\{\boldsymbol{r_i}, \boldsymbol{p_i}\})}$$

avec A un préfacteur tenant compte des symétries du système, N le nombre d'atomes du système et $\beta = (k_B T)^{-1}$, où k_B est la constante de Boltzmann et T la température. Par commodité, le préfacteur et le domaine d'intégration ne seront plus précisés. Soit $\boldsymbol{R}$ le vecteur à $3N$ composantes des positions des N atomes du système et $\boldsymbol{P}$ celui des impulsions, de sorte que $\mathcal{H}(\{\boldsymbol{r_i}, \boldsymbol{p_i}\})$ se note plus simplement $\mathcal{H}(\boldsymbol{R}, \boldsymbol{P})$. Cependant je n'expliciterai plus systématiquement la dépendance du hamiltonien en ses variables. Notons $d\boldsymbol{R} = \prod_i d\boldsymbol{r_i}$ et $d\boldsymbol{P} = \prod_i d\boldsymbol{p_i}$.

Avec ces conventions, la fonction de partition s'écrit de manière plus concise :

$$\mathcal{Z} = \int d\boldsymbol{R} d\boldsymbol{P} e^{-\beta \mathcal{H}}$$

La fonction de partition "code" les états microscopiques d'un système à l'équilibre et les grandeurs macroscopiques d'intérêt se relient à elle de différentes manières.

Poids de Boltzmann

A l'équilibre, la fréquence d'occurrence d'un état caractérisé par des positions $\boldsymbol{R}^*$ et des impulsions $\boldsymbol{P}^*$ est décrite par sa densité de probabilité :

$$\rho(\boldsymbol{R}^*, \boldsymbol{P}^*) = \frac{e^{-\beta \mathcal{H}(\boldsymbol{R}^*, \boldsymbol{P}^*)}}{\mathcal{Z}}$$

également appelée "poids de Boltzmann".

Grandeurs macroscopiques

Une observable $\mathcal{O}$ du système, expérimentalement mesurable, est la moyenne de la valeur qu'elle prend sur les états microscopiques du système selon la densité de probabilité ci-dessus. On a :

$$\begin{aligned} \mathcal{O}_{exp} = \langle \mathcal{O} \rangle &= \int d\boldsymbol{R} d\boldsymbol{P}\ \mathcal{O}(\boldsymbol{R}, \boldsymbol{P})\ \rho(\boldsymbol{R}, \boldsymbol{P}) \\ &= \int d\boldsymbol{R} d\boldsymbol{P}\ \mathcal{O}\ \frac{e^{-\beta \mathcal{H}}}{\mathcal{Z}} \end{aligned}$$

Energie libre

L'énergie libre F d'un système se relie à sa fonction de partition par la relation :

$$F = -\beta^{-1} \ln \mathcal{Z}$$

Gaz parfait

Le gaz parfait est un modèle important de mécanique statistique car analytiquement résoluble. Dans ce système, les molécules n'interagissent pas, de telle manière que le hamiltonien se réduit à sa composante cinétique. On a alors :

$$\begin{aligned}
\mathcal{H}(\boldsymbol{R},\boldsymbol{P}) &= \sum_i \frac{p_i^2}{2m_i} \\
\mathcal{Z}_{gaz\ parfait} &\propto \int (\prod_i d\boldsymbol{p_i}) e^{-\beta \sum_i \frac{p_i^2}{2m_i}} \\
&\propto \prod_i \int d\boldsymbol{p_i} e^{-\beta \frac{p_i^2}{2m_i}}
\end{aligned}$$

5.2.2 Première méthode de calcul de F : Estimation de $\langle e^{+\beta(\mathcal{H}-E_c)} \rangle$

En notant E_c l'énergie cinétique d'un système, la quantité $\langle e^{+\beta(\mathcal{H}-E_c)} \rangle$ se relie directement à son énergie libre F. On a en effet :

$$\begin{aligned}
\langle e^{+\beta(\mathcal{H}-E_c)} \rangle &= \int d\boldsymbol{R}\ d\boldsymbol{P}\ e^{+\beta(\mathcal{H}-E_c)} \times \frac{e^{-\beta\mathcal{H}}}{\mathcal{Z}} \\
&= \frac{1}{\mathcal{Z}} \int d\boldsymbol{R}\ d\boldsymbol{P}\ e^{-\beta E_c} \\
&= \frac{\mathcal{Z}_{\text{gaz parfait}}}{\mathcal{Z}}
\end{aligned}$$

où $\mathcal{Z}_{\text{gaz parfait}}$ est la fonction de partition du gaz parfait associé au système. Cette fonction de partition est connue analytiquement. On a donc au final :

$$\begin{aligned}
F &= -\beta^{-1} \ln \mathcal{Z} \\
&= \beta^{-1} \ln \langle e^{+\beta(\mathcal{H}-E_c)} \rangle + F_{gaz\ parfait}
\end{aligned} \tag{5.1}$$

Ainsi, le calcul par simulation de $\langle e^{+\beta(\mathcal{H}-E_c)} \rangle$ donne directement la valeur de F. Cette méthode n'est en fait pas utilisable en pratique et n'a qu'un intérêt théorique. En effet, le terme $\langle e^{+\beta(\mathcal{H}-E_c)} \rangle$ est dominé par les configurations où le potentiel $\mathcal{H} - E_c$ est maximal et ce sont également les configurations les plus rares étant donné que leur poids de Boltzmann est proportionnel à $e^{-\beta\mathcal{H}}$. Ces configurations ne sont en général jamais échantillonnées en des temps de simulation raisonnables et cette méthode échoue à donner la bonne valeur de F dans des cas très simples, comme par exemple un système

de 200 molécules d'eau.

On ne connaît en fait pas de méthode pratique permettant de calculer directement l'énergie libre d'un système par la simulation. Il est seulement possible de calculer des différences d'énergie libre lors d'évolution entre deux ou plusieurs états et c'est ce que font les trois méthodes présentées ci-dessous. Ces méthodes diffèrent entre elles par le paramètre de l'évolution : température, potentiel supplémentaire ou variable mécanique.

5.2.3 Dénaturation d'un ARN sous l'effet de la température : la REMD

"*Comment échantillonner des configurations rares à température ambiante mais communes à haute température ?* "

A température donnée, la différence d'énergie libre entre deux états se relie directement au rapport des fréquences auxquelles ils sont observés à l'équilibre. Ainsi, à température ambiante, il est théoriquement possible de mesurer la différence d'énergie libre entre l'état natif d'un ARN double brin et son état dénaturé en comptant le nombre d'occurrences de chacun de ces états dans une simulation suffisamment longue. L'état dénaturé étant très rare à température ambiante, cette procédure est inefficace car de très longs temps de simulations sont en fait requis. Ce comptage direct a été effectué une fois par un groupe disposant de moyens informatiques colossaux [98]. Dans cet article, Sorin *et al.* simulent une courte tête d'épingle à l'équilibre pendant 200 μs à l'aide de 40000 processeurs et observent seulement deux fois cette dénaturation.

L'idée de la REMD ("*replica exchange molecular dynamics*") [99] est d'obtenir plus rapidement des configurations dénaturées à température ambiante, correctement distribuées selon la statistique de Boltzmann, en conduisant en parallèle des simulations du même système à des températures plus élevées et en permutant périodiquement les configurations entre températures. L'état dénaturé est en effet le plus fréquent pour $T > T_d$ où T_d est la température de dénaturation de l'ARN considéré. En notant T_1 la température d'intérêt, la REMD consiste à pratiquer N simulations à N températures différentes $T_1 < T_2 < ... < T_d < ... < T_N$ en brassant les structures de manière à obtenir pour chaque simulation une trajectoire distribuée selon la statistique de Boltzmann relative à sa température. Concrètement, à intervalles de temps réguliers, les configurations obtenues à T_i et T_j sont permutées avec une certaine probabilité $P(i \leftrightarrow j)$ donnée par :

$$P(i \leftrightarrow j) = \min(1, \exp[(\beta_j - \beta_i)(U_i - U_j) + (\beta_j p_j - \beta_i p_i)(V_i - V_j)]) \qquad (5.2)$$

avec $\beta_i = (k_B T_i)^{-1}$ et U_i, p_i et V_i l'énergie, la pression et le volume de la configuration i au moment où l'interversion est tentée.

On peut montrer que ce choix de $P(i \leftrightarrow j)$, qui s'apparente à un critère de Métropolis généralisé, garantit la convergence de l'ensemble des trajectoires vers leur distributions boltzmaniennes.

Après convergence, l'énergie libre d'appariement $\Delta F_{app}(T_i)$ se déduit en calculant la fraction P_i^{app} des configurations où les brins sont appariés dans la trajectoire i et en utilisant :

$$\Delta G_{app}(T_i) = -\beta_i^{-1}[\ln P_i^{app} - \ln(1 - P_i^{app})] \tag{5.3}$$

La REMD permet ainsi d'échantillonner correctement à température ambiante les configurations dénaturées rares. Cependant, son coût de calcul peut paraître excessif à plus d'un titre.

Désapparier deux brins d'ARN complémentaires est la simulation par REMD la plus simple que nous puissions imaginer pour calculer des énergies de dipaires. Pour observer la dénaturation de ce système, il est nécessaire de choisir une température maximale T_N supérieure à la température de dénaturation T_d à laquelle se séparent les brins. Il est notoire que les modèles d'énergie actuels de biopolymères surestiment largement cette température T_d car ils ne sont optimisés que pour des températures proches de la température ambiante : ceci accroît en premier lieu l'intervalle de température à couvrir. De plus, pour des ARNs double brin de quelques paires de bases, T_d est en général bien supérieure à 100°C, température à laquelle l'eau bout sous pression atmosphérique. Les modèles d'eau actuels ne permettent pas de rendre compte d'une transition liquide-vapeur, ce qui invalide la vraisemblance physique des trajectoires simulées à ces niveaux de températures. Il faut donc les considérer uniquement comme de coûteux intermédiaires de calculs nécessaires à l'obtention de la distribution de configurations recherchée à température ambiante : il n'y a aucune information thermodynamique fiable à en tirer.

En outre, pour pouvoir prétendre avoir engendré à température ambiante une trajectoire représentative comportant des configurations où les brins d'ARN sont séparés, il faut que le brassage entre les différents niveaux de température soit bon. Il est d'usage de chercher à obtenir des probabilités d'échange $P(i \leftrightarrow i+1)$ de l'ordre de 25% et cela fixe en retour le nombre de trajectoires intermédiaires à simuler entre la température minimale et la température maximale.

En effet, on peut estimer la valeur de $P(i \leftrightarrow i+1)$ en remplaçant les énergies par leurs valeurs moyennes :

$$\begin{aligned}(\beta_{i+1} - \beta_i)(U_i - U_{i+1}) &\approx (\beta_{i+1} - \beta_i)(\langle U \rangle_{\beta_i} - \langle U \rangle_{\beta_{i+1}}) \\ &\approx -(\beta_{i+1} - \beta_i)^2 \left.\frac{d\langle U \rangle}{d\beta}\right|_{\beta_i} \end{aligned} \tag{5.4}$$

Et en première approximation, en notant N_{ddl} le nombre de degrés de libertés du système :

$$\langle U \rangle_{T_i} \approx -N_{ddl} k_B T_i = -\frac{N_{ddl}}{\beta_i} \tag{5.5}$$

$$\left.\frac{d\langle U \rangle}{d\beta}\right|_{\beta_i} \approx \frac{N_{ddl}}{\beta_i^2} \tag{5.6}$$

Pour que la quantité (5.4) soit d'ordre 1, il faut donc avoir

$$(\beta_{i+1} - \beta_i) \propto \frac{\beta_i}{\sqrt{N_{ddl}}} \tag{5.7}$$

soit :

$$(T_{i+1} - T_i) \propto \frac{T_i}{\sqrt{N_{ddl}}} \tag{5.8}$$

Le nombre de températures intermédiaires requis est donc de l'ordre de $\sqrt{N_{ddl}}$, ce qui limite encore la taille des systèmes pouvant être efficacement pris en charge par la REMD.

