Simulations et expériences sur le repliement de l'ARN : prédictions statistiques des pseudonoeuds in silico et réalisation de commutateurs ARN par transcription in vitro

Alain Xayaphoummine

Thèse de Doctorat
de l'Université Louis PASTEUR
STRASBOURG I

présentée par
Alain XAYAPHOUMMINE

pour obtenir le grade de
Docteur de l'Université STRASBOURG I

Spécialité : Physique

Simulations et expériences sur le repliement de l'ARN : prédictions statistiques des pseudonœuds *in silico* et réalisation de commutateurs ARN par transcription *in vitro*.

Soutenue le 18 juin 2004 devant le jury composé de :

M. Jean-Louis VIOVY	(rapporteur externe)
M. Thomas SIMONSON	(rapporteur externe)
M. Jörg BASCHNAGEL	(président du jury)
M. Benoît MASQUIDA	(examinateur)
M. Didier CHATENAY	(directeur)
M. Hervé ISAMBERT	(co-directeur)

Remerciements

Je tiens à remercier Didier Chatenay qui m'a accueilli au sein du laboratoire de dynamique des fluides complexes. Didier est un homme entier, amoureux de la science et qui partage son enthousiasme débordant à tous ceux qui ont la chance de le côtoyer.

Je remercie Hervé Isambert, qui m'a encadré et guidé tout au long de ce travail. Je le remercie pour sa confiance et pour la liberté qu'il m'a laissées.

Je remercie les rapporteurs de cette thèse Jörg Baschnagel, Jean-Louis Viovy, Thomas Simonson pour la rapidité avec laquelle ils ont lu mon manuscrit et l'intérêt qu'ils ont portés à mon travail. Merci également à Benoît Masquida qui a pris très à cœur son rôle d'examinateur.

J'aimerais remercier l'ensemble des membres du laboratoire de dynamique de fluides complexes, pour les bons moments passé ensembles. En particulier, Fabrice Thalmann avec qui j'ai beaucoup interagi. Jérôme Robert, qui m'a accueillit dans le groupe de biophysique expérimentale. Carlos Marques si simple dans ses relations et si pertinents dans ses conseils . Nicolas Rivier toujours de bonne humeur et ses petites histoires de physique qui ont égayé des journées de travail parfois longues et fastidieuses.

Sans mes deux stagiaires, Thomas Bucher et Myriam Benisty, je n'aurais certainement pas pu obtenir d'interface web aussi sécurisé et fonctionnel, ni pu tester autant de technique de gels de polyacrylamides. Je remercie leur patience, et leur travail exemplaire qu'ils ont fourni lors de leur stage sous ma direction.

J'ai beaucoup appris au contact d'autres thésards et j'aimerais les remercier ici pour leurs conseils et les cafés échangés. Je citerais Thierry Marchal, Sébastien Harlepp, Nicolas Douarche et Christian Rick.

Je n'oublierais pas les aides permanentes reçues du personnel administratif et technique. Nadia Bouaouina et Josianne Haccoun les deux secrétaires du laboratoire, qui ont grandement facilité toutes les opérations administratives courantes. Patrick Allgayer, et Alain Steyer pour leur cuve d'électrophorèse " maison ". Marc Basler pour ses astuces d'électronique. Et je n'oublierais pas Annie Picchinenna qui m'a beaucoup aidé dans la phase de clonage en bactérie.

Jean-Claude Massot, Céline Xayaphoummine et Thierry Renault, et Céline Pocquet qui ont relu attentivement tout ou partie de ce manuscrit. Merci à tout ce temps qu'ils ont consacré à redonner un peu de rigueur à mon écriture souvent lourde.

Enfin, je remercie ma femme, pour m'avoir soutenue et encouragé tout au long de ces années et pour m'avoir donnée la joie d'être papa d'une adorable petite fille, Aurélie.

Table des matières

Première partie

Introduction

Introduction générale

Contexte général

Les polymères sont des macromolécules constituées par un arrangement de blocs élémentaires. Ces blocs, autrement appelés *monomères*, sont reliés entre eux par des liaisons covalentes pour former la *chaîne*. Les propriétés physiques des homopolymères sont essentiellement déterminées par la nature du monomère, la longueur de la séquence et la nature de la jonction entre les monomères.

La majeure partie des biopolymères sont des "hétéropolymères", leur séquence est construite à l'aide d'un ensemble fini de monomères chimiquement différents. Les protéines sont par exemple constituées à l'aide de 20 briques élémentaires appelées acides aminés. Les acides nucléiques ADN ou ARN ne font appel qu'à quatre nucléotides différents : l' Adénine (notée A), la Guanine (G), la Cytosine (C), et la Thymine (T) remplacée par l'Uracile(U) dans l'ARN. La physique des hétéropolymères reprend les comportements des chaînes et des jonctions de monomères, mais l'effet physique essentiel est influencé par la succession des monomères le long de la séquence.

Les interactions entre les résidus permettent aux polymères de se replier sur eux-mêmes. Si l'attraction entre les résidus l'emporte sur l'affinité entre la chaîne et le solvant, les homopolymères et les hétéropolymères se compactent en général sous la forme de pelote de "conformation aléatoire". Dans le cas spécifique des biopolymères, cet arrangement est guidé par les interactions spécifiques entre les monomères qui guident le repliement vers une conformation généralement "unique" compacte appelée "conformation native". Lors du processus de

réarrangement, les biopolymères tendent à minimiser l'exposition de leurs parties hydrophobes avec le solvant.

La formation de la conformation la plus stable est guidée par le principe de moindre énergie libre. Pour l'ADN et l'ARN, la structure d'équilibre contrebalance la perte d'énergie entropique due à l'appariement des paires de base par l'énergie de "stacking" qui corespond à la sommation des deux contributions énergétique que sont les énergies d'appariement et d'empilement des paires de base. Cet arrangement est stabilisé par les liaisons hydrogènes qui se forment entre ces bases complémentaires positionnées en vis-à-vis appartenant au même brin (ARN ou ADN simple brin) ou à un brin complémentaire (ADN double brin, ARN antisens) [70].

Les structures natives des biopolymères sont biologiquement fonctionnelles. Il est donc à la fois essentiel pour les cellules qu'ils adoptent leur conformation native mais aussi que celle-ci soit atteinte dans un laps de temps biologiquement compatible. D'un point de vue théorique, on distingue deux aspects dans la recherche de la structure native. Le premier aspect est la prédiction correcte de la structure native. Le second est la modélisation correcte de la dynamique du processus de repliement.

Les expériences de repliement de molécules d'ARN *in vitro*, montrent que certaines molécules peuvent rester piégées dans des conformations alternatives non fonctionnelles [57, 61, 81]. Une partie d'entre elles pouvant être renaturées en interagissant de manière non spécifique avec des protéines. Ces expériences *in vitro* qui tentent de comprendre les processus *in vivo*, démontrent l'importance du design de la séquence [58, 69]. Bien que la molécule puisse finir par trouver sa structure native, la cinétique de repliement est parfois accélérée par certaines protéines qui empêchent la formation de structures métastables fortement piégées. Mais au-delà du seul codage de la structure fonctionnelle, la séquence des bases semble aussi détenir des informations essentielles au bon déroulement de la cinétique de repliement. C'est cette propriété des séquences d'ARN qui nous a intéressé dans cette thèse.

Aspects Biologiques de l'ARN

Tout au long de l'évolution, la nature a développé le principe d'organisation hiérarchique. Quelques éléments simples sont utilisés pour créer de plus grosses unités, qui à leur tour sont uti-

lisées comme briques élémentaires. Chaque niveau d'organisation ayant de nouvelles propriétés propres à l'organisation.

Le rôle de l'ADN est de contenir tout le patrimoine génétique dépendant des gènes hérités exprimés ou non (génotype). Quant aux protéines, elles assurent le bon fonctionnement cellulaire en accord avec les conditions externes. Les ARN peuvent endosser toutes ces fonctions ; ils peuvent stocker l'information génétique (ARN des virus), sont l'intermédiaire entre l'ADN et la protéine (ARN messagers) et présentent des activités catalytiques ou enzymatiques (ribozymes). Selon Spiegelman [76], le phénotype des ARN peut être relié à leur configuration spatiale, ainsi les ARN présentent la propriété d'exprimer à la fois le génotype et le phénotype sur le même support. Cette fonctionnalité multiple est la pierre angulaire de l'hypothèse d'un monde tout ARN aux tous premiers instants de la vie [22]

Dans le fonctionnement cellulaire simplifié, les portions codantes de l'ADN sont transcrites en ARN messager (ARNm) par la protéine ARN polymérase. Les ARN messagers sont ensuite traduit en protéines par la machinerie du ribosome. Ces dernières peuvent interagir de manière spécifique sur le double brin ADN pour réguler la transcription (exemple du Lac Z). La traduction du code génétique contenant quatre unités élémentaires -A,U(T),G,C- en séquence d'acide aminé (20 différents) se fait par l'intermédiaire d'un ARN fonctionnel, l'ARN de transfert (ARNt). La combinatoire de trois acides nucléiques (codon) permet le codage de 64 mots différents soit suffisamment pour coder de manière redondante les 20 acides aminés différents. Lors de la traduction les tRNA sont admis au cœur du ribosome qui les décharge de leur acide aminé pour l'accrocher à la chaîne de la séquence de la protéine. Les ARN messagers sont pris en tenaille entre les deux sous-parties principales du ribosome. La petite sous-unité vient reconnaître l'ARN messager, puis la grosse sous unité contenant les sillons de guidage des ARN de transfert vient ensuite coiffer le tout. Le processus fait intervenir les ARN ribosomaux (ARNr) qui catalysent notamment la formation de la liaison peptidique.

Ces dernières années, le monde des ARN a connu une effervescence particulière liée à la découverte de beaucoup d'autres types d'ARN fonctionnel. Ces ARN non-messager qui incluent les ARN ribosomaux et de transfert, se révèlent d'une remarquable diversité structurale et fonctionnelle. Avec la fin du décryptage du génome humain et le fait que 98% des séquences sont non codantes chez les eucaryotes multicellulaires évolués, il apparaît que l'expression tempo-

relle des gènes et les évènements post transcriptionnels doivent être finement orchestrés, pour exprimer toute la variété phénotypique entre les espèces et les individus. Depuis les années quatre-vingt, d'autres types de fonctions ont été attribuées aux ARNs ; fonctions enzymatiques et catalytiques, utilisées dans la maturation des ARN messagers et de transfert. Les ARN des introns de groupe I et II permettent l'auto-épissage des introns et la ligation entre exons des ARN messagers, l'ARN de la RNase P mature les pré-ARN de transfert, les *small nuclear RNA* sont requis lors d'étapes conduisant au découpage des précurseurs des ARN messagers. Chaque famille d'ARNnm partage la même signature structurale (séquence ou corrélation d'appariement de base dans la structure native). Chaque élément d'une famille se distingue l'un de l'autre par une variabilité de ces éléments structuraux leur procurant une spécificité de substrat ou éventuellement de fonction. Deux familles se distinguent actuellement des autres : les *small nucleolar RNA* (ARNsno) et les *micro RNA* (ARNmi). Ces deux familles ciblent les ARN à travers leur complémentarité de séquence. Les ARNsno participent à la modification post transcriptionelle des ARN ribosomaux. Quant aux ARNmi ils se lient aux ARNm des gènes dont ils régulent l'expression, soit au travers d'un contrôle de la traduction, soit par l'activation de la dégradation du complexe ARNmi-ARNm, au même titre que les ARN d'interférence *siARN*, mais selon des processus distincts. Une dernière famille joue un rôle prépondérant dans la dégradation des transcrits avortés d'ARN messagers. Les *tmRNA* permettent de relâcher un ribosome engagé dans un ARN messager qui ne contient pas de codon stop. Ils interagissent en lieu et place de l'ARN de transfert et servent de cible pour les protéines impliquées dans la dégradation de la séquence tronquée d'ARN messager ainsi que de celle de la séquence d'acide aminé qu'il a produit. Tous ces ARNnm sont directement fonctionnels à travers leur structure ARN et agissent comme riborégulateurs dans un grand nombre de processus qui requièrent une reconnaissance spécifique d'un autre acide nucléique.

Aspect fonctionnel des pseudonœuds

Les configurations tridimensionnelles de certaines molécules d'ARN présentent des imbrications d'hélices dénommées pseudonœuds (figure 1). Dans le repliement tridimensionnel, les pseudonœuds apparaissent la plupart du temps comme des interactions locales et induisent des

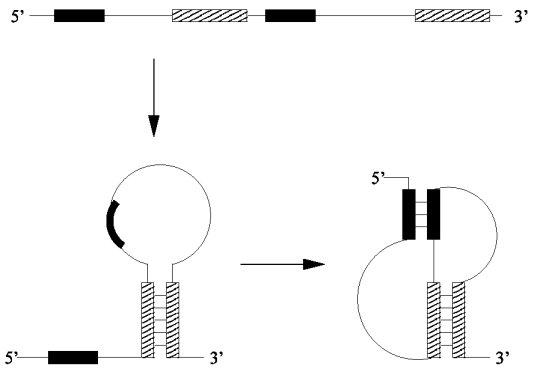


FIG. 1 – Pseudonœud

rôles différents selon leur position et la nature de l'ARN. Les différentes fonctionnalités induites par la formation de pseudonœuds peuvent être regroupées en trois grandes familles.

- **Actions sur la traduction** : en début de séquence, les pseudonœuds servent de cible pour des protéines de contrôles dans les ARN messagers [27]. Dans plusieurs virus, l'expression de la protéine de *replicase* est contrôlée par le processus de décalage négatif ou positif de la fenêtre de codage (*ribosomal frame-shifting et in-frame read-through*), qui se traduit par le recul ou l'avance d'un nucléotide le long de la séquence de l'ARN messager. Dans le second cas, le pseudonœud permet de passer au-dessus du codon d'arrêt de transcription AUG(X) en décalant la fenêtre de lecture sur le codon (A)UGX. La position de ce type de pseudonœuds est essentielle pour le bon fonctionnement de cette propriété [7, 10, 87]. La présence des trois pseudonœuds de l'ARN ribosomal du 16S semble essentielle pour l'interaction avec les protéines de structures. Et l'un d'entre eux est connu comme étant essentiel dans le maintien de l'ARN de transfert et le site de reconnaissance du codon.
- **Action sur la structure** : les pseudonœuds de cœur servent de poutrelle d'architecture pour les ARN enzymatiques. Ceci a été démontré sur le cas de l'ARN de la RNAse P [24], les introns de groupe I [58] et le HDV [18]
- **Action sur la réplication** : beaucoup de virus de plante utilisent un pseudonœud pour se replier dans une structure tridimensionnelle proche de celle des ARN de transfert [45].Cette ressemblance structurale permet d'en usurper les propriétés biologiques.

Les enzymes de reconnaissance des ARN de transfert interagissent indifféremment avec ces structures et leur fonction en est détournée [45]. Le pseudonœud en fin de séquence permet aussi au virus de la mosaïque du tabac de mimer une queue poly(A) stabilisant l'ARN messager et augmentant l'expression du gène d'un facteur 100 [19].

Pertinence de l'approche cinétique

Les biopolymères trouvent leur structure fonctionnelle au cours d'un repliement spontané ou à l'aide de molécules chaperones. Pour la grande majorité d'entre eux, cette cinétique dure de l'ordre de quelques fractions de seconde à quelques minutes dans les conditions physiologiques de température et de salinité. Et pourtant, la recherche exhaustive de la configuration native, parmi l'ensemble des configurations possibles prendrait plusieurs millions d'années, en supposant une équiprobabilité des conformations. Pour résoudre le problème, il existe en fait une inégalité dans le choix entre les différentes configurations. Cette inégalité reflète la tendance naturelle qu'a le système à préférer les structures plus stables. Ce biais naturel permet de s'approcher rapidement de la structure native ou tout du moins d'un minimum énergétique local. Le paysage énergétique des structures possibles peut être visualisé comme un relief Alpin. Cette analogie permet d'imaginer plusieurs solutions au problème de repliement, menant d'une configuration excitée à la structure native. Au lieu d'une vision figée du repliement empruntant toujours le même chemin, la représentation en termes de paysage énergétique permet d'imaginer des processus beaucoup plus stochastiques. Chaque chaîne polymère suit une histoire cinétique différente en empruntant des structures intermédiaires différentes, tout en atteignant en fin de compte, la ou les configuration(s) native(s). Notez que dans ce travail, les intéractions avec des molécules chaperone ne sont pas directement modélisées mais peuvent être simuler *ad hoc* en contraignant la formation de cetaines hélices

Tout au long du parcours, des interactions non-natives et natives sont mises en compétition pour permettre au système de trouver les hélices de la conformation native. Le repliement obéit à deux forces contraires, l'interaction locale entre bases hydrophobes proches voisines, minimisant leur interaction avec le solvant en adoptant une structure d'empilement, et les interactions électrostatiques globales dues au squelette phosphaté fortement chargé négativement. L'impos-

sibilité de résoudre toutes les interactions en même temps, mène à un système énergétiquement frustré. Quant au paysage énergétique, il exprime cette propriété par son aspect montagneux. Au cours de l'évolution, les séquences des ARN fonctionnels ont pu être optimisées afin de réduire au mieux les frustrations du système et ainsi faciliter la recherche de l'optimum. Néanmoins, une hélice de cinq paires de bases peut facilement atteindre des énergies de stabilisation de l'ordre de 41 kJ/mol (10 kcal/mol) alors que l'énergie d'activation thermique kT est de l'ordre de 2.5 kJ/mol (0.6 kcal/mol). La comparaison entre l'énergie thermique kT et l'énergie de dissociation d'une hélice formée nous permet d'affirmer qu'il est relativement aisé de piéger des ARN dans des configurations non natives. Cet effet est d'autant plus important que la séquence est longue, et donc que la longueur et le nombre d'hélices en compétition augmentent. Au cours d'un repliement, ces hélices non-natives mais particulièrement stables sont énergétiquement favorisées, et peuvent guider la molécule vers un état piégé, métastable et non fonctionnel. L'accès à la configuration native, se fait alors le long d'un processus lent, les hélices non natives devant être dégrafées pour laisser place à des configurations plus stables [69].

Au cours du processus de repliement co-transcriptionnel, nous avons posé par hypothèse que la molécule se replie selon une cascade d'intermédiaires formés d'hélices natives et non-natives. L'ordre d'apparition des bases et la fréquence de synthèse, régulent la formation des hélices en augmentant régulièrement le nombre d'hélices en compétition. Néanmoins, pour placer une nouvelle hélice, il faut la plupart du temps commencer par dégrafer tout ou partie d'un certain nombre d'hélices déjà formées. L'inégalité qui en résulte, guide la molécule à travers le paysage énergétique, en favorisant les chemins les plus probables. Dans ce schéma, la séquence des bases contient toute l'information sur l'ordre d'apparition des hélices en compétition, et donc, la stabilité du ou des chemins qui mènent à la structure native.

Organisation du manuscrit

Ce travail s'articule autour de trois grands thèmes : prédiction numérique, approche expérimentale et applications publiques.

– **Prédiction numérique**. Dans cette partie seront discutés les différents types d'approches communément utilisées ainsi que notre modélisation de la cinétique du repliement des

structures d'ARN avec pseudonœuds. Comme souligné précédemment, il est aisé de piéger la molécule dans des conformations métastables, et de manière plus générale, il n'est pas improbable que certaines parties de l'espace des configurations soient sans cesse revisitées. Pour accélérer l'exploration de l'espace des structures, je présenterai un modèle simple d'accélération exacte de la cinétique de repliement qui s'appuie sur un réseau d'états précédemment visités au cours de la simulation. Cette méthode généralisable à tous les systèmes frustrés, permet de visiter à coup sûr un nouvel état à chaque itération tout en intégrant la cinétique de manière exacte sur l'ensemble des chemins qui y mènent. Cette méthode permet de plus de calculer de manière exacte toutes les moyennes des observables d'intérêt du système. L'intégration sur l'ensemble des trajectoires dynamiques de repliement nous fait perdre la connaissance de la dynamique locale du système interne au cluster, mais permet d'effectuer l'équivalent de la moyenne des chemins de repliement suivit par une infinité de systèmes équivalents. Ce "coarse-graining" de la dynamique locale, est à cheval entre la description exhaustive des transitions locales et une approche thermodynamique du problème. J'utiliserai cette méthode pour déterminer entre autres une distribution dans la statistique de présence de pseudonœuds dans les structures ARN de séquences aléatoires et biologiques.

– **Approche expérimentale**. L'étude de la cinétique de repliement fait apparaître expérimentalement et numériquement l'importance de la séquence elle-même dans le repliement des molécules d'ARN. Le paysage énergétique est directement modulé par cette succession de résidus, et s'il est possible de trouver un ensemble de séquences codant pour une même configuration, une séquence a aussi la possibilité de coder pour plusieurs fonctions. Au cours de l'évolution, les mutations ont sans doute permis au système de s'enrichir en fonctionnalités tout comme d'optimiser les séquences pour un repliement efficace. Au-delà de l'information essentielle des appariements des bases dans la configuration native, d'autres informations semblent aussi codables le long de la séquence des bases. Je m'intéresserai au cas particulier du codage du chemin de repliement co-transcriptionnel c'est-à-dire au chemin de repliement au cours de la synthèse ADN $\rightarrow$ ARN.
– **Applications publiques**. Il est courant dans le monde de la bioinformatique de rendre son programme accessible au plus grand nombre, soit en le diffusant comme code libre,

soit en l'interfaçant sur le WEB. Chaque approche à ses partisans, de notre côté, nous avons opté pour une interface WEB. Dans cette courte partie, nous présenterons le site de ***KineFold*** et son organisation. Nous présenterons aussi l'outil ***RNAMovies*** *with pseudoknots*, version adaptée de ***RNAMovies*** pour la visualisation des animations de cinétique de repliement de molécules d'ARN avec pseudonœuds.

Deuxième partie

Simulation numérique

Chapitre 1

Introduction

1.1 Structure des ARN

Les acides nucléiques sont composés d'une succession de nucléotides. Le nucléotide se décompose en trois parties : *le sucre, la base et le groupe phosphoryl*. L'élément central, *le sucre*, est formé par un pentose sous forme cyclique (furanose) : le ribose pour l'ARN, le 2' désoxyribose pour l'ADN (cf. figure 1.1).

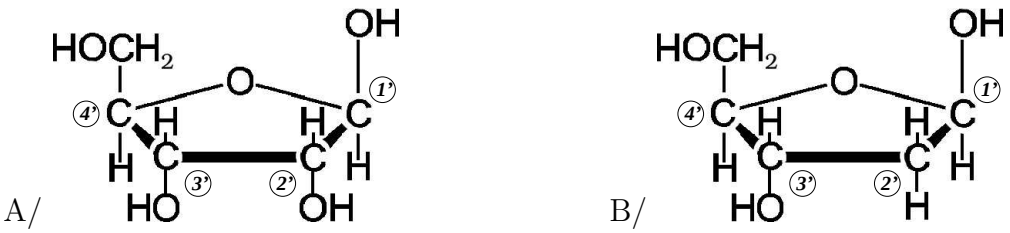


FIG. 1.1 – A/ribose et B/désoxyribose. Les deux molécules ne diffèrent que par l'oxydation de la fonction alcool du carbone 2'.

Une base organique est reliée au sucre par le carbone 1', soit une base purine adénine (**A**) ou guanine (**G**)(cf figure 1.2) soit pyrimidine cytosine (**C**), thymine (**T**) dans l'ADN ou uracile (U) dans l'ARN (cf figure 1.3).

FIG. 1.2 – Bases purine : **A**dénine,**G**uanine

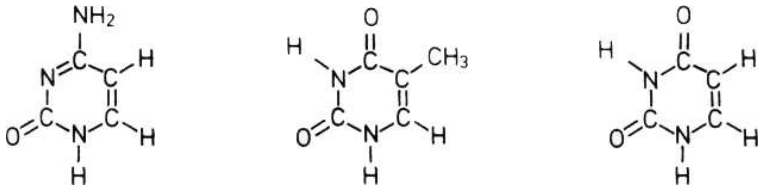


FIG. 1.3 – Bases pyrimidine : **C**ytosine,**T**hymine, **U**racile

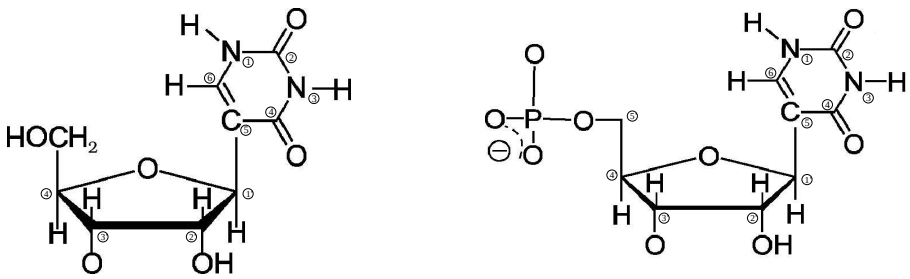


FIG. 1.4 – Nucléoside et nucléotide

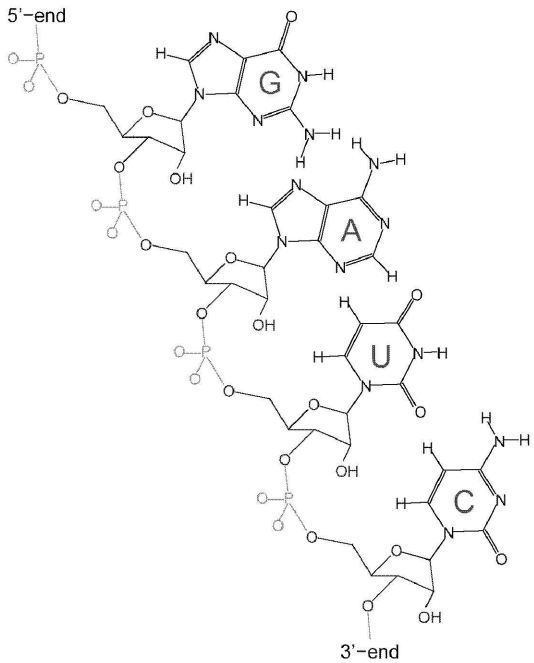


FIG. 1.5 – Chaîne d'ARN. En notation réduite, 5'-GAUC-3', ou encore GAUC

L'ensemble sucre-base forme le *nucléoside* (cf. figure 1.4). Par l'action d'un acide phosphorique la fonction alcool du carbone 5' est estérifiée pour constituer le *nucléotide* (cf. figure 1.4).

La polymérisation se fait par l'estérification d'un alcool situé en 3' d'un nucléotide par le phosphate d'un autre nucléotide, la liaison covalente établie porte le nom de 3'-5' phosphodiester. En raison de l'asymétrie de la liaison, la chaîne obtenue est polaire. Dans la notation conventionnelle, la succession des bases qui forment la séquence, est notée de l'extrémité 5' vers l'extrémité 3' (cf. figure 1.5). La séquence des bases, composée des quatre lettres A,U,G et C, forme ce que l'on appelle la **structure primaire**.

Les cycles aromatiques ont tendance à naturellement s'empiler pour minimiser leur surface de contact avec l'eau, entraînant une structuration hélicoïdale de la chaîne polynucléotidique [70]. la figure 1.6 montre l'empilement de quelques couples de bases.

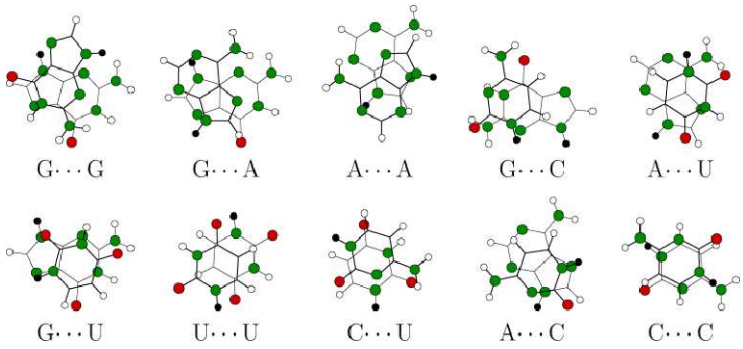


FIG. 1.6 – Empilement naturel obtenu pour quelques couples de bases.

En présence de solvant polaire, la contribution principale à cet effet est d'ordre entropique. En solution, les molécules d'eau se structurent autour des cycles aromatiques, ce qui réduit le nombre de configurations possibles de leurs réseaux de liaisons hydrogène. Suite à l'empilement, la surface exposée au solvant est réduite et les molécules d'eau sont relâchées, ce qui augmente l'entropie globale du système et donc diminue l'énergie libre.

La chaîne structurée longitudinalement peut interagir latéralement par les interactions électrostatiques des terminaisons hydroxyles des bases. Il existe vingt-huit manières différentes d'arranger spatialement deux nucléotides isolés pour former des liaisons hydrogènes. Mais pour des raisons d'encombrement stérique imposées par la structure hélicoïdale de la chaîne, seules

deux paires de bases sont stables pour l'ADN : les paires Watson-Crick $A = U$ (qui forment deux liaisons hydrogène) et $G \equiv C$ (trois liaisons hydrogène). Pour les ARN, certains appariements non canoniques apparaissent dans certaines structures de ribozyme. L'ensemble des possibilités connues sont référencées dans [86]. Mais, on ne rencontre couramment que les appariements Watson-Crick et un troisième arrangement, l'appariement non canonique $G = U$.

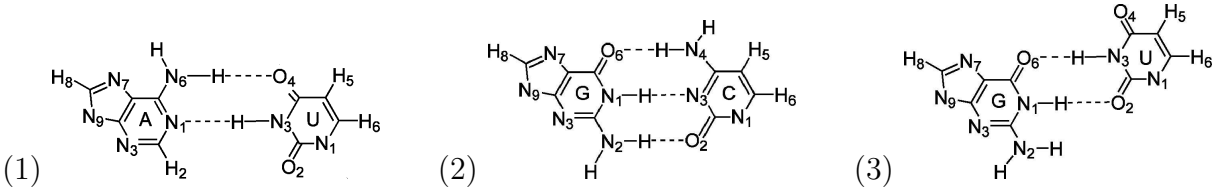


FIG. 1.7 – Les paires Watson-Crick (1)$A = U$, (2)$G \equiv C$ et l'appariement non canonique (3)$G = U$

La double hélice obtenue pour le double brin ADN se présente principalement sous la forme B. Le pas de l'hélice est de 3,4 nm et l'on dénombre 10 paires de bases par tour. Les cycles aromatiques des bases sont parallèles entre eux, et les deux brins complémentaires de l'hélice sont antiparallèles.

A la différence de l'ADN (génomique) double brin, l'ARN transcript se présente sous forme de simple brin. En abscence d'un second brin complémentaire, les molécules d'ARN se replient sur elles-mêmes. Le long de la séquence de la molécule d'ARN, il est possible de trouver une multitude de régions complémentaires et antiparallèles. Les hélices se forment en coudant la chaîne. Ces appariements locaux donnent lieux à de courtes hélices droites très compactes dénommées hélice A. La figure 1.8 montre la différence entre l'hélice A de l'ARN et l'hélice standard B de l'ADN.

FIG. 1.8 – Hélice B de l'ADN à droite et hélice A de l'ARN à gauche

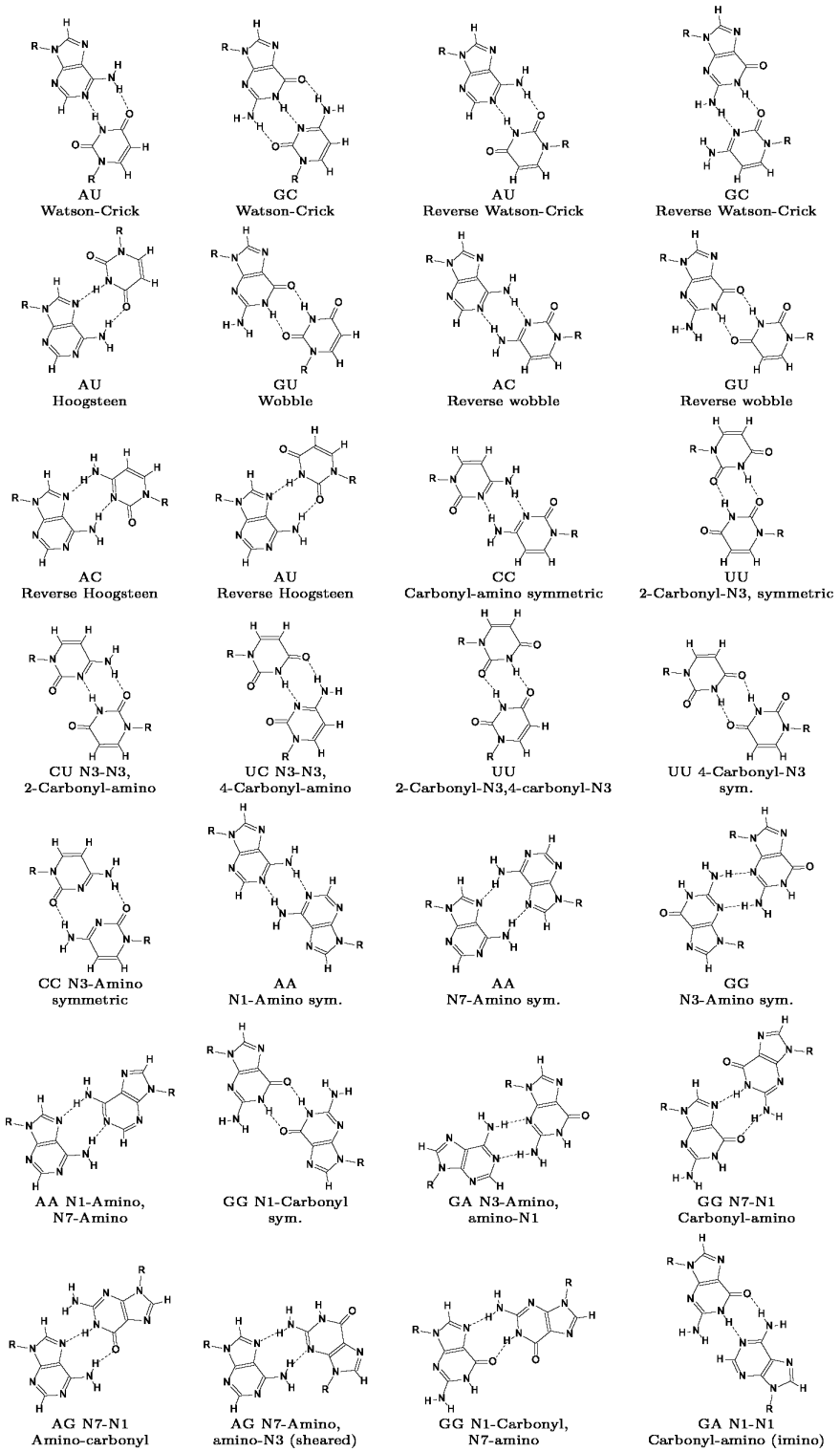


FIG. 1.9 – Vingt-huit manières différentes d'arranger spatialement deux nucléotides pour former des liaisons hydrogènes [70]

La structure obtenue est une altenance de simples brins souples et doubles brins extrêmements rigides donnant lieu à une structure enchevêtrée dite **structure secondaire**. On peut ainsi définir la **structure secondaire** de l'ARN par l'ensemble des couples de bases (i,j) appariées qui donnent lieu à la formation des hélices.