Ainsi, la principale limite de cette méthode est son coût de calcul.

Considérons l'exemple étudié dans l'article *"Molecular Dynamics Simulation of the Structure, Dynamics, and Thermostability of the RNA Hairpins uCACGg and cUUCGg"* [100], où la REMD est utilisée de manière canonique. Dans cet article, le système étudié est une courte tête d'épingle de quatre bases de long fermée par une paire Wobble ou Watson-Crick, en solution. Les auteurs ont trouvé une température de dénaturation d'environ 400K et ont eu recours pour leur étude à 48 simulations en parallèle couvrant une plage de température allant de 297K à 495K. La convergence a été jugée satisfaisante au bout d'un temps simulé de 30 ns pour chaque trajectoire, ce qui représente un temps simulé cumulé de 1,44 μs.

Pour mesurer une énergie de dipaire par REMD, il me paraît inévitable de considérer un système de deux brins complémentaires contenant la dipaire d'intérêt. Afin que la mesure ne soit pas biaisée par des effets de bords, la dipaire d'intérêt doit être située au centre de l'ARN double brin et chaque brin devrait au moins avoir une longueur de 6 bases. Ainsi, ce système minimal (deux brins d'ARN complémentaires de 6 bases de long) est déjà bien plus stable que la tête d'épingle étudiée dans l'article. Sa température de dénaturation sera plus élevée et sa simulation nécessitera plus de températures intermédiaires. Etant donné qu'il y a 21 énergies de dipaires différentes à mesurer, il faudra un temps de simulation cumulé de l'ordre de la milliseconde et je ne dispose pas de la puissance de calcul nécessaire. Il est en particulier inconcevable d'aller plus loin en cas de succès, à savoir mesurer des énergies associées aux renflements et courtes boucles internes.

En conclusion, la REMD est mal adpatée au problème de la mesure de l'énergie libre de dipaires. Sa finalité est d'obtenir des trajectoires distribuées selon une statistique correcte à différentes températures et elle est notamment utilisée pour étudier des changements conformationnels subtils dépendant de la température. Quand seule la différence d'énergie libre à une température donnée nous intéresse, le coût à payer est trop élevé.

5.2.4 Perturbation du système sous l'effet d'un nouveau potentiel : le "*umbrella sampling*"

Comment échantillonner un sous-espace de configurations rares caractérisé par un certain paramètre d'orde χ ?

Soit un paramètre d'ordre $\chi(\boldsymbol{R}, \boldsymbol{P})$. On cherche à déterminer l'énergie libre associée au sous-espace de configurations où ce paramètre prend une valeur χ_1. Cette énergie libre $F(\chi_1)$ s'exprime :

$$\begin{aligned} F(\chi_1) &= -\beta^{-1} \ln \mathcal{Z}(\chi_1) \\ &= -\beta^{-1} \ln \int d\boldsymbol{R}\, d\boldsymbol{P}\, \delta(\chi(\boldsymbol{R}, \boldsymbol{P}) - \chi_1) e^{-\beta H(\boldsymbol{R}, \boldsymbol{P})} \\ &= -\beta^{-1} \ln \mathcal{Z} - \beta^{-1} \ln \int d\boldsymbol{R}\, d\boldsymbol{P}\, \delta(\chi - \chi_1) \frac{e^{-\beta H}}{\mathcal{Z}} \\ &= F - \beta^{-1} \ln \langle \delta(\chi - \chi_1) \rangle \end{aligned} \tag{5.9}$$

Si χ_1 est éloigné de la valeur d'équilibre $\langle \chi \rangle$, alors il est problématique d'estimer le terme $\langle \delta(\chi - \chi_1) \rangle$ par la simulation, les configurations correspondantes étant rares. Pour ce faire, l'idée de l'*umbrella sampling* [101] consiste à introduire un nouveau potentiel qui favorise les trajectoires présentant des valeurs de χ proches de χ_1 : par exemple $V(\chi) = K(\chi - \chi*)^2$, avec $K > 0$ et $\chi*$ proche de χ_1. Le hamiltonien devient $\mathcal{H}' = \mathcal{H} + V$ et une simulation avec ce nouveau potentiel permet de calculer la variable d'intérêt en remarquant que :

$$\begin{aligned} \langle \delta(\chi - \chi_1) \rangle_{\mathcal{H}} &= \frac{\int d\boldsymbol{R}\, d\boldsymbol{P}\, \delta(\chi - \chi_1) e^{-\beta \mathcal{H}}}{\int d\boldsymbol{R}\, d\boldsymbol{P}\, e^{-\beta \mathcal{H}}} \\ &= \frac{\int d\boldsymbol{R}\, d\boldsymbol{P}\, \delta(\chi - \chi_1) e^{-\beta(\mathcal{H} + V - V)}}{\int d\boldsymbol{R}\, d\boldsymbol{P}\, e^{-\beta(\mathcal{H} + V)}} \frac{\int d\boldsymbol{R}\, d\boldsymbol{P}\, e^{-\beta(\mathcal{H} + V)}}{\int d\boldsymbol{R}\, d\boldsymbol{P}\, e^{-\beta(\mathcal{H} + V - V)}} \\ &= \frac{\langle \delta(\chi - \chi_1) e^{+\beta V} \rangle_{\mathcal{H}'}}{\langle e^{+\beta V} \rangle_{\mathcal{H}'}} \\ &= \frac{e^{+\beta V(\chi_1)} \langle \delta(\chi - \chi_1) \rangle_{\mathcal{H}'}}{\langle e^{+\beta V} \rangle_{\mathcal{H}'}} \end{aligned} \tag{5.10}$$

Le numérateur est dorénavant accessible par la simulation mais le problème s'est reporté sur le dénominateur comme on peut s'en convaincre par le même argument que donné en 5.2.2 : $\langle e^{+\beta V}\rangle_{\mathcal{H}'}$ est dominé par les configurations où le potentiel V est grand, alors que celles-ci sont rares car leur poids de Boltzmann est proportionnel à $e^{-\beta\mathcal{H}'} = e^{-\beta\mathcal{H}}e^{-\beta V}$. Cependant, le dénominateur –indépendant de χ_1– peut-être éliminé en considérant une autre valeur du paramètre d'ordre χ_2 bien échantillonnée par le même potentiel. En effet on a en utilisant (5.9) et (5.10) :

$$\begin{aligned} F(\chi_2) - F(\chi_1) &= -\beta^{-1} \ln \frac{\langle\delta(\chi-\chi_2)\rangle_{\mathcal{H}}}{\langle\delta(\chi-\chi_1)\rangle_{\mathcal{H}}} \\ &= V(\chi_1) - V(\chi_2) - \beta^{-1} \ln \frac{\langle\delta(\chi-\chi_2)\rangle_{\mathcal{H}'}}{\langle\delta(\chi-\chi_1)\rangle_{\mathcal{H}'}} \end{aligned}$$

où tous les termes sont maintenant accessibles par simulation.

Ainsi, on peut calculer toute différence d'énergie libre $F(\chi_f) - F(\chi_0)$ en choisissant suffisamment de valeurs χ_j et de potentiels V_j intermédiaires, tels que les V_j permettent de bien échantillonner les espaces de configurations correspondants à χ_j et χ_{j+1}.

Le *umbrella sampling* permet également d'obtenir le profil d'énergie libre $F(\chi)$. En effet, deux simulations avec les potentiels V_j et V_{j+1} nous permettent d'obtenir les profils d'énergie libre $F_j(\chi)$ et $F_{j+1}(\chi)$ sur des intervalles $[a_j, b_j]$ et $[a_{j+1}, b_{j+1}]$, chacun à une constante près qui vaut respectivement $\ln\langle e^{+\beta V_j}\rangle_{H_j}$ et $\ln\langle e^{+\beta V_{j+1}}\rangle_{H_{j+1}}$. En ayant choisi les potentiels V_j et V_{j+1} de manière à ce que leurs intervalles de valeurs bien échantillonnés se recouvrent (ie : $b_j > a_{j+1}$), on obtient simplement le profil d'énergie sur $[a_j, b_{j+1}]$ en "raccordant" F_j et F_{j+1}, c'est à dire en éliminant les constantes sans calcul en choisissant une valeur $\chi* \in [a_{j+1}, b_j]$ et en imposant $F_j(\chi*) = F_{j+1}(\chi*)$. Des méthodes statistiques automatisées, comme la *weighted histogram analysis method* (WHAM, [102]) permettent de réaliser cette opération de manière optimale.

Cette puissante méthode de dynamique moléculaire est abondamment utilisée dans la littérature mais je n'ai pas trouvé de manière de l'adapter au problème du calcul d'énergie de dipaires. En effet, il me semble que l'unique manière de procéder consiste à se donner un système contenant des liaisons Watson-Crick, puis à ouvrir certaines d'entre elles par un potentiel bien choisi.

La question sous-jacente à laquelle je me suis inévitablement heurté dans ma réflexion a alors été : *qu'est ce qu'une liaison Watson-Crick ouverte ?*

Une liaison Watson-Crick peut se définir facilement. Par exemple, pour A=U, on peut la définir comme l'ensemble des configurations où les deux liaisons hydrogène caractéristiques de cette paire existent, c'est à dire lorsque les distances entre (U)H3..N1(A) et (U)O4..H6(A) sont bien de l'ordre de $r_{LH} = 2$Å. Les configurations où la paire est ouverte sont alors les configurations où ces distances *ne* sont *pas* proches de r_{LH}. Or, avec la méthode du *umbrella sampling*, on ne peut pas définir *négativement* un état d'intérêt. On ne peut s'intéresser qu'à des transitions entre états associés à des valeurs données du paramètre d'ordre. S'intéresser seulement aux configurations où les distances précitées valent par exemple 4Å serait une réduction excessive de l'état "paire ouverte''. Ce problème ne se poserait pas s'il existait une caractérisation expérimentale de ce qu'est une paire Watson-Crick s'ouvrant spontanément (du point de vue conformationnel) mais le fait que cet état soit forcément instable rend utopique l'espoir de l'observer.

Dans la littérature, le cas du retournement de base est fréquemment étudié ([103], [104]). En toute généralité, un axe est défini de manière à ce qu'une rotation autour de celui-ci permette de faire sortir une base à l'extérieur de l'hélice. La diversité des manières de faire illustre ce problème de définition : le choix de l'axe, le choix de faire tourner les deux bases ou seulement une et, dans ce dernier cas, le choix de cette base, le traitement du reste de la structure (les liaisons voisines sont-elles maintenues ?) sont autant de problématiques qui démontrent la multiplicité de l'état "paire ouverte". En conséquence, ces simulations ne peuvent pas permettre de calculer le paramètre *thermodynamique* d'énergie de dipaire.

Ainsi, je n'ai pas pu mettre en place de simulation de *umbrella sampling* à cause d'inextricables problèmes de définition. La méthode suivante permet de les contourner.

5.2.5 Transformation mécanique du système : l'intégration thermodynamique

"*Comment calculer la variation d'énergie libre au cours d'une transformation mécanique et continue du système ?* "

Soit λ la variable réactionnelle d'une telle transformation, telle que l'état de départ corresponde à la valeur $\lambda = 0$ et l'état final à $\lambda = 1$. En notant $\mathcal{H}_\lambda$ les hamiltoniens du système au cours de sa transformation, on a :

$$\begin{aligned} F_{final} - F_{initial} &= F(\lambda = 1) - F(\lambda = 0) = \int_0^1 d\lambda \frac{dF_\lambda}{d\lambda} \\ &= -\beta^{-1} \int_0^1 d\lambda \frac{d \ln \mathcal{Z}_\lambda}{d\lambda} \\ &= -\beta^{-1} \int_0^1 d\lambda \frac{1}{\mathcal{Z}_\lambda} \frac{d\mathcal{Z}_\lambda}{d\lambda} \end{aligned}$$

Or :

$$\begin{aligned} \frac{d\mathcal{Z}_\lambda}{d\lambda} &= \frac{d}{d\lambda} \int d\boldsymbol{R}\, d\boldsymbol{R}\, e^{-\beta \mathcal{H}_\lambda(p,q)} \\ &= -\beta \int d\boldsymbol{R}\, d\boldsymbol{P}\, \frac{d\mathcal{H}_\lambda}{d\lambda} e^{-\beta \mathcal{H}_\lambda} \end{aligned}$$

Donc :

$$\begin{aligned} F_{final} - F_{initial} &= \int_0^1 d\lambda \int d\boldsymbol{R}\, d\boldsymbol{P}\, \frac{d\mathcal{H}_\lambda}{d\lambda} \frac{e^{-\beta \mathcal{H}_\lambda}}{\mathcal{Z}_\lambda} \\ &= \int_0^1 d\lambda\, \langle \frac{d\mathcal{H}_\lambda}{d\lambda} \rangle_{\mathcal{H}_\lambda} \end{aligned} \tag{5.11}$$

Les relations ci-dessus requièrent seulement que le paramètre λ n'influe pas sur la température (*ie* $\frac{d\beta}{d\lambda} = 0$) et que les $\mathcal{H}_\lambda$ soient bien dérivables par rapport à λ. Il n'y a pas de contrainte donnée sur la nature de la transformation. En particulier, elle n'est pas tenue d'être physiquement réaliste, ce qui laisse de la place pour l'imagination.

Application de l'intégration thermodynamique au calcul de l'énergie libre d'hydratation d'une molécule [105]

Cette liberté fait de l'intégration thermodynamique un outil efficace pour calculer l'énergie libre d'hydratation d'une petite molécule c'est-à-dire la différence d'énergie libre résultant de l'ajout de cette molécule dans de l'eau. En effet, l'intégration thermodynamique permet d'éviter de simuler l'immersion de cette molécule par cette astuce : il suffit de créer directement dans la solution une molécule analogue inerte (c'est-à-dire sans interaction avec son milieu) et de lui conférer progressivement les propriétés d'interaction de la molécule désirée. Concrètement, en dynamique moléculaire les atomes sont représentés par une masse m, une charge électrique q et deux paramètres C_6 et C_{12} intervenant dans le potentiel Lennard-Jones. L'analogue inerte de la molécule étudiée est tout simplement une molécule de même topologie dont les atomes n'interagissent pas avec le milieu, c'est-à-dire pour lesquels q, C_6 et C_{12} ont été fixés à 0. L'intégration thermodynamique consiste alors à faire tendre ces paramètres vers leurs valeurs réelles q^r, C_6^r et C_{12}^r. Cela peut se faire par exemple avec le paramétrage suivant :

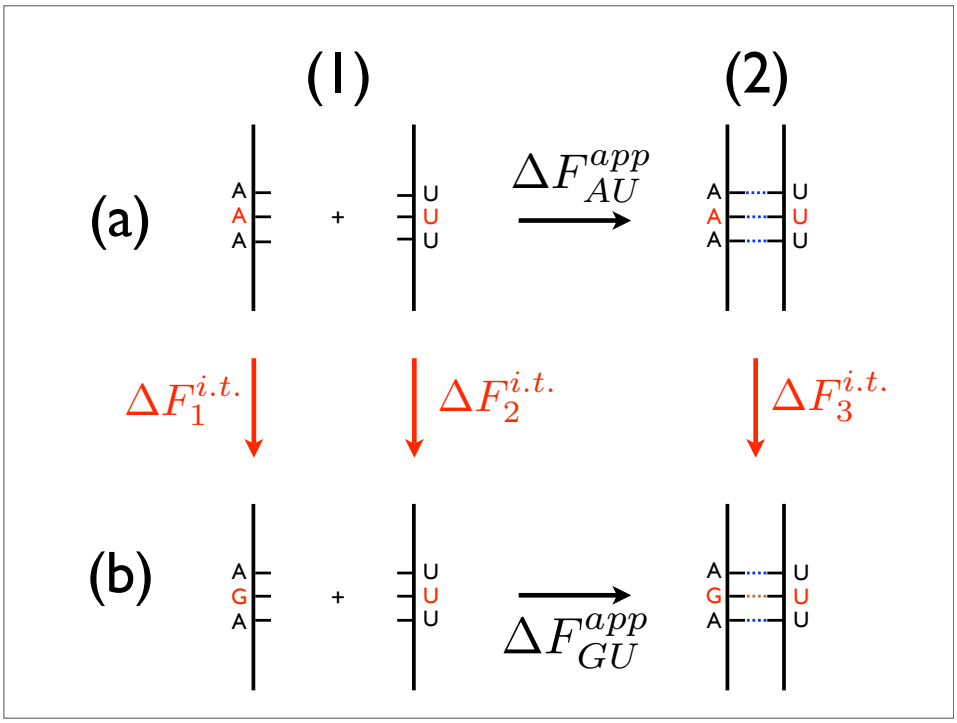
$$\begin{aligned} q(\lambda) &= \lambda q^r \\ C_6(\lambda) &= \lambda C_6^r \\ C_{12}(\lambda) &= \lambda C_{12}^r \end{aligned}$$

Pour $\lambda = 0$ la molécule n'interagit pas avec le milieu. Pour $\lambda = 1$ ses paramètres ont les bonne valeurs. La différence d'énergie libre entre ces deux états, calculée par intégration thermodynamique, correspond bien à l'énergie libre d'hydratation.

5.3 Application de l'intégration thermodynamique au calcul d'énergie libre d'appariement de deux brins d'ARN

5.3.1 Principe

L'intégration thermodynamique est une méthode astucieuse permettant *d'immerger une molécule sans simuler son immersion*, qui est un processus dynamique complexe. De même, elle peut être adaptée pour *calculer des énergies d'appariement sans avoir à désapparier de brins*. Profitant de la ressemblance des pyrimidines C et U entre elles et des purines A et G entre elles, il est possible de continûment transformer l'une en l'autre en gardant en place des brins d'ARN appariés pour déduire toutes les énergies de dipaires à une constante près. Voici un exemple en clarifiant le principe :



Pour connaître la différence d'énergie libre de formation entre une paire A–U et une paire G–U, il faut pouvoir calculer la quantité $\Delta F_{AU}^{app} - \Delta G_{GU}^{app}$, ces deux termes étant associés respectivement aux réactions (a) et (b) du schéma précédent.

L'intégration thermodynamique permet de calculer la différence d'énergie libre associée à la transformation d'une guanine en une adénine dans les réactions (1) et (2).

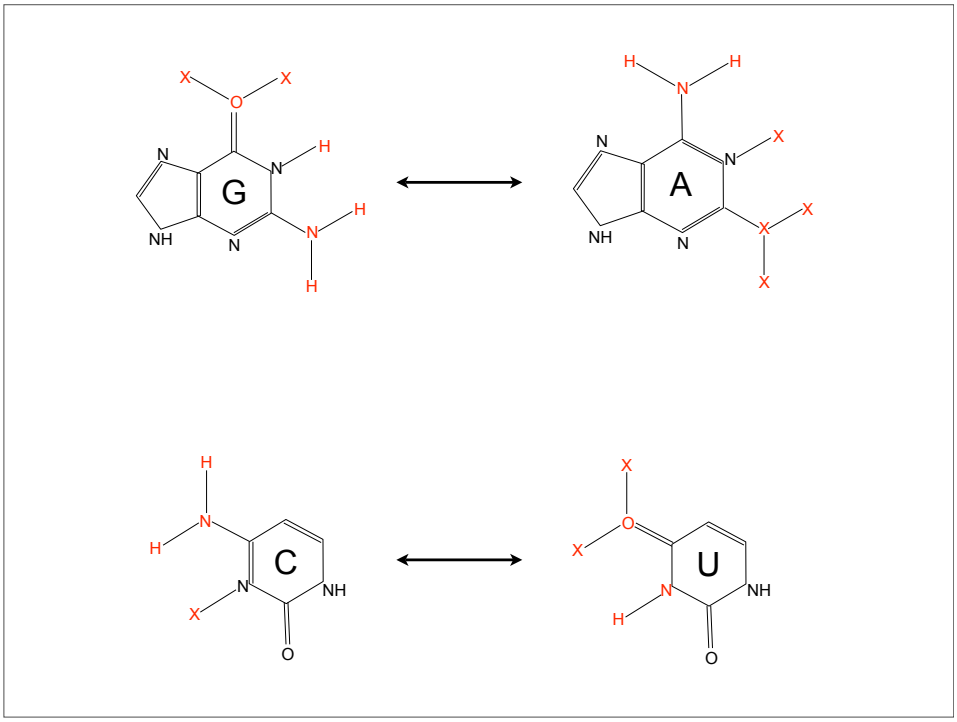
L'énergie libre ne dépendant pas du chemin réactionnel suivi, on peut conclure par :

$$\Delta F_{AU}^{app} - \Delta F_{GU}^{app} = \Delta F_1^{i.t.} + \Delta F_2^{i.t.} - \Delta F_3^{i.t.} \tag{5.12}$$

Les termes de droite sont tous calculables par simulation et les complexes simulations d'appariement (a) et (b) sont évitées.

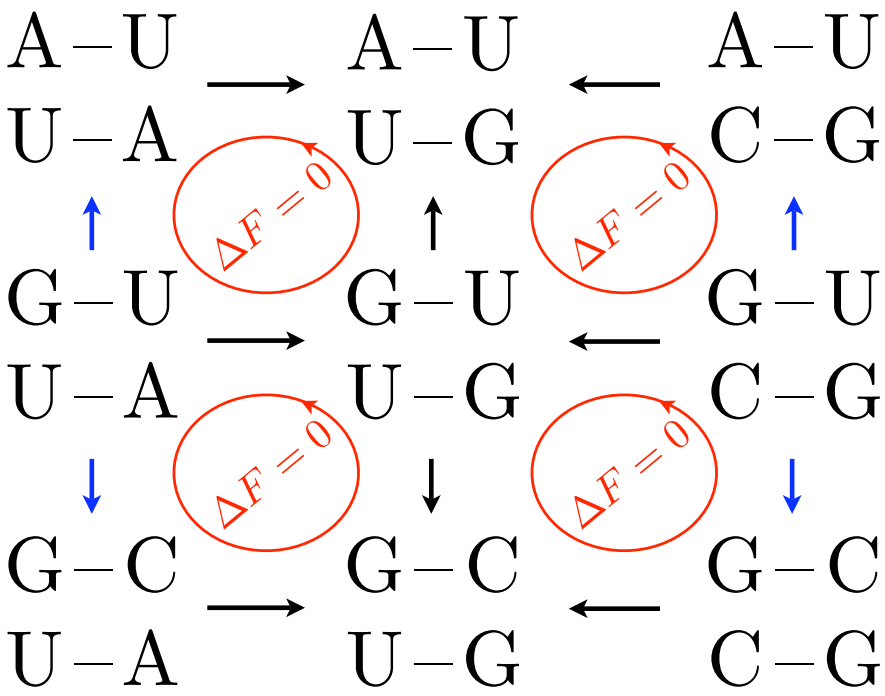
5.3.2 Transformation des purines et des pyrimidines

Comme mentionné en 4.1.1, les contraintes informatiques de la dynamique moléculaire imposent que la transformation d'une molécule en une autre garde constant le nombre de liaisons covalentes. Cela se réalise en créant un nouveau type (noté 'X') d'atome inerte, c'est à dire dont les paramètres q^X, C_6^X et C_{12}^X sont nuls. Cet atome est utilisé comme cible pour les atomes devant disparaître au cours d'une transformation, ce qui permet de maintenir les liaisons covalentes. On "prolonge" ainsi les topologies de chaque nucléotide de manière à ce que G et C puissent se transformer en respectivement A et U par de simples transformations atomiques, comme illustré par le schéma suivant :