Il semblerait au vue des courbes de fusion, que la structuration tridimensionnelle fasse intervenir un ensemble d'interactions de plus faible énergie. Il a été montré que la **structure tertiaire** fait intervenir, entre autres, des interactions non canoniques à trois bases et des interactions médiées par les molécules d'eau et les ions. En particulier, les ions divalents, comme le Mg^{2+}, interagissent avec les sites de coordination de l'ARN. La présence ou l'absence de tels sels conditionne la fonction d'un ARN. Cette structuration est donc essentielle pour le bon fonctionnement biologique des ARN fonctionnels.

1.2 Motifs dans les ARN

La structure tridimensionnelle d'un ARN peut être vue comme une architecture modulaire composée de briques élémentaires dénommées motifs [80, 52]. Certains motifs appartiennent de droit à la structure secondaire, d'autres sont classés parmi les interactions tertiaires.

La structure secondaire est strictement définie par les relations mathématiques suivantes, proposées par Waterman [84]. En indiçant chaque nucléotide dans l'ordre de la séquence : une structure secondaire S est formellement définie par l'ensemble de toutes les paires de bases indicées (i, j) avec $i < j$ telles que :

1. une base i ne peut être appariée à plus d'une base j ;
2. pour deux couples de bases appariées $(i, j), (k, l)$, les positions relatives des bases suivent la loi suivante : si $i < k < j$ alors $i < l < j$.

La première loi implique que chaque nucléotide ne peut s'apparier qu'une et une seule fois. La seconde loi interdit les appariements s'entrecroisant. Une structure secondaire est donc uniquement formée d'appariements soit emboîtés, soit branchés.

1.2.1 Motifs élémentaires

En suivant les règles de Waterman, la structure secondaire est décomposable en une succession de motifs élémentaires emboîtés les uns à la suite des autres. On en distingue quatre :

1. La tige, qui correspond à une hélice A. Elle est elle-même composée d'une succession d'empilementappariement (*stacking*) de paires de bases consécutives ;
2. La boucle terminale, fermée par une paire de bases apparié ;
3. La boucle interne, symétrique ou non, fermée par deux paires de bases ;
4. Les boucles multiples qui relient plusieurs hélices entre elles.

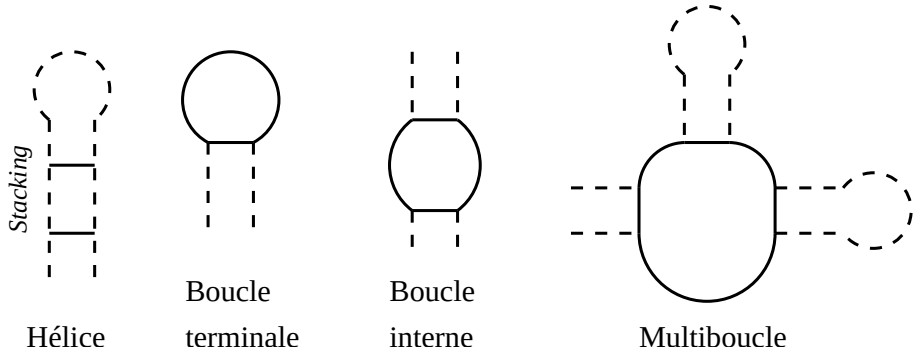


FIG. 1.10 – Motifs de la structure secondaire

Les lois de Waterman couplées à la description de la structure secondaire comme somme de motifs élémentaires ont permis de créer des algorithmes extrêmement efficaces pour résoudre rapidement le problème de prédiction de structures secondaires d'ARN [94, 48, 54, 83, 30, 47, 31]

1.2.2 Motif pseudonœud

Les pseudonœuds sont des motifs constitués d'au moins deux hélices. Ils résultent de l'interaction d'une boucle libre d'une structure secondaire avec un brin libre complémentaire situé en dehors de la boucle. L'interaction boucle-boucle (pseudonœud I) est la plus connue, tout comme le pseudonœud de Pleij (pseudonœud H) [64], et figure 1.11.

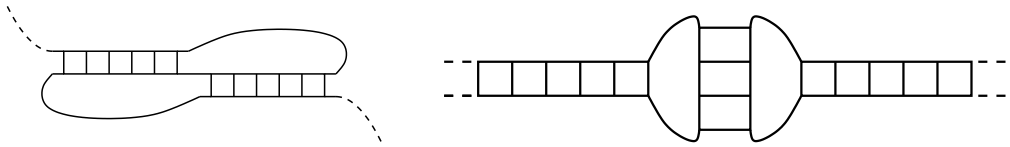


FIG. 1.11 – Deux exemples de pseudonœuds : le pseudonœud H et le pseudonœud I

La présence de pseudonœuds sur une structure d'ARN ne correspond pas a priori à un ensemble unique d'hélices ; il y a une part d'arbitraire dans le choix des hélices que l'on définit comme pseudonœuds. En se formant, les pseudonœuds confèrent une rigidité tridimensionnelle à l'ensemble de la structure d'un ARN. De plus pour des raisons stériques, la taille des hélices est finie, et les brins libres les reliant sont de dimensions minimales imposés. Ces contraintes de longueur découlent d'une analyse géométrique du motif particulier [64] : extrémités à relier et diposition des sillons profond et étroit des hélices du pseudonœud.

Dans l'état actuel de nos connaissances, les éléments énergétiques majeurs de la formation des motifs à pseudonœuds sont identiques à ceux qui sont utilisés dans le cas des motifs simples : un coût entropique dû à la conformation de la chaîne, et un gain d'énergie qui découle de l'appariement empilement des paires de bases dans chacune des hélices. Ainsi, il semble possible d'inclure ce type de motif dans un algorithme de prédiction de structure secondaire. Néanmoins, les approches classiques dites de programmation dynamique utilisées pour résoudre les structures secondaires sans pseudonœud, deviennent alors difficiles à mettre en œuvre, coûteuses en temps de calcul et en ressources système [68, 2].

Chapitre 2

Prédiction de structures secondaires d'ARN

A la différence de l'ADN (génomique) double brin, l'ARN transcript se présente essentiellement sous forme de simple brin. La séquence de l'ARN n'étant pas auto-complémentaire, la molécule cherche à optimiser les appariements de ses bases pour stabiliser la molécule. Ces appariements locaux conduisent au repliement de la molécule dans sa structure secondaire. Contrairement au repliement tertiaire qui fait intervenir l'environnement salin de la molécule, la structure secondaire est totalement définie par la combinatoire des appariements des bases de la séquence. De plus, expérimentalement, il a été montré que le repliement des molécules d'ARN suit un processus hiérarchique : la structure primaire (séquence) code pour la structure secondaire (appariement local des bases), elle-même, support du repliement tertiaire [13, 38, 80]. La structure secondaire de la molécule apporte la contribution principale à l'énergie de stabilisation [4] et aux contraintes stériques qui guident le repliement tridimensionnel. C'est pourquoi, chercher à prédire cette structure secondaire peut s'avérer à la fois techniquement possible et biologiquement intéressant. Néanmoins, la prédiction du repliement des molécules d'ARN reste un des challenges de la bioinformatique. En effet, même en se limitant aux interactions les plus simples, les appariements de paires de bases Watson-Crick et G=U, le traitement exact des motifs pseudonœud, n'est pas intégré à la majeure partie des codes existants.

La majeure partie des algorithmes de prédiction de structures secondaires se base sur les lois d'appariement de Waterman et exclus donc de facto la possibilité de former des pseudo-nœuds. Pour prédire la conformation la plus stable, ils associent un poids à chaque structure selon un critère donné : nombre maximum d'appariements [29, 54], minimum de l'énergie libre [94]. lorsque le nombre de séquences homologues est suffisant, les méthodes de covariations peremttent de déterminer les hélices principales de la strutures secondaire[35, 36]. La plupart de ces codes ne traitent pas des pseudonœuds ou les incluent dans un second temps d'analyse, comme une correction à la structure secondaire trouvée. Sur le même principe de recherche d'optimisation, Rivas et Eddy [68] ont écrit un code de repliement incluant certains motifs à pseudonœuds, de complexité algorithmique d'ordre $O(n^6)$ en temps et $O(n^4)$ en mémoire pour une séquence de n bases, en généralisant l'approche de M. Zuker. Tatsuya Akutsu présente une solution au problème à l'aide d'un algorithme d'ordre $O(n^4)$ en temps en généralisant l'approche d'appariement maximum [2]. Ces deux approches restent mathématiquement complexes et nous ne les traiterons pas ici. Une synthèse des différentes méthodes de prédiction *ab initio* et de leurs applications est donnée dans les références [78, 30].

2.1 Prédiction d'appariement : modèles de base

La manière la plus simple de prédire une structure secondaire consiste à lister tous les brins complémentaires d'une séquence donnée puis de former toutes les combinaisons possibles d'hélices compatibles et de calculer l'énergie totale de chaque structure. Un tel code exhaustif n'est pas applicable à de longues molécules, néanmoins, il a été utilisé par Pipas et McMahon [63] pour la prédiction de structures secondaires d'ARN de transfert.

L'algorithme de Martinez [46] attribue à chaque hélice, un poids statistique proportionnel à sa constante d'équilibre donnée par son poids de Boltzmann. La molécule est repliée par un algorithme Monte-Carlo. A chaque pas du programme, une hélice compatible avec la structure courante est choisie aléatoirement jusqu'à ce qu'il ne soit plus possible d'ajouter une quelconque hélice. Cette méthode ne tient pas compte des effets de déstabilisation des parties simples brins. Elle suppose aussi que les structures de plus basses énergies soient constituées statistiquement des hélices les plus stables compatibles.

2.2 Algorithmes de programmation dynamique

Ce type d'algorithme récursif permet de trouver de manière certaine une solution au problème de repliement de molécule d'ARN sans pseudonœuds. Il se déroule en deux étapes : le remplissage d'une matrice de score qui sauvegarde à chaque pas la meilleure évolution pour le point donné, et la reconstitution (aussi appelée *backtracking*) qui recherche le meilleur chemin le long de la matrice de score. Au meilleur chemin est associé la solution au problème posé. Dans le cas du repliement de structures secondaires de molécules d'ARN, ce chemin correspond à la meilleure structure obtenue pour le jeu de paramètres que l'on s'est fixé.

2.2.1 Principe et application pour l'algorithme d'appariement maximum

L'algorithme d'appariement maximum *"maximum matching"*, est l'application directe des lois de Waterman. Il identifie la structure secondaire la plus stable, à celle dont le nombre de bases appariées est la plus grande. Le premier algorithme qui résout le problème a été proposé par Nussinov et Jacobson [53]. Dans cette approche, on admet par hypothèse que les appariements sont décomposables en une somme d'empilement de paires de bases qui n'interagissent pas entre elles.

La matrice de score, $M(i,j)$, compte le nombre maximum des appariements qu'il est possible de former le long de la sous-séquence délimitée par les $i^{\text{ème}}$ et $j^{\text{ème}}$ bases. L'élément de matrice général s'écrit :

$$\begin{aligned} M(i,j) &= \max\left\{ M(i,j-1), \max_{i\leq l\leq j-1}\left\{ [M(i,l-1)+1+M(l+1,j-1)]\rho(a_l,a_j)\right\}\right\} \\ &\quad \text{avec, } a_l, a_j \in \{A,U,C,G\} \\ &\quad \text{et, } \rho(a_l,a_j) = \begin{cases} 1 \ : \text{ si } a_l \text{ et } a_j \text{ s'apparient,} \\ 0 \ : \text{ sinon} \end{cases} \end{aligned} \tag{2.1}$$

Les deux possibilités évoquées dans l'expression sont représentées sur la figure 2.2.1.

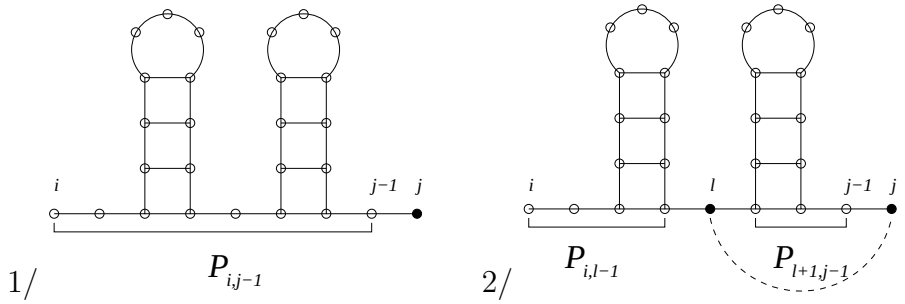


FIG. 2.1 – 1/ La nouvelle base (j) ne forme pas d'appariement, le nombre des bases appariées dans le segment $[i, j]$ est donc égal au nombre des bases appariées dans le segment $[i, j-1]$. 2/ La base l s'apparie avec la nouvelle base j. Le nombre de paires de bases est la somme de toutes les bases appariées dans les segments $[i, l-1]$ et $[l+1, j-1]$ plus un : la nouvelle paire de bases (l, j)

Remplissage de la matrice

Les éléments de matrice M(i,j) sont calculés de proche en proche en remplissant la matrice sous-diagonale par sous-diagonale en commençant par la première sous-diagonale et en remontant vers l'élément $M(1, n)$. Au final, le nombre maximum des appariements dans la structure est stocké dans l'élément de matrice $M(1, n)$, ou n est la longueur de la séquence.

Reconstruction ou *backtracking*

Pour reconstruire la meilleure structure, on remonte le long de la matrice à la recherche des meilleurs appariements locaux (méthode dite de *backtracking*). A partir du segment $[1, n]$, les sous-segments $[1, l-1]$ et $[l+1, n-1]$ qui génèrent la valeur $M(1, n)$ sont recherchés et l'appariement (l, n) est stocké dans la liste des paires de bases trouvées.

Si la base n n'est pas appariée, la séquence est réduite d'une unité et la recherche s'effectue sur le nouveau segment $[1, n-1]$.

Si la base j est appariée avec la base l, alors l'inégalité suivante est vérifiée :

$$N_{pb} + \Big(M[i, l-1] + 1 + M[l+1, j-1]\Big) + \sum_{p,q} M[p, q] \geq M[1, n] - \delta \qquad (2.2)$$

Avec N_{pb} le nombre des paires de bases déjà trouvées, $(M[i, l-1] + 1 + M[l+1, j-1])$ le nombre des paires de bases dans l'intervalle $[i, j]$, et $(\sum_{p,q} M[p, q])$ la somme des paires de

bases résultantes dans les intervalles $[p, q]$ non encore étudiés.

La procédure est itérée tant qu'il reste des sous-segments non nuls. En fin de procédure, tous les appariements correspondant à la meilleure structure sont stockés dans la liste des appariements. A partir de ces appariements, l'énergie de la structure obtenue est calculée en sommant la contribution individuelle de chaque stacking et de chaque brin libre de la configuration.

La complexité algorithmique croit comme $O(n^3)$ en temps et $O(n^2)$ en mémoire.

Exemple

Pour illustrer la méthode, nous avons effectué le remplissage et la reconstruction manuelle de la matrice de score pour la molécule théorique donnée figure 2.2. La figure 2.3 explique l'obtention de la valeur de la cellule [9,17].

Remplissage

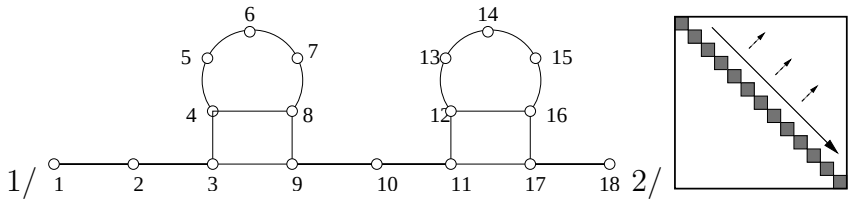


FIG. 2.2 – 1/molécule test. 2/Remplissage de la matrice : le long de chaque sous-diagonale en remontant de la seconde diagonale vers l'élément $M(1, n)$. Le remplissage de la matrice est ici relativement rapide : les seules bases appariées sont les couples [4,8], [3,9], [12,16], et [11,17]. L'expression 2.1 se réduit donc le plus souvent à $M[i, j] = M[i, j - 1]$, représenté par le caractère "—".

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	#	0	0	0	0	0	0	1	2	2	2	2	2	2	2	3	4	4
2		#	0	—	—	—	—	1	2	—	—	—	—	—	—	3	4	4
3			#	0	—	—	—	1	2	—	—	—	—	—	—	3	4	4
4				#	0	—	—	1	—	—	—	—	—	—	—	2	3	3
5					#	0	—	—	—	—	—	—	—	—	—	1	2	2
6						#	0	—	—	—	—	—	—	—	—	1	2	2
7							#	0	—	—	—	—	—	—	—	1	2	2
8								#	0	—	—	—	—	—	—	1	2	2
9									#	0	—	—	—	—	—	1	2	2
10										#	0	—	—	—	—	1	2	2
11											#	0	—	—	—	1	2	2
12												#	0	—	—	1	—	1
13													#	0	—	—	—	0
14														#	0	—	—	0
15															#	0	—	0
16																#	0	0
17																	#	0
18																		#

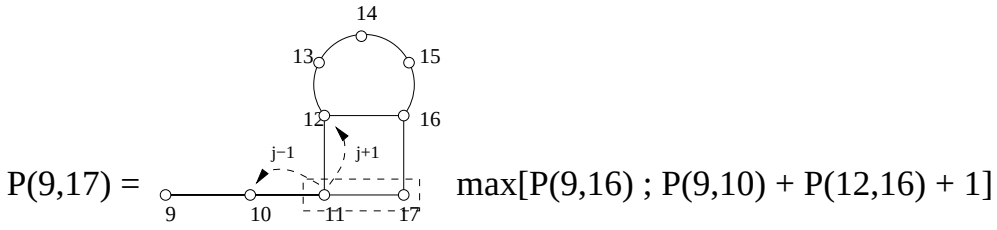


FIG. 2.3 – L'élément P(9,17), est calculé à partir des éléments P(9,10) et P(12,16), puisque le couple de bases [11,17] et apparié.

Remontée

La remontée commence de l'élément $P(1, 18)$ et traverse la matrice en direction de la seconde diagonale. A chaque appariement trouvé, se forme deux branches. La suite des éléments retenus

est :

$$P(1,18) \rightarrow P(1,17) \rightarrow \boxed{\begin{array}{c}\text{app.}\\ [11,17]\end{array}} \begin{cases} P(1,10) \rightarrow P(1,9) \rightarrow \boxed{\begin{array}{c}\text{app.}\\ [3,9]\end{array}} \begin{cases} P(1,2) \\ P(4,8) \rightarrow \boxed{\begin{array}{c}\text{app.}\\ [4,8]\end{array}} \end{cases} \rightarrow P(5,7) \rightarrow P(5,6) \\ P(12,16) \rightarrow \boxed{\begin{array}{c}\text{app.}\\ [12,16]\end{array}} \rightarrow P(13,15) \rightarrow P(13,14) \end{cases}$$

L'algorithme suit l'une des bifurcations jusqu'à arriver sur un élément de la seconde diagonale. Puis reprend à la dernière bifurcation et ainsi de suite. Lors de la remontée dans les branches $P(1,9)$ et $P(1,2)$ les éléments de l'équation 2.2 prennent les valeurs suivantes :

$$\begin{aligned} P(1,9) \quad &: \; N_{bp} = 1 \; , \; M[1,2] + 1 + M[4,8] = 0 + 1 + 1 \; , \; \sum_{p,q} M[p,q] = M[12,16] = 1 \\ P(1,2) \quad &: \quad N_{bp} = 2 \quad , \quad \sum_{p,q} M[p,q] = M[4,8] + M[12,16] = +1 \end{aligned}$$

2.2.2 Algorithme MFold

L'algorithme MFold présenté par Zuker [94, 95], utilise une estimation plus réaliste de la stabilité thermodynamique de la structure. L'énergie libre de la configuration est décomposée en la somme des contributions de chacun des motifs élémentaires, principalement : boucles libres, appariements-empilements des bases, et mésappariements terminaux des boucles internes et terminales. A ceci s'ajoute : une correction pour les boucles multiples qui tend à déstabiliser ce type de structure, et une correction pour les bases pendantes adjacentes à la première paire de bases d'une hélice dans les boucles multiples et les boucles terminales.

La comparaison se fait donc sur la minimisation obtenue par l'appariement ou non de la dernière base ajouté j avec n'importe quelle base libre l de la séquence. L'énergie libre de la meilleure structure secondaire formée à partir du segment de séquence $[1,j]$ est stockée dans l'élément de matrice F_j^5. F_j^5 minimise l'expression suivante :

$$F_j^5 \quad = \quad \min_l \left\{ F_{j-1}^5 \; , \; \left(F_{l-1}^5 + C[l,j] + d_{l-1}^5[l,j] + d_{j+1}^3[l,j] \right) \right\} \tag{2.3}$$

où, l'on reconnaît par analogie avec l'équation 2.1, l'égalité des termes F_j^5 et F_{j-1}^5, si les bases l et j ne sont pas appariées, et l'énergie correspondante à la sous-structure obtenue si les bases j et l, s'apparient. Le terme $C[l,j]$ contient toute l'information énergétique de la sous-structure définie sur le segment $[l,j]$ et les termes $d_{l-1}^5[l,j]$ et $d_{j+1}^3[l,j]$ correspondent aux correctifs d'énergie apportés par les deux bases $l-1$ et $j+1$ non appariées qui suivent l'hélice terminée par la paire de bases $[l,j]$. Le terme $C[l,j]$ contient tous les apports énergétiques de la sous-structure : boucles libres, boucles internes, empilements des bases, mésappariements terminaux et heuristiques associés aux boucles multiples. A chaque type de motif est associé une expression de son énergie libre. L'évaluation du terme F_j^5 fait donc intervenir un ensemble de calculs intermédiaires (non détaillés ici), stockés dans des matrices spécifiques.

Backtracking

A la fin de la procédure, l'élément F_n^5 contient l'énergie libre minimale associée à la séquence. Le processus de *backtracking* permet ensuite de remonter le long de la matrice pour identifier la structure obtenue.

De même que dans le cas de la recherche d'appariements maximums, le processus commence du segment le plus long $[1,n]$ pour aller vers les segments les plus courts. Pour tout segment $[i,j]$, la procédure consiste à identifier toutes les paires de bases immédiatement intérieures à i et j. Si la base j n'est pas appariée alors :

$$F_j^5 = F_{j-1}^5 \tag{2.4}$$

et la recherche se poursuit sur le segment $[1,j-1]$. Si elle est appariée, la recherche se poursuit le long de la sous-structure en identifiant la base k qui vérifie l'égalité suivante :

$$F_j^5 = F_{k-1}^5 + C[k,j] + d_{k-1}^5[k,j] + d_{j+1}^3[k,j] \tag{2.5}$$

La recherche se poursuit le long des matrices intermédiaires pour identifier les éventuelles multiboucles résultantes afin de retenir les différentes hélices raccordées. Enfin, le long de chaque hélice, les bases appariées et la taille des boucles terminales et internes sont identifiées selon la même méthode à l'aide des matrices appropriées.

Structures suboptimales

En 1989, Zuker introduit un algorithme récursif pour rechercher toutes les structures suboptimales [93]. Pour ce faire, on relie par l'esprit les deux extrémités de la séquence pour la recirculariser. Ainsi présenté, toute paire de bases (i, j) découpe la structure en deux sous-parties, une "interne" et une "externe" au segment $[i, j]$. Sur chaque sous-partie, il est possible de définir le meilleur repliement et évaluer son énergie libre. La somme des deux énergies définissant l'énergie libre de la structure complète.

Les deux matrices résultantes sont calculées récursivement. La stratégie consiste ensuite à déterminer les appariements qui mènent à une énergie totale proche de l'énergie minimum. Cette recherche peut être faite de manière exhaustive [89]. Le nombre de configurations obtenues grossit exponentiellement avec l'écart à la structure de plus basse énergie, et la plupart de ces structures sont quasi-identiques.

Pour ne représenter qu'un sous-ensemble de configurations visuellement différentes, MFold utilise un critère de distance topologique entre les structures. Pour toutes paires de bases (i, j) de la première structure et toute paire de bases (k, l) de la seconde, la distance minimale d qui sépare les appariements doit vérifier les inégalités suivantes :

$$|i - k| \leq d \text{ et } |j - l| \leq d \tag{2.6}$$

Le filtre sélectionne les structures d'*intérêt*, mais cette condition *arbitraire* ne donne aucune garantie quant à trouver les structures suboptimales les plus représentatives.

2.2.3 Algorithme thermodynamique de McCaskill

Dans les types d'analyses ci-dessus, l'énergie est associée à la prédiction des appariements des régions complémentaires d'ARN. Les énergies de stabilisation apportées par l'appariement des bases et la déstabilisation due aux boucles libres sont additionnées pour évaluer l'énergie libre totale de la molécule. Une autre manière de procéder consiste à prédire les régions appariées en se basant sur la thermodynamique [48]. La probabilité d'appariement d'une région d'énergie libre ΔG est associée à son poids de Boltzmann $[e^{-\Delta G/kT}]$. L'utilisation croisée des probabilités et de la programmation dynamique permet de prédire la structure la plus stable ainsi que d'identifier les appariements qui contribuent le plus à cette stabilisation.

La clé de la méthode repose sur le calcul de la fonction de partition Q du système. Elle est obtenue à partir de l'évaluation de la fonction partielle et auxiliaire $Q^b(i,j)$ qui résulte de la somme des probabilités de former la structure incluant la paire appariée de base (i,j). L'élément de matrice $Q(i,j)$ est obtenu en tenant compte de tous les éléments pour lesquels les bases i et j ne sont pas appariées. Les éléments de matrices $Q^b(i,j)$ et $Q(i,j)$ sont calculés récursivement. Au final, l'élément de matrice $Q(1,n)$ contient la fonction de partition totale du système.

L'énergie libre d'une structure $F(S)$ est donnée par la somme des énergies libres F_L des éléments qui la compose, et la fonction de partition est donnée par la somme des poids de Boltzmann de chaque structure soit : $Q = \sum_S e^{-[F(S)/kT]}$. A l'aide de cette notation, nous pouvons exprimer les éléments de matrices $Q^b(i,j)$ et $Q(i,j)$.

$$\begin{aligned} Q^b(i,j) &= \sum_L e^{-[F_L/kt]} \prod_{(h,l)\in L} Q^b(h,l) \\ Q(i,j) &= 1.0 + \sum_{\substack{h,l \\ i\leq h<l\leq j}} Q(i,h-1)Q^b(h,l) \end{aligned} \tag{2.7}$$

Avec $Q(i,i) = 1.0$ et $Q(i,i+1) = 1.0$. Les différents éléments de matrices sont obtenus récursivement par programmation dynamique.

Lorsque la fonction de partition du système est établie, il est immédiat de calculer la probabilité d'existence d'une structure particulière : $P(S) = \frac{1}{Q(1,n)} e^{-[F(S)/kt]}$. Mais, au-delà de cette application, cette méthode permet d'accéder simplement à toute la variété des structures et des appariements possibles en s'intéressant à la probabilité $P_{h,l}$ d'appariement d'une paire de bases particulière. Par définition cette probabilité est donnée par : $P_{h,l} = \sum_{S\ni(h,l)} P(S)$. Si (h,l) n'est pas une paire de bases particulière, c'est-à-dire, si (h,l) ne ferme pas une hélice, alors la probabilité s'exprime simplement sous la forme suivante :

$$P_{h,l} = \frac{Q(1,h-1)\, Q^b(h,l)\, Q(l+1,n)}{Q(1,n)} \tag{2.8}$$

Le terme général est la somme des probabilités des quatre possibilités d'appariements : (h,l) apparié dans une hélice, ou fermant une boucle interne, terminale, ou multiple. Le résultat est donné sous forme de tableau dont les cases sont plus ou moins grisées selon le logarithme de la probabilité correspondante (P_{ij}), pour les indices $i < j$. La partie basse du tableau est utilisée pour afficher la structure optimale.

Le calcul de la fonction de partition permet d'accéder à toutes les observables thermodynamiques, en particulier le calcul de la chaleur spécifique C_p, et d'éditer une courbe de fusion pour les structures secondaires.

2.3 Algorithme cinétique

2.3.1 Introduction

Au lieu de s'attaquer au problème en prenant en compte toute la combinatoire, nous pouvons amplement réduire la complexité en modélisant la dynamique locale de la chaîne. En effet, il a été démontré expérimentalement qu'une séquence d'ARN se replie à travers une succession de structures intermédiaires en quasi-équilibre [62, 57], suivant un processus stochastique. La dynamique du repliement des ARN peut donc être modélisée comme une succession de transitions élémentaires dont la cinétique est décrite par une loi d'Arrhenius pour le franchissement de la barrière des états de transition. A chaque pas, le programme calcule toutes les transitions vers les structures proches voisines et choisit aléatoirement un état vers lequel transiter. L'approche cinétique modifie le regard que l'on a sur le système. La notion de distance entre les états du système ne s'exprime plus comme une différence absolue en énergie mais comme un rapport de taux de transition (cf figure 2.4).

2.3.2 Principe de l'algorithme *KineFold*

Approche de la dynamique de repliement

Les transitions élémentaires de la dynamique correspondent à la formation ou à la dissociation d'une hélice dans la structure courante. Le taux de transition d'une structure à l'autre est donné par la loi d'Arrhenius $k = k^0 \cdot \exp(-\Delta G/kT)$. k^0 représente la dynamique locale de nucléation de l'hélice sur quelques bases, et kT, l'énergie thermique. Au travers de l'évaluation de la différence d'énergie entre les structures, nous rendons compte de la dynamique moléculaire sous-jacente. L'algorithme de Monte-Carlo cinétique qui en découle nous permet de suivre la cinétique de repliement et la relaxation de la molécule à temps long.

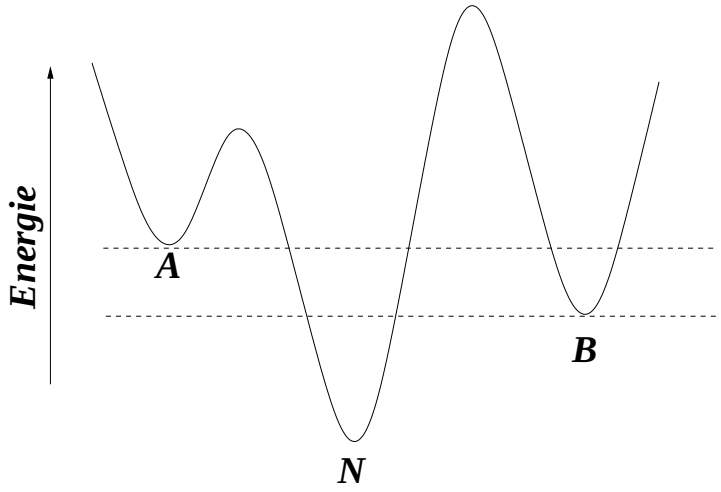


FIG. 2.4 – Thermodynamique *versus* réaction cinétique. Thermodymaniquement l'état B est plus *proche* de l'état natif N, alors que cinétiquement, l'état N est plus accessible en venant de l'état A.