5.3.3 Planification des simulations

Toutes les énergies de dipaires Watson-Crick et Wobble se calculent à une constante près en réalisant une succession de transformations selon ce schéma :

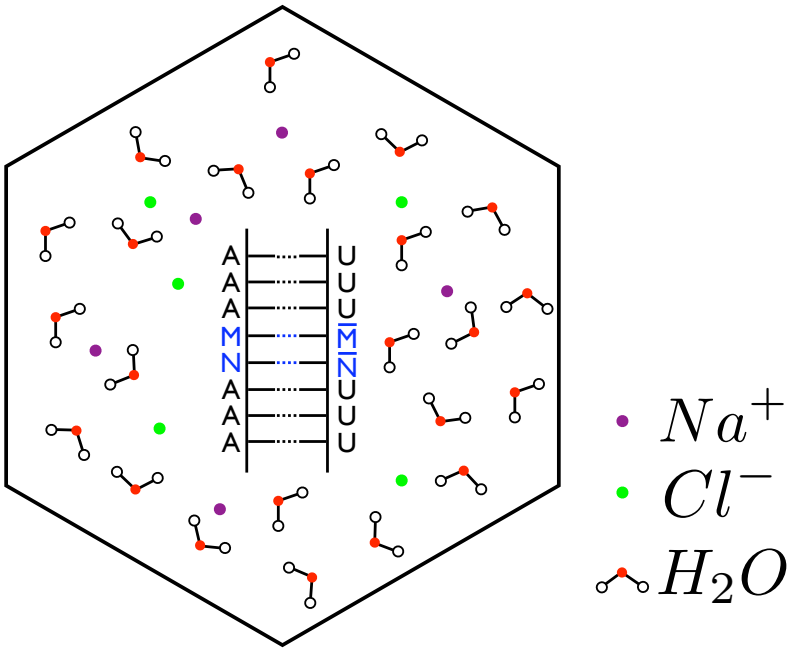


Le calcul des différences d'énergie libre associées aux flèches noires permet d'obtenir toutes ces énergies de dipaires à une même constante près. Simuler les flèches bleues est facultatif à cet effet mais, du fait que celles-ci ferment des cycles le long desquels on doit avoir $\Delta F = 0$, leur calcul fournit un test très efficace de la cohérence des valeurs obtenues et il doit être réalisé à ce titre.

Le schéma ci-dessus ne concerne que les dipaires de type $\begin{pmatrix}\text{Pur–Pyr}\\ \text{Pyr–Pur}\end{pmatrix}$. Une opération similaire doit donc être également réalisée pour ceux de type $\begin{pmatrix}\text{Pur–Pyr}\\ \text{Pur–Pyr}\end{pmatrix}$ et $\begin{pmatrix}\text{Pyr–Pur}\\ \text{Pur–Pyr}\end{pmatrix}$. Au final toutes les énergies de dipaires seront données à trois constantes près qu'il sera facile de paramétrer par des méthodes d'optimisation *a posteriori* sur les bases de données.

5.3.4 Système simulé

Le système simulé générique est un ARN double brin complémentaire placé dans une cellule dodécahédrique remplie de molécules d'eau et d'ions Na^+ et Cl^- à une concentration de 1M.



Chaque brin fait 8 bases de long : par défaut un brin est composé de 8 adénines et l'autre de 8 uraciles. La dipaire d'intérêt $\begin{pmatrix} M-\overline{M} \\ N-\overline{N} \end{pmatrix}$ est placé au centre de l'ARN double brin. Les nucléotides destinés à muter au cours de l'intégration thermodynamique sont modifiés par l'ajout adéquat d'atomes inertes X ainsi qu'expliqué précédemment. La structure initiale de l'ARN double brin est construite suivant le modèle de double hélice référencé *1qcu* dans la PDB.

5.4 Détails techniques de l'intégration thermodynamique

5.4.1 Description de $\mathcal{H}_\lambda$

L'intégration thermodynamique a été implémentée dans GROMACS. J'explicite dans cette section la dépendance en λ des différents potentiels constituant $\mathcal{H}_\lambda$.

Cas général

Les atomes mutants sont modifiés linéairement suivant le schéma :

$$\begin{aligned} m^\lambda &= (1-\lambda)m^0 + \lambda m^1 \\ q^\lambda &= (1-\lambda)q^0 + \lambda q^1 \end{aligned}$$

Les paramètres intervenant dans les potentiels d'interactions liées sont transformés de la même manière. Dans le cas de potentiels harmoniques, notés en toute généralité $V(x) = \frac{k}{2}(x - x_{ref})^2$, il s'agit de k et x_{ref}, qui deviennent :

$$\begin{aligned} k^\lambda &= (1-\lambda)k^0 + \lambda k^1 \\ x^\lambda_{ref} &= (1-\lambda)x^0_{ref} + \lambda x^1_{ref} \end{aligned}$$

Tous les potentiels d'interaction, à l'exception du potentiel Lennard-Jones, gardent la même forme en tenant compte de ces nouvelles dépendances en λ :

$$\begin{aligned} V^\lambda_{Coulomb}(i,j) &= \frac{q^\lambda_i q^\lambda_j}{\varepsilon r_{ij}} \\ V^\lambda_{harmonique} &= \frac{k^\lambda}{2}(x - x^\lambda_{ref})^2 \\ V^\lambda_{dihedre\ propre} &= k^\lambda(1 + \cos(n_\phi \phi - \phi^\lambda_{ref})) \\ E^\lambda_c &= \frac{p^2}{2m^\lambda} \end{aligned}$$

Le potentiel Lennard-Jones constitue un cas à part.

Cas du potentiel Lennard-Jones

Supposons qu'à l'instar des autres potentiels le potentiel Lennard-Jones devienne simplement :

$$\begin{aligned} V_{LJ}^{\lambda}(r) &= \frac{C_{12}^{\lambda}}{r^{12}} - \frac{C_6^{\lambda}}{r^6} \\ \text{avec :} \quad C_6^{\lambda} &= (1-\lambda)C_6^0 + \lambda C_6^1 \\ C_{12}^{\lambda} &= (1-\lambda)C_{12}^0 + \lambda C_{12}^1 \end{aligned}$$

Cette forme conduit à de grandes difficultés de calcul dans le cas de la création ou annihilation de particules. Plaçons-nous par exemple dans le cas de la création de particule, c'est-à-dire où $C_6^0 = 0$ et $C_{12}^0 = 0$. On a alors $V_{LJ}^{\lambda} = \lambda V_{LJ}^1$ et :

1. $V_{LJ}^{\lambda}(r)$ tend simplement vers 0 $\forall r \neq 0$ quand $\lambda \to 0$
2. $\frac{dV_{LJ}^{\lambda}}{d\lambda}(r)$ est indépendant de λ et $\forall \lambda > 0$, $\frac{dV_{LJ}^{\lambda}}{d\lambda}(0) = V_{LJ}^1(0) = +\infty$

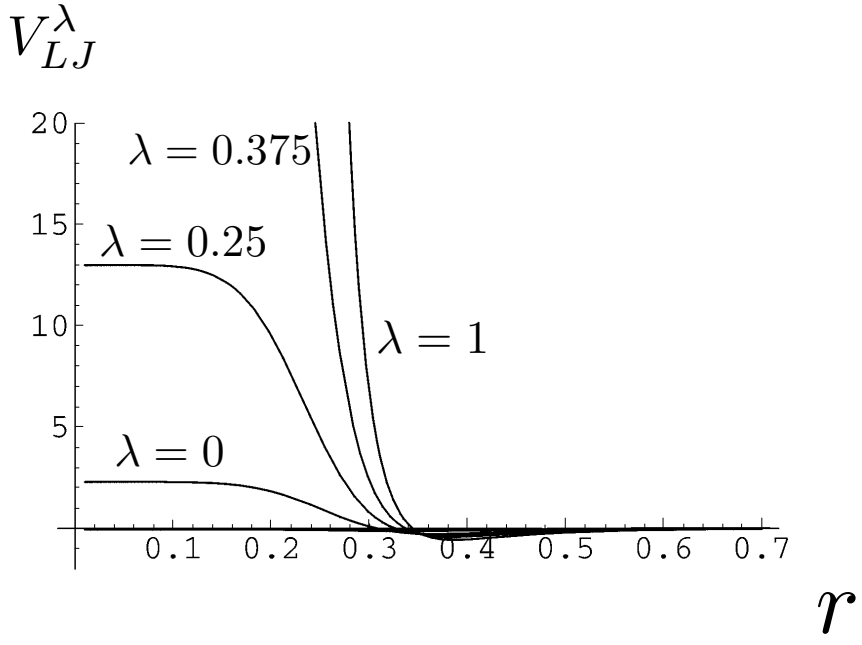
Le point (1) signifie que l'effet répulsif du potentiel Lennard-Jones dit de "noyau dur" perd en amplitude lorsque λ tend vers 0 de sorte que le système peut explorer de plus en plus facilement des configurations où r s'approche de 0. Or $\frac{dV_{LJ}^{\lambda}}{d\lambda}$ ne dépend pas de λ et n'est donc pas atténué lorsque λ tend vers 0. Comme ce terme est divergent en $r = 0$ pour tout λ, il en résulte que $\langle \frac{dV_{LJ}^{\lambda}}{d\lambda} \rangle_{\lambda}$ peut varier sur plusieurs ordres de grandeur lorsque λ tend vers 0.

Ce phénomène a été observé sur des exemples simples [106]. Le calcul de l'intégrale (5.11) est alors peu convenable car son évaluation requiert un nombre très coûteux de simulations pour compenser les erreurs de mesure induites par la forte variation de $\langle \frac{dV_{LJ}^{\lambda}}{d\lambda} \rangle_{\lambda}$ à l'approche de $\lambda = 0$.

Il est donc nécessaire de trouver une autre dépendance en λ, non linéaire. Parmi les différentes possibilités, l'approche dite de noyau mou ("*soft-core*") s'est imposée en pratique. Elle consiste en ce paramétrage :

$$\begin{aligned} V_{LJ}^{\lambda}(r) &= \lambda V_{LJ}^{1}(R_1^{\lambda}(r)) \\ \text{avec :} \quad R_i^{\lambda}(r) &= (\alpha C_6^i (1-\lambda)^p + r^6)^{\frac{1}{6}} \qquad \alpha > 0,\ p > 0 \end{aligned}$$

Ainsi, pour $\lambda = 0$ et $\lambda = 1$, V_{LJ}^{λ} a bien la forme attendue, tandis que la divergence en $r = 0$ lorsque $\lambda \to 0$ disparaît. α et p sont deux paramètres choisis de manière à ce que le profil obtenu soit le plus régulier possible, afin que les erreurs inhérentes à la discrétisation de l'intégrale soient les plus faibles possible. En général, on choisit $\alpha = 1$ et $p = 2$. Voici quelques tracés de V_{LJ}^{λ} pour différentes valeurs de λ illustrant l'effet de noyau mou :



Lorsque s'opèrent simultanément créations et annihilations de particules, le potentiel Lennard-Jones utilisé est :

$$V_{LJ}^{\lambda}(r) = \lambda V_{LJ}^{1}(R_1^{\lambda}(r)) + (1-\lambda) V_{LJ}^{0}(R_0^{1-\lambda}(r))$$

5.4.2 Autres paramètres de la simulation

Réaliser une simulation de dynamique moléculaire requiert le réglage d'une centaine d'autres paramètres. La liberté de choix et le manque d'études comparatives systématiques de leur conséquences justifient partiellement que la dynamique moléculaire soit parfois qualifiée d'"art" [107]. Dans la littérature, on peut observer une grande diversité de réglages des paramètres principaux de la dynamique moléculaire et ceux-ci sont rarement justifiés.
L'objet de cette section est de donner un aperçu de la difficulté du choix des modèles d'eau et des modèles d'énergie (également appelés "champs de force" dans ce contexte) pour l'ARN. Il y a quatre principaux champs de force utilisés pour simuler les biopolymères (ADN, ARN et protéines) : GROMOS, AMBER, OPLS et CHARMM, chacun ayant plusieurs versions. Il y a aussi une grande variété de modèles d'eau disponibles, les plus populaires étant TIP3P, SPC, SPC/E, TIP4P, TIP4P-EW et TIP5P.

Quelle combinaison choisir pour le problème de calcul d'énergie libre de dipaires d'ARN en solution ?

Modèles d'eau et interactions à longue portée

Les modèles d'eau ont été développés indépendamment des champs de force. Chacun de ces modèles d'eau est le résultat d'une optimisation selon un critère spécifique mais, au final, aucun ne permet de rendre compte simultanément des propriétés thermodynamiques, cinétiques et électrostatiques de l'eau à température ambiante.

Ces modèles diffèrent en nature. Dans TIP3P [108], SPC et SPC/E [109] l'eau est modélisée par deux atomes d'hydrogène et un atome d'oxygène affectés de charges électriques partielles. Les modèles TIP4P [108] et TIP4P-EW [110] rajoutent à ces trois atomes une quatrième particule, sans masse mais possédant une charge pour mieux reproduire le dipôle de l'eau. TIP5P quant à lui modélise l'eau par 5 particules. Dans tous ces modèles, la géométrie de la molécule d'eau est fixée, ce qui implique qu'elle n'est pas polarisable.

L'optimisation des paramètres effectuée pour ces différents modèles est aussi associée à d'autres réglages, notamment la manière de traiter les "interactions à longue portée". Cette dernière est une nécessité imposée par des contraintes informatiques. En effet, la complexité des algorithmes de dynamique moléculaire est proportionnelle au nombre

de paires d'atomes en interaction dans le système. La plupart des petits systèmes d'intérêt en solution requièrent un nombre d'atomes de l'ordre de 10^4 mais il serait bien trop coûteux de calculer à chaque pas de temps les énergies des 4.5×10^7 paires associées. Il convient donc en pratique de fixer une limite supérieure à la portée des interactions non-liées (Coulomb et Lennard-Jones) pour restreindre autant que possible le nombre réel de paires d'atomes à considérer. En s'imposant une valeur r_{max} au-delà de laquelle les interactions non liées à longue portée doivent être nulles, plusieurs manières de procéder existent :

- tronquer $V_{Coulomb/LJ}$ à r_{max} ("*cut-off*")
- remplacer le potentiel $V_{Coulomb/LJ}(r)$ par $f(r) \times V_{Coulomb/LJ}(r)$, où $f(r)$ est une fonction décroissante valant 1 en $r = 0$ et 0 pour $r > r_{max}$ ("*shift*")
- conserver $V_{Coulomb/LJ}(r)$ à l'identique sur un intervalle $[0, r_1]$ avec $r_1 < r_{max}$ et appliquer l'opération précédente sur $[r_1, r_{max}]$ ("*switch*")

Ces trois méthodes induisent une altération plus ou moins brutale du potentiel $V_{Coulomb/LJ}(r)$. Dans le cas où les conditions au bord sont périodiques (c'est-à-dire où chaque face de la cellule communique avec la face opposée), une quatrième méthode, la "sommation d'Ewald", permet de calculer correctement et efficacement les interactions coulombiennes à longue portée. Sans détailler techniquement, la sommation d'Ewald décompose l'interaction coulombienne en un terme de courte portée calculé directement et un terme à longue portée que les conditions périodiques permettent de calculer efficacement en passant par sa transformée de Fourier. Il en existe d'efficaces implémentations (*"particle-mesh Ewald"*, PME [111]). Dans ce cas, l'emploi de conditions périodiques nécessite de mettre la molécule d'intérêt en solution dans suffisamment d'eau pour être écrantée de l'influence de ses propres images périodiques. Enfin, citons aussi la méthode RF ("*Reaction field*") qui estime l'interaction coulombienne au-delà de r_{max} en supposant que le milieu se comporte comme un champ homogène.

Les modèles d'eau évoqués précédemment ont ainsi été développés en liaison avec un certain traitement des interactions à longue portée. TIP3P et TIP4P utilisent par exemple un *switch* de l'interaction coulombienne entre 7.5Å et 8.5Å alors que TIP4P-EW est un reparamétrage de TIP4P en utilisant PME. Voici un tableau extrait de [112] qui illustre comment différents modèles reproduisent la constante de diffusion D, la permittivité diélectrique ϵ_0, la densité ρ et l'énergie potentielle E_{pot} de l'eau en fonction du traitement des interactions coulombiennes à longue portée (PME, *switch*, *switch2*, *shift*, *cut-off*, *RF*).

model	cutoff	$D\,10^5$ cm^2 s^{-1}	ϵ_0	ρ (g/L)	E_{pot} (kJ/mol)
expt		2.3	78.5	997	−41.7
TIP3P	PME	5.76(0.03)	92(4)	970.9(0.2)	−39.882(0.002)
	switch	4.26(0.11)	102(5)	1004.3(0.2)	−40.705(0.003)
	Switch2	5.65(0.15)	103(5)	987.3(0.2)	−40.133(0.002)
	Shift	5.8(0.2)	101(5)	981.5(0.2)	−39.823(0.002)
	Cut-Off	3.88(0.02)	48(1)	1000.8(0.2)	−41.698(0.003)
	RF	3.72(0.02)	54(3)	991.0(0.3)	−41.318(0.003)
TIP4P	PME	3.73(0.02)	49(2)	980.4(0.2)	−41.282(0.002)
	switch	2.65(0.02)	53(2)	1026.5(0.2)	−42.293(0.003)
	switch2	3.53(0.08)	52(2)	998.4(0.2)	−41.528(0.003)
	shift	3.78(0.04)	51(2)	990.3(0.2)	−41.172(0.003)
	cutoff	3.9(0.5)	41(1)	1010.9(0.3)	−41.19(0.01)
	RF	3.02(0.03)	72(3)	981.2(0.2)	−40.282(0.005)
TIP5P	PME	2.95(0.05)	88(7)	969.3(0.2)	−40.232(0.006)
	switch	2.72(0.04)	87(6)	988.6(0.3)	−40.353(0.005)
	switch2	2.75(0.07)	89(6)	983.5(0.3)	−40.521(0.005)
	shift	2.94(0.06)	89(6)	980.4(0.2)	−40.139(0.005)
	cutoff	4.48(0.09)	43(1)	991.2(0.2)	−42.359(0.006)
	RF	4.38(0.05)	65(3)	974.3(0.2)	−41.610(0.003)
SPC	PME	4.29(0.04)	67(3)	963.9(0.2)	−41.535(0.003)
	switch	3.24(0.01)	72(4)	1005.5(0.3)	−42.375(0.006)
	switch2	4.11(0.06)	69(3)	982.3(0.2)	−41.794(0.003)
	shift	4.27(0.11)	62(3)	973.9(0.2)	−41.452(0.003)
	cutoff	2.9(0.2)	43(1)	1014.5(0.3)	−47.55(0.01)
	RF	2.71(0.04)	77(4)	996.0(0.2)	−46.668(0.004)
SPC/E	PME	2.70(0.04)	62(4)	986.5(0.2)	−46.618(0.004)
	switch	2.11(0.02)	74(5)	1027.8(0.3)	−47.414(0.005)
	switch2	2.55(0.01)	76(5)	1003.6(0.2)	−46.873(0.004)
	shift	2.71(0.13)	78(5)	995.7(0.2)	−46.515(0.009)

On s'aperçoit de la grande variabilité des résultats et que ceux-ci, du moins pour les grandeurs estimées ci-dessus, peuvent être qualitativement meilleurs en utilisant un traitement des interactions à longue portée différent de celui avec lequel le modèle a été développé (par exemple, considérer les performances de TIP3P avec *switch2* et *cut-off*).

Champs de force

Les champs de force ont été développés, chacun selon sa propre méthode, de manière à reproduire des données expérimentales comme les géométrie, spectres vibrationnels ou énergies conformationnelles de petites molécules. Historiquement, ces modèles ont d'abord été développés pour les protéines, ensuite pour l'ADN et enfin pour l'ARN. Chacune des optimisations de ces modèles a été effectuée en utilisant un certain modèle d'eau : ainsi CHARMM et AMBER s'appuient sur TIP3P, GROMOS sur SPC et OPLS sur TIP4P. Il n'y a actuellement pas eu de mises à jour consécutives aux progrès

ultérieurement effectués sur les modèles d'eau (comme l'évolution de TIP4P en TIP4P-EW et SPC en SPC/E) mais rien n'exclut que ces champs de force soient effectivement plus performants avec des modèles d'eau autres que ceux avec lesquels ils ont été établis.

La comparaison systématique des aptitudes de ces champs de force à reproduire une variété de données simples, comme certaines propriétés structurelles ou des énergies d'hydratation, est toujours un sujet d'actualité. Concernant l'ADN, l'accent a été mis sur la bonne reproduction de sa forme régulière canonique "B" mais la validité des modèles concernant l'étude de structures non canoniques, comme les renflements, n'a jamais été définitivement établie, bien que certains travaux présentent des résultats encourageants. La validation de ces modèles pour l'ARN est encore plus rare, surtout en ce qui concerne les mesures d'énergies libres d'hydratation.

L'étude comparée des énergies libres d'hydratation des acides aminés, en fonction des modèles d'eau et des modèles d'énergie, est un travail relativement récent [113] et montre que les différentes combinaisons de modèles surestiment globalement ces dernières. Dans cet article, Shirts et *al.* ont également montré que le modèle d'eau le plus performant de ce point de vue était TIP3P-MOD, une version de TIP3P rarement utilisée et reparamétrée pour reproduire l'énergie libre d'hydratation du méthane. Ce dernier point étaye la recommandation générale qui veut que, pour un problème donné, il vaille mieux utiliser les jeux de paramètres qui ont été optimisés pour le problème qui s'en rapproche le plus. Il n'y a pas encore de "modèle universel" qui permettrait d'avoir des performances du niveau de l'état-de-l'art sur toutes les questions classiques étudiées dans la littérature.

Pour le problème qui m'intéresse, l'idéal serait de disposer d'un modéle d'eau et d'un champ de forces ayant été conjointement optimisés pour reproduire des énergies libres d'appariement d'acides nucléiques ou au moins leurs énergies libres d'hydratation. De tels modèles ne sont pas disponibles. De plus, l'étude de Shirts et *al.* n'a, à ma connaissance, pas encore eu d'équivalent avec les ribonucléotides. Il me semble donc que la question du choix du modèle d'eau et du champ de forces pour mon problème reste ouverte et ce sera le travail de simulation que j'entreprends qui, en cas de succès, permettra *a posteriori* d'établir la validité de tel ou tel choix de modèles pour calculer des énergies libres d'appariement.

J'ai finalement décidé d'utiliser comme champ de force la paramétrisation 1999 de AMBER, AMBER99 [114], en conjonction avec le modèle d'eau TIP4P-EW, combinaison déjà utilisée dans [100].

Ce dernier choix implique d'utiliser la méthode PME pour traiter les interactions coulombiennes à longue portée et donc d'utiliser des conditions aux bords périodiques.