L'évaluation de l'énergie libre se décompose en deux sous-parties : l'énergie de stacking, donnée par la somme des énergies de formation des hélices, et le coût entropique des configurations des sections de chaîne simple brin déterminé par le nombre de configurations accéssible pour la chaîne polymère maintenue à ces deux extrémités. En absence de pseudonœuds, la somme des contributions énergétiques de chaque brin libres et de chaque hélice correspond à l'énergie libre totale de la configuration. Mais contrairement au cas des structures secondaires classiques, le coût entropique des pseudonœuds ne peut pas être simplement évalué à l'aide de la somme des entropies des brins libres [94]. Pour quantifier le coût entropique des pseudonœuds, nous modélisons les couplages d'orientation des hélices impliquées en identifiant la structure à un réseau de bâtonnets rigides reliés entre eux par des ressorts gaussien.

Une structure avec pseudonœud se décompose selon trois motifs indépendants : le brin libre, l'hélice, le pseudonœud. Nous utilisons l'approche classique qui consiste à sommer les contributions individuelles pour obtenir l'énergie libre totale. Les différents motifs sont obtenus par un découpage de la structure en sous-domaines quasi-indépendants appelés *"nets"* [34]. Les "nets" ne contiennent qu'un seul type de motif : brin libre, hélice ou pseudonœud et sont reliés entre eux par des sections simple brin ou double brins qui contraignent la structure globale.

Nous pouvons ainsi obtenir l'énergie libre de chaque éléments indépendant de la structure. Lors du regroupement de tous les éléments indépendants quelques hélices peuvent être ajustées pour minimiser l'énergie totale de la structure [34].

Par cette approche de la dynamique de repliement, les pseudonœuds sont naturellement intégrés à la structure, contrairement aux codes classiques décrits ci-dessus.

Algorithme cinétique de base

Avant chaque transition, nous recherchons *toutes* les évolutions possibles de la structure courante a, en lui ajoutant ou en lui retirant une hélice. Pour chaque structure l obtenue, nous calculons le taux de transition k_{la} associé au passage de la structure courante a vers la configuration l.

$$k_{la} = k^0 \cdot exp(-\Delta G_{l \leftarrow a}/kT)$$

La différence d'énergie libre $\Delta G_{l \leftarrow a}$ correspond au coût entropique pour mettre les deux brins complémentaires en vis-à-vis pour former une hélice, et à la perte d'énergie libre d'appariement si l'on dissocie une hélice.

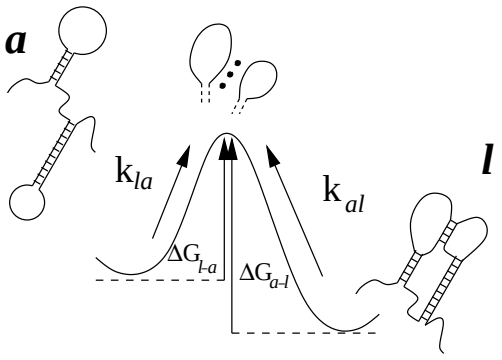


FIG. 2.5 – Représentation schématique de la transition de a vers l et inversement. La hauteur de la barrière $\Delta G_{l \leftarrow a}$ correspond au coût entropique dû pour mettre les deux brins complémentaires en vis-à-vis.

La probabilité p_{la} d'effectuer cette transition depuis la structure a est donnée par le ratio entre k_{la} et la somme de tous les taux de transition vers chacune des structures accessibles

depuis a :

$$p_{la} = \frac{k_{la}}{\sum_{<l>} k_{la}}$$

La transition à effectuer est choisie stochastiquement et le nouvel état choisi devient l'état courant. Le temps de vie physique τ_a de la structure courante est évalué comme l'inverse de la somme de tous les taux de transition pour nucléer ou dissocier une nouvelle hélice. La probabilité p_{la} et le temps de vie de l'état a , s'écrivent alors sous la forme :

$$\frac{1}{\tau_a} = \sum_{<l>} k_{la} \qquad \Rightarrow \qquad p_{la} = k_{la} \cdot \tau_a$$

A la transition, l'insertion de la nouvelle hélice implique généralement d'ajuster les extrémités des hélices avec lesquelles elle entre en compétition. Par souci d'efficacité numérique, ceci est effectué suivant un principe de minimisation locale des énergies libres d'appariement en zippant ou dézippant les hélices en compétition autour d'un noyau de nucléation de 2 ou 3 bases. Toutefois, après chaque transition, les positions des extrémités de toutes les hélices en compétition sont optimisées globalement pour minimiser l'énergie libre totale de la molécule.

Chapitre 3

Accélération de la cinétique [90]

3.1 Introduction

Le problème dans l'algorithme cinétique de base est que dans la majeure partie de la cinétique, le système reste dans les mêmes sous-espaces de configuration, et la probabilité d'explorer de nouveaux états décroît rapidement au fur et à mesure que le nombre de structures déjà connues augmentent. Les techniques d'accélération numérique courantes sont basées sur des méthodes thermodynamiques qui scrutent le paysage énergétique [5, 9]. Or, dans le cas du repliement des molécules d'ARN, nous avons pu constater que le nombre typique de structures visitées est très inférieur au nombre de structures possibles. C'est pourquoi les accélérations de type thermodynamique peuvent s'avérer plus coûteuses en temps de calcul que la simulation cinétique qu'elles sont censées améliorer. Pour accélérer la cinétique, nous proposons une méthode qui tire précisément profit du caractère piégé de la dynamique locale du système. A la base, cet algorithme choisit stochastiquement une structure encore non visitée, reliée à la structure courante par une suite de structures connues gardées en référence (typiquement ≤ 100). La cinétique intermédiaire est alors moyennée *exactement* sur toutes les successions possibles de transitions entre ces configurations connues gardées en référence. Le point essentiel de cette méthode est qu'à chaque pas nous visitons de manière sûre un nouvel état du système,

différent de ceux qui sont gardés en référence. Le facteur d'accélération, que nous définissons comme le rapport entre le nombre total moyen de transitions intégrées et le nombre de nouvelles structures effectivement calculées, varie de 10 à 10^5 selon le degré de piégeage des ARN étudiés.

Dans le processus Markovien de sauts aléatoires de la structure a vers la structure b, les temps de vie de l'état, les probabilités d'échappement et le taux de transition direct de l'état a vers l'état b sont reliés (voir algorithme cinétique de base plus bas). Le principe de la méthode est basé sur le calcul des probabilités d'évolution de l'état a vers l'état b lorsqu'on tient compte explicitement de l'ensemble des chemins discrets $\mathcal{C}_m^{\mathcal{A}}$, s'appuyant sur l'ensemble $\mathcal{A}$ des configurations déjà visitées gardées en référence. La nouvelle matrice des probabilités de transitions $P^{\mathcal{A}}$ somme les poids statistiques de *tous* les chemins $\mathcal{C}_{ba}^{\mathcal{A}}$ entre deux états a et b de $\mathcal{A}$, où le poids statistique d'un chemin particulier est donné par le produit des probabilités de transition directe entre les états intermédiaires consécutifs le long de ce chemin. Dans la nouvelle expression des probabilités de transitions, la transition singulière $(i \to j)$ peut être vue comme un processus stochastique localement renormalisé, passant de l'échelle locale à la taille du cluster.

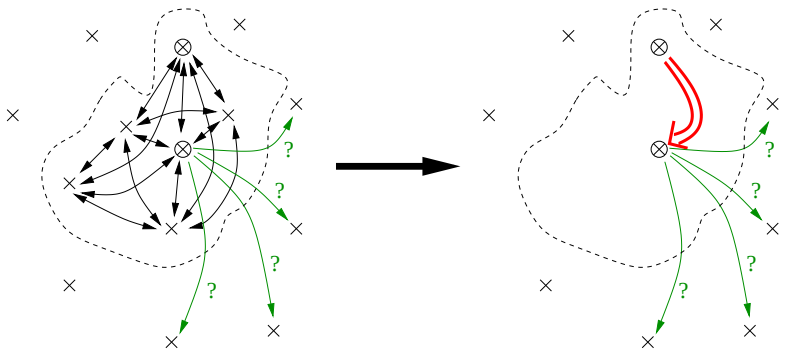


FIG. 3.1 – Schématiquement, l'idée de l'algorithme revient à remplacer toutes les transitions possibles (en noire) qui mènent de l'état initial vers l'état de sortie par une unique transition moyenne et directe (double flèche rouge). A partir de cet état de sortie, le système choisira stochastiquement parmi les états accesibles proches voisins n'appartenant pas au cluster, un nouvel état à visiter.

3.2 Principe du dépiégeage cinétique exact

Application à un système avec deux états de référence.

Nous allons dans un premier temps nous intéresser aux trajectoires dynamiques d'un système ne contenant que deux états de référence a et b distincts. Les relations établies dans ce cas simple nous permettront d'obtenir par analogie les relations matricielles dans le cas général.

Probabilité d'échappement

Dans la suite, nous noterons la probabilité de transiter directement de l'état a vers l'état b par p_{ba}, de même, nous noterons la probabilité de transition de l'état b vers l'état a par p_{ab}. La notation matricielle de droite à gauche pour l'ordre des indices sera utilisée tout au long du chapitre.

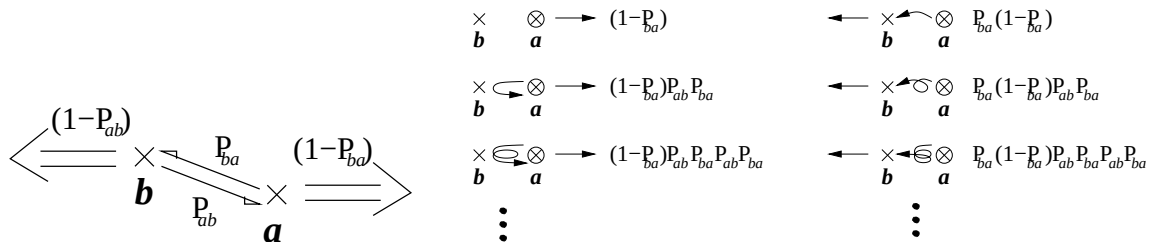


FIG. 3.2 – Schéma de principe des probabilités d'échappement des deux états. Expression des probabilités d'échapper directement, après un aller-retour, deux allers-retour, etc ...

Par définition la probabilité de s'échapper directement de l'état a vers n'importe quel état différent de l'état b est $p_a^e = (1 - p_{ba})$. De même, la probabilité de faire un aller-retour entre l'état a et l'état b avant de s'échapper par l'état a est égal à la probabilité de transiter de a vers b puis de b vers a et enfin de a vers n'importe quel état différent de b soit :

$$\begin{aligned} & (1 - p_{ba})(p_{ab}p_{ba}) \\ p_{aa}^e \quad = \quad & p_a^e(p_{ab}p_{ba}) \end{aligned} \tag{3.1}$$

Cette expression contient deux types de termes, le terme p_a^e qui traduit la probabilité de s'échapper directement de l'état a vers un état proche voisin différent de b d'une part et d'autre

part, le terme $(p_{ab}p_{ba})$ qui traduit la probabilité d'effectuer un aller-retour. La généralisation à n allers-retours est immédiate. Finalement, la probabilité moyenne de s'échapper de l'état a en étant initialement en a est donnée par la relation suivante :

$$\begin{aligned} P^e_{aa} &= (1-p_{ba}) + (1-p_{ba})(p_{ab}p_{ba}) + \cdots + (1-p_{ba})(p_{ab}p_{ba})^n + \ldots \\ &= (1-p_{ba}) \sum_{m=0}^{\infty} (p_{ab}p_{ba})^m \end{aligned} \tag{3.2}$$

La série géométrique est de raison inférieure à un, ce qui permet de réécrire l'expression (3.2) sous la forme finale suivante :

$$P^e_{aa} = \frac{(1-p_{ba})}{1-(p_{ab}p_{ba})} \tag{3.3}$$

La probabilité moyenne de s'échapper de l'état b en étant initialement en a est simplement donnée par la probabilité de transiter de l'état a vers l'état b multiplié par la probabilité moyenne de sortir de l'état b en étant initialement en b soit :

$$\begin{aligned} P^e_{ba} &= p_{ba}P^e_{aa} \\ &= \frac{p_{ba}(1-p_{ab})}{1-(p_{ba}p_{ab})} \end{aligned} \tag{3.4}$$

Finalement, la somme des expressions de P^e_{aa} et de P^e_{ba} qui traduit la probabilité de s'échapper de l'état a par n'importe quel chemin est bien normalisée comme il se doit :

$$P^e_{aa} + P^e_{ba} = 1 \tag{3.5}$$

Probabilité de présence

A partir des expressions de la probabilité P^e_{yx} d'échappement par l'état y en étant initialement en x nous pouvons déduire les probabilités de présence P_{yx} pondérées sur tous les chemins qui mènent de l'état y à l'état x en empruntant uniquement les états connus $\{x, y\}$ du système. Cette probabilité est simplement reliée à probabilité d'échappement P^e_{yx} par la relation suivante :

$$\begin{aligned} P^e_{yx} &= \sum_{l \neq \{x,y\}} p_{ly}P_{yx} = (1-p_{xy})P_{yx} \\ P^e_{yx} &= p^e_y P_{yx} \end{aligned} \tag{3.6}$$

où x et y représentent un des états de référence $\{a, b\}$, et $p_y^e = (1-p_{xy})$ représente la probabilité de transition directe de l'état y vers n'importe quel état l différent des états de références a et b.

Par analogie avec les équations (3.3) et (3.4), les expressions des probabilités P_{aa} et P_{ba} pondérées sur tous les chemins qui mènent de l'état a vers lui-même, et de l'état a vers b, en empruntant uniquement les états connus $\{a, b\}$ s'écrivent :

$$\begin{aligned} P_{aa} = \sum_{m=0}^{\infty} (p_{ab}p_{ba})^m &= \frac{1}{1-(p_{ab}p_{ba})} \\ P_{ba} = \sum_{m=0}^{\infty} p_{ba}(p_{ab}p_{ba})^m &= \frac{p_{ba}}{1-(p_{ab}p_{ba})} \end{aligned} \tag{3.7}$$

Les résultats se réexpriment sous la forme matricielle suivante, avec P^e la matrice des probabilités d'échappement et P la matrice des probabilités de présence :

$$P^e = \begin{pmatrix} \frac{(1-p_{ba})}{1-(p_{ab}p_{ba})} & \frac{p_{ab}(1-p_{ba})}{1-(p_{ab}p_{ba})} \\ \frac{p_{ba}(1-p_{ab})}{1-(p_{ba}p_{ab})} & \frac{(1-p_{ab})}{1-(p_{ba}p_{ab})} \end{pmatrix} \qquad P = \begin{pmatrix} \frac{1}{1-(p_{ab}p_{ba})} & \frac{p_{ab}}{1-(p_{ab}p_{ba})} \\ \frac{p_{ba}}{1-(p_{ba}p_{ab})} & \frac{1}{1-(p_{ba}p_{ab})} \end{pmatrix} \tag{3.8}$$

Temps d'échappement

Sur le même principe nous pouvons écrire le temps moyen mis par le système pour s'échapper d'un des deux états. Notons τ_a le temps de vie de l'état a et τ_b le temps de vie de l'état b. Ces temps de vie représentent le temps moyen mis par le système pour transiter de l'état courant vers n'importe quel autre état accessible. Le temps moyen mis pour s'échapper des

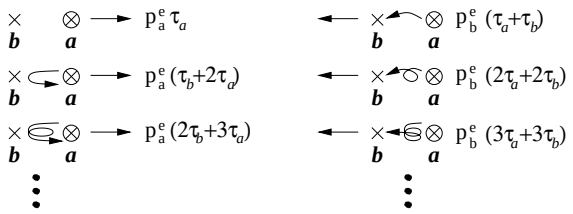


FIG. 3.3 – Schéma de principe des temps d'échappement d'un des deux états et expression des temps moyens pondérés correspondant après un aller-retour et deux allers-retour.

états a, b en étant initialement parti de l'état a est donc donné par la somme des temps de vie

de chacun des états visités pondérée du poids du chemin correspondant qui mène de l'état a vers un état extérieur. Pour un aller-retour, le temps moyen pondéré est égal à la somme des temps de vie de chacun des états visités, multiplié par la probabilité d'effectuer cet aller-retour, soit : $(\tau_a + \tau_b + \tau_a)(p_{ab}p_{ba})p_a^e$. De même le temps pondéré de deux aller-retours est donné par : $(\tau_a+\tau_b+\tau_a+\tau_b+\tau_a)(p_{ab}p_{ba})^2p_a^e$. Le temps d'échappement pondéré sur tous les chemins possibles s'écrit donc sous la forme suivante :

$$\begin{aligned} \bar{\tau}_a &= p_a^e\tau_a + p_b^e p_{ba}(\tau_b + \tau_a) + p_a^e p_{ab}p_{ba}(\tau_a + \tau_b + \tau_a) + \cdots \\ \Rightarrow \bar{\tau}_a &= p_a^e\tilde{P}_{aa}\{t\} + p_b^e\tilde{P}_{ba}\{t\} \end{aligned} \tag{3.9}$$

Les matrices $\tilde{P}_{aa}\{t\}$ et $\tilde{P}_{ba}\{t\}$ représentent les temps moyens pondérés des chemins qui mènent de a vers l'un des deux états respectivement a, b :

$$\begin{aligned} \tilde{P}_{aa}\{t\} &= \sum_{m=0}^{\infty}(p_{ab}p_{ba})^m\big[\tau_a + m(\tau_a + \tau_b)\big] \\ \tilde{P}_{ba}\{t\} &= \sum_{m=0}^{\infty}p_{ab}(p_{ba}p_{ba})^m\big[(m+1)(\tau_a + \tau_b)\big] \end{aligned} \tag{3.10}$$

Généralisation

Probabilité de présence et d'échappement

Sur un ensemble A de n états de référence, les expressions des probabilités de présence et d'échappement sont reliées aux poids statistiques $W^{C_m^A}$ de chaque chemin C_m^A de support A et de longueur m : $W^{C_m^A} = \prod^{C_m^A} p_{lk}$ où les indices k et l courent sur les états consécutifs le long du chemin C_m^A qui débute par l'état i et finit dans l'état j. A partir de l'expression de chaque chemin pondéré nous pouvons écrire par analogie avec les équations (3.7), l'expression de chaque élément de matrice P_{ji}^A, qui représente la probabilité pondérée d'être dans l'état j en étant parti de l'état i :

$$P_{ji}^A = \sum_{m:j\leftarrow i}^{C^A} W^{C_m^A} = \sum_{m:j\leftarrow i}^{C^A}\left(\prod_{j\leftarrow i}^{C_m^A} p_{lk}\right) \tag{3.11}$$

La sommation porte sur l'ensemble des chemins C^A de toutes les longueurs m qui débutent de l'état i et finissent sur l'état j.

A partir de cette définition nous pouvons écrire la probabilité de sortir du cluster A des états de référence par l'état j : $p_j^{eA} = 1 - \sum_{<k>}^{A} p_{kj}$, où la sommation porte sur tous les états $< k >$ plus proches voisins de l'état j qui appartiennent au cluster d'état A. Nous retrouvons les expressions de la probabilité de s'échapper du cluster des états de référence (3.6) ainsi que la normalisation (3.5) :

$$P_{ji}^{eA} = p_j^{eA} P_{ji}^{A} \tag{3.12}$$

$$1 = \sum_{j}^{A} p_j^{eA} P_{ji}^{A} \quad \forall i \in A \tag{3.13}$$

Ainsi l'algorithme de dépiégeage cinétique exact procède en deux étapes à partir d'un état i du cluster d'état connu A. Dans un premier temps, nous choisissons un état j du cluster à partir duquel nous allons nous échapper. Ce choix est effectué aléatoirement avec la probabilité $p_j^{eA} P_{ji}^{A}$ de suivre le chemin moyen correspondant. La seconde étape consiste à choisir parmi tous les états accessibles l n'appartenant pas au cluster, l'état vers lequel le système va transiter. Ce choix ne doit plus se faire stochastiquement selon la loi p_{kj} mais selon la loi conditionnelle p_{kj}/p_j^{eA} pour tenir compte du fait que nous avons déjà intégré dans un premier temps la probabilité de s'échapper de l'état j (p_j^{eA}) lors du choix aléatoire de l'état de sortie.

Moyenne des observables du système

A partir des expressions des probabilités moyennées sur l'ensemble des chemins possibles, nous pouvons aisément définir toutes les moyennes de toutes les observables d'intérêt du système. Nous allons centrer la discussion sur l'observable moyenne du temps mis par le système pour s'échapper du cluster. Introduisons la matrice $\mathcal{T}[P^A]_{ji}\{t\} = \tilde{P}^A\{t\}$, transformée temporelle du temps écoulé dans le cluster de la matrice initiale P^A.

Comme dans le cas à deux états, le temps d'échappement du cluster le long **d'un** chemin C_m^A est simplement donné par la somme des temps de vie de chacun des états visités consécutivement : $\sum^{C_m^A} \tau_h$. Dans l'approche de moyenne exacte de la cinétique d'évolution du système le long des états de référence, nous nous intéressons à **tous** les chemins qui mènent de l'état i vers l'état j en visitant uniquement des états appartenant au cluster A. Le poids temporel de chaque chemin est donné par le temps cumulé pondéré par la probabilité d'effectuer

le chemin particulier : $\left(\sum^{C_m^A} \tau_h\right) \prod^{C_m^A} p_{lk}$. Le temps pondéré du chemin moyen de déplacement dans le cluster entre les états i et j est donc donné par la somme sur tous les chemins C^A des temps cumulés pondérés du poids de chacun des chemins :

$$\tilde{P}_{ji}^A\{t\} = \sum_{m:j\leftarrow i}^{C^A} \left[\left(\sum_{j\leftarrow i}^{C_m^A} \tau_h \right) \prod_{j\leftarrow i}^{C_m^A} p_{lk} \right] \tag{3.14}$$

En sommant sur tous les états j appartenant au cluster A, nous sommons sur tous les chemins qui partent de l'état i et sortent du cluster. Cette expression est équivalente au temps de moyen $\bar{t}_i^A$ d'échappement du cluster en étant initialement en i :

$$\bar{t}_i^A = \sum_j^A p_j^{eA} \tilde{P}_{ji}^A\{t\} \tag{3.15}$$

Le temps moyen particulier, $\bar{t}_{ji}^A$, correspondant au temps d'échappement du cluster par l'état particulier j quant à lui est donné par :

$$\bar{t}_{ji}^A = \frac{p_j^{eA} \tilde{P}_{ji}^A\{t\}}{p_j^{eA} P_{ji}^A} = \tilde{P}_{ji}^A\{t\}/P_{ji}^A \tag{3.16}$$

Sur le même principe nous pouvons introduire pour toutes les observables x_i la matrice de poids $\tilde{P}^A\{x\}$ appropriée, qui rend compte de la valeur moyenne de l'observable x_i pondérée du poids de chaque chemin emprunté par le système sur le cluster A des états de référence. En particulier, la longueur des chemins empruntés est donnée par la matrice $\tilde{P}^A\{\ell\}$. En comptant chaque transition directe est de longueur unitaire ($\ell = 1$), la longueur moyenne des chemins le long des états de référence du cluster A entre les états i et j est donnée par l'élément de matrice suivant :

$$\tilde{P}_{ji}^A\{\ell\} = \sum_{m:j\leftarrow i}^{C^A} \left[\left(\sum^{C_m^A} 1 \right) \prod^{C_m^A} p_{lk} \right] \tag{3.17}$$

Et l'expression de la longueur moyenne $\bar{\ell}_{ji}^A$ du chemin menant de l'état i à l'état j est donnée par :

$$\bar{\ell}_{ji}^A = \tilde{P}_{ji}^A\{\ell\}/P_{ji}^A \tag{3.18}$$

Physiquement, cette longueur moyenne $\bar{\ell}^A_{ji}$ représente le nombre moyen de transitions élémentaires qu'un algorithme classique aurait effectué pour passer de l'état initial i à l'état final j appartenant au cluster A. La mesure de la longueur moyenne $\bar{\ell}^A_{ji}$ est indépendante du système considéré, de plus, le calcul étant exact, l'algorithme peut être utilisé continûment.

De même, nous pouvons sur le même schéma définir toutes les matrices associées aux moyennes temporelles d'observables y, $\tilde{P}^A\{yt\}$. En particulier la moyenne temporelle de l'énergie des états visités le long de tous les chemins pondérés C^A qui mènent de l'état i à l'état j et sa matrice associée s'écrivent :

$$\bar{E}^A_{ji} = \tilde{P}^A_{ji}\{Et\}/\tilde{P}^A_{ji}(t) \tag{3.19}$$

$$\tilde{P}^A_{ji}\{Et\} = \sum_{m:j\leftarrow i}^{C^A} [(\sum^{C^A_m} E_h t_h) \prod^{C^A_m} p_{lk}] \tag{3.20}$$

3.3 Fusion de deux clusters disjoints.

Pour l'instant nous avons obtenu les expressions des probabilités et des différents types d'observables moyennées le long d'un cluster d'état connu A. Nous allons dans cette partie, écrire les expressions des matrices P^C et $\tilde{P}^C$ résultantes de la fusion de deux clusters d'état A et B disjoints. Par hypothèse, pour chaque cluster A et B, les matrices P^A, P^B des probabilités de présence, et toutes les matrices $\tilde{P}^A$, $\tilde{P}^B$ des observables sont connues. Seule la caractérisation de la transition directe d'un cluster à l'autre reste à définir. La fusion des deux ensembles A et B est similaire à la fusion de deux états i et j, à ceci près que les relations des probabilités de transition ne sont plus scalaires, mais matricielles.

Aller simple d'un cluster à l'autre

Probabilité moyenne de présence

Introduisons les matrices de transfert T^{BA} et T^{AB}, où chaque élément de matrice $T^{BA}_{ji} = p_{ji}$ est la probabilité de passer d'un état i du cluster A vers un état j du cluster B en une transition directe, et réciproquement pour T^{AB}_{ij}. Si les états i de A et j de B ne sont pas proches voisins, les éléments de matrice T^{BA}_{ji}, T^{AB}_{ij} sont nuls.

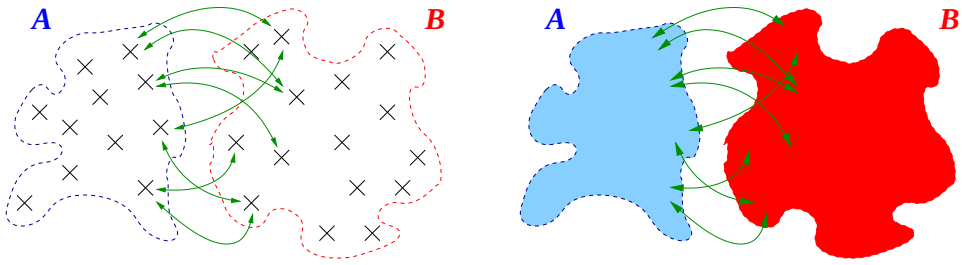


FIG. 3.4 – Chaque cluster peut être imaginé comme un super-état. Les relations de transition directe entre les super-états A et B sont nécessairement matricielles, puisqu'il y a plusieurs manières de transiter de l'ensemble A vers l'ensemble B.

A l'aide de cette nouvelle matrice nous pouvons exprimer toutes les probabilités de transition unique de l'état i du cluster A vers l'état j du cluster B, sous la forme : $\mathcal{P}_{ij}^{BA} = (P^B T^{BA} P^A)_{ji}$. La probabilité $\mathcal{P}_{ji}^{BA}$ représente l'analogue de la probabilité p_{ab} pour les super-états A et B. La probabilité de sortir par l'état j du cluster B en partant de l'état i du cluster A est donnée par analogie avec le cas à deux états par :

$$\begin{aligned} p_{ba}^e = (1-p_{ab})p_{ba} \quad &\rightleftharpoons \quad \mathcal{P}_{ji}^{eBA} = \left(1 - \sum_{\{i\in A\}} \mathcal{P}_{ij}^{AB}\right)\mathcal{P}_{ji}^{BA} \\ &\Rightarrow \quad \mathcal{P}_{ji}^{eBA} = \left(1 - \sum_{\{i\in A\}} (P^A T^{AB} P^B)_{ji}\right)(P^B T^{BA} P^A)_{ji} \end{aligned} \tag{3.21}$$

L'analogie s'arrête là, puisque l'expression $(P^B T^{BA} P^A)_{ji}$ est immédiatement l'expression du chemin moyen emprunté par le système pour passer de l'état i du cluster A vers l'état j du cluster B en transitant de A vers B une et une seule fois.

Moyenne des observables

Intéressons-nous à l'expression d'un seul de ces chemins particuliers C_l^{BA} qui mène de l'état i appartenant au cluster A vers l'état j appartenant au cluster B. Ce chemin est composé des deux sous-chemins C_m^B et C_n^A qui ont pour support respectivement les clusters B et A et de la transition directe c_{ba}^{BA} de l'état a de sortie du cluster A vers l'état b d'entrée du cluster B.

L'expression de ce chemin et la probabilité d'effectuer un tel chemin sont données par :

$$C_l^{BA} = C_m^B + c_{ba}^{BA} + C_n^A \quad (l = m + 1 + n) \tag{3.22}$$

$$p_{ji}^{BA} = (\prod_{j\leftarrow b}^{C_m^B} p_{lk}) p_{ba} (\prod_{a\leftarrow i}^{C_n^A} p_{l'k'}) \tag{3.23}$$

En sommant cette dernière expression sur tous les sous-chemins C_m^B et C_n^A, nous définissons le poids statistique du chemin pondéré qui mène de l'état i du cluster A vers l'état j de B en transitant une et une seule fois de A vers B au travers de la transition directe p_{ba} de l'état $(a \in A)$ vers l'état $(b \in B)$:

$$[\sum_{m:j\leftarrow b}^{C^B} (\prod_{j\leftarrow b}^{C_m^B} p_{lk})] p_{ba} [\sum_{n:a\leftarrow i}^{C^A} (\prod_{a\leftarrow i}^{C_n^A} p_{l'k'})] = (\prod_{j\leftarrow b}^{B} p_{lk}) p_{ba} (\prod_{a\leftarrow i}^{A} p_{l'k'}) \tag{3.24}$$

De même, nous pouvons définir la contribution au temps moyen d'échappement de l'union de A et B de ce chemin particulier :

$$\left(\sum_{j\leftarrow b}^{B} t_h + \sum_{a\leftarrow i}^{A} t_{h'}\right) \prod_{j\leftarrow b}^{B} p_{lk} \cdot p_{ba} \cdot \prod_{a\leftarrow i}^{A} p_{l'k'} = \tag{3.25}$$

$$\left(\sum_{j\leftarrow b}^{B} t_h \prod_{j\leftarrow b}^{B} p_{lk}\right) p_{ba} \prod_{a\leftarrow i}^{A} p_{l'k'} + \prod_{j\leftarrow b}^{B} p_{lk}\ p_{ba} \left(\sum_{a\leftarrow i}^{A} t_{h'} \prod_{a\leftarrow i}^{A} p_{l'k'}\right)$$

Ce qui se résume en notation matricielle par l'expression suivante :

$$\mathcal{T}[P^B T^{BA} P^A]\{t\} = \mathcal{T}[P^B]\{t\}\ T^{BA} P^A + P^B T^{BA} \mathcal{T}[P^A]\{t\}. \tag{3.26}$$

De manière identique nous pouvons exprimer les matrices $\mathcal{T}[P^B T^{BA} P^A]\{x\}$ et $\mathcal{T}[P^B T^{BA} P^A]\{yt\}$:

$$\begin{aligned} \mathcal{T}[P^B T^{BA} P^A]\{x\} &= \mathcal{T}[P^B]\{x\}\ T^{BA} P^A + P^B T^{BA} \mathcal{T}[P^A]\{x\} \\ \mathcal{T}[P^B T^{BA} P^A]\{yt\} &= \mathcal{T}[P^B]\{yt\}\ T^{BA} P^A + P^B T^{BA} \mathcal{T}[P^A]\{yt\}. \end{aligned} \tag{3.27}$$

Ces expressions définissent l'opération $\mathcal{T}$ comme équivalente à une opération de différentiation classique applicable à n'importe quelle combinaison de matrice de probabilité pondérée. Selon cette définition la différentiation des matrices de transitions directes T^{BA} et T^{AB} est nulle : $\mathcal{T}[T^{BA}]_{ij} = 0 = \mathcal{T}[T^{AB}]$.

Généralisation

A l'aide des expressions dérivées ci-dessus et en s'appuyant sur la discussion faite pour le cas à deux états simples nous pouvons expliciter les expressions des matrices P^C et $\tilde{P}^C$ pour l'union C des deux ensembles disjoints A et B. Comme nous l'avons précédemment signalé, nous pouvons considérer les ensembles A et B comme des super-états.

Définissons les deux matrices $P^{Ab} = P^A T^{AB}$ et $P^{Ba} = P^B T^{BA}$ de probabilité de pouvoir s'échapper de n'importe quel état de l'ensemble A (respectivement B) en commençant par transiter directement d'un état b (respectivement a) de l'ensemble B (A) vers l'ensemble A (B). Nous pouvons exprimer les matrices P^{Ab} et P^{Ba} selon le schéma 3.5.