Intégrateur

La méthode proposée pour calculer des énergies libres d'appariement par intégration thermodynamique requiert seulement le calcul de $\langle \frac{d\mathcal{H}_\lambda}{d\lambda} \rangle$ pour plusieurs valeurs de λ. Les aspects cinétiques des fluctuations du système à l'équilibre ne sont pas importants. Ainsi, ce n'est pas l'équation newtonienne du mouvement qui a été utilisée mais plutôt l'équation de Langevin. Celle-ci inclut en plus un terme de frottement et un processus stochastique de sorte que les trajectoires ne sont pas physiquement réalistes mais convergent plus rapidement vers l'ensemble Boltzmannien. Pour l'atome i :

$$m_i \frac{d\boldsymbol{p}_i}{dt} = -m_i \xi_i \boldsymbol{p}_i + \boldsymbol{F}_i(\boldsymbol{r}_i) + \zeta_i(t) \tag{5.13}$$

où ξ_i est un coefficient de frottement et ζ_i est un bruit gaussien vérifiant :

$$\langle \zeta_i(t) \zeta_j(t+s) \rangle = 2 m_i \eta_i k_B T \delta(s) \delta_{ij} \tag{5.14}$$

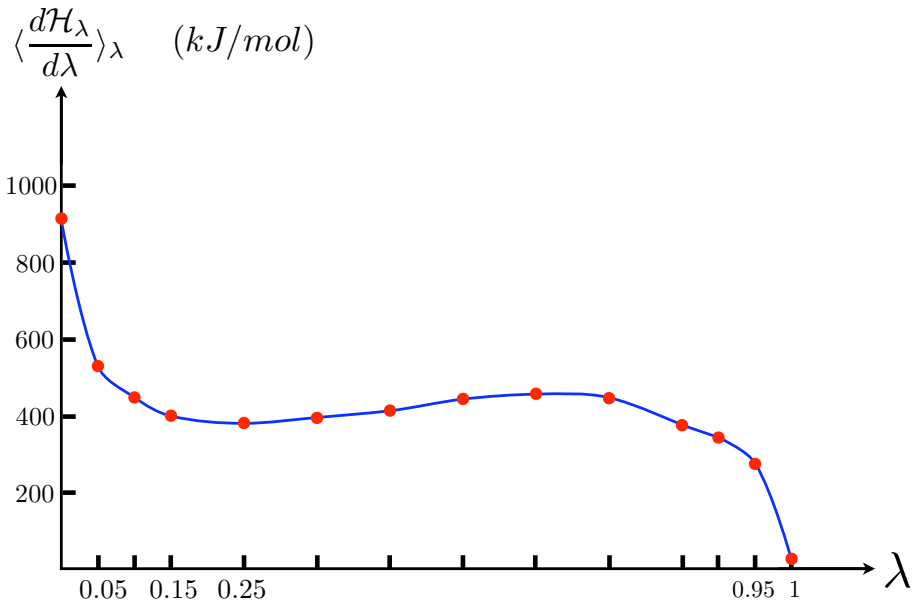
Autres réglages

Une simulation de dynamique moléculaire avec GROMACS requiert de nombreux autres réglages. Par exemple, la géométrie des molécules d'eau étant fixée, une procédure spéciale est requise pour mouvoir ces molécules sans perturber cette géométrie. L'algorithme SETTLE a ainsi été utilisé.

Ces nombreux autres réglages ne seront pas présentés ici et ont le plus souvent été choisis de manière standard.

5.5 Résultats

Les différentes transformations indiquées dans la figure "Planification des simulations" ont été réalisées. Le profil général de $\langle \frac{d\mathcal{H}_\lambda}{d\lambda} \rangle_\lambda$ observé pour ces transformations est le suivant :



Profil obtenu pour la transformation $\begin{pmatrix} G-C \\ C-G \end{pmatrix} \rightarrow \begin{pmatrix} G-C \\ U-G \end{pmatrix}$

Ces courbes ont été discrétisées en calculant $\langle \frac{d\mathcal{H}_\lambda}{d\lambda} \rangle$ aux points $\lambda = 0, 0.05,$ $0.1, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.9, 0.95, 0.95, 1$. Pour chacun de ces points, le temps simulé a été de 3 ns et seules les dernières 2 ns ont été conservées pour calculer ces valeurs moyennes. L'intégrale (5.11) correspond à l'aire de la surface en dessous de la courbe.

Il s'est avéré que la condition "$\Delta F = 0$" qui doit être respectée par les cycles indiqués sur la figure "Planification des simulations" n'a été satisfaite pour aucun d'entre eux. Les résultats ne sont donc pas cohérents et sont inutilisables. Plusieurs explications peuvent être avancées :

- la quantité $\langle \frac{d\mathcal{H}_\lambda}{d\lambda} \rangle$ est difficile à mesurer car elle est petite par rapport à l'énergie totale du système (qui comprend en général environ 5000 molécules d'eau). La convergence de cette quantité est difficile à déterminer.
- Dans toutes les transformations considérées, une paire Watson-Crick devient une paire Wobble ou inversement. Or ces deux types de paires n'ont pas la même géométrie. Passer de l'une à l'autre requiert un réarrangement qui ne se fait pas toujours complètement dans les temps de simulation considérés.
- Pour chaque valeur de λ, la simulation a été initialisée en utilisant la configuration finale obtenue pour la valeur de λ précédente. Par exemple, la configuration initiale de la simulation correspondant à $\lambda = 0.5$ est la configuration obtenue au bout des 3 ns de simulation de $\lambda = 0.4$. En procédant ainsi, il est possible qu'une anomalie non détectée puisse se transmettre d'une simulation à l'autre.

Une analyse des résultats qui ne sera pas présentée a montré que ces différents problèmes se retrouvaient dans les trajectoires simulées. Il n'est évidemment pas exclu que d'autres problèmes non mentionnés surviennent. Après avoir fait ce constat, ces travaux de simulation ont été abandonnés car ils demandent trop de temps de calcul et trop de travail pour une seule personne.

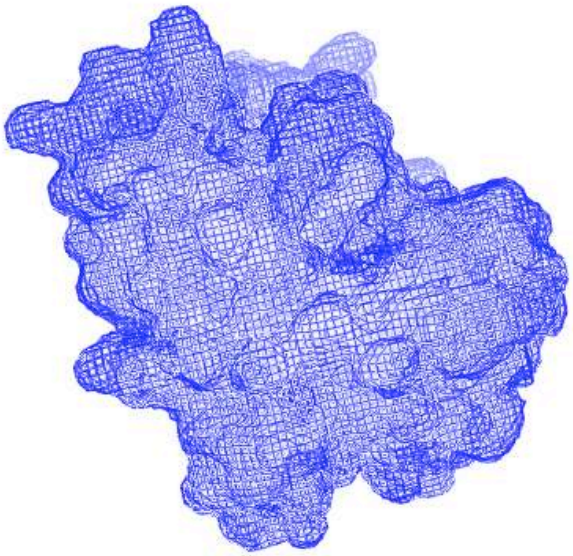
5.5.1 Un “à-côté”

Dans l'article [115], Azuara *et al.* présentent un nouveau modèle d'eau fondé sur un approfondissement de l'équation de Poisson-Boltzmann incluant le moment dipolaire de l'eau (équation GPBL). Les propriétés de ce modèle ont été comparées à celles d'un modèle d'eau explicite à l'aide d'une simulation que j'ai effectuée.

Le but de cette simulation était d'étudier la distribution moyenne des molécules d'eau autour de la protéine référencée 1TIM dans la PDB. Cette protéine a été placée dans une cellule cubique de 7.35 nm complétée de 12548 molécules d'eau SPC/E. La géométrie de cette protéine extraite de la PDB a été maintenue fixe tout au long de la simulation en appliquant à chacun de ses atomes des potentiels harmoniques les contraignant dans leur position initiale. Ce sont les trajectoires des molécules d'eau autour de cette architecture rigide qui font l'intérêt de la simulation. Trois simulations de 8 ns ont été effectuées en utilisant le champ de forces GROMOS-53A6. Les trajectoires ont été analysées en découpant la cellule en une grille de maille $a = 2.1$Å ou $a = 1.4$Å et en calculant selon un protocole standard décrit dans [116] :

- $d(\boldsymbol{r})$: la densité, c'est à dire l'occupation moyenne de chaque maille par des molécules d'eau.
- $\vec{P}(\boldsymbol{r})$: l'orientation moyenne de ces dipôles pour chaque maille.

  Ces deux cartes ont été comparées aux solutions de l'équation GPBL, ce qui a permis de mettre en évidence des différences de nature entre ces modèles. L'équation produit des cartes plus homogènes que la dynamique moléculaire où des pics de densité sont bien plus marqués. Les autres résultats de ces analyses sont présentés dans l'article.



Une ligne de niveau de la carte $d(\boldsymbol{r})$

Conclusion

Ainsi donc, dans ce chapitre, il a été exposé une méthode de calcul d'énergie libre de de dipaires par dynamique moléculaire. Bien que son principe soit théoriquement correct, la définition précise de simulations soulève plusieurs autres problèmes techniques et conceptuels qui n'ont certainement pas tous de réponse définitive en l'état actuel du développement de la dynamique moléculaire. Bien qu'elle n'ait pas été présentée dans ce chapitre, la complexité de l'analyse des volumineuses données engendrées par la dynamique moléculaire ne doit pas non plus être occultée. La somme de toutes ces difficultés explique au moins partiellement mon incapacité à obtenir des résultats cohérents dans les temps que j'ai consacrés à ces travaux.

Chapitre 6

Conclusion

Dans les travaux présentés dans cette thèse, les deux grandes problématiques de la prédiction de structures secondaires ont été étudiées.

Un examen des deux méthodes principalement utilisées pour calibrer les différents paramètres du modèle d'énergie libre a soulevé plusieurs interrogations. L'interprétation des valeurs expérimentales repose sur une simplification qui a été démontrée comme excessive dans le cas de l'ADN et il y a un ensemble d'indices qui laissent penser qu'il en est de même pour l'ARN. Les méthodes s'appuyant sur l'étude des bases de données ne s'assurent pas suffisamment de la qualité de l'information que ces dernières emmagasinent : beaucoup de structures ne sont pas vérifiées expérimentalement et il y a de nombreuses erreurs d'annotation.

MC, un nouveau paramétrage, est ainsi proposé en rationalisant toutes les étapes de l'optimisation. La base de données choisie est une base de données d'ARNt à la structure secondaire sure et la présence de bases chimiquement modifiées est prise en compte. Au final, 98.5% des 413 séquences utilisées sont expliquées par ce nouveau paramétrage. De plus, la qualité de la prédiction de structures avec pseudo-noeuds se voit aussi significativement améliorée. Cependant, n'ont été calibrés que les paramètres que cette base de données permettait de calibrer. D'autres, comme les paramètres des boucles internes et renflements, ne sont que passablement estimés. De futurs travaux doivent donc s'intéresser aux carences de ce paramétrage. Tout d'abord, les raisons pour lesquelles MC ne permet pas une prédiction de 100 % sur la base de donnée utilisée doivent être trouvées. La base de données étant irréprochable, les 1,5 % manquants contiennent une information d'une valeur certaine. Ensuite, la même démarche doit être poursuivie pour paramétrer renflements et boucles internes. Ceci signifie compiler des