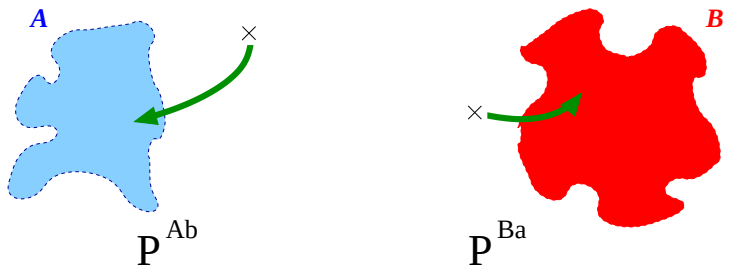


FIG. 3.5 – Expression schématique des matrices P^{Ab} et P^{Ba}

La matrice P^C est obtenue comme étant la sommation infinie sur toutes les possibilités d'allers-retours entre les ensembles d'états A et B. Par analogie avec les expressions (3.7), l'expression de la probabilité d'être dans un état du cluster A ou B en étant parti d'un état du cluster A ou B pondéré sur tous les allers-retours possibles est donnée par les relations suivantes, avec I la matrice identité :

$$\begin{aligned} Q^{AA} &= \left[I + P^{Ab}P^{Ba} + (P^{Ab}P^{Ba})^2 + \cdots\right] P^A = L^A P^A \\ Q^{BA} &= P^{Ba} L^A P^A \\ Q^{BB} &= \left[I + P^{Ba}P^{Ab} + (P^{Ba}P^{Ab})^2 + \cdots\right] P^B = L^B P^B \\ Q^{AB} &= P^{Ab} L^B P^B \\ \text{avec} \quad L^A &= \left[I - P^{Ab}P^{Ba}\right]^{-1} \quad \text{et} \quad L^B = \left[I - P^{Ba}P^{Ab}\right]^{-1} \end{aligned} \tag{3.28}$$

L'expression de la matrice P^C est simplement donnée par :

$$P^C = \begin{pmatrix} Q^{AA} & Q^{AB} \\ Q^{BA} & Q^{BB} \end{pmatrix} \tag{3.29}$$

De même pour chaque observable, nous pouvons définir les matrices $\tilde{P}^{Ab} = \tilde{P}^A T^{AB}$ et $\tilde{P}^{Ba} = \tilde{P}^B T^{BA}$. Les matrices $\tilde{P}^C$ qui représentent les valeurs moyennes du cheminement du système pour chacune des observables, sont obtenues par simple "différentiation" de la matrice P^C :

$$\begin{aligned} \mathcal{T}[P^C] = \quad \tilde{P}^C &= \begin{pmatrix} \tilde{Q}^{AA} & \tilde{Q}^{AB} \\ \tilde{Q}^{BA} & \tilde{Q}^{BB} \end{pmatrix}, \text{ avec} \\ \tilde{Q}^{AA} &= \tilde{L}^A P^A + L^A \tilde{P}^A \\ \tilde{Q}^{BA} &= \tilde{P}^{Ba} L^A P^A + P^{Ba} \tilde{L}^A P^A + P^{Ba} L^A \tilde{P}^A \\ \tilde{Q}^{BB} &= \tilde{L}^B P^B + L^B \tilde{P}^B \\ \tilde{Q}^{AB} &= \tilde{P}^{Ab} L^B P^B + P^{Ab} \tilde{L}^B P^B + P^{Ab} L^B \tilde{P}^B \\ \tilde{L}^A &= L^A \big(\tilde{P}^{Ab} P^{Ba} + P^{Ab} \tilde{P}^{Ba} \big) L^A \\ \text{et} \quad \tilde{L}^B &= L^B \big(\tilde{P}^{Ba} P^{Ab} + P^{Ba} \tilde{P}^{Ab} \big) L^B \end{aligned} \tag{3.30}$$

Les équations 3.30 et 3.29 sont valables quelle que soit la taille n et m des ensembles A et B. En particulier les matrices P^C et $\tilde{P}^C$ peuvent être calculées récursivement à partir des N états de référence du système et des $2N$ matrices scalaires (1×1) $P^i = [1]$ et $\tilde{P}^i\{x\} = [x_i]$ où l'indice i court sur l'ensemble $\{1, N\}$ et x_i représente la valeur de l'observable d'intérêt pour l'état i. En regroupant les états deux par deux puis 4 par 4, etc. ... à l'aide des équations (3.30) et (3.29), nous obtenons les matrices P^C et $\tilde{P}^C$ en $\mathcal{O}(N^3)$ opérations constituées de multiplications et d'inversions de matrice. Il est toutefois possible d'améliorer ensuite la mise à jour des matrices P^C et $\tilde{P}^C$ avec un algorithme en $\mathcal{O}(n^2)$

3.4 Algorithme en $\mathcal{O}(n^2)$

Le processus d'échappement se fait algorithmiquement en deux étapes : une première étape qui concerne la renormalisation du chemin sur le cluster, et une seconde qui traduit la transition

directe d'un état connu vers un nouvel état. La première étape met à profit le développement des expressions ci-dessus qui simule **toutes** les évolutions possibles du système le long des états de référence A, en **une seule** opération, et renvoie le comportement moyen du système le long du cluster des états A. La seconde étape, est effectuée par la partie classique de l'algorithme de calcul de transition directe entre deux états proches voisins. Au final, le nouvel état est intégré au cluster des états de référence.

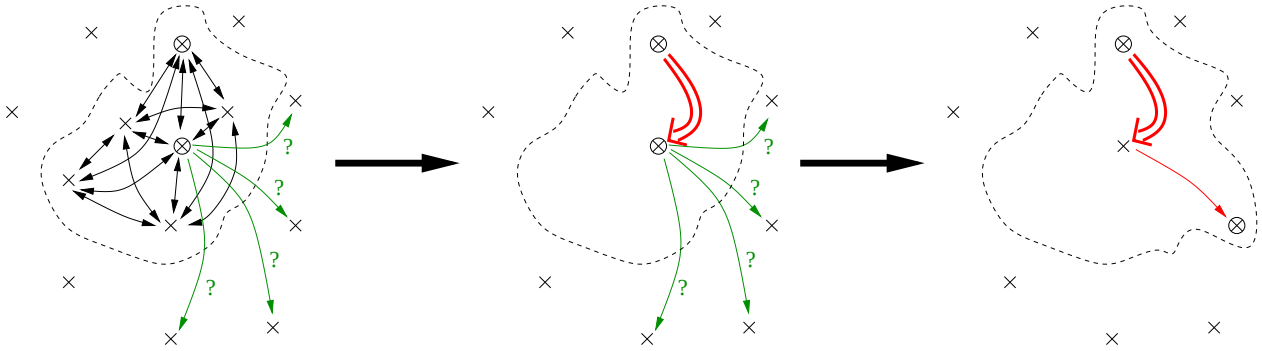


FIG. 3.6 – La première partie de l'algorithme choisit un état de sortie et mesure la probabilité moyenne de l'atteindre en empruntant tous les chemins possibles. Ce processus peut être vu comme une transition directe renormalisée sur le cluster (double flèche rouge). Le système choisit ensuite un état de sortie parmi les possibles (flèches vertes). L'état choisi (simple flèche rouge) est ensuite intégré au cluster

Pour des raisons matérielles et d'efficacité de l'algorithme, il n'est pas possible de stocker une grande quantité d'états. Les matrices P^C et $\tilde{P}^C$ ne peuvent donc pas dépasser une certaine limite dépendante de la capacité de mémoire et de calcul de l'ordinateur. Lorsque cette limite est atteinte, nous devons procéder à la dissociation d'un état du cluster avant de pouvoir intégrer le nouvel état visité. L'ensemble de ces opérations peut être fait avec un algorithme en $\mathcal{O}(n^2)$, comme nous allons le démontrer dans la section suivante. Nous réexprimerons aussi, les probabilités d'échappement $p_{lj}P^A_{ji}$ ainsi que toutes les observables du système en tenant compte du fait que le choix de l'état j de sortie du cluster est effectué antérieurement à la mise à jour des matrices.

Mise à jour des matrices P^C et $\tilde{P}^C$.

Relation de *Shermann-Morrison*.

Dans la partie traitant de la fusion de deux clusters distincts, nous avons suivi un procédé de calcul des matrices P^C et $\tilde{P}^C$, par calcul récursif sur les N états de référence à chaque modification de l'ensemble C des états gardés en mémoire. Or, dans le cas particulier où seul un état visité est à ajouter ou retirer de l'ensemble C, les sous-clusters A et B sont de taille $(n = N - 1)$ et $(m = 1)$. Les matrices P^B et L^B se réduisent à des scalaires (matrices 1×1) pour l'état unique B. De même les matrices T^{AB}, P^{Ab} et $\tilde{P}^{Ab}$ et les matrices T^{BA}, P^{Ba} et $\tilde{P}^{Ba}$ deviennent des vecteurs respectivement de taille $(n \times 1)$ et $(1 \times n)$. Nous pouvons donc utiliser la relation de *Shermann-Morrison* [79] pour mettre à jour les matrices P^C et $\tilde{P}^C$. Appliquée à la matrice L^A, la relation de *Shermann-Morrison* s'exprime sous la forme suivante :

$$\begin{aligned} L^A &= \left[I - P^{Ab} \otimes P^{Ba}\right]^{-1} = \left[I - u \otimes v\right]^{-1} \\ &= I + u \otimes v + u \otimes v \cdot u \otimes v + u \otimes v \cdot u \otimes v \cdot u \otimes v + \cdots \\ &= I + u \otimes v\left[1 + v \cdot u + (v \cdot u)^2 + (v \cdot u)^3 + \cdots\right] \\ L^A &= I + \frac{P^{Ab} \otimes P^{Ba}}{(1 - P^{Ab} \cdot P^{Ba})} \end{aligned} \tag{3.31}$$

Le calcul de l'inversion de la matrice $I - P^{Ab} \otimes P^{Ba}$ nécessite donc :

- $n - 1$ additions et n multiplications pour calculer le produit scalaire $P^{Ab} \cdot P^{Ba}$, et une opération d'addition supplémentaire pour le calcul de $\alpha^{-1} = (1 - P^{Ab} \cdot P^{Ba})$
- n opérations de multiplication pour le calcul de $\alpha^{-1} \cdot P^{Ab}$
- n^2 opérations de multiplication pour le calcul de la matrice $\mathcal{P} = \alpha^{-1} \cdot P^{Ab} \otimes P^{Ba}$
- n opérations d'addition pour le calcul de l'expression finale $L^A = I + \alpha \cdot \mathcal{P}$

Soit un total de $(n^2 + 2n)$ opérations de multiplications et $2n$ opérations d'additions. Cette relation permet donc de calculer L^A à l'aide d'un algorithme en $\mathcal{O}(n^2)$.

De manière identique pour toute matrice M de taille $n \times n$, les matrices $L^A M$ s'expriment sous la forme :

$$L^A M = M + \frac{P^{Ab} \otimes (P^{Ba} M)}{(1 - P^{Ab} \cdot P^{Ba})} \tag{3.32}$$

Et de même, le calcul de $L^A\,M$ est obtenu avec un algorithme en $\mathcal{O}(n^2)$ opérations, en déroulant les opérations selon le schéma suivant : $P^{Ba}M$ puis $P^{Ab} \otimes (P^{Ba}M)$.

Ajout d'un état.

Ajouter un état au cluster revient à fusionner les deux ensembles A et B. Les expressions des éléments de matrice P^C et $\tilde{P}^C$ résultantes de la fusion des deux ensembles ont été obtenues précédemment (équations 3.29 et 3.30). L'union des deux ensembles A et B peut être schématisée comme suit :

$$P^A \cup P^B = \left(\begin{array}{c|c} P^A & P^{Ab} \\ (n \times n) & (n \times 1) \\ \hline P^{Ba}\,(1 \times n) & P^B\,(1 \times 1) \end{array}\right) \implies P^C = \left(\begin{array}{c|c} L^A P^A & P^{Ab} L^B P^B \\ \hline P^{Ba} L^A P^A & L^B P^B \end{array}\right) \tag{3.33}$$

Chaque élément de matrice peut être calculé à l'aide de la relation de *Shermann-Morrison*. Les éléments des matrices $\tilde{Q}^{BA}$ et $\tilde{Q}^{AB}$ sont calculés à l'aide de l'expression (3.31) en remplaçant M par les matrices $(n \times n)$ $\tilde{P}^{Ab} \otimes P^{Ba}$ et $P^{Ab} \otimes \tilde{P}^{Ba}$.

Retrait d'un état.

Retirer un état du cluster, revient à exprimer les matrices originales P^C et $\tilde{P}^C$ comme résultantes de l'association de l'état B au cluster A. Après avoir choisi l'état B à retirer du cluster C, il suffit d'identifier les colonnes Q^{AB}, Q^{BA} et leur intersection Q^{BB} pour obtenir toutes les informations sur l'état B. Schématiquement cela revient à effectuer les opérations suivantes :

$$P^C = \left(\begin{array}{c|c} & \\ & \\ \hline & \end{array}\right) \implies P^A \cup P^B = \left(\begin{array}{c|c} P^A & P^{Ab} \\ \hline P^{Ba} & P^B \end{array}\right)$$

Nous cherchons à exprimer les matrices initiales P^A et $\tilde{P}^A$, définies par les équations 3.28, en fonction des vecteurs Q^{AB}, Q^{BA}, T^{AB} et T^{BA} et des matrices Q^{AA}, Q^{BB}, $\tilde{P}^B$, et $\tilde{Q}^{AA}$, connues.

Le vecteur P^{Ab} est obtenu en inversant les équations (3.28) : $P^{Ab} = Q^{AB}/Q^{BB}$, et par définition, le vecteur P^{Ba} se réduit à $P^{Ba} = T^{BA}$, puisque la probabilité $P^B = [1]$ d'être en B est certaine. A partir des matrices P^{Ba}, P^{Ab}, nous pouvons définir la matrice L^A :

$$[L^A]^{-1} = I - P^{Ab} \otimes P^{Ba} = I - \frac{Q^{AB} \otimes T^{BA}}{Q^{BB}} \tag{3.34}$$

Les matrices P^A et $\tilde{P}^A$ s'expriment simplement sous la forme suivante :

$$\begin{aligned} Q^{AA} &= L^A P^A \quad \Longleftrightarrow \quad P^A = [L^A]^{-1} Q^{AA} \\ \tilde{Q}^{AA} &= L^A \Big[\tilde{P}^A \big(I + T^{AB} \otimes Q^{BA} \big) + \frac{\tilde{P}^B}{Q^{BB}} Q^{AB} \otimes Q^{BA} \Big] \\ \Leftrightarrow \quad \tilde{P}^A &= \Big([L^A]^{-1} \tilde{Q}^{AA} - \frac{\tilde{P}^B}{Q^{BB}} Q^{AB} \otimes Q^{BA} \Big) \big(I + T^{AB} \otimes Q^{BA} \big)^{-1} \end{aligned} \tag{3.35}$$

En appliquant la relation de *Shermann-Morrison* pour inverser la matrice $I + T^{AB} \otimes Q^{BA}$:

$$\big(I + T^{AB} \otimes Q^{BA} \big)^{-1} = \Big(I - \frac{T^{AB} \otimes Q^{BA}}{1 - T^{AB} \cdot Q^{BA}} \Big)$$

nous obtenons finalement :

$$\begin{aligned} P^A &= \Big(I - \frac{Q^{AB} \otimes T^{BA}}{Q^{BB}} \Big) Q^{AA} \\ \tilde{P}^A &= \Big[\Big(I - \frac{Q^{AB} \otimes T^{BA}}{Q^{BB}} \Big) \tilde{Q}^{AA} - \frac{\tilde{P}^B}{Q^{BB}} Q^{AB} \otimes Q^{BA} \Big] \Big(I - \frac{T^{AB} \otimes Q^{BA}}{1 - T^{AB} \cdot Q^{BA}} \Big) \end{aligned} \tag{3.36}$$

Toutes les opérations peuvent être effectuées en $\mathcal{O}(n^2)$ opérations, nous pouvons donc retirer l'état B du cluster C en $\mathcal{O}(n^2)$ opérations.

Conclusion.

Le cluster des états peut donc être continuellement mis à jour en utilisant alternativement les équations (3.28, 3.30) et (3.36) à l'aide d'un schéma en $\mathcal{O}(n^2)$ opérations.

Stabilité de la méthode.

Les relations (3.28, 3.30) et (3.36) sont continûment utilisées pour mettre à jour les informations sur le cluster. Sur un grand nombre d'opérations, nous pouvons estimer que les erreurs numériques et la faible instabilité de l'algorithme fait doucement dériver les résultats. Pour réduire cette dérive due au mauvais conditionnement des matrices, nous divisons chaque matrice par l'élément le plus grand (en lieu et place de la valeur propre la plus grande). Cette légère modification nous permet d'obtenir des dérives numériques très modérées pour un nombre de mises à jour n inférieur à 300. Au-delà, pour éviter que le système ne dérive trop au cours du temps, il suffit de recalculer de manière exacte les matrices P^A et $\tilde{P}^A$ à partir des n états isolés en $\mathcal{O}(n^3)$ opérations, toutes les $n^{\text{ième}}$ opérations de mise à jour. Le schéma global incluant le recalcul exact reste d'ordre $\mathcal{O}(n^2)$.
Ci-dessous, le graphique de l'évolution de l'erreur $\epsilon = \sum_i^A (\sum_j^A p_j^{eA} P_{ji}^A - 1)^2$ au cours du temps, pour une molécule fortement piégée :

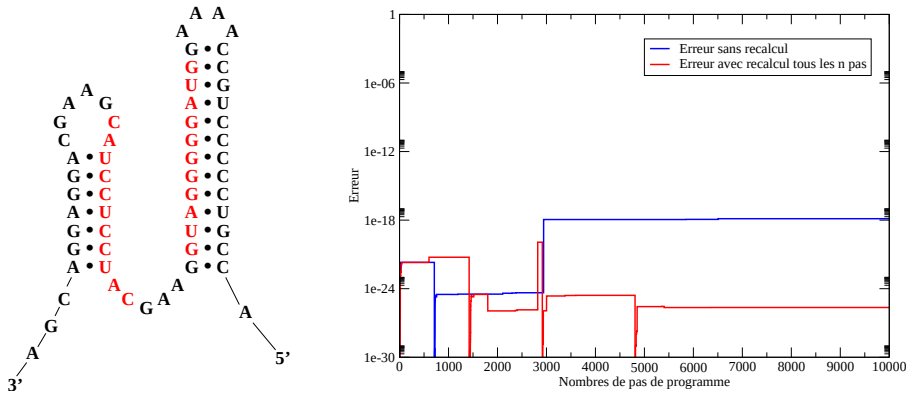


FIG. 3.7 – Variation de l'erreur en fonction du nombre de pas du programme, au cours d'une simulation de dépiégeage de la molécule de droite. L'erreur ϵ est quasi nulle tout au long de la simulation, pour une mise à jour tous les 300 pas.

Choix de l'état à retirer.

Lorsque le cluster atteint sa taille limite, il faut commencer par retirer un état du cluster avant de pouvoir ajouter le nouvel état visité. En principe, le choix de cet état peut se faire

de façon arbitraire. Néanmoins, en retirant l'un des états statistiquement les plus visités, nous avons remarqué que l'efficacité de la méthode diminue fortement. En effet, retirer un tel état, revient à réduire en partie le nombre effectif de chemins et par voie de conséquence, cela réduit l'efficacité nette de la méthode.

Nous avons choisi de retirer l'état $j^\star$ qui présente la fréquence de sortie la plus faible en étant initialement en i. C'est à dire l'état $j^\star$ qui vérifie l'équation suivante :

$$\frac{1}{t^A_{j^\star i}} = \min_j{}^A \left(\frac{1}{t^A_{ji}} \right) \tag{3.37}$$

Il est possible qu'un autre choix soit plus judicieux en fonction du système dynamique étudié.

3.5 Résultat de l'algorithme d'accélération

L'accélération de la cinétique se mesure par le rapport de la longueur effective moyenne du chemin suivi ($\bar{l} = \sum l_{ij}$) sur le nombre de pas de programme informatique. Ce rapport nous donne l'accélération théorique accessible, indépendante du matériel utilisé. L'accélération expérimentale constatée est donnée par le rapport du temps CPU dépensé à faire la simulation avec et sans la méthode d'accélération. Ce second rapport tient compte des limitations physiques du matériel informatique utilisé. En particulier des tailles des différents registres de mémoires CPU et de mémoire vive, ainsi que de tous les temps de latence dus aux flux d'informations échangées entre les registres.

Nous avons testé l'efficacité du cluster à différentes tailles limite et pour différentes molécules. La figure (3.8) montre les résultats obtenus pour l'accélération de la simulation de la cinétique de molécules fortement piégés, aléatoires et biologiques, représentés sur la figure (3.9). Pour chaque type de molécule, la courbe en tiré plein représente l'accélération attendue, et la courbe en trait plein, l'accélération obtenue. Les simulations sont effectuées sur une machine standard architecturée autour d'un processeur AMD MP Athlon cadencé à 1 Ghz avec 128 ko de cache de second niveau et 528 Mo de mémoire vive. Les facteurs d'accélération obtenus varient d'un facteur 5 à 10^5 selon la longueur et le type de molécule. Les deux courbes d'accélération théoriques et expérimentales sont très proches pour les petites tailles de cluster ($n \leq 50$). Au-

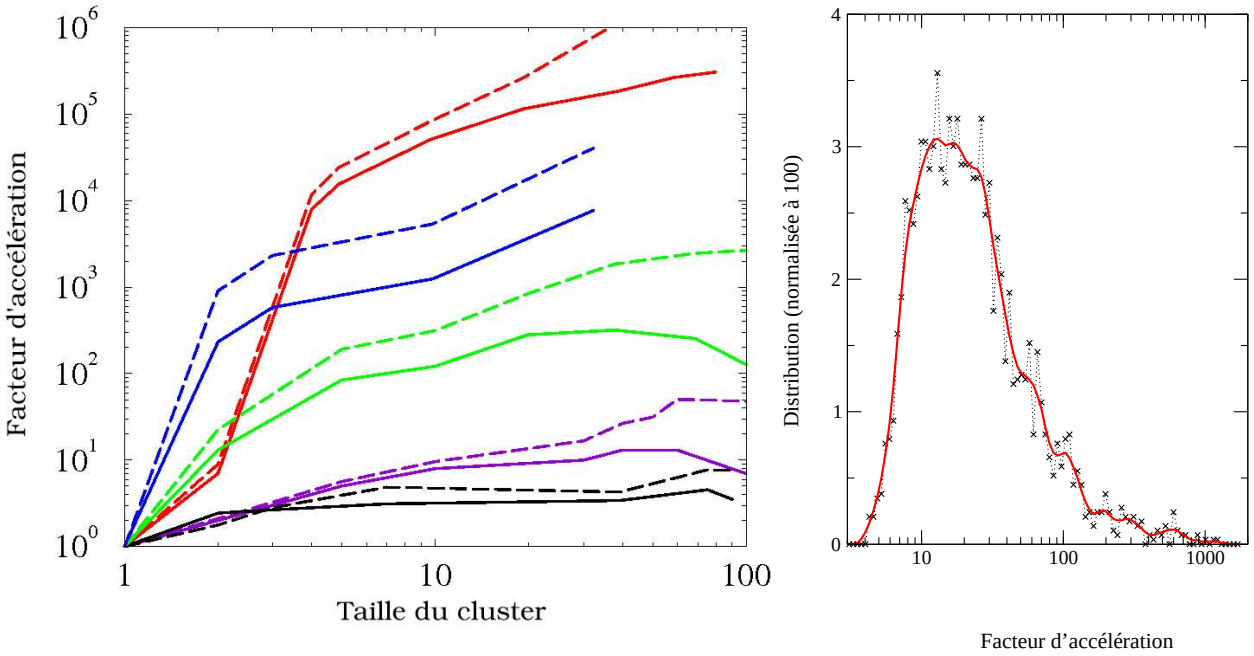


FIG. 3.8 – **A gauche** : accélération théorique et expérimentale obtenues par la présente méthode d'accélération de la dynamique stochastique. **i/** En rouge : molécule bistable de 67 nucléotides de long (espace des structures engendré par une combinatoire de 37 hélices possibles). **ii/** En bleu : molécule bistable de séquence réverse de la molécule (38 hélices possibles). **iii/** En vert : ribozyme du virus de l'hépatite delta (84 hélices possibles).**iv/** En violet : accélération moyenne pour un ensemble de séquences aléatoires de 100 nucléotides de long et composées à 50% de base G et C. **v/** En noir : ribozyme d'intron de Groupe I (894 hélices possibles).**A gauche** : distribution du facteur d'accélération obtenu pour l'ensemble de séquences aléatoires de 100 bases et 50% de G/C pour un nombre d'états de références fixé à 40 unités.

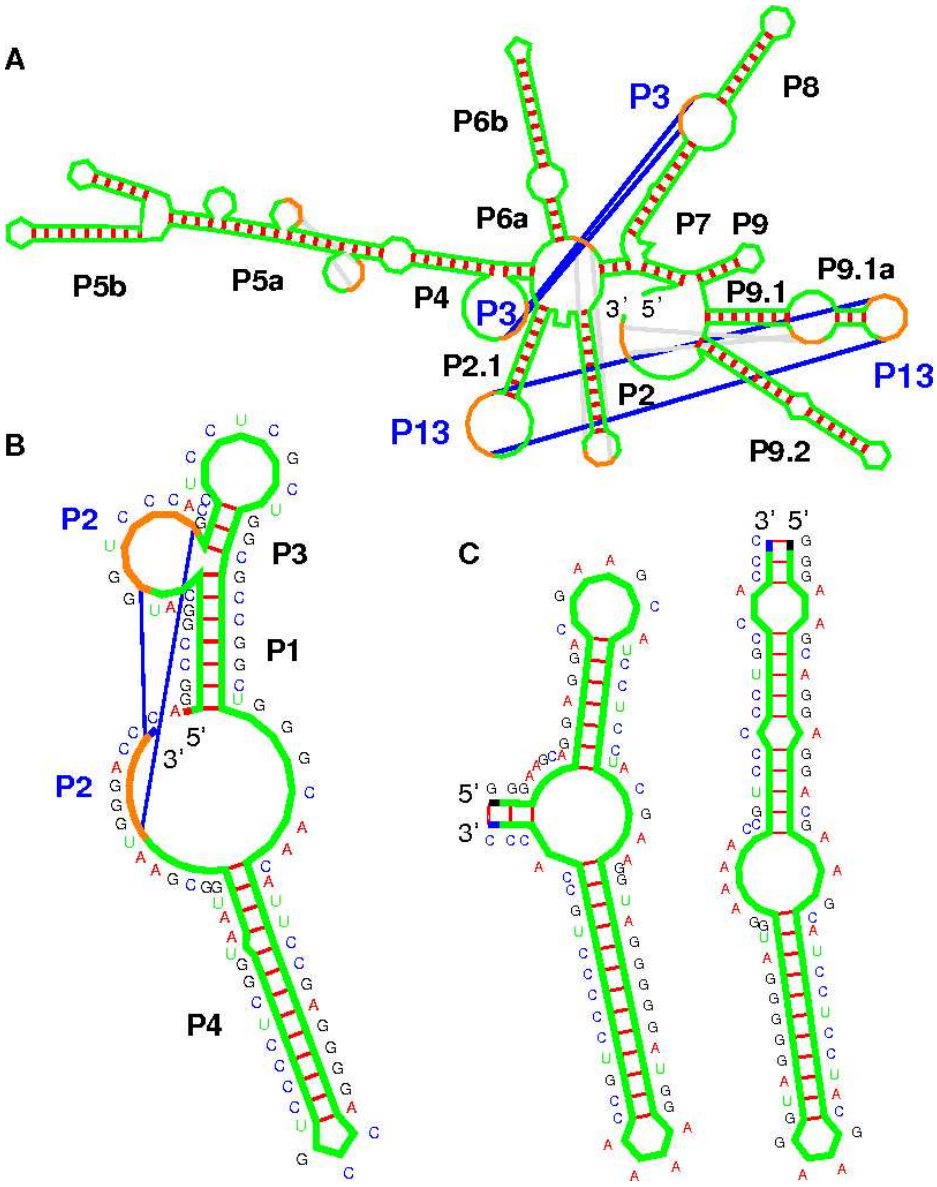


FIG. 3.9 – Prédiction des structures ARN avec l'algorithme d'accélération. Les structures sont dessinées à l'aide de la version adaptée aux pseudonœuds de **RNAMovies** [16]. **A/** Intron de groupe I de Tetrahymena 394 bases de long. **B/** Ribozyme du virus de l'hépatite delta : 88 bases de long.**C/** Les deux structures bistables et fortement piégées d'un ARN artificiel de 67 bases de long.

delà de cette limite, le temps de calcul de la mise à jour des matrices P^C et $\tilde{P}^C$ réduit l'efficacité de la méthode.
Les courbes *i* et *ii* représentent les facteurs d'accélération obtenus pour un couple de molécules bistables fortement piégées. Le système est placé initialement dans l'une ou l'autre des deux configurations bistables. La courbe d'accélération présente deux régimes distincts. Pour des faibles tailles de cluster, la pente de la courbe est très forte, le système n'a que très peu de structures en mémoire et les choix de sortie pointent essentiellement vers des structures appartenant au puits de potentiel créé par la structure bistable initiale. Au-delà d'une taille critique, le cluster contient la majeure partie des états de plus basse énergie, et le système peut ainsi scruter le haut de la barrière énergétique et basculer dans l'autre puits. Le cluster peut ainsi évoluer en englobant alternativement l'un ou l'autre des puits de potentiel. Le cluster englobe ainsi deux sous-clusters connectés par des chemins de faible poids comparé au poids des chemins ayant pour support l'un ou l'autre des sous-clusters.

La méthode d'accélération est donc particulièrement adaptée aux petits systèmes facilement piégés. Néanmoins, pour des systèmes de grande taille comme l'intron de Group I, des facteurs d'accélération non négligeables peuvent être obtenus.

La performance de l'algorithme est machine dépendant, sur des processeurs grand public de dernière génération (Pentium 4, Opteron, G5), l'existence de cache de second niveau plus grand, permet sans doute d'améliorer la limite supérieure de la taille du cluster. De manière complémentaire le passage d'un compilateur standard (ici GCC 2.8) à un compilateur spécifique au processeur utilisé (GCC 3.2) permet d'obtenir des gains d'accélération moyens de l'ordre de 30% (mesure expérimentale). Dernier point, le passage d'une architecture 32 bits (Pentium, Athlon) pour le codage des données à une architecture 64 bits (Opteron, G5) couplé à un compilateur spécifique (GCC 3.2) permet d'obtenir des gains de performance pour des calculs matriciels de l'ordre de 2 [20].

3.6 Remarques sur la physique du cluster

Dans la simulation classique de transition directe de proche en proche, nous suivons exactement un unique chemin de repliement. Dans la simulation avec accélération, à chaque pas

du programme nous effectuons une marche aléatoire, moyennée sur l'ensemble des chemins possibles qui ont pour support les états du cluster.

Cet algorithme d'accélération de la dynamique de la stochastique est exact au sens où l'algorithme de base donnerait des résultats moyens statistiquement identiques mais après beaucoup plus de calculs (ce qui implique un temps de calcul plus long et une dérive numérique plus importante). Cette méthode de simulations stochastiques accélérées ne suppose pas que le système soit à l'équilibre (comme pour les approches Monte-Carlo classiques) mais les matrices de probabilités introduites sont l'analogue dynamique de la fonction de partition d'équilibre : il s'agit de "fonctions de partition" sur toutes les trajectoires dynamiques parmi un ensemble d'états de références. Cet algorithme permet en particulier d'obtenir les observables du système (moyennes temporelles etc. ...) par différentiation des matrices de probabilité.

Chapitre 4

Résultats numériques

4.1 Introduction

L'algorithme d'accélération nous permet d'accéder à toutes les valeurs moyennées des observables du système. En particulier, nous avons étudié la proportion de pseudonœuds présente dans les structures secondaires d'ARN.

4.2 Validation de l'algorithme

Pour valider l'algorithme, nous avons replié un ensemble de séquences tests non biologiques de séquences de bases tirées aléatoirement et de différentes longueurs, selon les deux méthodes accessibles : repliement cinétique sans accélération par le cluster et avec accélération par le cluster. Le critère de convergence est donné par le pourcentage de structures identiques (mêmes appariements) entre les deux méthodes, lorsque les structures sont relaxées. Par définition, nous avons supposé qu'une structure était relaxée lorsque le système ne trouve pas de structure d'énergie plus basse pour trois simulations indépendantes lorsqu'on augmente le temps de simulation d'un facteur dix.

KineFold dans sa version classique, a été validé sur la prédiction entre autres de la structure

du ribozyme de l'ARN viral de l'hépatite delta. A l'aide de son extension, nous avons pu replier un intron de groupe I et vérifier le bon accord entre les structures prédites et connues (80% des paires de bases préditent correctement dont les pseudonœuds connus). La figure (3.9) montre les conformations obtenues pour ces deux molécules. Une dizaine d'autres structures connues avec pseudonœuds ont aussi été préditent correctement avec cette approche de simulation stochastique [34].

4.3 Relaxation des structures

La cinétique de relaxation ne présente aucun point d'arrêt, le système évolue continuellement. Nous devons donc définir un critère d'arrêt de la simulation qui rende compte de la possible relaxation de la molécule, dans sa configuration biologiquement active. Cette structure peut différer de la meilleure structure thermodynamiquement parlant si le système rencontre un état métastable particulièrement piégé. A l'arrêt de la simulation, l'état de plus basse énergie courant représente l'état optimum atteint au cours de la simulation, et pour que cet état soit considéré comme un bon candidat pour représenter l'état relaxé du système nous avons défini une méthode simple de discrimination du temps minimal de relaxation.
Par définition nous appèlerons temps de relaxation $\tau_r = t_{inst} - t_{opt}$ le temps qui sépare l'instant courant t_{inst} de la simulation de l'instant où le système a trouvé l'état optimum courant t_{opt}. Dès qu'un nouvel état plus stable est rencontré, l'ancien état optimum et son instant correspondant sont remplacés par le nouvel état optimum et l'instant présent. Nous devons donc définir une limite inférieure et une borne d'arrêt pour ce temps de relaxation.

Plus une séquence est longue, plus le temps de relaxation est grand. De même, plus la proportion de base G/C augmente, plus le temps de relaxation est grand. Le premier temps de relaxation à définir est donc celui obtenu pour un ensemble de molécules de séquences aléatoires courtes (ici 50 nucléotides) et ne contenant pas de bases G/C. Puis de proche en proche nous définissons le temps nécessaire à la relaxation de séquences aléatoires de différentes longueurs (100 et 150 nucléotides) et de proportions fixées en base G/C (0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%). Chaque ensemble de séquences de longueur n est constitué de 100 à 3000 séquences indépendantes, générées par n^2 permutations aléatoires d'une séquence déterminée contenant

la proportion fixée de couples de bases G/C et A/U.

Le temps moyen de relaxation est obtenu en moyennant les temps de relaxation de chaque molécule. Pour chaque séquence, le temps de relaxation τ_r doit représenter au moins 80% du temps total t_{cin} de la cinétique, et de plus, ce temps doit être supérieur à un temps minimum estimé *ad-hoc* t_{min}. Pour les séquences de 50 bases et 0% de G/C le temps minimal est fixé à 250 ms.
Nous avons ajouté un critère d'arrêt pour mettre fin à la simulation de séquences "pathologiques" ayant des difficultés à relaxer. Si le temps total t_{cin} de la cinétique simulée est a fois supérieur au temps minimum de relaxation, si l'efficacité moyenne du cluster ($< \bar{\ell}^A_{ji} >_{ji}$ 3.18) est inférieure à une limite que l'on se fixe et si le temps de relaxation est inférieur à m% du temps total de la cinétique, alors, la simulation est arrêtée. Typiquement, en phase de test $a = 20$, $m = 75\%$ et l'efficacité minimale du cluster est fixée à un facteur 10.
En conclusion, les critères d'arrêts sont les suivants :

Relaxation de la structure : $\tau_r / t_{cin} \geq 0.8$

Réjection de la séquence : $t_{cin} \geq a \cdot t_{min} \quad \& \quad \tau_r / t_{cin} \geq m \quad \& \quad emove/move \leq 10$

De plus, pour être considérée comme relaxée, une séquence doit atteindre la même structure d'énergie minimum pour trois simulations indépendantes consécutives, sinon elle est rejetée.

Lorsque toutes les séquences d'un ensemble ont été repliées, nous obtenons le temps moyen de relaxation pour la longueur et la proportion en base G/C donné, ainsi que la proportion des séquences qui ont été rejetées. Le temps moyen de relaxation est considéré comme validé si la proportion de séquences avortées est inférieure à 10%. Le résultat est donné dans le graphique 4.1

Nous nous sommes basés sur ces résultats numériques pour définir un heuristique $f(l, \%GC)$ d'estimation du temps minimum de relaxation :

$$f(l, \%GC) = \frac{(l+50)}{4} \exp^{(l+50)} (\%GC)^{1.85}$$

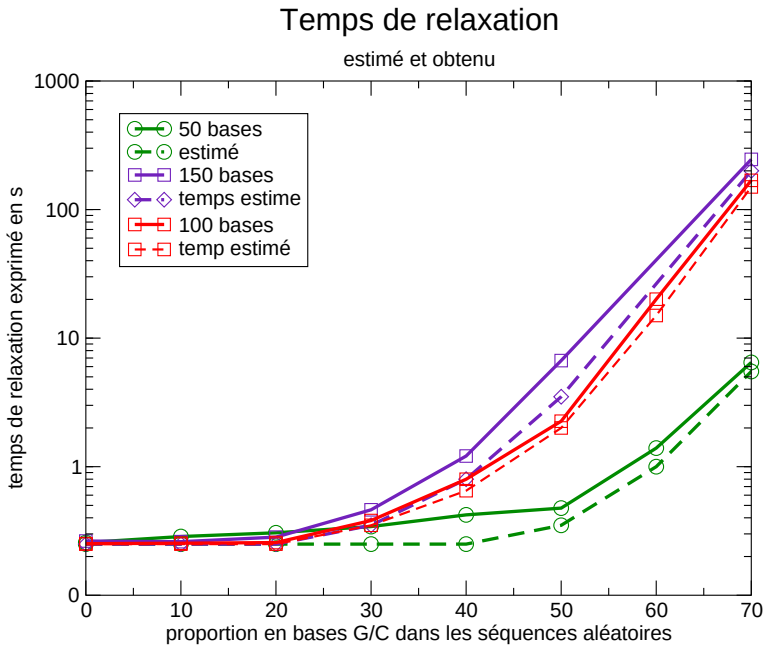


FIG. 4.1 – Evolution du temps de relaxation pour des séquences de tailles 50, 100 et 150 bases et de proportion en base G/C variant de 0% à 70% par pas de 10%.

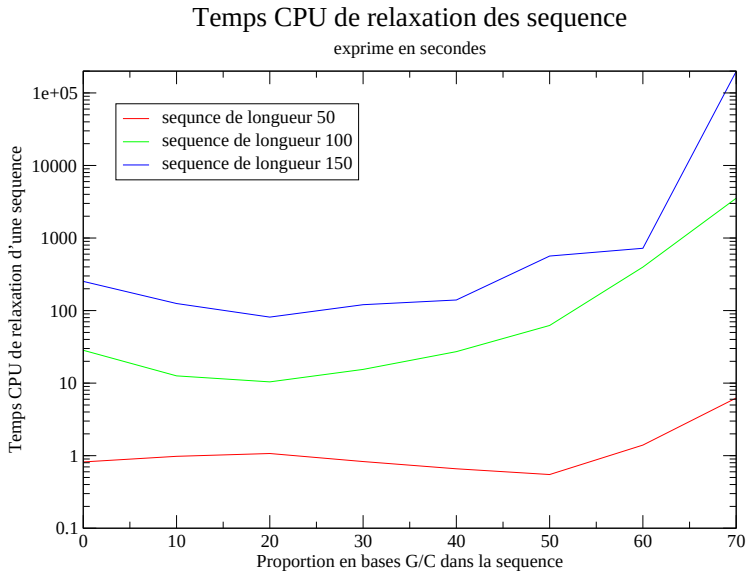


FIG. 4.2 – Evolution du temps moyen CPU utilisé pour relaxer une séquence de taille 50, 100 et 150 bases et de proportion en base G/C variant de 0% à 70% par pas de 10%. temps obtenu sur un Athlon Mp 1.2 Ghz, avec une taille de cluster fixée à 35 éléments.

4.4 Proportion de pseudonœuds dans des séquences aléatoires courtes

Notre approche de la prédiction de structure secondaire par la simulation de la cinétique de repliement nous permet de suivre la dynamique des transitions entre les états métastables. Le traitement exact de la contribution énergétique des pseudonœuds ajouté aux contributions classiques des coûts entropiques des boucles terminales et des énergies libres d'appariement-empilement de bases, nous permet de prédire naturellement des structures avec pseudonœuds. Le code classique de repliement a été préalablement testé sur un ensemble de molécules test [34], puis les mêmes molécules ont été repliées à l'aide de l'algorithme d'accélération. Après nous être convaincus de l'efficacité de la méthode, nous nous sommes intéressés à quantifier la proportion de pseudonœuds dans les structures secondaires d'ARN.

Un pseudonœud est un motif qui fait intervenir au moins deux hélices imbriquées. Nous définissons les hélices à pseudonœuds comme les hélices à enlever pour revenir à une structure secondaire qui vérifie les lois d'appariement de Watermann [84]. Cette définition est en fait ambiguë puisque au minimum un pseudonœud inclut deux hélices. Il existe donc deux possibilités et deux structures en arbre qui résolvent ce pseudonœud. Pour lever la dégénérescence des structures non nouées, nous choisissons d'enlever les hélices énergétiquement les plus faibles. La proportion de pseudonœuds est alors définie comme étant la somme minimale des longueurs des hélices à enlever de la structure pour résoudre les pseudonœuds divisée par la somme des longueurs de toutes les hélices présente dans la structure étudiée.

4.4.1 Détection de motif pseudonœuds

La détection de motifs pseudonœuds revient à identifier les hélices à retirer de la configuration pour retrouver une structure dont tous les appariements vérifient les lois de Watermann. Notons que pour un motif pseudonœud, le nombre d'hélices participant au motif est au moins égal à deux. Nous aurons donc à *choisir* les hélices à retirer. Pour nous rapprocher le plus possible des codes de prédiction de structures secondaires classiques, nous avons décidé de retirer systématiquement l'hélice ou les hélices qui minimisent les changements dans la configuration.

Ci-dessous, nous décrivons un algorithme qui à partir d'une structure partiellement ou totalement repliée à l'aide de l'algorithme ***KineFold*** , détermine de manière univoque, les hélices à retirer pour résoudre les motifs pseudonœuds. L'idée principale de l'algorithme est d'identifier, itérativement, le maximum de la structure arborescente, vérifiant les lois d'appariement de Watermann, puis de retirer de la liste des hélices restantes, une hélice parmi celles qui structurent le motif pseudonœud. Le choix de l'hélice à exclure peut se baser sur le nombre d'appariements minimum à enlever ou, sur l'énergie minimale de l'hélice à ouvrir.

Dans un premier temps, l'algorithme lit la structure en suivant son squelette de l'extrémité $5'$ vers l'extrémité $3'$. L'identifiant de chaque brin d'hélice visité est empilé dans une pile contenant tous les identifiants de brins à traiter (voir schéma 4.3).
Dans cette pile, si deux identifiants contigus représentent les deux parties complémentaires de la même hélice, alors, les deux brins sont retirés de la pile. On peut imager le processus comme un tetris unicolonne, où chaque bloc correspond à un brin apparié de la structure et arrive dans l'ordre de lecture de la séquence. Si les briques s'emboîtent, elles sont décimées, sinon, elles sont empilées. En fin de lecture, la pile résultante de la décimation contient la ou les sous-structures incluant au moins un couple d'hélices ne vérifiant pas la seconde loi d'appariement de Watermann (voir figure 4.4).
Chaque couple de brins ne vérifiant pas la seconde loi d'appariement définit un intervalle le long de la séquence. Chaque intervalle contient au moins un brin d'hélice. Nous associons à chaque couple, un poids proportionnel à la somme des longueurs ou des énergies des brins contenus dans l'intervalle. Le couple dont le poids est le plus proche de zéro désigne la section qui contient les identifiants des hélices à exclure. Les brins correspondants sont retirés de la pile et les hélices sont stockées sur la liste des hélices simplifiant un motif pseudonœud.
L'opération de décimation et d'identification est réitérée tant que la pile n'est pas vide. Ci-dessous, nous illustrons le processus de décimation et d'identification sur quelques exemples simples.

La méthode décrite ci-dessus assure de tuer en une passe toute structure ou sous-structure strictement arborescente, c'est-à-dire, vérifiant les lois d'appariement de Waterman[84]. En effet (voir le schéma 4.3) :

– pour des hélices branchées, les brins complémentaires se suivent les uns les autres, la pile

ne contient jamais plus d'une seule référence et à chaque nouveau brin visité, elle est vidée.

– pour des hélices imbriquées, les identifiants sont enfilés tant que l'on ne visite pas le brin complémentaire de la tige boucle-terminale. Dès que celle-ci est identifiée, la pile est vidée de proche en proche.

Si la structure contient au moins un couple d'hélices telles que leurs brins complémentaires s'alternent, l'algorithme doit effectuer deux passages au moins pour résoudre la configuration (voir le schéma 4.3).

4.4.2 Résultats

A l'aide de l'algorithme d'accélération, nous pouvons suivre les moyennes temporelles de n'importe quelle observable du système, en particulier, la proportion de pseudonœuds présents dans les structures visitées. Nous nous sommes donc intéressés à la proportion de pseudonœuds présents dans les configurations visitées lorsque la structure a relaxé. Le résultat obtenu reflète donc la proportion moyenne en pseudonœuds attendus pour une population de molécules relaxées maintenues dans un bain à 37°C. Par définition, la proportion en pseudonœuds d'une configuration donnée, est calculée comme le rapport du nombre de bases appariées qu'il faut retirer de la configuration pour retrouver une structure vérifiant les lois d'appariement de Waterman [84], divisé par le nombre total de bases appariées dans la configuration nouée (figure 4.4.2).

L'analyse statistique a été faite selon le schéma suivant : nous avons réinitialisé un ensemble de séquences aléatoires de longueurs différentes 50, 100 et 150 nucléotides selon le même schéma que celui utilisé pour l'étude du temps de relaxation des structures. Les critères d'arrêt de chaque "run" sont identiques à ceux qui sont utilisés dans l'étude ci-dessus. De plus, chaque séquence est repliée trois fois selon trois trajectoires indépendantes, et la structure est admise comme relaxée si et seulement si la simulation a abouti, et si et seulement si au moins deux des trois structures finales obtenues sont identiques. Les taux de réussite varient de 90% à 100%. La proportion moyenne en pseudonœuds est remise à zéro dès qu'un nouveau minimum est atteint. Ainsi, la proportion sauvegardée est représentative de la population à l'équilibre. Seuls les runs ayant thermalisé sont pris en compte pour l'exploitation graphique des résultats. Notez que

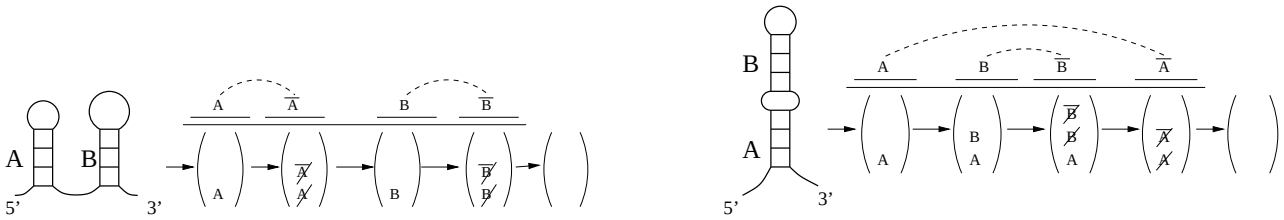


FIG. 4.3 – Illustration de l'évolution de la pile des hélices à traiter pour les cas d'hélices branchées (à gauche), et imbriquées (à droite). Lorsque la partie complémentaire de l'hélice est visitée, les deux brins complémentaires sont décimés. En fin de lecture, la pile est vide, reflétant l'inexistance de motif pseudonœuds dans la structure.

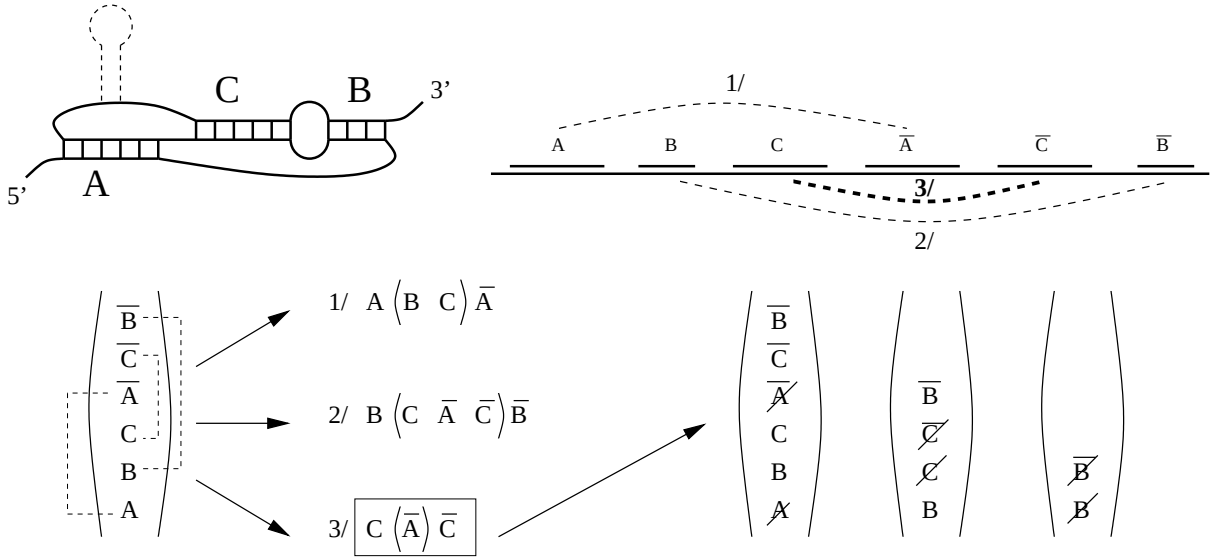


FIG. 4.4 – Illustration de l'évolution de la pile des hélices à traiter pour le cas d'une structure contenant un pseudonœud. En trait pointillé, la sous-partie strictement arborescente décimée lors du remplissage de la pile, elle est aussi retirée de l'analyse pour simplifier le schéma. La pile contient tous les brins résultant de l'analyse de décimation. Parmi les trois couples identifiés comme contenant des brins d'hélices simplifiant le pseudonœud. Le couple $C(A)\bar{C}$, est celui qui contient le minimum de brins. L'hélice constituée des brins $(A;\bar{A})$ est retirée de la pile et l'analyse de décimation est réitérée sur la liste simplifiée. En retirant l'hélice A, la structure résultante est strictement arborescente, le second passage vide donc la pile. Notez que l'algorithme choisit de retirer le minimum d'hélices pour résoudre le motif, en effet, nous aurions aussi pu choisir de retirer les hélices B et C pour retrouver une structure arborescente.

compte tenu des critères utilisés, les distributions ne prenant en compte que les configurations à l'équilibre et les distributions obtenues sur l'ensemble des configurations repliées sont très proches.

La figure 4.4.2, représente l'évolution moyenne et l'écart-type du pourcentage de pseudonœuds en fonction de la proportion des bases GC/AU pour les différentes longueurs 50, 100 et 150 nucléotides(figure 4.6).

En moyenne, les structures présentent près de 10 à 20% de pseudonœuds excepté pour les séquences de 50 nucléotides n'ayant que des bases A et U. C'est le seul cas pour lequel il n'y a quasiment jamais de pseudonœuds. Au-delà de 50% de bases G/C dans la séquence, la proportion moyenne de pseudonœuds est très proche quelle que soit la longueur des séquences aléatoires étudiées. De plus pour les séquences de 100 et 150 nucléotides, le pourcentage de pseudonœuds en fonction de la proportion de bases GC/AU est quasi-identique. Ceci suggère qu'en moyenne, la proportion de pseudonœuds suit une courbe limite indépendante de la longueur et la composition en bases des séquences aléatoires étudiées.

Les distributions de proportion de pseudonœuds sont très étalées, elle varient de quelques pourcents à près de 30% de bases appariées dans les hélices à retirer selon la longueur et la proportion en bases G/C dans la séquence.
La distribution obtenue avec les séquences de 50 bases, montre un double comportement : à faible taux en bases G/C, la formation d'un pseudonœud a un coût entropique trop élevé pour qu'il se forme. Au fur et à mesure que la proportion en bases G/C augmente, la stabilité de l'appariement, permet la formation de motif pseudonœuds. La séquence étant très courte, le nombre d'appariements impliqués dans la formation de motif pseudonœuds est rapidement prépondérant, d'où la formation d'une bosse centrale dans la distribution. En notant que la proportion de bases appariées est de l'ordre de 65%. Ce nombre de bases à retirer est de l'ordre de 5, soit une hélice unique.
La distribution obtenue pour les séquences de 100 bases fait charnière entre le comportement obtenu avec des séquences courtes et les séquences plus longues. A faible taux de bases G/C, le comportement se rapproche de celui de séquences courtes. Le coût entropique de formation de pseudonœuds ne permet pas de refermer la structure sur elle-même. Par contre, avec l'augmentation de la probabilité d'appariement de bases G/C, les configurations ont une tendance

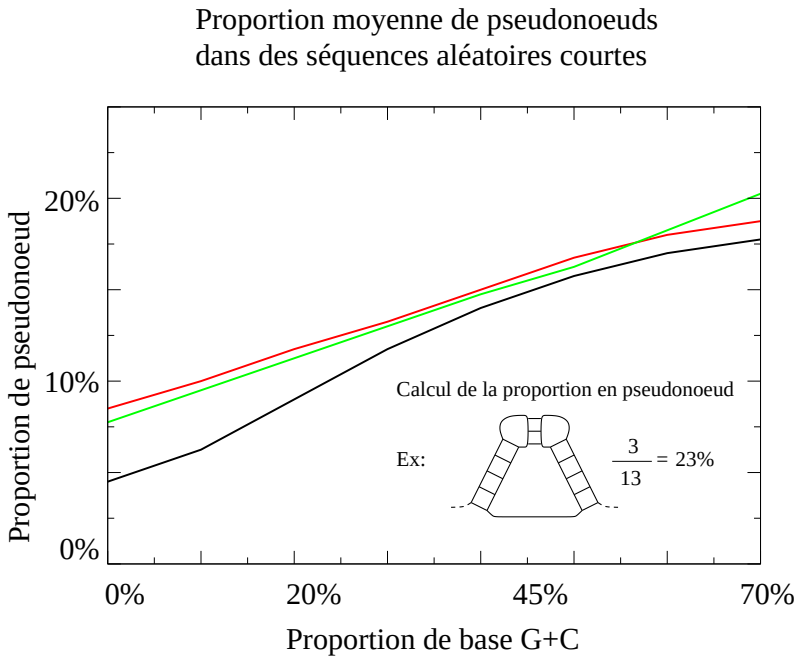


FIG. 4.5 – Proportion de pseudonœuds dans des séquences de longueurs et de distribution en base GC/AU différentes.

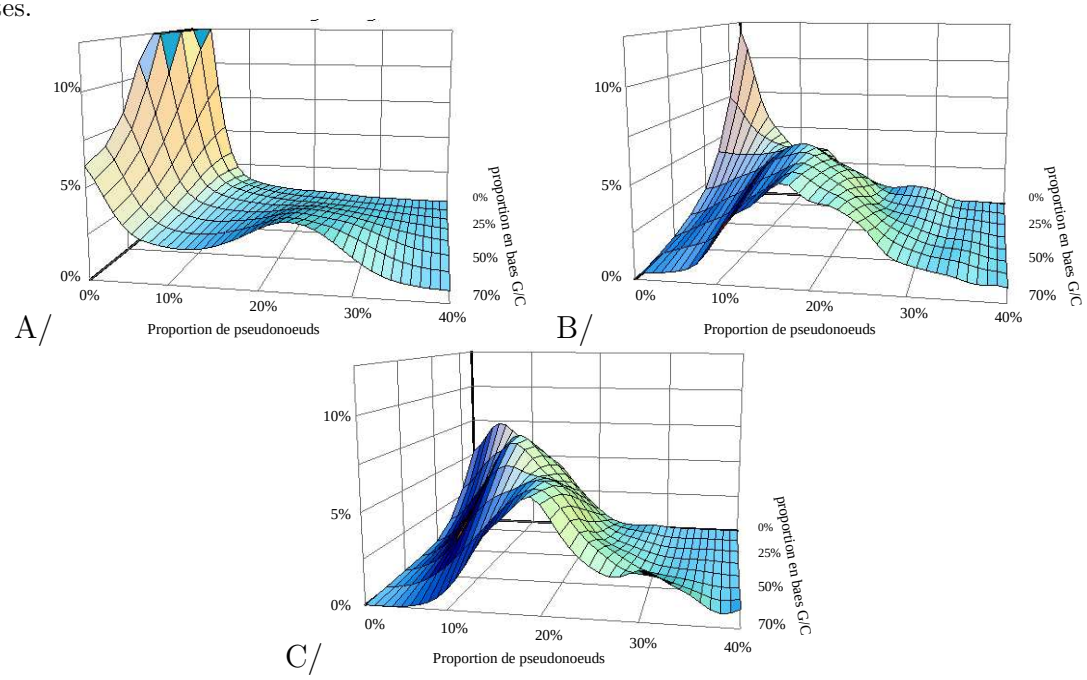


FIG. 4.6 – Distribution des pseudonœuds dans des séquences de longueurs et de distribution en bases GC/AU différentes. **A :** 50 base ; **B :** 100 bases ; **C :** 150 bases

naturelle à former des motifs noués.
Pour des séquences "longues" (150 bases), la longueur relative de la séquence permet en moyenne de former des pseudonœuds quelle que soit la proportion de bases G/C. Par comparaison avec les séquences "courtes", il apparaît que des pseudonœuds plus "délocalisés" sont favorables à la stabilisation de la molécule. Néanmoins, puisqu'en moyenne, il semble que la proportion de pseudonœuds n'évolue pas entre des séquences de 100 et 150 bases, il se peut que cette "délocalisation" reste tout de même limitée, c'est-à-dire qu'il existerait une taille caractéristique pour les interactions entres boules et brins libres.

Nous avons comparé ces résultats avec le repliement de section d'ARN messagers de 150 bases de long issus du génome d'*Escherichia Coli* et de *Saccharomyces cerevisiae (yeast)*. Les sections sont prises en milieu de gène. Pour évaluer le caractère aléatoire ou déterministe des séquences, nous avons généré deux autres ensembles à partir des séquences génomiques :

- le premier ensemble est obtenu en permutant aléatoirement les bases de chacune des séquences, pour obtenir de vraies séquences aléatoires mais au contenu en bases identiques.
- les séquences du second vérifient les corrélations obtenues par une analyse markovienne d'ordre 2 des séquences génomiques. Elles ont un contenu en bases différent, mais vérifient les statistiques de codon de l'ensemble de référence.

Les distributions obtenues (figure 4.7) pour les différents ensembles de séquences "génomiques" montrent le même comportement de distribution très étalée pour la proportion en pseudonœuds, et se comportent comme celles qui sont obtenues pour l'ensemble des séquences aléatoires de composition en bases G/C identiques. La représentation en proportion cumulée (en dessous de chaque graphique), montre que les "accidents" ne sont a priori pas significatifs du point de vue de la statistique. L'écart entre les courbes et de l'ordre de $1/\sqrt{n}$, avec n le nombre de séquences. Ces résultats indiquent que les séquences génomiques présentent un fort potentiel de formation de motif pseudonœuds qui peut être mis à profit pour moduler la structure tridimensionnelle de l'ARN messager.

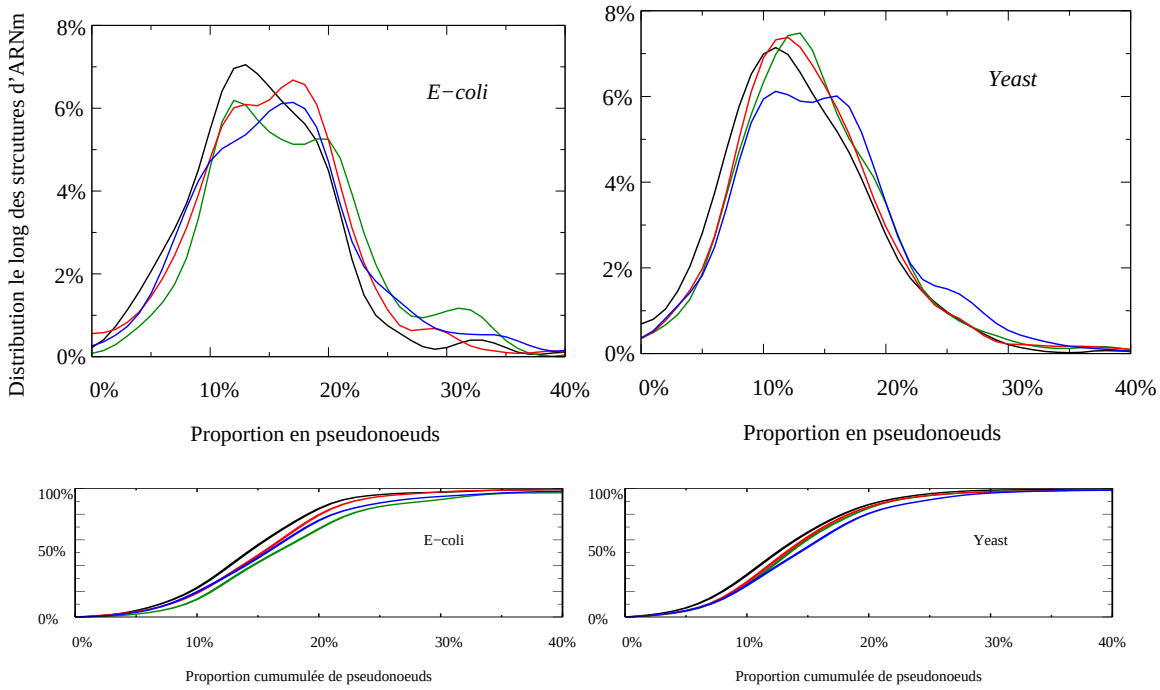


FIG. 4.7 – Distribution de pseudonœuds le long de structures obtenues par le repliement de séquences génomiques tronquées d'ARN messager de 150 bases de long. A gauche : *Escherichia Coli* et à droite *Saccharomyces cerevisiae (yeast)*. **Rouge** : séquences génomiques ; **Verte** séquences permutées (même composition en bases) ; **Noire** séquences markoviennes basées sur la statistique de codon ; **Bleue** : séquences purement aléatoires avec une proportion en base G+C de 50% (à gauche) et 40%(à droite), correspondant aux proportions moyennes de bases G+C contenues dans les séquences génomiques d'*Escherichia Coli* et *Saccharomyces cerevisiae (yeast)*

4.5 Quantification de l'erreur en négligeant les pseudonœuds

L'étude de la proportion de pseudonœuds dans les structures de séquences aléatoires, n'informe pas directement sur l'erreur réelle commise par les prédictions classiques des structures non nouées. Pour comparer quantitativement les structures obtenues en autorisant ou non la formation de motifs pseudonœuds, nous avons défini une distance basée sur la différence d'appariement entre les structures relaxées avec et sans pseudonœuds.

La manière la plus simple de définir une distance entre deux structures de séquence identique, est de compter le nombre de paires de bases qui différent entre elles. Nous avons simplement généralisé cette mesure à la distance entre deux ensembles de structures pour tenir compte non pas de la seule configuration relaxée, mais des trois plus basses structures obtenues. Nous comparons ainsi les probabilités que la paire de bases (i, j) soit appariée dans l'un et l'autre des deux ensembles.

La probabilité d'appariement $< P(i,j) >_\zeta$ sur l'ensemble ζ est simplement donnée par la somme des certitudes d'appariement de la paire de bases (i, j) sur une des configurations appartenant à ζ pondérée par son poids de Boltzmann. L'expression de la probabilité pondérée s'exprime sous la forme suivante :

$$< P(i,j) >_\zeta \quad = \quad \sum_{k=0}^{n} \frac{e^{-\beta E_k}}{Z_\zeta} \delta_\zeta^k(i,j), \qquad Z_\zeta = \sum_{k=0}^{n} e^{-\beta E_k} \tag{4.1}$$

$$\text{avec} \qquad \delta_\zeta^k(i,j) = \begin{cases} 1 \ : \text{ si } i \text{ et } j \text{ sont appariées dans} \\ \qquad \text{la structure } k \text{ appartenant à } \zeta, \\ 0 \ : \text{ sinon} \end{cases}$$

$$\text{et} \quad \text{où} \quad k \text{ indice le numéro de la configuration étudiée}$$

La distance normalisée entre les deux ensembles est donnée par la différence de probabilité d'appariement entre les deux ensembles ζ et ζ' :

$$D \quad = \quad \sum_{(i,j)} \frac{|< P(i,j) >_\zeta - < P(i,j) >_{\zeta'}|}{\sum_{(i,j)} \{1/2(< P(i,j) >_\zeta + < P(i,j) >_{\zeta'} + | < P(i,j) >_\zeta - < P(i,j) >_{\zeta'} |)\}} \tag{4.2}$$

En se limitant à une seule configuration par ensemble et en admettant que la seule différence entre les deux configurations est la présence d'hélices fermant les pseudonœuds, la distance D est égale à la proportion de pseudonœuds dans la structure[1].

Le graphique 4.8 donne l'évolution de la distance moyenne entre les deux ensembles pour une longueur et une proportion en base G/C donnée. Contrairement au cas de la proportion de pseudonœuds, plus la longueur de la séquence augmente et plus la distance entre les structures augmente. Mais cet effet moyen n'est absolument pas caractéristique de l'effet constaté. Au regard des distributions obtenues pour les distances, nous ne pouvons que conclure sur le fait que les ensembles peuvent être très semblables comme extrêmement différents. Seule certitude, pour un faible taux de bases G/C, le système à du mal à former des pseudonœuds, ce qui se traduit par une distance moyenne entre les ensembles proche de zéro.