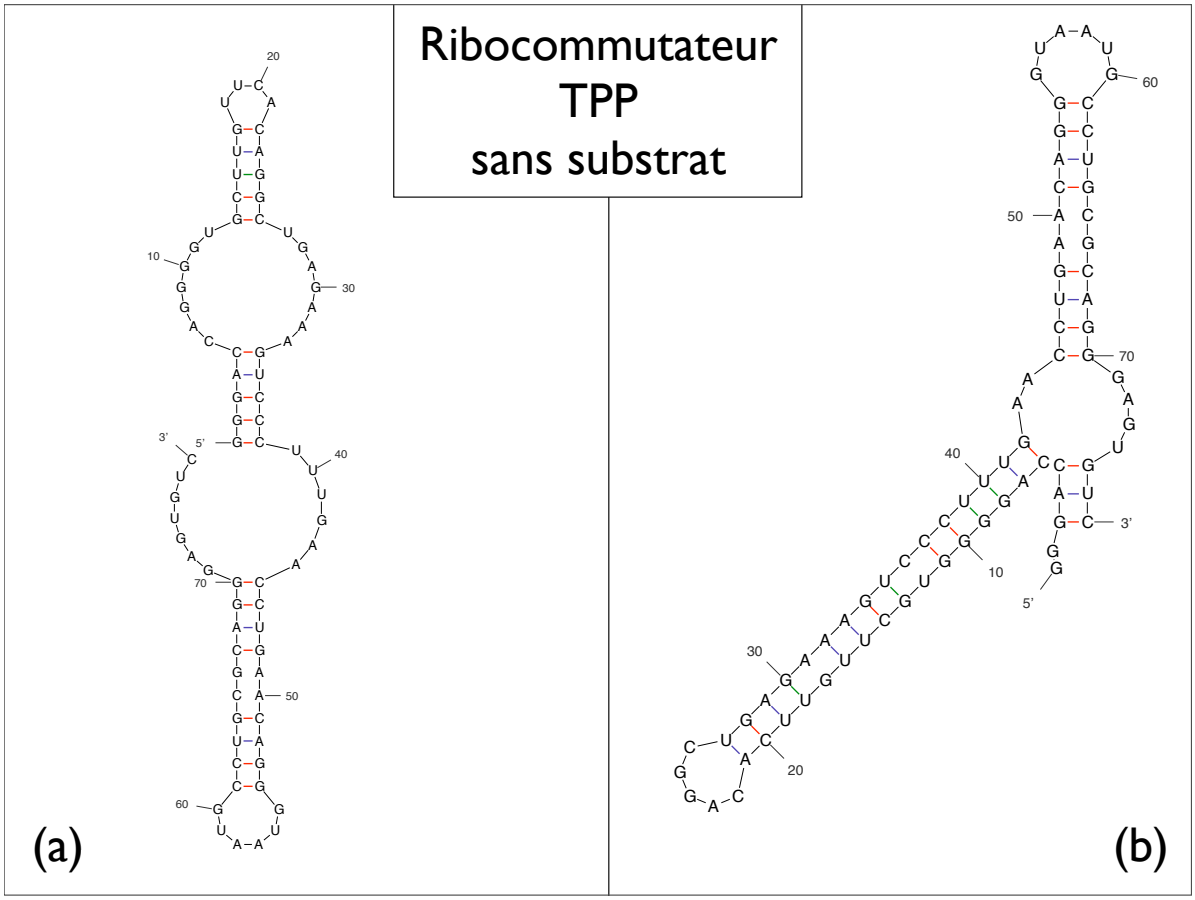
données expérimentales sures et se poser la question de la forme et du nombre des paramètres que ces données permettront d'estimer correctement.

TT2NE, un nouvel algorithme de prédiction de structures secondaires avec pseudo-nœuds, a été détaillé. TT2NE se différencie des algorithmes actuellement disponibles en étant le seul à pouvoir proposer n'importe quelle topologie de pseudo-nœuds tout en garantissant de trouver le minimum d'énergie libre. Cette approche porte ses fruits sur les séquences testées puisque TT2NE améliore clairement l'état de l'art représenté par les algorithmes McQfold et HotKnots. La réussite de TT2NE tient à la fois à l'algorithme, au modèle d'énergie et au traitement des pseudo-nœuds utilisés. Une nouvelle classification des pseudo-nœuds a en effet été présentée et sa pertinence a été justifiée quantitativement et qualitativement. Cette classification fondée sur le genre, un nombre entier de nature topologique, permet de caractériser efficacement les pseudo-nœuds observés dans la nature. TT2NE utilise cette classification pour restreindre sa recherche aux structures les plus vraisemblables, ce qui lui permet de faire des calculs exacts dans une certaine limite de taille. Les principales erreurs de prédiction commises par TT2NE sont dues à une difficulté liée au pseudo-noeuds qui n'a pas été prise en compte dans ces travaux : les contraintes stériques. Contrairement aux structures secondaires planaires qui sont toujours stériquement possibles en vertu de leur nature arborescente, le pseudo-nœud est un motif tridimensionnel contraint. Il est ainsi possible qu'un pseudo-nœud dessiné sur un diagramme ne corresponde à aucune structure réelle car la rigidité de l'ARN empêche d'effectuer toutes les contorsions requises. Une compréhension fine des contraintes stériques est ainsi nécessaire pour mieux prédire les pseudo-nœuds. Lorsque cette source d'erreur aura été maîtrisée, les capacités de TT2NE devront être développées pour pouvoir prendre en charge des séquences plus longues car il est actuellement limité à des séquences de 150 bases de long sur un simple processeur. Pour se faire, une première idée est d'implémenter TT2NE sur une architecture parallèle, opération à laquelle son principe –l'exploration d'un graphe– se prête bien. Il est cependant aussi judicieux de chercher de nouvelles stratégies d'optimisation de l'exploration par toute sorte d'analyses préliminaires à l'image de la méthode de séparation et d'exploration progressives déjà présentée.

Les différents travaux présentés dans cette thèse reposent sur la même philosophie : se donner les moyens d'avoir une précision de prédiction de 100% dans des cas simples, ARNt et courts pseudo-noeuds. L'objectif n'est pas de pouvoir revendiquer les meilleurs résultats sur les quelques 3500 structures secondaires répertoriées dans les bases de données, sur lesquelles MC et TT2NE n'ont souvent même pas été testés. Mon but a été plutôt de recenser les faiblesses de notre compréhension du repliement de l'ARN et de

concevoir des manières d'y pallier. Mon sentiment est qu'il faut être en mesure de pouvoir expliquer individuellement chaque échec si on veut trouver une manière de tous les résoudre. Cela implique entre autres l'examen parfois fastidieux de nombreuses séquences et erreurs de prédiction. De nouvelles voies ont été ainsi partiellement explorées et les résultats préliminaires présentés dans cette thèse laissent espérer qu'à l'avenir elles nous rapprocheront encore plus du but.

Pour conclure, considérons l'exempe du ribocommutateur TPP. La structure de cet ARN lié à son substrat a été résolue [117]. Sa structure sans substrat ne l'est pas encore mais le groupe du Professeur Doniach y travaille. En attendant les résultats, voici les prédictions de structures obtenues avec les modèles MC et Turner99 :



(a) Meilleure structure secondaire obtenue avec Turner99
(b) Meilleure structure secondaire obtenue avec MC

Bibliographie

[1] N.B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(04) :499–512, 2001.

[2] HF Noller, V. Hoffarth, and L. Zimniak. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science*, 256(5062) :1416–1419, 1992.

[3] R. Giegé, M. Frugier, and J. Rudinger. tRNA mimics. *Current opinion in structural biology*, 8(3) :286–293, 1998.

[4] JHA Nagel, AP Gultyaev, K. Gerdes, and CWA Pleij. Metastable structures and refolding kinetics in hok mRNA of plasmid R1. *RNA*, 5(11) :1408–1418, 1999.

[5] C. Yin, F.N. Salloum, and R.C. Kukreja. A Novel Role of MicroRNA in Late Preconditioning : Upregulation of Endothelial Nitric Oxide Synthase and Heat Shock Protein 70. *Circulation Research*, 104(5) :572, 2009.

[6] B.P. Lewis, C.B. Burge, and D.P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1) :15–20, 2005.

[7] S. Sassen, E.A. Miska, and C. Caldas. MicroRNA : implications for cancer. *Virchows Archiv*, 452(1) :1–10, 2008.

[8] J. Johansson, P. Mandin, A. Renzoni, C. Chiaruttini, M. Springer, and P. Cossart. An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell*, 110(5) :551–561, 2002.

[9] W. Winkler, A. Nahvi, and R.R. Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910) :952–956, 2002.

[10] R.F. Gesteland, J.F. Atkins, and H. Storchova. *The RNA world*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY :, 2006.

[11] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183) :51–55, 2008.

[12] P. Brion and E. Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1) :113–137, 1997.

[13] S.L. Heilman-Miller, D. Thirumalai, and S.A. Woodson. Role of counterion condensation in folding of the Tetrahymena ribozyme. I. Equilibrium stabilization by cations. *Journal of Molecular Biology*, 306(5) :1157–1166, 2001.

[14] R.B. Lyngso and C.N.S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4) :409–427, 2000.

[15] M. Zuker, J.A. Jaeger, and D.H. Turner. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic acids research*, 19(10) :2707, 1991.

[16] K.J. Doshi, J.J. Cannone, C.W. Cobaugh, and R.R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5(1) :105, 2004.

[17] D.H. Mathews and D.H. Turner. Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3) :270–278, 2006.

[18] JS McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29 :1105–1119, 1990.

[19] E.A. Schultes and D.P. Bartel. One sequence, two ribozymes : Implications for the emergence of new ribozyme folds. *Science*, 289(5478) :448–452, 2000.

[20] G. Vesnaver and KJ Breslauer. The contribution of DNA single-stranded order to the thermodynamics of duplex formation. *Proceedings of the National Academy of Sciences*, 88(9) :3569–3573, 1991.

[21] J.A. Holbrook, M.W. Capp, R.M. Saecker, M.T. Record Jr, et al. Enthalpy and Heat Capacity Changes for Formation of an Oligomeric DNA Duplex : Interpretation in Terms of Coupled Processes of Formation and Association of Single-Stranded Helices†. *Biochemistry*, 38(26) :8409–8422, 1999.

[22] D. Porschke, OC Uhlenbeck, and FH Martin. Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers*, 12(6) :1313–1335, 1973.

[23] PN Borer, B. Dengler, I. Tinoco Jr, and OC Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *Journal of molecular biology*, 86(4) :843, 1974.

[24] J.M. Blose, M.L. Manni, K.A. Klapec, Y. Stranger-Jones, A.C. Zyra, V. Sim, C.A. Griffith, J.D. Long, and M.J. Serra. Non-Nearest-Neighbor Dependence of

the Stability for RNA Bulge Loops Based on the Complete Set of Group I Single-Nucleotide Bulge Loops†. *Biochemistry*, 46(51) :15123–15135, 2007.

[25] 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(16) :4844–4851, 1997.

[26] N. Sugimoto, R. Kierzek, S.M. Freier, and D.H. Turner. Energetics of internal GU mismatches in ribooligonucleotide helixes. *Biochemistry*, 25(19) :5755–5759, 1986.

[27] T. Xia, J. SantaLucia Jr, 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(42) :14719–14735, 1998.

[28] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5) :911–940, 1999.

[29] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19) :7287–7292, 2004.

[30] P.J. Mikulecky and A.L. Feig. Heat capacity changes in RNA folding : application of perturbation theory to hammerhead ribozyme cold denaturation. *Nucleic acids research*, 32(13) :3967, 2004.

[31] P. Wu, S. Nakano, and N. Sugimoto. Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation, 2002.

[32] J.C. Takach, P.J. Mikulecky, and A.L. Feig. Salt-dependent heat capacity changes for RNA duplex formation. *Journal of the American Chemical Society*, 126(21) :6530, 2004.

[33] T.V. Chalikian, J. Volker, G.E. Plum, and K.J. Breslauer. A more unified picture for the thermodynamics of nucleic acid duplex melting : a characterization by calorimetric and volumetric techniques. *Proceedings of the National Academy of Sciences*, 96(14) :7853–7858, 1999.

[34] D.H. Mathews and D.H. Turner. Experimentally derived nearest-neighbor parameters for the stability of RNA three-and four-way multibranch loops. *Biochemistry*, 41(3) :869–880, 2002.

[35] K.J. Doshi, J.J. Cannone, C.W. Cobaugh, and R.R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5(1) :105, 2004.

[36] M. Andronescu, A. Condon, H.H. Hoos, D.H. Mathews, and K.P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13) :i19, 2007.

[37] C.B. Do, D.A. Woods, and S. Batzoglou. CONTRAfold : RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14) :e90–e98, 2006.

[38] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6) :446–454, 1999.

[39] R.D. Dowell and S.R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC bioinformatics*, 5(1) :71, 2004.

[40] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. Rfam : An RNA family database. *Nucleic Acids Research*, 31(1) :439, 2003.

[41] M. Sprinzl and K.S. Vassilenko. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 33 :D139–D140, 2005.

[42] J. Pan, D. Thirumalai, and S.A. Woodson. Folding of RNA involves parallel pathways. *Journal of molecular biology*, 273(1) :7–13, 1997.

[43] M. Helm, H. Brule, F. Degoul, C. Cepanec, JP Leroux, R. Giege, and C. Florentz. The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial tRNA. *Nucleic acids research*, 26(7) :1636, 1998.

[44] S.R. Morgan and P.G. Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *The Journal of Chemical Physics*, 105 :7152, 1996.

[45] G.R. Björk, J.M.B. Durand, T.G. Hagervall, R. Leipuvien, H.K. Lundgren, K. Nilsson, P. Chen, Q. Qian, and J. Urbonavičius. Transfer RNA modification : influence on translational frameshifting and metabolism. *FEBS letters*, 452(1-2) :47–51, 1999.

[46] T.G. Hagervall, S.C. Pomerantz, and J.A. McCloskey. Reduced misreading of asparagine codons by Escherichia coli tRNALys with hypomodified derivatives of 5-methylaminomethyl-2-thiouridine in the wobble position. *Journal of molecular biology*, 284(1) :33–42, 1998.

[47] J. Ofengand and C. Henes. The Function of Pseudouridylic Acid in Transfer Ribonucleic Acid. *Journal of Biological Chemistry*, 244(22) :6241–6253, 1969.

[48] M. Helm, R. Giege, and C. Florentz. A Watson-Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of human mitochondrial tRNALys. *Biochemistry*, 38(40) :13338–13346, 1999.

[49] JJ Dalluge, T. Hashizume, AE Sopchik, JA McCloskey, and DR Davis. Conformational flexibility in RNA : the role of dihydrouridine. *Nucleic acids research*, 24(6) :1073, 1996.

[50] A. Beniaminov, E. Westhof, and A. Krol. Distinctive structures between chimpanzee and humanin a brain noncoding RNA. *RNA*, 14(7) :1270, 2008.

[51] M.A. Rosenblad, J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB : Signal recognition particle database. *Nucleic acids research*, 31(1) :363, 2003.

[52] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic acids research*, 31(13) :3450, 2003.

[53] A. Lescoute and E. Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12(1) :83–93, 2006.

[54] G. Thill, M. Vasseur, and N.K. Tanner. Structural and sequence elements required for the self-cleaving activity of the hepatitis delta virus ribozyme. *Biochemistry*, 32(16) :4254–4262, 1993.

[55] MY Kuo, L. Sharmeen, G. Dinter-Gottlieb, and J. Taylor. Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. *Journal of Virology*, 62(12) :4439–4444, 1988.

[56] A.R. Ferré-D'Amaré, K. Zhou, and J.A. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395(6702) :567–574, 1998.

[57] T.R. Cech, A.J. Zaug, P.J. Grabowski, et al. In vitro splicing of the ribosomal RNA precursor of Tetrahymena : involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27(3 Pt 2) :487–496, 1981.

[58] P.L. Adams, M.R. Stahley, M.L. Gill, A.B. Kosek, J. Wang, and S.A. Strobel. Crystal structure of a group I intron splicing intermediate. *Rna*, 10(12) :1867–1887, 2004.

[59] K. Rietveld, R. Van Poelgeest, CWA Pleij, JH Van Boom, and L. Bosch. The tRNA-like structure at the 3'terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Research*, 10(6) :1929, 1982.

[60] A. Belkum, J.P. Abrahams, C.W.A. Pleij, and L. Bosch. Five pseudoknots are present at the 204 nucleotides long 3'noncoding region of tobacco mosak virus RNA. *Nucleic Acids Research*, 13(21) :7673, 1985.

[61] D. Matsuda and T.W. Dreher. The tRNA-like structure of Turnip yellow mosaic virus RNA is a 3'-translational enhancer. *Virology*, 321(1) :36–46, 2004.

[62] Leendert Bosch Edwin B. Dam, Cornelius W. A. Pleij. RNA pseudoknots : Translational frameshifting and readthrough on viral RNAs. *Virus genes*, 4(2) :121–136, 1990.

[63] O. Namy, S.J. Moran, D.I. Stuart, R.J.C. Gilbert, and I. Brierley. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature*, 441(7090) :244–247, 2006.

[64] RB Lyngsø, M. Zuker, and CNS Pedersen. An improved algorithm for RNA secondary structure prediction [Technical Report RS-99-15], 1999.

[65] H. Orland and A. Zee. RNA folding and large N matrix theory. *Nuclear Physics, Section B*, 620(3) :456–476, 2002.

[66] G. t Hooft. A planar diagram theory for strong interactions. *Nuclear Physics B*, 72(3) :461–473, 1974.

[67] M. Pillsbury, H. Orland, and A. Zee. Steepest descent calculation of RNA pseudoknots. *Physical Review E*, 72(1) :11911, 2005.

[68] WJ Melchers, JG Hoenderop, HJ Bruins Slot, CW Pleij, EV Pilipenko, VI Agol, and JM Galama. Kissing of the two predominant hairpin loops in the coxsackie B virus 3'untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *Journal of virology*, 71(1) :686–696, 1997.

[69] TC Gluick and DE Draper. Thermodynamics of folding a pseudoknotted mRNA fragment. *Journal of molecular biology*, 241(2) :246–262, 1994.

[70] FHD Van Batenburg, AP Gultyaev, CWA Pleij, J. Ng, and J. Oliehoek. PseudoBase : a database with RNA pseudoknots. *Nucleic Acids Research*, 28(1) :201, 2000.

[71] G. Vernizzi, H. Orland, and A. Zee. Enumeration of RNA structures by matrix models. *Physical review letters*, 94(16) :168103, 2005.

[72] T.C. Gluick, R.B. Gerstner, and D.E. Draper. Effects of Mg2+, K+, and H+ on an equilibrium between alternative conformations of an RNA pseudoknot. *Journal of molecular biology*, 270(3) :451–463, 1997.

[73] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of RNA structures. *Journal of Molecular Biology*, 379(4) :900–911, 2008.

[74] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1) :68–82, 1978.

[75] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1) :133–148, 1981.

[76] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13) :3406, 2003.

[77] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19) :7287–7292, 2004.

[78] IL Hofacker, W. Fontana, PF Stadler, LS Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f "ur Chemie/Chemical Monthly*, 125(2) :167–188, 1994.

[79] T.R. Einert, P. Näger, H. Orland, and R.R. Netz. Impact of loop statistics on the thermodynamics of RNA folding. *Physical Review Letters*, 101(4) :48103, 2008.

[80] Van Batenburg FHD Gultyaev AP and Pleij CWA. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, 250(1) :37–51, 1995.

[81] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285(5) :2053–2068, 1999.

[82] R.M. Dirks and N.A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry*, 24(13) :1664–1677, 2003.

[83] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC bioinformatics*, 5(1) :104, 2004.

[84] H. Isambert and E.D. Siggia. Modeling RNA folding paths with pseudoknots : application to hepatitis delta virus ribozyme. *Proceedings of the National Academy of Sciences*, 97(12) :6515–6520, 2000.

[85] J. Ruan, G.D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1) :58–66, 2004.

[86] J. Ren, B. Rastegari, A. Condon, and H.H. Hoos. HotKnots : heuristic prediction of RNA secondary structures including pseudoknots. *Rna*, 11(10) :1494–1504, 2005.

[87] D. Metzler and M.E. Nebel. Predicting RNA secondary structures with pseudoknots by MCMC sampling. *Journal of Mathematical Biology*, 56(1) :161–181, 2008.

[88] C. Papanicolaou, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. *Nucleic Acids Research*, 12(1 Pt 1) :31–44, 1984.

[89] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. Introduction to algorithms, 2001.

[90] K. Rietveld, R. Van Poelgeest, CWA Pleij, JH Van Boom, and L. Bosch. The tRNA-Like structure at the 3'terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Research*, 10(6) :1929, 1982.

[91] A. Belkum, J.P. Abrahams, C.W.A. Pleij, and L. Bosch. Five pseudoknots are present at the 204 nucleotides long 3'noncoding region of tobacco mosak virus RNA. *Nucleic Acids Research*, 13(21) :7673, 1985.

[92] I. Barrette, G. Poisson, P. Gendron, and F. Major. Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic Acids Research*, 29(3) :753, 2001.

[93] C. Chiaruttini, M. Milet, and M. Springer. A long-range RNA-RNA interaction forms a pseudoknot required for translational control of the IF3-L35-L20 ribosomal protein operon in Escherichia coli. *EMBO JOURNAL*, 15(16) :4402–4413, 1996.

[94] PR Wills. Potential pseudoknots in the PrP-encoding mRNA. *Journal of theoretical biology*, 159(4) :523, 1992.

[95] B. Felden, C. Massire, E. Westhof, J.F. Atkins, and R.F. Gesteland. Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Research*, 29(7) :1602, 2001.

[96] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, and H.J.C. Berendsen. GROMACS : fast, flexible, and free. *Journal of Computational Chemistry*, 26(16) :1701–1718, 2005.

[97] T.E. Cheatham III and M.A. Young. Molecular dynamics simulation of nucleic acids : Successes, limitations, and promise. *Biopolymers*, 56(4) :232–256, 2000.

[98] E.J. Sorin, Y.M. Rhee, B.J. Nakatani, and V.S. Pande. Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. *Biophysical journal*, 85(2) :790–803, 2003.

[99] U.H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3) :140–150, 1997.

[100] A. Villa, E. Widjajakusuma, and G. Stock. Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg. *J. Phys. Chem. B*, 112(1) :134–142, 2008.

[101] G.M. Torrie and J.P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling : application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett*, 28(4) :578–581, 1974.

[102] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. *Journal of Computational Chemistry*, 13(8) :1011–1021, 1992.

[103] N.K. Banavali and A.D. MacKerell. Free energy and structural pathways of base flipping in a DNA GCGC containing sequence. *Journal of molecular biology*, 319(1) :141–160, 2002.

[104] E. Giudice and R. Lavery. Nucleic acid base pair dynamics : the impact of sequence and structure using free-energy calculations. *J. Am. Chem. Soc*, 125(17) :4998–4999, 2003.

[105] TP Straatsma and HJC Berendsen. Free energy of ionic hydration : Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *The Journal of Chemical Physics*, 89(9) :5876–5886, 1988.

[106] J.W. Pitera and W.F. van Gunsteren. A comparison of non-bonded scaling approaches for free energy calculations. *Molecular Simulation*, 28(1) :45–65, 2002.

[107] DC Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[108] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2) :926–935, 1983.

[109] HJC Berendsen, JR Grigera, and TP Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24) :6269–6271, 1987.

[110] H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations : TIP4P-Ew. *The Journal of chemical physics*, 120(20) :9665–9678, 2004.

[111] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald : An Nlog (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12) :10089–10092, 1993.

[112] D. van der Spoel and P.J. van Maaren. The origin of layer structure artifacts in simulations of liquid water. *J. Chem. Theory Comput*, 2(1) :1–11, 2006.

[113] M.R. Shirts and V.S. Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of chemical physics*, 122(13) :134508, 2005.

[114] J. Wang, P. Cieplak, and P.A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules ? *Journal of Computational Chemistry*, 21(12), 2000.

[115] C. Azuara, H. Orland, M. Bon, P. Koehl, and M. Delarue. Incorporating dipolar solvents with variable density in Poisson-Boltzmann electrostatics. *Biophysical Journal*, 95(12) :5587–5605, 2008.

[116] J. Higo and M. Nakasako. Hydration structure of human lysozyme investigated by molecular dynamics simulation and cryogenic X-ray crystal structure analyses : on the correlation between crystal water sites, solvent density, and solvent dipole. *Journal of computational chemistry*, 23(14) :1323–1336, 2002.

[117] S. Thore, M. Leibundgut, and N. Ban. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science*, 312(5777) :1208–1211, 2006.