Si l'on autorise la molécule à former des motifs pseudonœuds, les configurations de plus base énergies obtenues peuvent être du tout au tout semblables ou très différentes. Donc, il ne suffit pas en général de simplement retirer (ou d'ajouter) les hélices à pseudonœuds pour retrouver la structure arborescente (ou la structure nouée) qui correspond à la configuration relaxée. Le fait de retirer les hélices à pseudonœuds, implique un réarrangement moyen qui modifie fortement la configuration obtenue, bien au-delà du simple réarrangement local des doubles brins retirés.

4.5.1 Dénaturation mécanique d'ARN

KineFold a été adapté à la prédiction de courbes de force-extension pour comprendre les mécanismes de dénaturation mécanique de structures secondaires de molécule d'ARN [26]. Dans le principe, l'expérience consiste à liguer un ARN structuré natif à deux brins ADN, l'un greffé

[1]Dans ce cas, les expressions se réécrivent sous la forme suivante :

$$|< P(i,j) >_{\zeta} - < P(i,j) >_{\zeta'}| = n_{psd}$$

$$1/2(< P(i,j) >_{\zeta} + < P(i,j) >_{\zeta'} + | < P(i,j) >_{\zeta} - < P(i,j) >_{\zeta'} |) = 1/2([n_{psd} + n_h] \; + \; n_h + |n_{psd}|)$$

soit $D = \dfrac{n_{psd}}{n_h + n_{psd}}$

avec n_{psd} : nombre d'appariements appartenant aux hélices à pseudonœuds
n_h : nombre d'appariements de la structures sans pseudonœuds

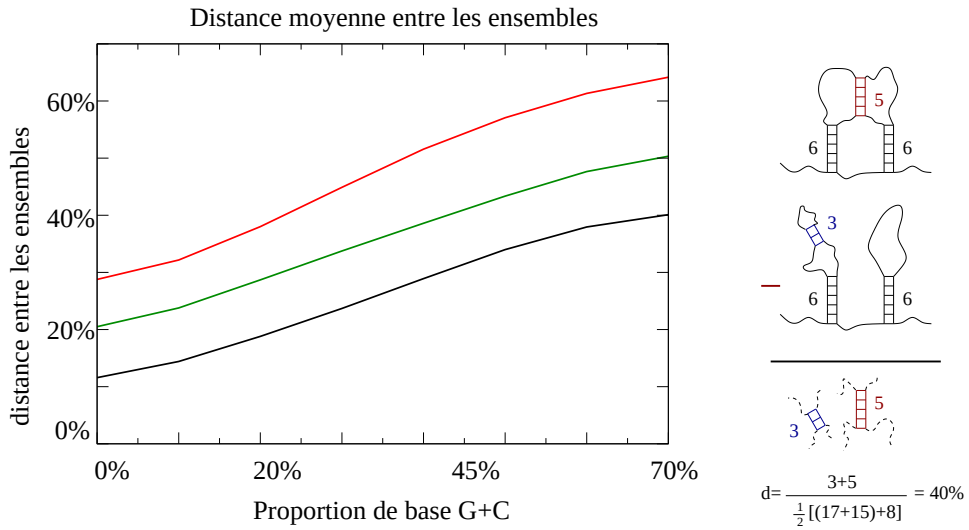


FIG. 4.8 – Evolution de la distance moyenne entre les séquences relaxées avec et sans pseudonœuds. Contrairement à la moyenne en pseudonœuds, plus la longueur de la séquence augmente et plus la distance entre les structures avec et sans pseudonœuds augmente. **Noir** pour des longueurs de 50 bases ; **Vert** pour des longueurs de 100 bases ; **Rouge** pour des longueurs de 150 bases.

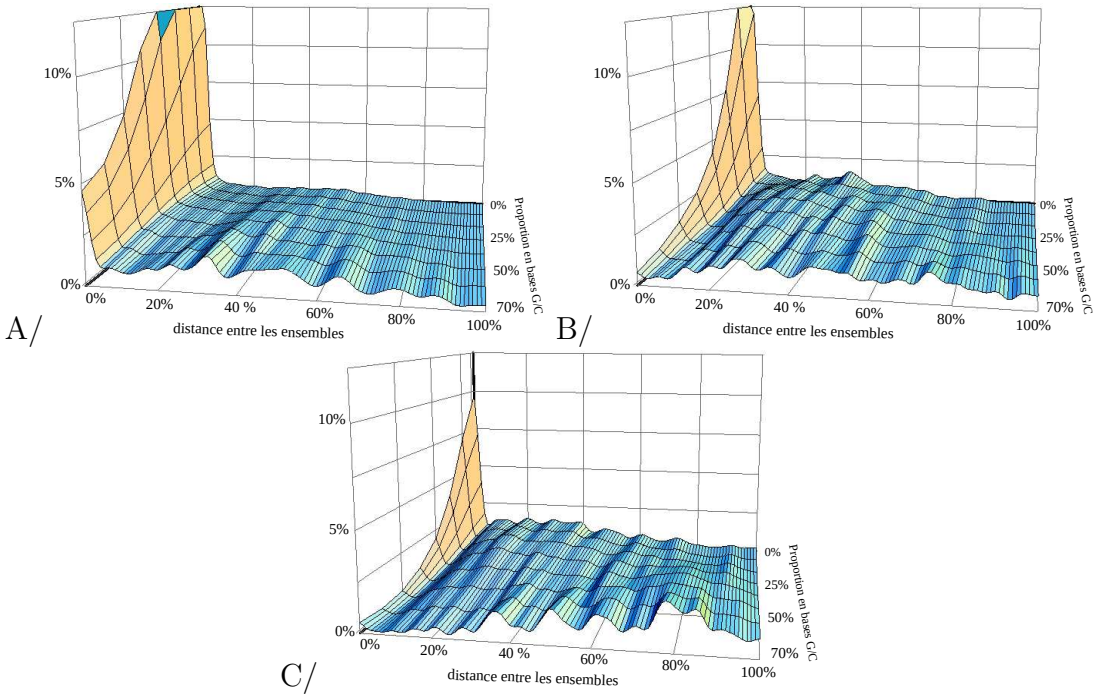


FIG. 4.9 – Distribution des distances entre les séquences de longueurs et de distribution en base GC/AU différentes. **A :** 50 bases ; **B :** 100 bases ; **C :** 150 bases. Les distributions des distances sont quasiment plates : les structures avec et sans pseudonœuds sont totalement décorrélées.

sur une surface fonctionnalisée, l'autre attaché à une bille diélectrique piégée par une pince optique. En déplaçant la paroi, le système est mis sous tension(cf. 4.10).

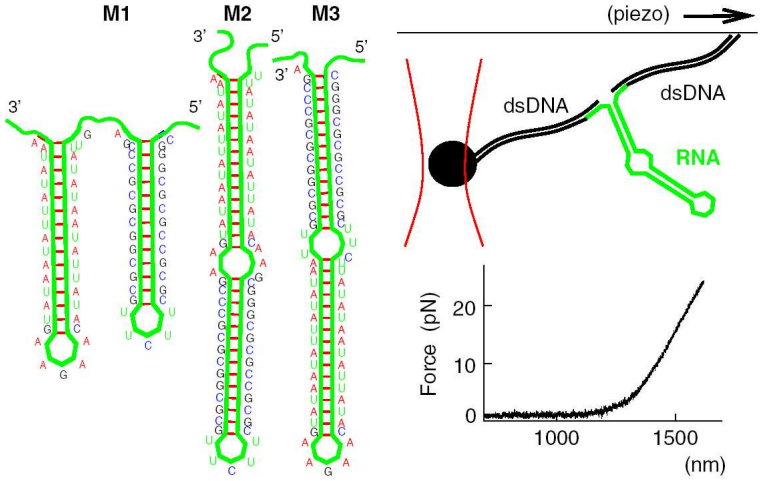


FIG. 4.10 – Trois ARN artificiels pour lesquels les signaux de dénaturations sont suffisament simple pour être exploités sans l'aide de l'outil informatique. Schéma de principe du montage de dénaturation mécanique. L'ARN est maintenu entre deux bras ADN accrochés pour l'un à une surface mobile, pour l'autre à une bille d'or pincée optiquement.

La force induit une dénaturation de la molécule en dégrafant les bases appariées. La courbe de force-extension présente alors des sauts caractéristiques de la relaxation de la tension par l'ouverture d'hélices sous tension. Après dénaturation totale, la molécule peut être renaturée en relachant progressivement la tension. On peut alors obtenir quelques courbes de dénaturation/renaturation avec la même molécule (cf. 4.11).

A l'aide de l'algorithme ***KineFold*** , il est possible de simuler ce type d'expérience. Les bras ADN double brin et la bille sont modélisés comme des ressorts idéaux couplés à une relaxation visqueuse typique obtenue à partir des données expérimentales. La tension appliquée est induite par la contrainte de distance bout à bout appliquée à la chaîne (cf. 4.11).

Dans un premier temps, la prédiction numérique a été comparée avec les résultats expérimentaux obtenus pour de courtes molécules synthétiques de structures connues. Les courbes de force-extensions obtenues, sont très similaires à celles qui sont obtenues expérimentalement

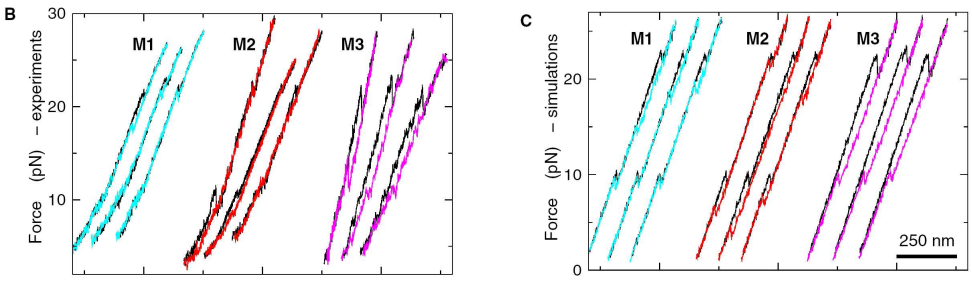


FIG. 4.11 – A gauche, courbes expérimentales de la dénaturation mécanique. A droite, courbes obtenues par simulations numériques de la dénaturation mécanique.

et rendent bien compte de l'aspect d'hystérèse entre l'aller et le retour. La concordance des positions de début et fin et l'amplitude du cycle d'hystérèse conforte l'idée que l'approche cinétique employée permet effectivement de prédire la dynamique de l'ARN de l'échelle de la seconde à la minute. Le couplage entre la courbe obtenue et la structure correspondante est immédiat et permet de comprendre les mécanismes de réarrangement qui conduisent à l'existence de ce cycle d'hystérèse (cf 4.12).

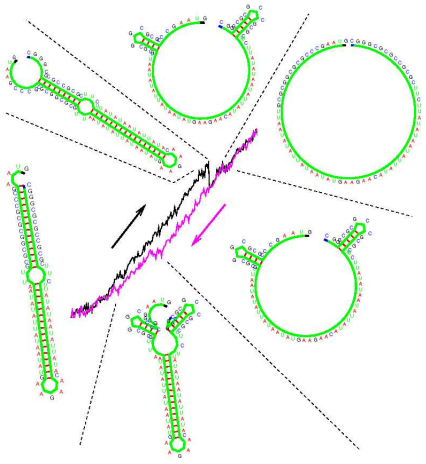


FIG. 4.12 – Représentation des structures de la molécule M1 au cours du cycle de traction-relaxation. Le cycle d'hysthérèse s'explique par l'existence d'une structure métastable qui se forme au cours de la relaxation.

L'application à des molécules plus longues, ici au 16S, permet de tester l'hypothèse avancée

selon laquelle pour les longues molécules, les réarrangements de la molécule sous tension lissent les événements internes en s'ajustant continuellement à la tension imposée [50, 21, 42]. La simulation a été effectuée en tenant compte d'un nombre croissant d'hélices potentielles, et en utilisant comme structure initiale la configuration native obtenue par phylogénétique (cf 4.13).

La simulation disqualifie de suite la possibilité de ne faire intervenir que les hélices natives. En effet, les courbes de force-extension simulées ne sont ni reproductibles ni proches de la courbe expérimentale. Pour s'en approcher au mieux, il faut à la fois permettre la formation d'hélices non natives et la possibilité de réarranger localement la molécule à l'aide de ces hélices au cours de la traction.

Cette étude apporte à la fois à notre code de prédiction numérique et à la compréhension de la dynamique hors équilibre des molécules d'ARN. La validation par l'expérience, nous permet de conclure que l'approche utilisée par ***KineFold*** est cohérente avec la réalité, et que l'approche cinétique rend effectivement compte de processus physiques reproductibles. Du point de vue de l'expérience, ***KineFold*** apporte un regard nouveau sur les processus de réarrangements internes au cours de la tension. En particulier, l'étude montre que ceux-ci ne permettent pas de lisser complètement les signaux de réponse à la force comme attendu. De plus, les réarrangements sont essentiellement locaux, ce qui permet de garder intacte la majeure partie de la structure interne.

Ma contribution à cette partie se réduit à l'obtention de la visualisation graphique du chemin de repliement. Les modifications apportées au programme **RNAMovies** sont expliquées dans le chapitre 8 page 135

4.6 Conclusion

Les résultats de cette étude sont assez inattendus puisqu'ils réfutent l'idée couramment admisse selon laquelle il y aurait peu de pseudonœuds dans les structures d'ARN. Nos résultats montrent que les distances ainsi évaluées sont typiquement deux fois plus grandes que la proportion de pseudonœuds pour une même longueur et une même composition GC/AU des séquences. Ceci semble indiquer que l'enlèvement ou l'ajout de pseudonœuds dans les structures peut induire un réarrangement important de celles-ci. Ceci implique qu'il est sans doute ardu de com-

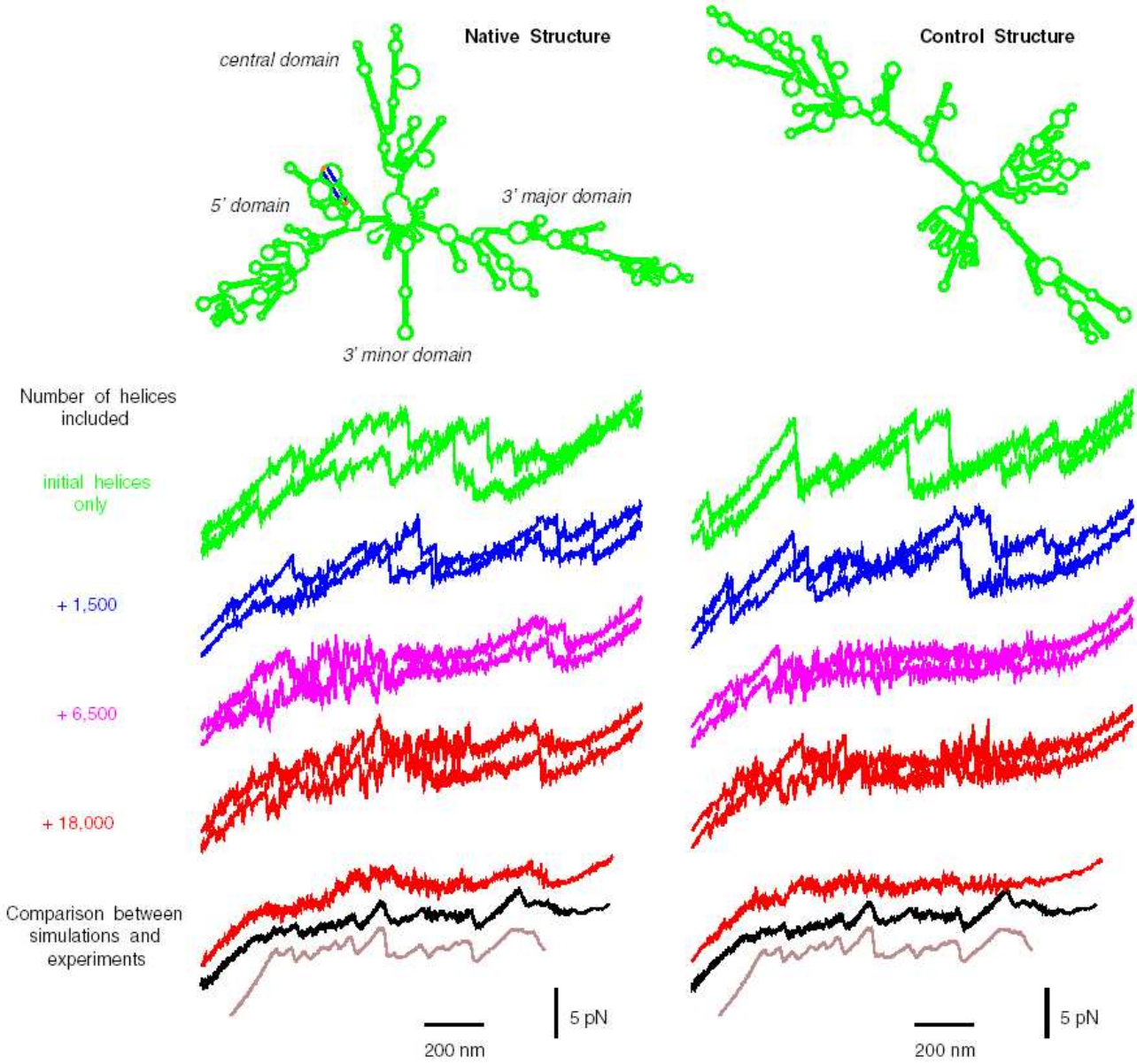


FIG. 4.13 – Simulations des courbes de force-extention de l'ARN ribosomal 16S d'*E. Coli* pour la structure native connue (à gauche) et obtenue par minimisation numérique de l'énergie libre (à droite) par l'algorithme de Zuker. Deux simulations sont représentés pour chaque ensemble de conditions pour illustrer la reproductibilité. En vert : seules les hélices natives peuvent se former. Bleu, magenta et Rouge : ajout d'hélices de tailles et d'énergies libres plus petite. La comparaison avec les courbes expérimentales en noir et brun démontre la nécessité de faire intervenir des hélices non-natives qui ont la possibilité de se former au cours de la dénaturation mécanique. Les courbes numériques obtenues avec la structure connue se rapproche qualitativement plus des courbes expérimentales.

mencer par prédire la meilleure configuration vérifiant les lois d'appariement de Watermann puis dans un second temps d'analyse, permettre la formation de motifs pseudonœuds.

Troisième partie

Approche expérimentale de la cinétique de repliement de molécules d'ARN

Chapitre 5

Introduction

5.1 Contexte général

5.1.1 Séquences biologiques et information

Le *dogme central* introduit par Francis Crick à la fin des années 50, souligne le flot d'informations de l'ADN qui contient le code génétique, à l'ARN, copie intermédiaire, jusqu'aux protéines, qui sont les principales molécules fonctionnelles de la cellule.

ADN → ARN → protéine.

Nous savons maintenant que les ARN sont aussi des molécules fonctionnelles [59]. Si les ARn sont catalytiques , ont parle de ribozymes.

Par analogie, le *dogme moléculaire* relie les différentes étapes de la constitution de la molécule fonctionnelle. Dans le cas d'un ARN, la séquence des bases définit l'ensemble des appariements intramoléculaires possibles. La combinaison de ces paires de bases constitue la structure secondaire. La fonction résulte de l'organisation tridimensionnelle de cette ossature. Schématiquement le *dogme moléculaire* s'énonce de la manière suivante : une séquence se replie dans une structure qui exprime une fonction.

une séquence → une structure → une fonction.

Pour une séquence, la structure qui exprime la fonction est unique. Le *dogme moléculaire* peut être réexprimé sous la forme suivante : une séquence code pour une structure fonctionnelle.

une séquence → une structure fonctionnelle.

La structure fonctionnelle n'est pas la seule conformation possible pour la molécule. Pour une séquence donnée, le nombre d'hélices qui peuvent se former croît comme le carré de la longueur de la chaîne du polynucléotide. L'ensemble des combinaisons d'hélices compatibles, engendre un espace de structures métastables qui grandit exponentiellement avec la longueur de la séquence (en se limitant aux pseudonœuds les plus simples) [88]. A partir d'une séquence donnée, il est donc possible d'accéder à une multitude de structures. De même, pour une conformation donnée, la mutation couplée d'une paire de base ne modifie généralement pas la structure. Une structure peut être atteinte par une multitude de séquences. Un étude analytique basée sur la théorie des graphes menée par Haslinger [28, 27] démontre que les 4^n séquences de longueur de chaîne n, engendrent un nombre de structures secondaires sans pseudonœuds proportionnel à 1.86^n. En y intégrant les structures qui présentent les conformations les plus courantes de pseudonœuds, le sous-espace des structures croît comme 2.64^n [28, 27]. Le nombre total de structures engendrées par les 4^n séquences est plus petit que le nombre total de séquences. Potentiellement, la quantité d'informations contenue le long d'une séquence est donc beaucoup plus grande que celle a priori nécessaire pour définir sa ou ses structure(s) d'équilibre.

5.1.2 Réseaux de mutation neutres : évidence d'une grande adaptabilité des ARN

Au cours de l'évolution, ces degrés de libertés supplémentaires des structures primaires ont sans doute permis aux molécules d'ARN d'évoluer tout en gardant leur fonctionnalité. Lors de la réplication du génome, les erreurs de copies induisent des mutations ponctuelles dans les séquences (substitution, insertion, délétion d'un nucléotide). La pression de sélection permet aux séquences "mutantes" les mieux adaptées à l'environnement de remplacer les séquences "naturelles". Au cours de l'évolution, l'adaptation des espèces aux nouveaux environnements, est sans doute facilitée par la plastique des séquences des molécules constitutives (ARNs, protéines).

Ainsi, pour les ribozymes, une même fonction présente sur différents organismes peut être portée par des séquences nucléotidiques différentes.

Mathématiquement, il est possible de caractériser les différences entre toutes ces séquences en utilisant la distance de Hamming. En deux mots, pour les ARN, cette métrique compte le nombre minimum d'opérations d'édition (substitutions, insertions, ou délétions) qu'il faut effectuer pour passer d'une séquence S_A^n de n nucléotides à une séquence S_B^m de m nucléotides. Mais au-delà de la simple constatation, il est possible de chercher numériquement, l'ensemble des séquences S^m de longueur m fixée qui se replient dans une structure donnée. On caractérise alors les relations de similitude entre les différentes séquences à l'aide de la distance de Hamming. Dans ce cas, la taille de la chaîne étant fixée, la distance de Hamming se réduit au nombre minimum de substitutions qu'il faut effectuer pour passer de la séquence S_i^m à la séquence S_j^m. Il a été démontré que l'ensemble S^m forme un ou plusieurs réseaux de séquences deux à deux proches voisines, c'est-à-dire de distance unitaire au sens des distances de Hammings, appelés *réseaux neutres*. Néanmoins, pour un réseau neutre donné, les séquences les plus distantes peuvent être extrêmement différentes. Par exemple, seuls sept nucléotides sont strictement conservés sur l'ensemble des séquences d'intron de groupe I, alors que la structure secondaire incluant le cœur du ribozyme est préservée [40].

Des analyses théoriques basées sur l'utilisation de code de prédiction de structures secondaires suggèrent que différents réseaux neutres peuvent s'entrelacer et s'approcher aussi près que possible. C'est-à-dire qu'il existe presque toujours des séquences appartenant à deux réseaux neutres distincts qui ne diffèrent que d'une seule base [67]. Cette proximité dans les séquences ne se retrouve pas forcément sur les structures auxquelles chaque réseau neutre est associé [67]. Jusqu'aux travaux de Erik Schultes et David Bartel, il n'y avait pas d'évidence expérimentale d'une possible intersection entre différent réseau neutre biologique. C'est-à-dire, la possibilité pour une séquence de se replier indifféremment dans plusieurs structures biologiquement fonctionnelles [72]. Leur résultat démontre qu'il est possible, par mutation neutre, de trouver de nouvelles fonctionnalités pour les ribozymes. De plus, cette expérience vient renforcer l'idée de la possible existence d'un monde tout ARN antérieur à la formation des protéines.

En conclusion, l'étude de l'évolution des génotypes (séquences) et de leurs phénotypes associés,a prouvé tant théoriquement qu'expérimentalement qu'une structure native est accessible

à partir d'une multitude de séquences et qu'une séquence peut de même coder pour plusieurs fonctions. En conséquence, la majeure partie des degrés de liberté de la structure primaire, à contrario d'être contraints au seul codage de la structure fonctionnelle, laisse possiblement accès au codage d'informations supplémentaires, notamment sur la dynamique de repliement.

5.1.3 Modulation de la cinétique par le design de la séquence

Si du point de vue de la thermodynamique, une structure donnée peut être atteinte par une multitude de polynucléotides, d'un point de vue de la dynamique, les cinétiques de repliement de ces différentes chaînes d'acides nucléiques présentent de grandes disparités. Dès lors que le nombre de résidus augmente, le temps caractéristique nécessaire à la molécule pour trouver sa conformation active peut varier dans un large domaine d'intervalle de temps, typiquement de quelques millisecondes à plusieurs minutes. De plus, le chemin cinétique n'est pas unique et une même séquence peut suivre plusieurs chemins, comme démontré sur différentes séquences de groupe I de *Tetrahymena* [62]. Les structures intermédiaires rencontrées au cours du repliement agissent comme des pièges cinétiques locaux qui sont à l'origine de la différence des temps de relaxation de la molécule dans les différents chemins de repliement [61]. Certaines structures particulièrement stables, sont même identifiées à la structure relaxée et interprétées comme étant le résultat d'un mauvais repliement [57, 60]. La mutation de nucléotides spécifiques s'avère donc une méthode efficace pour déstabiliser une structure intermédiaire cinétiquement piégée comme démontré sur un ensemble de séquences mutantes de groupe I de *Tetrahymena* [81]. Néanmoins, cette modification le long de la séquence des bases entraîne une modification du paysage énergétique, et peut engendrer d'autres pièges et d'autres voies de cheminement.

Au lieu de s'efforcer de déstabiliser thermodynamiquement une conformation particulièrement piégée, Tao Pan, Xiangwang Fang et Tobin Sosnick, ont démontré en permutant circulairement la séquence d'un long ARN structuré en domaines, qu'il était possible d'éviter que ses états piégés ne se forment pendant la transcription [60]. En effet, au cours de la synthèse, le repliement des différents domaines peut être gêné par des appariements inter-domaines qui forment des structures intermédiaires particulièrement stables et bloquent la nucléation *in vitro* de la structure fonctionnelle. Dans ce cas, une permutation circulaire de la séquence des bases permet

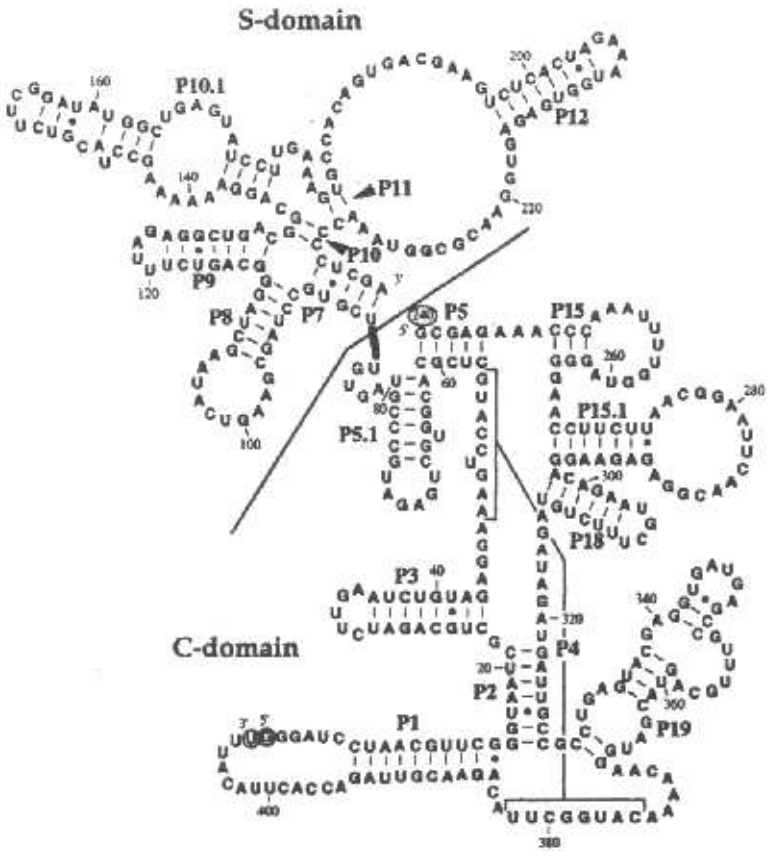


FIG. 5.1 – Structure secondaire de l'ARN de la RNAse P permuté en position 240. Les extrémités naturelles sont indiquées par les nucléotides encerclés. Schématiquement, la séquence passe du format $C_{begin} - S - C_{end}$ à $C_{all} - S$ ce qui lui permet de se replier plus facilement et donc plus rapidement lors de sa transcription.

de commencer la synthèse par l'extrémité 5' du domaine de son choix, ce qui évite la majeure partie des interactions non natives et accélère considérablement la vitesse du repliement. Sur l'exemple de l'ARN de la RNase P de *B. subtilis* (fig. 5.1), les domaines catalytiques (C) et de reconnaissance spécifique (S) peuvent se replier indépendamment l'un de l'autre et sont fonctionnels peu de temps après leur synthèse respective [41]. Alors que lors de la synthèse *in vitro* de la molécule complète naturelle, l'interaction induite entre les deux domaines conduit à la formation de structures intermédiaires fortement piégées. Par permutation circulaire, la séquence est réorganisée de telle sorte que l'ensemble du domaine C soit transcrit dans un premier temps, avant la synthèse de la séquence du domaine S (cf. figure 5.1). La chaîne synthétique permutée se replie quinze fois plus rapidement que la séquence naturelle. Dans ce contexte, le repliement co-transcriptionnel apparaît comme un moyen d'éviter certaines structures piégées indésirables. La modulation de la cinétique étant obtenue ici en conservant la séquence des bases mais en permutant l'ordre de leur apparition. L'hypothèse de Pan et collaborateurs associe donc l'information cinétique au décodage des sous-domaines dans les molécules d'ARN. Dans ce schéma, la modulation de la cinétique est principalement associée à la propriété d'appariements non-natifs entre des séquences codant pour différents domaines.

5.2 Hypothèses sur repliement co-transcriptionnel

Les évolutions par mutations neutres ont démontré au cours du temps l'aptitude d'une molécule à s'adapter à son environnement. Les mutations aléatoires agissent comme une "force" stochastique dans l'espace des séquences. A chaque mutation, la séquence est modifiée d'une base ce qui correspond à une dérive aléatoire d'une distance unité dans l'espace des séquences. En l'absence de sélection, toutes les séquences sont accessibles et les 4^n séquences devraient cœxister. Avec la pression de sélection, seules les séquences qui codent pour la bonne fonction ont une chance de survie. Les séquences finalement sélectionnées présentent une aptitude particulière (rapidité de repliement, taux d'efficacité de la fonction,...) qui leur permet de supplanter toutes les autres. Les études sur le repliement de molécules d'intron de groupe I démontrent que la séquence contient effectivement de l'information de type cinétique sur leur chemin de repliement, puisqu'il est possible de modifier la vitesse moyenne de repliement de molécules

en mutant quelques bases sélectionnées. Néanmoins, il n'est pas forcément nécessaire de faire appel aux mutations ponctuelles pour modifier la cinétique de repliement. L'expérience de permutation de la séquence de l'ARN de RNAse P, nous l'a prouvé. Selon l'hypothèse du groupe de Sosnick, une façon particulièrement efficace d'éviter de former des structures piégées est de réorganiser la séquence de telle manière que la synthèse de la molécule commence à coup sûr par l'extrémité de l'un des domaines natifs. Dès lors que chaque domaine natif est indépendant, son repliement le serait aussi. Ainsi, la séquence totale ne serait qu'une succession de modules structuraux indépendants. Au cours de la synthèse, les interactions entre domaines sont supposées réduites au minimun, puisque chaque domaine totalement synthétisé a atteint sa conformation définie stable. La molécule ainsi synthétisée évite de se piéger dans des structures intermédiaires particulièrement stables. Il est intéressant à noter que ce scénario est parfaitement indépendant du sens de synthèse. Nous pouvons mener le même raisonnement en imaginant une polymérase synthétisant la molécule en sens inverse, c'est-à-dire de l'extrémité $3'$ vers l'extrémité $5'$. La séquence lue en sens réverse se présente aussi comme une chaîne de séquences de domaines natifs et indépendants. Selon l'hypothèse du groupe de Sosnick, elle devrait donc se replier par domaines. En fin de synthèse, la structure obtenue devrait être la même que celle obtenue en synthétisant la molécule dans le sens conventionnel.

L'expérience de permutation nous permet de conclure que de l'information de type cinétique est distribuée le long de la séquence, mais ne permet pas d'expliquer la manière dont cette information est codée. Les expériences de mutations sélectives prouvent que les structures intermédiaires jouent un rôle majeur dans la dynamique. Si l'on regarde la permutation comme un ensemble de mutations ponctuelles, la permutation doit modifier la cascade d'intermédiaires qui guide le repliement de la molécule. La disposition adéquate de la séquence modifie la cinétique de repliement en modifiant la cascade d'intermédiaires. Selon cette nouvelle hypothèse, l'information cinétique ne serait donc pas codée au niveau macroscopique des domaines indépendants, mais serait plus proche de la cinétique de nucléation des hélices intermédiaires. En particulier, les transitions entre les différentes structures métastables devraient être régulées par l'ordre d'apparition des différentes hélices non natives. La stabilité intrinsèque de ces hélices joue le rôle de médiateur de l'information cinétique. Lors de la synthèse, l'ordre d'apparition des bases module la formation des hélices et donc la formation des structures à venir. De ce fait, le sens

de lecture de la structure primaire devrait influer sur le déroulement de l'histoire cinétique. En traduisant la séquence dans le sens naturel de l'extrémité 5' vers l'extrémité 3', la cascade d'hélices intermédiaires est différente de celle obtenue si une polymérase théorique avait traduit la séquence de l'extrémité 3' vers l'extrémité 5'. En d'autres termes, la stabilité des hélices *primo* formées influe sur la probabilité de formation des hélices potentielles qui naissent tout au long de la synthèse de la chaîne. En fin de synthèse, les probabilités de formation des différentes hélices diffèrent selon le sens de lecture de la séquence. Selon cette hypothèse, en fin de synthèse, les structures secondaires formées peuvent être potentiellement différentes. A temps long, les deux molécules hypothétiques, obtenues par synthèse $5' \rightarrow 3'$ et $3' \rightarrow 5'$, doivent relaxer dans des configurations d'équilibre identiques.

5.3 Stylistique des séquences test

Nous avons conçu une séquence artificielle qui peut adopter deux structures stables (figure 5.2). La première structure présente deux hélices indépendantes, la seconde deux hélices imbriquées. La séquence suit les règles énoncées par le groupe de Sosnick ; elle est décomposable en sous-séquences codantes pour des domaines indépendants qui se replient rapidement et indépendamment. Et, la séquence est organisée telle que la polymérase commence par transcrire l'un des deux sous-domaines. Pour notre séquence, les domaines sont réduits à leur plus simple expression : ils ne contiennent qu'une seule hélice.

Dans le schéma de repliement par blocs indépendants, chaque bloc a le temps de trouver sa conformation native et en fin de synthèse, nous devrions avoir la même structure quel que soit le sens de transcription. La structure obtenue est la structure "branchée" qui présente les deux hélices indépendantes. Après avoir relaxé, la molécule occupe en revanche statistiquement ses deux configurations de basse énergie, la configuration "cigare" en hélices imbriquées et la configuration "branchée".

Pour tester notre hypothèse de codage du chemin de repliement co-transcriptionnel le long d'une séquence d'ARN, nous avons codé, en plus de ces deux états d'équilibre, une cascade d'intermédiaires différents selon le sens de synthèse (figure et légende 5.2). Lorsque la molécule est synthétisée dans le sens conventionnel, de l'extrémité $5' \rightarrow 3'$, elle se replie dans la confor-

mation "branchée". Au contraire, si cette même molécule pouvait être transcrite dans le sens inverse, de l'extrémité $3' \rightarrow 5'$, la cascade des structures intermédiaires attendue ne serait pas la même et guiderait la molécule vers la configuration en hélices imbriquées. Cette différence de chemin de repliement repose sur les stabilités relatives entre les hélices en compétition au cours de la transcription. Les quatre hélices natives (P_{1A}, P_{1B} et P_{2A}, P_{2B}) qui forment les deux structures stables, sont deux à deux incompatibles (cf figure 5.2).
Au cours de la transcription en sens conventionnel (respectivement en sens inverse), l'hélice P_{1A} (respectivement P_{1B}) se forme dans un premier temps. Puis entre en compétition avec l'hélice P_{2A} (séquence en rouge sur la figure) dès que cette dernière peut essayer de nucléer (flèches vertes). L'hélice P_{1A} étant plus stable que l'hélice P_{2A}, celle-ci résiste, et guide le repliement vers la structure #1 pour une transcription naturelle. A l'inverse pour une hypothèse $3' \rightarrow 5'$, l'hélice P_{1B} étant moins stable que l'hélice P_{2A}, elle lui cède sa place, et la molécule est conduite alors dans la structure #2 en fin de synthèse. A temps long, la molécule relaxe comme il se doit dans les deux configurations stables quel que soit le sens de synthèse.

Ces scénarios hypothétiques de transcription d'une même séquence dans ses deux directions ne sont malheureusement pas réalisables à l'aide d'enzymes biologiques. Pour s'en rapprocher, il faut imposer des contraintes de symétrie sur les hélices des molécules d'ARN que l'on souhaite étudier comme nous allons le voir ci-dessous.

Du point de vue de la séquence des bases, transcrire la séquence directe $5' - ABCD - 3'$ dans le sens réverse ressemble à une transcription de la séquence réverse $5' - DCBA - 3'$ dans le sens conventionnel. En inversant la polarité de la séquence, nous modifions les énergies de stacking de la structure. En effet l'énergie d'empilement des paires de bases est dépendant de l'orientation, et suit en général les relations suivantes pour des paires de bases différentes (ici G/C et A/U) :

$$\begin{matrix} 5'-\overrightarrow{G\,A}-3' \\ 3'-\overleftarrow{C\,U}-5' \end{matrix} \equiv \begin{matrix} 3'-\overleftarrow{A\,G}-5' \\ 5'-\overrightarrow{U\,C}-3' \end{matrix} \neq \begin{matrix} 5'-\overrightarrow{G\,A}-3' \\ 3'-\overleftarrow{C\,U}-5' \end{matrix}$$

Par conséquent, de manière générale, les hélices obtenues en renversant l'orientation de la séquence, sont d'énergies différentes.

$$\begin{matrix} 5'-\overrightarrow{G\,A\,C\,U}-3' \\ 3'-\overleftarrow{C\,U\,G\,A}-5' \end{matrix} \neq \begin{matrix} 5'-\overrightarrow{U\,C\,A\,G}-3' \\ 3'-\overleftarrow{A\,G\,U\,C}-5' \end{matrix}$$

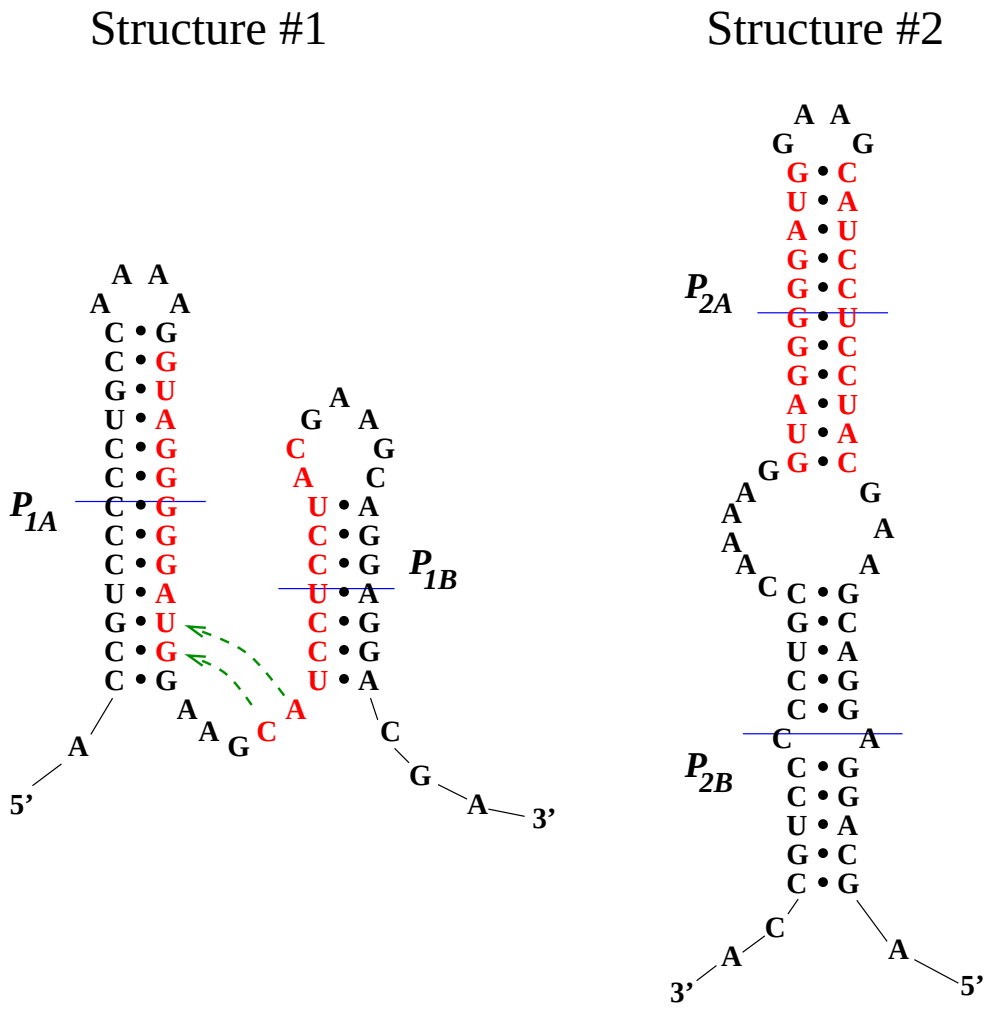


FIG. 5.2 – Exemple d'une molécule bistable. Deux conformations stables obtenues avec une séquence théorique. Chaque structure d'équilibre s'appuie sur deux hélices (P_{1A}/P_{1B} et P_{2A}/P_{2B}) palindromique et incompatibles. Lors d'une transcription inverse ($3' \rightarrow 5'$), l'hélice P_{1b} entre en compétition avec l'hélice P_{2A} (séquence en rouge). Les flèches vertes indiquent un processus de nucléation de l'hélice P_{2A} à partir des nucléotides libres. La formation de l'hélice P_{2A} fait basculer la molécule de sa conformation stable #1 vers sa conformation stable #2.

Les molécules directes et réverses ne représentent donc pas en général un couple de molécules modèles pour la simulation de transcription dans le sens conventionnel et inverse.

Pour résoudre ce problème, nous avons conçu des molécules bistables dont toutes les hélices des deux structures sont palindromiques. Par symétrie, tout empilement de bases communes $\begin{smallmatrix}5'-\overrightarrow{G\ A}-3'\\3'-\overleftarrow{C\ U}-3'\end{smallmatrix}$ est associé par son empilement symétrique $\begin{smallmatrix}5'-\overrightarrow{A\ G}-3'\\3'-\overleftarrow{U\ C}-3'\end{smallmatrix}$ au sein de la même hélice. Le schéma général de chaque hélice est ainsi donné par :

$$\begin{smallmatrix}5'-\overrightarrow{G\ A\ C\ U\cdots U\ C\ A\ G}-3'\\3'-\overleftarrow{C\ U\ G\ A\cdots A\ G\ U\ C}-5'\end{smallmatrix} \equiv \begin{smallmatrix}3'-\overleftarrow{G\ A\ C\ U\cdots U\ C\ A\ G}-5'\\5'-\overrightarrow{C\ U\ G\ A\cdots A\ G\ U\ C}-3'\end{smallmatrix} \equiv \begin{smallmatrix}5'-\overrightarrow{G\ A\ C\ U\cdots U\ C\ A\ G}-3'\\3'-\overleftarrow{C\ U\ G\ A\cdots A\ G\ U\ C}-5'\end{smallmatrix}$$

La symétrie du palindrome assure que l'énergie libre des hélices est indépendante du sens de transcription. En imposant que toutes les hélices qui composent les structures de plus basses énergies soient palindromiques, nous nous assurons que les énergies des structures stables obtenues en transcrivant les séquences directes et réverses soient équivalentes deux à deux.

En conclusion, pour chaque molécule (1 ou 2), la bistabilité est obtenue en ajustant les énergies des structures (#1 et #2) grâce au choix de leurs appariements. De plus pour s'assurer que les énergies des structures deux à deux similaires soient les mêmes lorsqu'elles sont obtenues en synthétisant la séquence directe d'une part et la séquence réverse d'autre part, nous forçons toutes les hélices de ces structures à être palindromiques. Les molécules sont donc bistables et leurs structures sont d'énergie équivalente : les quatre structures obtenues sont ainsi d'énergie similaire (cf. figure 5.3).

Le codage du chemin cinétique suit les règles explicitées ci-dessus. Les hélices incompatibles sont synthétisées dans un ordre différent. Lors de la synthèse de la molécule directe, l'hélice la plus stable résiste et guide la molécule vers la structure #1. Dans la synthèse de la molécule réverse, la première hélice formée est plus faible que la grande hélice de la structure #2 et lui cède sa place, en fin de synthèse la molécule se replie dans la conformation #2 (cf. figure 5.4).

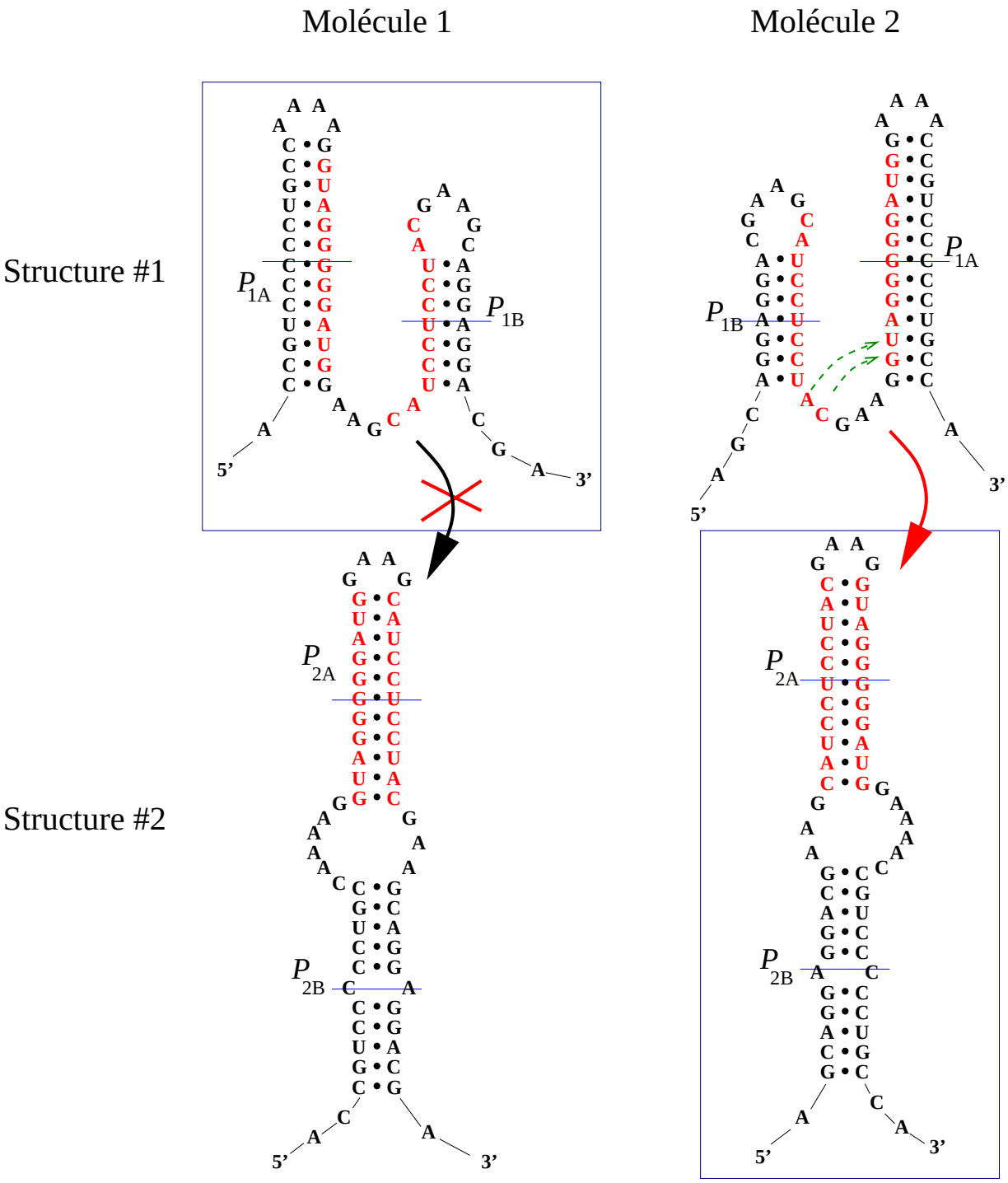
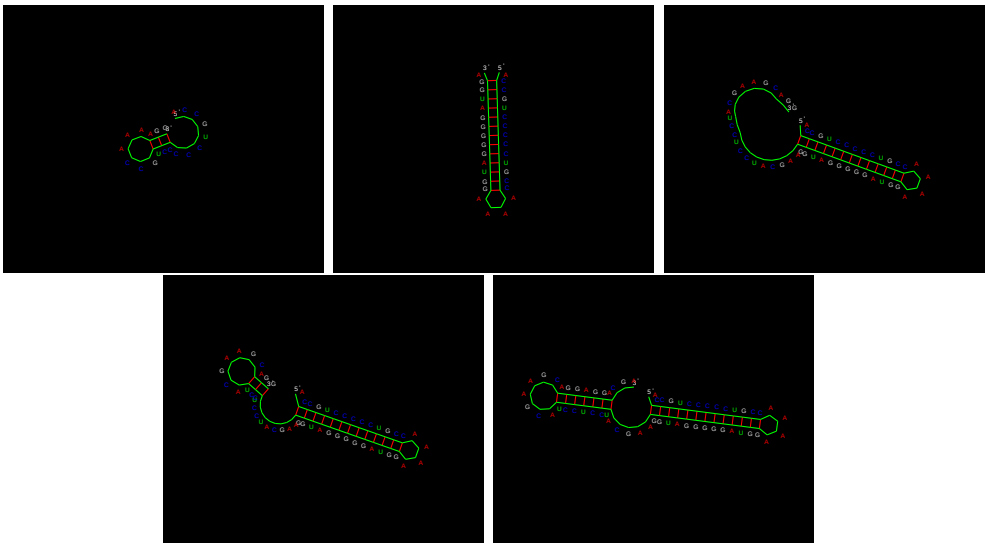


FIG. 5.3 – Les structures #1 et #2 de la molécule 1 et de la molécule 2 sont d'énergie similaire puisqu'elles sont composées des mêmes éléments. De plus, la structure #1 (resp. #2) de la molécule 1 et de la molécule 2 sont d'énergie identique, puisque leurs hélices sont palindromiques. Donc les quatre structures représentées : structures #1 et #2 des molécules 1 et 2, sont toutes d'énergie similaire. De plus lors de la synthèse de la molécule 1 la grande hélice *primo* formée résiste alors que dans le cas de la molécule 2, la petite hélice est dézippée suivant les flèches vertes pour former la grande hélice de la structure #2. En fin de synthèse, on s'attend à trouver les molécules 1 et 2 dans les configurations encadrées respectivement #1 et #2

Molécule directe



Molécule réverse

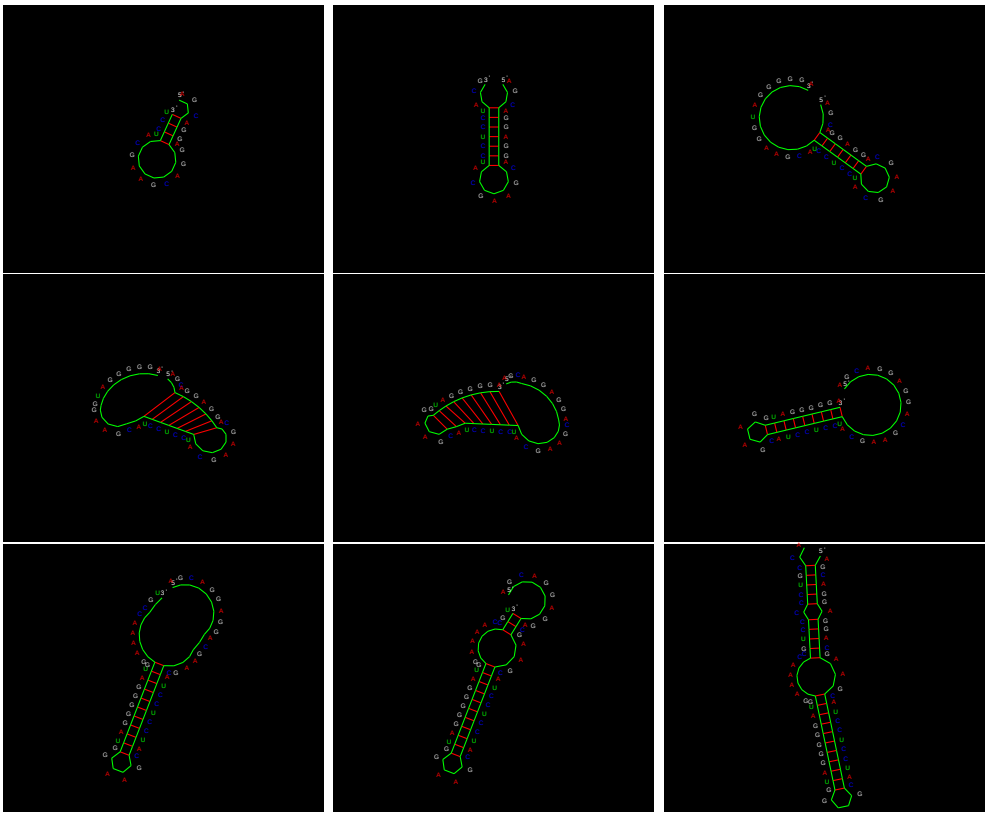


FIG. 5.4 – Instantané de la cinétique de co-repliement en cours de synthèse obtenue par simulation numérique calculée par ***KineFold*** et visualisé par **RNAMovies**. Pour la molécule directe, l'hélice P_{1A} résiste alors que, pour la molécule réverse, l'hélice P_{1B} cède sa place à l'hélice P_{2A}. Les structures intermédaires sont extrapolées par **RNAMovies**

Chapitre 6

Mise en œuvre expérimentale

6.1 Introduction

Le but de l'expérience est d'étudier la possibilité de coder de l'information sur la cinétique de repliement le long d'une séquence d'ARN. Nous avons défini un ensemble de règles théoriques qui permettent de générer des séquences ARN qui présentent les propriétés nécessaires pour tester nos idées. Le protocole expérimental translate l'idée papier en une vérification *in fine*. Il comporte donc deux étapes majeures. Dans un premier temps, la synthèse des séquences directe et réverse sont obtenues à partir d'un substrat ADN, pour rendre compte de la cinétique de repliement en cours de transcription. Puis la séparation des structures est obtenue sur gels d'électrophorèse. A temps court, une seule structure est attendue, alors qu'à l'équilibre thermodynamique les deux configurations devraient être présentes. La condition *sine qua none* est que le temps de relaxation soit beaucoup plus grand que le temps total de transcription (à 37°C) et de migration sur gel (à 10°C).

La séquence d'ADN qui code pour l'ARN désiré est synthétisée par une société spécialisée (IBA-NABS *Göttingen*, Allemagne). Nous amplifions ensuite le brin modèle par la méthode de clonage en bactéries pour obtenir un grand nombre de copies de la séquence initiale. Les molécules d'ARN sont ensuite transcrites à partir de ce substrat ADN En fin de transcription,

les structures obtenues sont séparées sur gel d'électrophorèse. Sous l'application d'un champ électrique, les molécules chargées négativement migrent en direction du potentiel positif. En présence d'obstacles aléatoires que constituent les fibres du gel, les molécules vont migrer à des vitesses différentes selon leur taille et leur structure. Dans notre cas, tous les ARN sont de tailles identiques et nous espérons pouvoir discriminer leur structure.

6.2 De l'ADN à l'ARN

La molécule d'ARN est obtenue par transcription d'un brin matrice en ADN linéaire par une enzyme l'ARN polymérase. Nous utilisons la T7 ARN polymérase issue du bactériophage du même nom pour transcrire nos substrats ADN en ARN. Les ARN polymérases ADN dépendante de bactériophage sont des protéines composées d'une unique sous-unité et sont extrêmement sensibles à leur séquence promotrice. Elles catalysent la réaction d'addition des nucléotides à partir d'un substrat ADN simple ou double brin, mais requièrent que la séquence promotrice soit dans tous les cas double brin. La synthèse se fait dans la direction $5' \rightarrow 3'$ du brin ADN "codant" et forme la séquence codante ARN par complétion du brin négatif de l'ADN. Les enzymes catalysent l'addition de nucléotides triphosphate à la suite de la chaîne et relâchent un groupement diphosphate en solution. La T7 ARN polymérase nécessite en outre la présence d'ions magnésium à une concentration variant de $60mMol/ml$ à $100mMol/ml$ pour être active. Habituellement, les sociétés commercialisant cette enzyme, fournissent le tampon salin adéquat pour garantir l'efficacité optimale de leur polymérase.

La synthèse débute à la première base suivant la séquence de démarrage spécifique à la polymérase, la boîte TATA pour la T7 ARN polymérase. Il est néanmoins recommandé d'insérer une courte séquence leader pour augmenter sensiblement le rendement de la transcription [1]. Pour obtenir des transcrits de taille déterminée, il existe deux méthodes. La première consiste à insérer un signal de fin de transcription spécifique en aval de la séquence à transcrire. Malheureusement, in vitro, la polymérase poursuit souvent la transcription et produit statistiquement des transcrits plus longs. La seconde, plus communément utilisée, consiste à transcrire la matrice ADN coupée par une enzyme, en aval de la séquence à transcrire. Lors de la transcription, la polymérase "tombe" en fin de séquence. Cette méthode est connue sous le nom de transcrip-

tion *run off* [49]. Il faut néanmoins prendre quelques précautions. En particulier se rappeler que la polymérase transcrit le brin complémentaire au brin codant. Ainsi, si la séquence laisse une extrémité 3' pendante, elle ne produira pas un ARN de séquence complète. L'idéal est d'utiliser une enzyme de restriction qui laisse une extrémité libre double brin (coupure franche). Néanmoins, il a été montré [85] que la T7 ARN polymérase ajoutait de façon non spécifique un résidu qui pouvait être un C dans 21% des cas, un A dans 18% des cas et un U dans 5%, à la fin de la chaîne ARN produite, en sus des bases de la séquence ADN transcrite.

Pour disposer d'autant de masse que nécessaire pour toutes nos expériences, nous allons dans un premier temps démultiplier le nombre de brins codants, par clonage.

6.2.1 Amplification d'un brin ADN

Le clonage de gène ou technologie d'ADN recombinant rassemble l'ensemble des techniques qui permettent de joindre entre eux deux segments ou plus d'ADN pour générer une unique molécule d'ADN capable de se répliquer de manière autonome lorsqu'elle est hébergée par un organisme hôte [15]. La séquence ADN générée, appelée vecteur, est constituée de trois éléments de base au moins :

- une origine de réplication qui permet de démultiplier les copies du vecteur dans la cellule hôte ;
- un gène de sélection codant par exemple pour une protéine de résistance à un antibiotique ;
- une région comportant un grand nombre de sites de restrictions uniques appelé *sites de polyclonage* qui permet d'ouvrir le vecteur à l'aide d'enzymes de restriction.

Il existe un grand nombre de vecteurs génériques qui comportent ces trois éléments. Pour insérer notre séquence à amplifier dans un vecteur générique, il suffit de digérer le vecteur à l'aide d'enzymes de restriction qui laissent des extrémités cohésives complémentaires avec celles de la séquence à insérer. On catalyse alors les jonctions (*ligation*) entre les brins hybrides à l'aide d'enzymes dites de ligase. Nous obtenons au final un nouveau vecteur basé sur le vecteur générique, mais qui comporte aussi notre séquence à cloner. Pour des inserts de petite taille, les vecteurs les plus utilisés sont les plasmides : des ADN circulaires de quelques milliers de bases qui utilisent la machinerie de réplication des bactéries.

Dans le principe, un clonage est très simple à réaliser. Le plasmide est d'abord coupé avec des enzymes de restriction puis recircularisé *in vitro*, par ligation, avec l'insert à amplifier. Le plasmide hybride est ensuite utilisé pour transformer des bactéries hôtes. Néanmoins en pratique, la difficulté majeure consiste à identifier les plasmides qui ont recircularisé avec l'insert de ceux qui se sont recircularisés sur eux-mêmes. Les quatre étapes majeures pour le clonage de l'ADN sont : la préparation du vecteur et de l'insert, la ligation des ADN hétérogènes, la transfection en bactéries, et enfin la discrimination des clones.

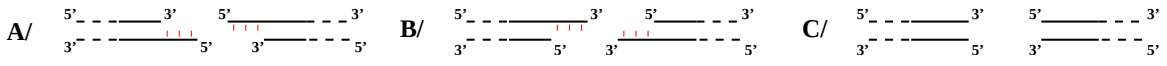


FIG. 6.1 – Différents types de digestion. **A/** digestion laissant des bases pendantes sur l'extrémité 5′. **B/** digestion laissant des bases pendantes sur l'extrémité 3′. **C/** digestion franche.

Les enzymes de restrictions reconnaissent de manière spécifique une séquence d'ADN double brin et coupent la double hélice sur un site déterminé adjacent ou contenu dans la séquence de reconnaissance. La digestion peut laisser des bases pendantes ou couper de façon franche la double hélice (cf. figure 6.1). Les sites de reconnaissance ont la propriété d'être auto-complémentaires. Lorsque l'enzyme digère le double brin sur le site de reconnaissance, elle laisse un brin qui peut se recombiner avec lui-même ou avec un brin voisin. A des fins de clonage, il est préférable de digérer le vecteur à l'aide de deux enzymes différentes qui ont pour cible des sites appartenant au site de polyclonage. En principe, le plasmide ne peut pas se recircularisers sans interagir avec un segment voisin. L'interaction peut avoir lieu avec l'insert à cloner ou avec son propre segment digéré par les enzymes.

La ligation entre l'insert et le plasmide implique la formation de nouvelles liaisons entre les phosphates résiduels localisés sur le brin 5′ et les parties hydroxyles adjacentes situées en 3′. Il peut se former au plus quatre ponts phosphodiesters, ou deux si les extrémités du vecteur sont préalablement déphosphorylées. Si les bases sont déphosphorylées, le plasmide recombiné présente sur chaque brin une entaille qui est réparée par la bactérie. La jonction est catalysée par l'action d'enzymes de ligation. Deux types de ligases sont souvent utilisés la ligase ADN issue d'*Escherichia Coli* et la ligase issue du bactériophage λ. La seconde est la plus couramment utilisée et présente la propriété de pouvoir lier des brins qui sont coupés de manière franche.

La ligation est une réaction bi-moléculaire dont la vitesse ne dépend que de la concentration en extrémités compatibles. En solution, l'ADN se comporte comme un polymère gaussien de taille de monomère équivalent b. La probabilité par unité de volume de trouver une extrémité à la surface varie comme $P = (\frac{3}{2\pi lb})^{\frac{3}{2}}$ où l et la longueur curviligne de l'ADN Quant à la concentration en bouts, elle est donnée par $E = 2\,N_0\,M$ avec N_0 le nombre d'Avogadro et M la concentration molaire de l'ADN Théoriquement lorsque $P = E$, le vecteur a autant de chance de réagir avec lui-même que d'ingérer la séquence à cloner. Si $P > E$, le plasmide réagit préférentiellement avec lui-même, et si $P < E$ la concatenation est favorisée. Si le plasmide est digéré par deux enzymes différentes, ses extrémités ne sont pas compatibles, seule la concaténation peut avoir lieu.

Il existe plusieurs techniques de transfert de plasmide en bactéries. La plus simple et la plus utilisée est la technique dite de *calcium compétence*. Les techniques de calcium compétence viennent des observations faites par Mandel et Higa [44] sur l'aptitude d'une bactérie à ingérer l'ADN bactériophage λ. Une culture d'*Escherichia Coli* traitée avec une solution de chlorure de calcium à la température de la glace puis brièvement chauffée (à $42°C$ pendant 90 secondes ou $37°C$ quelques instants selon les protocoles) a la propriété de pouvoir intégrer des brins ADN étrangers. La même méthode a été utilisée avec succès pour transfecter des plasmides [11].
Les plasmides sont mélangés avec les bactéries *compétentes* et adhèrent à leur surface. Le choc thermique permet à l'ADN de pénétrer dans la cellule. Le vecteur absorbé utilise alors la machinerie de réplication de la bactérie pour se répliquer d'une part et est transmis comme patrimoine génétique à sa déscendance d'autre part.

Les bactéries transformées sont cultivées en présence d'un agent de sélection, habituellement un antibiotique. Cette pression de sélection permet de s'assurer que le plasmide est bien transmis de génération en génération. La réplication du plasmide est un processus coûteux en temps et en énergie. En absence de pression de sélection, les bactéries n'ont plus besoin de garder le caractère de résistance et ne répliquent plus le plasmide, et par voie de conséquence aussi l'insert à amplifier.

Cette technique de clonage permet de récupérer quelques milligrammes de plasmide qui descendent tous d'un unique représentant de l'ADN hybride insert-plasmide.

6.2.2 Construction et amplification de notre brin matrice en ADN

Séquences ADN

La molécule d'ARN est obtenue par transcription d'un brin matrice ADN synthétisé par une société spécialisée puis amplifié en bactérie. La séquence modèle contient donc les différents éléments qui permettent l'initiation et la terminaison de la transcription, ainsi que l'amplification du brin modèle.

Nous avons choisi de transcrire notre séquence à l'aide de l'ARN polymérase T7. A notre séquence initiale qui code pour une molécule d'ARN, nous devons ajouter en amont le site promoteur et un triplet de bases G pour permettre une transcription efficace du brin codant. Nous devons ajouter à chaque extrémité de ce nouveau brin, des bras complémentaires aux sites de restriction du vecteur. Ces courtes séquences simple brin permettent la reconnaissance et la ligation entre les deux espèces. Après amplification, les copies de l'ADN hybride plasmide-insert sont circulaires. Pour être utilisé lors de la transcription en ARN, le complexe doit être linéarisé (mode de transcription *run off*). Nous devons prévoir un site de restriction franche supplémentaire situé au plus près de la fin de la séquence matrice, voir à cheval sur la séquence elle-même. Nous avons utilisé deux enzymes particulières pour effectuer ce travail : l'enzyme *Sma I* dans un premier temps et l'enzyme *Stu I* pour une seconde série de séquence. La figure 6.2 donne le schéma général de la séquence commandée. L'enzyme Sma I reconnaît le site $\begin{smallmatrix} 5'-\overrightarrow{CCC\ \ GGG}-3' \\ 3'-\overleftarrow{GGG\ \ CCC}-5' \end{smallmatrix}$, et coupe après le triplet de bases C. L'enzyme Stu I reconnaît le site $\begin{smallmatrix} 5'-\overrightarrow{AGG\ \ CCT}-3' \\ 3'-\overleftarrow{TCC\ \ GGA}-5' \end{smallmatrix}$ et coupe après le doublet de bases G. La figure 6.3 compare les modifications apportées par les contraintes pratiques sur la molécule théorique.

Choix du vecteur

Les plasmides de la famille pUC offrent un réel intérêt puisqu'ils se répliquent entre 500 et 700 fois au sein d'une même bactérie [71]. Ils apportent à la bactérie un gène de résistance à l'ampicilline et portent un segment d'ADN dérivé de l'opéron *lac* d'*Escherichia Coli* qui code pour le fragment amino-terminal du β-galactosidase (cf. figure 6.4). La synthèse de ce fragment (α) est induite par l'isopropylthio-β-D-galactoside (IPTG). En présence du fragment ω synthétisé par la molécule hôte, les deux parties se complémentent, on parle d'α-complémentation, et forme

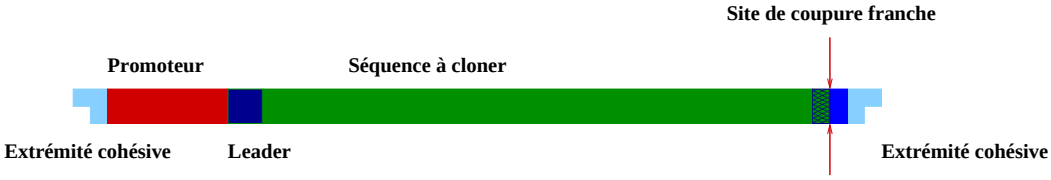


FIG. 6.2 – Séquence ADN modèle

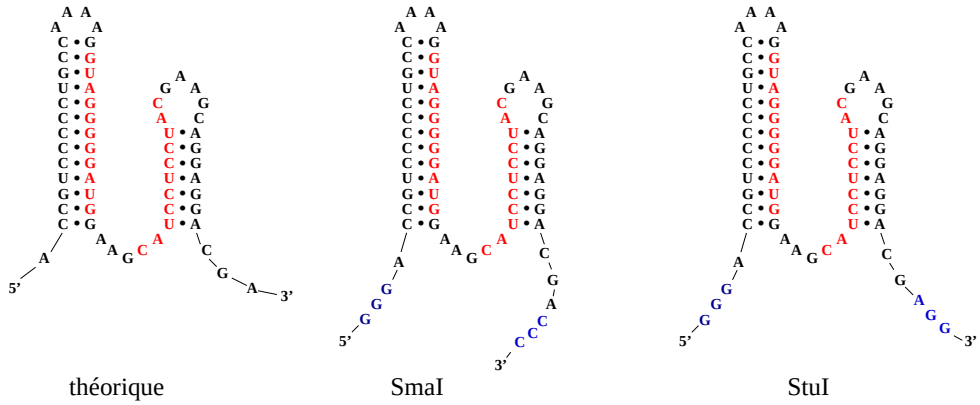


FIG. 6.3 – Comparaison entres les différentes structures théoriques et obtenues par transcription *in vitro*. Au centre séquence terminée par le site de reconnaissance SmaI. A droite séquence terminée par le site de reconnaissance StuI

la protéine active de β-galactosidase. Les bactéries exposées à l'IPTG prennent une teinte bleue lorsquelles dégradent le 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal). L'insertion de fragment ADN étranger dans le site de polyclonage inactive le fragment de β-galactosidase et abolit l'α-complémentation. Les bactéries qui ont ingéré le plasmide recombiné restent alors blanches. Ce test, aussi appelé test de *blanc-bleu*, s'effectue sur des bactéries étalées sur plaques de sélection qui contiennent l'antibiotique approprié et qui sont traitées en surface à l'IPTG et au X-gal. Il permet de discriminer rapidement entre les bactéries qui ont ingéré un plasmide recombiné de celles qui ont ingéré le vecteur initial. Quant aux bactéries qui n'ont pas ingéré de plasmide, elles meurent au contact de l'antibiotique.

Préparation de l'ADN et du vecteur

Les séquences ADN commandées sont des oligos simples brins. Dans un premier temps, nous devons hybrider les deux parties complémentaires de chaque séquence. Les deux brins sont mis en présence et chauffés à 90°C pendant cinq minutes puis refroidis à température ambiante. Sous l'effet de la chaleur, les simples brins ADN se dégrafent et perdent leur structure secondaire. Lors du refroidissement, les brins complémentaires s'hybrident, et nous obtenons les oligos double-brins.

Parallèlement à cette opération, le plasmide est digéré par les enzymes *kpn I* et *BamH I*. Les enzymes sont commandées à la société *New England Biolabs* et sont fournies avec la solution tampon et la protéine BSA purifiée.

Protocole de digestion avec les enzymes *Kpn I* ou *BamH I* :

dans un volume total de 50 μl. 1 μg de pUC19. 5 μl de tampon (concentration initiale 10x). 1 μl d'enzyme. 0.5 μl de BSA (concentration initiale 100x). complémenté avec de l'eau ultra pure. mis dans un bain thermostaté à 37°c pendant une nuit.

Entre chaque digestion, le plasmide est purifié, soit par la méthode de séparation phénol-chlorophorme, soit par la méthode de purification sur colonne. La purification phénol-chlorophorme est basée sur la sépararation des protéines et des acides nucléiques par démixtion des phases aqueuses et organiques. En présence de phénol, les protéines sont dénaturées et restent en suspension dans la phase organique alors que les acides nucléiques sont retenus dans la phase aqueuse. L'ajout de chloroforme favorise la démixtion et augmente l'effet dénaturant. Les acides

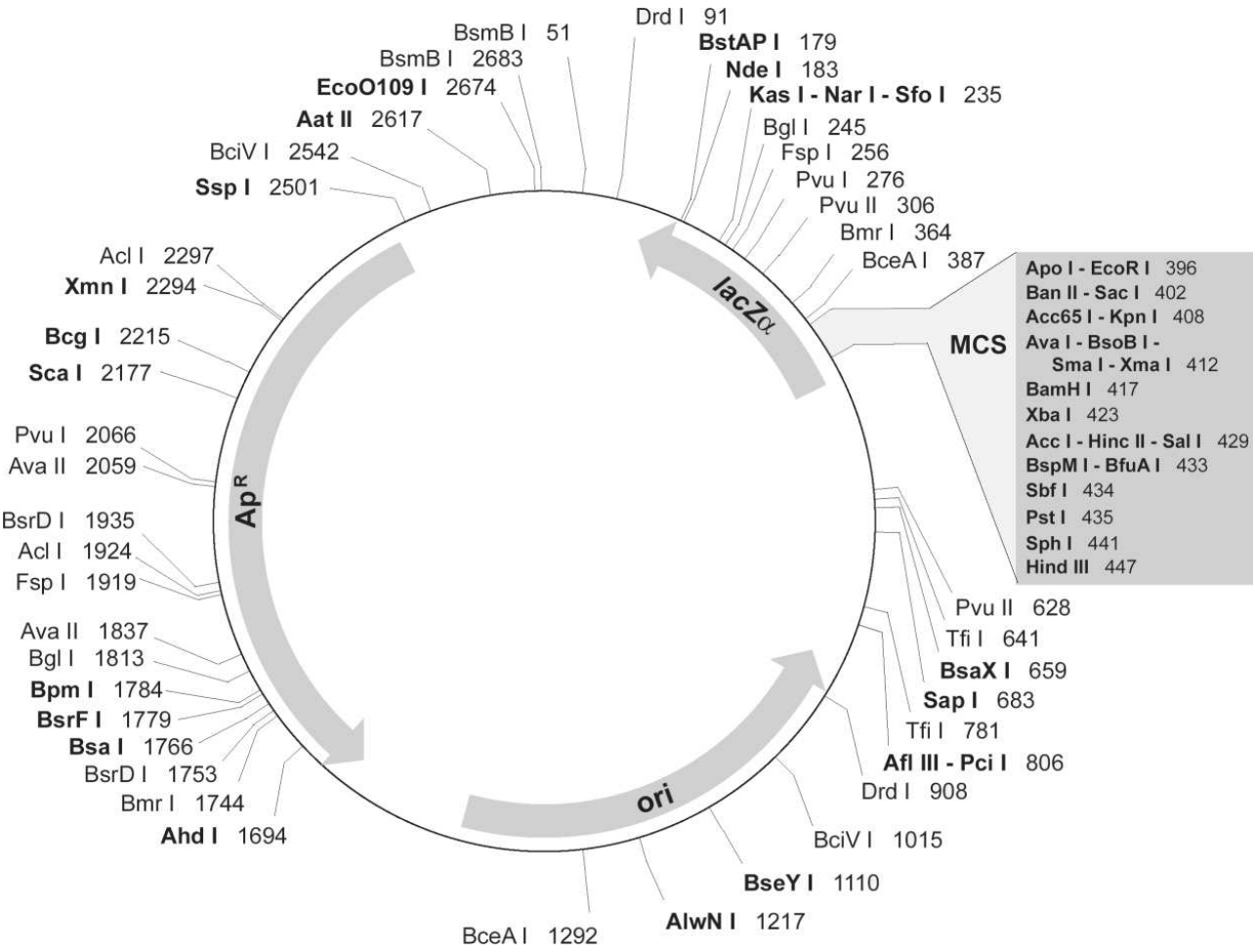


FIG. 6.4 – Cartographie du plasmide pUC19. *ori* : origine de réplication. *lacZ* α : gène codant pour le monomère α du β-D-galactoside. Ap^R : gène codant pour la résistance à l'ampiciline. *MCS* : site de multiclonage. Lors de l'insertion de la séquence à amplifier, le gène du *lacZ* α est fragmenté, ce qui neutralise la protéine produite

nucléiques sont ensuite condensés en présence de sels dans une solution d'éthanol. L'éthanol se comporte comme un mauvais solvant pour les acides nucléiques. A basse température, l'effet est démultiplié, et les polynucléotides ont tendance à s'agglomérer pour réduire au maximum leurs interactions avec le solvant.

Purification phénol-chlorophorme

Pour un volume de solution initiale à purifier. Ajouter un volume de phénol. Mélanger vigoureusement. Centrifuger une minute à 5000 tr/min. Retirer la phase organique. Mélanger et centrifuge à nouveau pour retirer le maximum de phénol. Ajouter un volume de chloroforme. Mélanger vigoureusement (*la solution démixe immédiatement en emportant les traces résiduelles de phénol.*). Centrifuger une minute à 5000 tr/min. Mélanger et centrifuger à nouveau pour retirer le maximum de chloroforme. Ajouter un large volume d'éthanol à 70% et 300 mM de NaOAC. Centrifuger au moins trois-quarts d'heure à 4°C à 15 000 tr/min. *En fin de centrifugation, l'ADN est concentré en une petite pelote blanche.* Aspirer le maximum de volume et faire sécher le culôt résultante. le culôt peut ensuite être reprise dans une solution d'eau ultra pure ou dans une solution tampon de pH basique.

La purification sur colonne est beaucoup plus rapide (environ cinq minutes). Le protocole dépend du kit utilisé. Dans notre cas, nous avons utilisé les colonnes de purification de Qiagen. Les filtres se chargent selon l'acidité du milieu. Dans un premier temps, ils retiennent les molécules d'ADN chargées négativement en surface. Puis la colonne est lavée pour retirer les protéines. L'ADN est enfin élué par une solution tampon de pH basique qui peut servir de solution de stockage.

Ligation

Pour la ligation intermoléculaire, une bonne fourchette de concentration en extrémités cohésives est comprise entre 7ng.base^{-1} et 21ng.base^{-1} selon le protocole proposé dans *Molecular Cloning* [71]. En réalité, à ces concentrations, notre plasmide a tendance à réagir avec d'autres brins linéaires de pUC19 et forme des multimères. Nous utilisons des concentrations variant de 1ng.μl^{-1} à 2ng.μl^{-1} de plasmide et une concentration relative en oligos variant de deux à dix oligos par plasmide (1ng.μl^{-1} de plasmide correspond à 3.5×10^9 plasmides). La solution stock d'oligos double brin est concentrée à 0.1nMol.μl^{-1} soit environ 6×10^{13} oligos.μl^{-1}.

Elle est donc diluée 10 000 fois pour pouvoir être utilisable. Cela donne le protocole de ligation suivant :

Protocole de ligation

volume total 10μl. 10 ng de plasmide pUC linéarisé. 1 μl de tampon de ligase (stock concentré 10x). 1 μl de ligase. 2.9 μl de solution d'oligos diluée (rapport 5 oligos pour 1 plasmide). complémenté avec de l'eau ultra pure. L'aliquote est placée dans un bain thermostaté à 16°C pendant deux heures.

Transfection

Pour transfecter les plasmides recircularisés, il faut au préalable rendre les bactéries *compétentes*. Les plasmides recircularisés sont ensuite mis en présence des cellules puis transfectés.

Protocole de calcium compétence

Inoculer un volume de 100 ml de LB [71] avec une souche de bactérie. Laisser pousser jusqu'à atteindre une colonisation d'environ 10^6 cellules par millilitre. Pour les souches *Escherichia Coli* $DH_5\alpha$ et GM2163, ceci correspond à une fourchette de densité optique à 600 nm de 0.4 à 0.6. Centrifuger les cellules et retirer le maximum de milieu nourricier LB. Resuspendre le culôt dans 10 ml de $CaCl_2$ à 0.1 M à la température de la glace. Centrifuger les cellules et retirer le maximum de liquide. Resuspendre le culôt dans 2 ml de $CaCl_2$ à 0.1 M à la température de la glace. Stocker la solution au réfrigérateur durant 24 à 48 heures avant de l'utiliser. Cette attente permet aux bactéries d'êtres plus réceptrices. Au-delà, elles perdent petit à petit leur propriété de *"compétence"*.

Transformation

L'ADN recircularisé est mis en présence de bactéries *compétentes*. 10μl de solution de plasmide pour 200μl de solution de cellule. Mixer la solution. Stocker sur glace pendant une demi-heure. Transférer les tubes à 42°C pendant 90 secondes. Transférer rapidement les aliquotes dans la glace. Laisser refroidir. Ajouter 800μl de milieu nouricier SOC [71] par tubes. Incuber les bactéries transfectées durant 45 minutes pour laisser le temps aux cellules d'exprimer le gène de résistance. Etaler 200μl de solution sur plaque de culture contenant l'antibiotique et traitée en surface au X-gal (40μl à 20 mg.μl^{-1}) et à l'IPTG (4μl à 200 mg.μl^{-1}). Mettre à 37°C jusqu'à apparition de colonies. (12 à 16 heures)

Sélection

Les bactéries qui ont ingéré le plasmide recircularisé avec l'insert restent blanches, alors que celles qui ont ingéré un plasmide recircularisé sur lui-même prennent une teinte bleue. Quant à celles qui n'ont pas ingéré de plasmide ou qui n'ont pas reçu de copie lors de la division cellulaire, elles meurent au contact de l'antibiotique. Pour sélectionner les colonies blanches, il est recommandé de stocker les plaques quelques instants au réfrigérateur à 4°C pour saturer la couleur des colonies bleutés.

Amplification et extraction

Une colonie blanche est utilisée pour inoculer une fiole de 100 ml de milieu nourricier LB. Lorsque le milieu est saturé, les plasmides sont extraits, soit selon les protocoles classiques soit à l'aide de kits spécialisés. Nous avons utilisé les deux méthodes. Le principe consiste à digérer les membranes cellulaires puis à précipiter les lipides dans une solution de surfactant. Le génome et les gros débris bactériens sont sédimentés par centrifugation. Dans le kit d'extraction comme le kit Genomed, le surnageant est ensuite filtré sur colonne. La solution finale contenant les plasmides est traitée à l'alcool pour concentrer l'ADN en pelote. Une fois séchée, le culôt est repris dans une solution d'eau ultra pure ou dans une solution tampon de pH basique.

Cette méthode de clonage amplifie un représentant unique de la séquence plusieurs milliards de fois.

6.2.3 Transcription en ARN

La séquence clonée se retrouve extraite et insérée dans un plasmide circulaire. Pour transcrire le brin matrice ADN en ARN il faut au préalable linéariser le plasmide à l'aide de l'enzyme de restriction choisi, ici SmaI ou StuI, qui reconnaît le site de coupure franche inséré le long de la séquence clonée. Après digestion, la solution est purifiée par la méthode de purification phénol-chlorophorme ou sur colonne.

La solution de plasmide linéaire est mise en présence de polymérase ARN T7 pour obtenir les molécules d'ARN. Nous avons utilisé les polymérases ARN T7 de la société *New England Biolabs*. Le protocole de transcription ainsi que la solution tampon est fourni avec l'enzyme. La

transcription se fait à 37°C et doit être en principe la plus brève possible pour que les molécules obtenues n'aient pas le temps de changer de configuration. En fin de synthèse, la solution est immédiatement stockée sur bloc de froid et mise au congélateur à -20°C.

6.2.4 Vérification des séquences clonées par séquençage direct

A chaque clonage, la séquence obtenue est séquencée par séquençage direct. Un grand nombre de séquences sont clonées par PCR en présence de nucléotide modifié. L'insertion de ces nucléotides a pour effet d'arrêter le processus de copie de la PCR. Les séquences tronquées sont ensuite séparées par gels d'électrophorèse capillaire. En sortie du capillaire, le nucléotide modifié est reconnu. La somme des signaux obtenus pour les quatre nucléotides intégré sur le grand nombre de séquences synthétisées par PCR, nous donne un diagramme de présence de chaque nucléotides pour chaque position de la séquence. Selon la proportion relative de chaque nucléotides pour chaque position de la séquence, il est possible de déduire la séquence dominante dans la solution.

6.3 Discrimination des structures, gels d'électrophorèse

L'électrophorèse désigne de manière générale la migration de particules sous l'effet d'un champ électrique. L'ADN et l'ARN étant des molécules chargées, l'électrophorèse sur gel s'est avérée une technique de choix pour séparer les fragments de polynucléotides d'après leur longueur.

La mobilité électrophorétique μ est souvent définie comme la balance des forces électrostatiques et visqueuses :

$$\mu = \frac{q}{\epsilon}$$

où q représente la charge effective de la chaîne qui dépend de la salinité et du type d'ions et de contre ions présents en solution, et ϵ est le cœfficient de friction obtenue en tenant compte des mouvements des ions de la solution.

Pour les molécules en chaîne, en l'absence de champ électrique, l'interaction hydrodynamique

entre segments domine la diffusion de translation. La diffusion est alors donnée par :

$$D = \frac{k_b T}{6\pi\eta R_H}\alpha N_0^{-\nu} \quad \nu = 0.6 \text{ en bon solvant}$$

où R_H est le rayon hydrodynamique de la chaîne, N_0 le nombre de monomères de la chaîne et η la viscosité du solvant.

En présence du champ électrique, l'interaction hydrodynamique est écrantée par le mouvement des contre ions. Le rapport q/ϵ devient indépendant de la taille des molécules.

Pour discriminer les différentes chaînes, il est nécessaire de contraindre leur déplacement par l'utilisation de gels.

La majeure partie des théories sur l'électrophorèse s'adresse à des chaînes de grandes longueurs devant leur longueur de persistance [75, 82]. Il n'y a pas un unique comportement de la chaîne contrainte par le gel, selon le rapport entre le rayon de giration de la molécule et la taille caractéristique des pores, les théories et les méthodes d'électrophorèses varient. Pour des tailles de pores grandes devant le rayon de giration de la molécule, le modèle dit de tamis moléculaire d'Ogston [55] prédomine. Lorsque le rapport s'inverse, la théorie de reptation introduite par de Gennes permet de décrire le mouvement de la chaîne dans le gel [39]. Compte-tenu de la faible taille de nos molécules et compte-tenu de la rigidité des hélices, nous pensons que le modèle de tamis moléculaire est plus adapté pour rendre compte du comportement de nos molécules en gels. Les structures migreraient donc librement dans chaque pore, puis seraient discriminées par leur capacité à se conformer à la géométrie locale du pore pour s'en échapper. Néanmoins, en ce qui concerne la mobilité des ADN simple brin, il n'y a pas de théories satisfaisantes permettant de décrire la mobilité de molécules structurées. Une étude expérimentale menée par Grainger, Lilley et leurs collaborateurs [23] donne du crédit à la relation dérivée par Lumpkin et Zimm [43] reliant la mobilité électrophorétique au carré du quotient de la longueur bout à bout sur la longueur de contour, moyennée sur l'ensemble des configurations de la molécule :

$$\mu = \frac{q}{\epsilon}\langle\frac{h_x^2}{L^2}\rangle$$

En dehors de la validation expérimentale de la relation macroscopique de la mobilité électrophorétique, ils montrent qualitativement qu'une excroissance non symétrique dans une hélice peut freiner la molécule lors de sa migration en gel. Dans notre cas, la conformation cigare présente une

excroissance non symétrique au milieu de la molécule, quant à l'autre conformation, les deux hélices sont reliées par un simple brin. Compte-tenu de la rigidité des hélices par rapport à la longueur curviligne de la molécule, les deux structures vont se comporter en gels comme deux bâtonnets joints, soit librement dans le cas de la structure branchée, soit élastiquement contraints dans le cas de la structure cigare. Nous nous attendons donc à une différence de mobilité électrophorétique entre les deux espèces. Cette différence de mobilité se traduit concrètement par l'apparition de deux bandes distinctes ; une par conformation. De plus, la bande la plus rapide devrait dans cette hypothèse correspondre à la structure la plus souple, c'est-à-dire à la structure branchée.

Il existe principalement deux familles de gels d'électrophorèses : les gels à base d'agarose et les gels à base de polyacrylamide. Les gels d'agarose sont faciles à manipuler et rapides à obtenir, mais ne présentent pas un haut pouvoir de séparation. Les gels de polyacrylamide par contre, permettent de discriminer entre des ADN hétéroduplexes qui ne présentent qu'un mésappariement et leur homologue homoduplexe, mais leur préparation et leur manipulation sont plus longues et plus délicates. De nouvelles techniques d'électrophorèses ont aussi vu le jour comme l'électrophorèse capillaire utilisée pour le séquençage [33] ou l'électrophorèse en présence d'obstacles cinétiques modulaires [12].

6.3.1 Gels d'agarose

Les gels d'agarose sont des gels physiques obtenus par le mélange de polymères d'agarose et d'une solution saline (dans notre cas du TAE -Tris Acétate EDTA-). L'agarose est un polymère linéaire extrait d'une algue. Les gels d'agarose sont préparés à des concentrations variant de 0.5% à 2% en masse par volume. La proportion d'agarose en solution définit le pouvoir de résolution du gel. Plus le gel est concentré, plus les chaînes courtes sont discriminées. Pour obtenir un gel d'agarose, il suffit de faire fondre la masse désirée d'agarose dans un volume de solution saline. Le gel est ensuite coulé à chaud dans un moule et se durcit en refroidissant. Lorsque le gel est pris en masse il est immergé à l'horizontal dans un bain de TAE. Il ne reste plus qu'à charger les puits par les solutions d'ARN ou d'ARN à étudier et mettre sous champ. Pour des gels d'agarose à 0.7%, le temps typique de migration pour obtenir une séparation effective entre les plasmides linéaires et circulaires et de l'ordre de 20 minutes, et pour séparer

l'ARN transcrit du plasmide une dizaine de minutes suffit.

Nous utilisons les gels d'agarose pour obtenir rapidement des informations sur la masse d'ARN transcrit ou sur la digestion effective des vecteurs circulaires.

6.3.2 Gels de polyacrylamide

Les gels de polyacrylamide sont des gels réticulés de monomères d'acrylamide (AA) et de bisacrylamide (BA). Il est à noter que les monomères d'acrylamide sont connus pour être cancérigènes. Les manipulations de ses réactifs sont donc à faire dans un environnement dédié, et l'on préfèrera les solutions stocks liquides et préalablement mixées, aux cristaux à dissoudre. De plus, lors de la manipulation de ce produit, il est préférable de porter des gants en vynil au lieu de gants en latex.

$$H_2C = CH - CO - CH_2$$
$$H_2C = CH - CO - NH - CH_2 - NH - CO - CH = CH_2$$

FIG. 6.5 – Formule semi développés de l'acrylamide (AA) et du bisacrylamide (BA)

Les paramètres de concentration sont définis de la façon suivante :

$$T = \frac{m_{AA+BA}}{vol_{solvant}} \quad C = \frac{m_{BA}}{m_{AA+BA}}$$

La proportion relative entre les deux espèces (noté C) fixe la porosité du gel, quant à la masse volumique totale en polymère (noté T), elle définit la taille caractéristique statistique des pores. A une concentration donnée en polymère, à T fixé, la variation de la proportion relative entre les espèces définit le taux de réticulation des chaînes. De plus, à faible ou forte concentration de bisacrylamide, la taille des pores est relativement homogène alors qu'à une proportion moyenne, la distribution est poly disperse. Pour nos expériences, les monomères sont mélangés à du TAE. La réticulation du gel est catalysée par du Tétra-Méthyl-Ethylène-Diamine (TEMED) et la réaction est initiée par l'ajout d'Ammonium Persulfate (APS). La solution est ensuite coulée entre deux plaques de verre espacées de quelques dixièmes de millimètre. Dans notre cas, nous avons utilisé des espaceurs de 0.5mm, et le gel est coulé à l'horizontal. A cette épaisseur, la

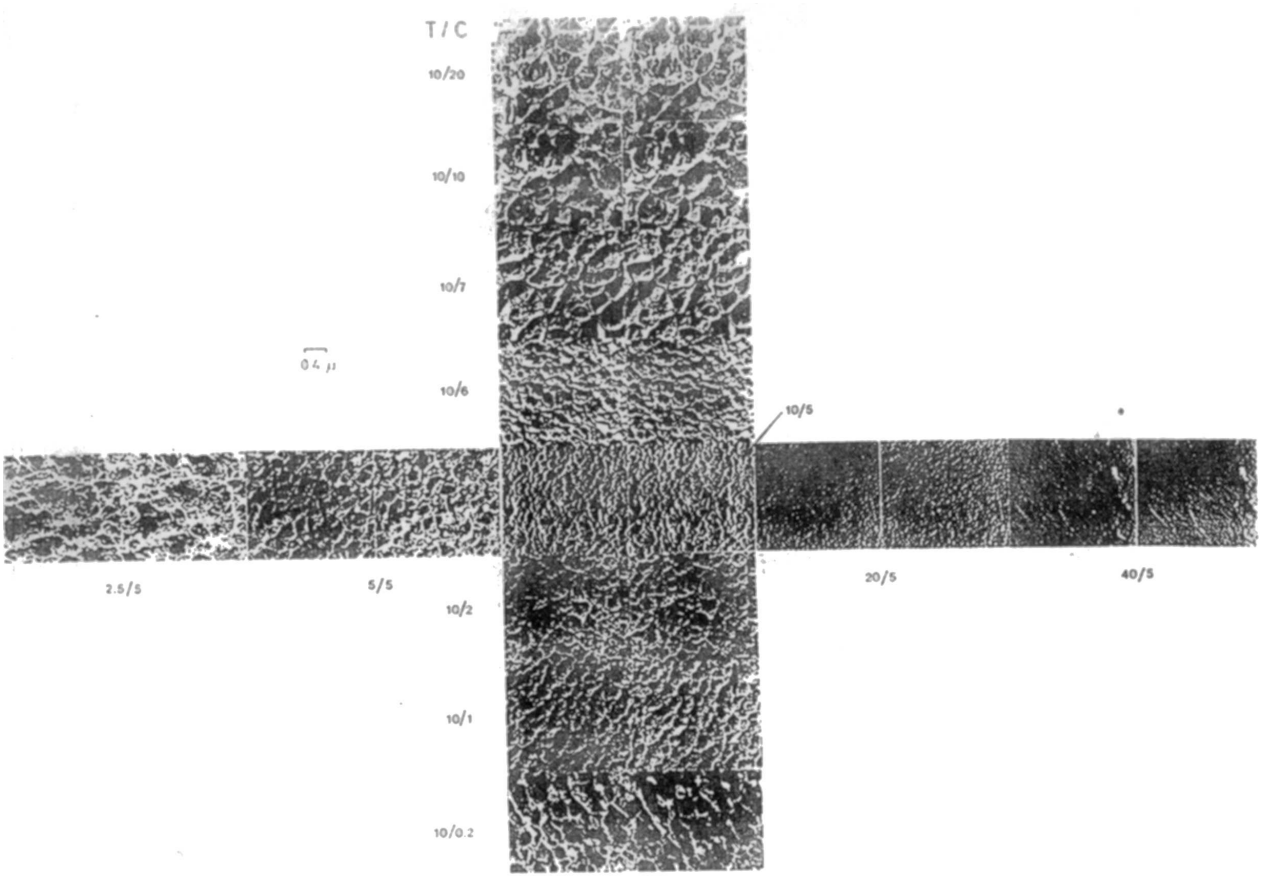


FIG. 6.6 – Image TEM de gels polyacrylamides obtenus en variant la concentration total d'acryl (T) et la concentration en bisacryl (C). Les valeurs sont exprimé sous la forme T/C [3].

solution s'étend entre les deux plaques par capillarité. Lorsque le gel est durci, il est monté à la verticale sur une cellule prévue à cet effet. Il est recommandé de laisser le gel gonfler dans le solvant au moins une nuit pour le stabiliser [65]. La sensibilité des gels de polyacrylamide permet de séparer les simples brins selon leur structure [74, 56]. Cette sensibilité est telle que l'on peut même dissocier des doubles brins hétéroduplexes comportant un unique mésappariement de leur homologue homoduplexe [51].

Gels de polyacrylamide utilisés

Les gels de polyacrylamide que nous utilisons sont des gels de concentration T=12% et concentration relative en monomère d'acrylamide/bisacrylamide de 29 pour 1. La solution initiale d'acrylamide bisacrylamide est fournie par la société Biorad. La solution stock de préparation pour gels de polyacrylamide est préparée dans un volume de 50 ml, selon le protocole suivant [14, 66] :

> Préparation solution stock pour gel de polyacrylamide
>
> > 15 ml d'acryl/bisacryl 29 :1 à 40%. 1 ml TAE 50x. 5 ml Hepes 660 mM. 1,7 ml Tris-HCl 1M. 250 μl $MgCl_2$ 2M. 10 μl EDTA 0,5 M. 26,5 ml eau ultra pure.
>
> le tout est stocké au réfrigérateur à 4°C.

Pour notre cellule, un volume de 14 ml de solution stock suffit. A ces 14 ml de préparation s'ajoutent 104 μl d'APS 10% et 4,9 μl de TEMED pour activer la réticulation.

Tampon de chargement

Les solutions d'ADN et d'ARN sont mélangées avec un volume dilué 10 fois de colorant (Xylène-cyanol) et de glycérol. Le glycérol augmente la densité de la solution et permet de s'assurer que la solution se dépose au fond des puits après chargement en gel.

6.3.3 Visualisation des polynucléotides

Pour visualiser nos chaînes ADN ou ARN nous utilisons un marqueur fluorescent, le bromure d'éthidium, qui vient s'intercaler entre les bases des parties double brins de nos molécules. Sous illumination ultraviolette, l'ADN se désexcite non radiativement par transfert d'énergie vers la

molécule d'éthidium bromure qui se désexcite de façon radiative dans le rouge. En gel d'agarose, l'intercalant est dilué dans la solution liquide. En fin de migration, le gel est immédiatement exploitable. En gel de polyacrylamide, le gel est dans un premier temps démoulé puis baigné dans une solution de TAE contenant de l'éthidium bromure (8μl d'éthidium bromure pour 200 ml de TAE) durant une dizaine de minutes avant d'être déposé sur plaque UV.

Pour visualiser l'avancement de l'électrophorèse, nous utilisons des colorants chargés dont la vitesse de migration est tabulée en fonction du type de gels et de la concentration relative en polymères.

En gel d'agarose, nous utilisons le bleu de bromophénol qui migre comme un ADN double brin de 70 paires de base dans un gel à 0.7%.

En gel de polyacrylamide, nous utilisons le xylène cyanol qui migre comme 30 paires de base dans un gel de polyacrylamide à T=12%.

6.3.4 ARN natifs et renaturés

Pour visualiser les structures d'ARN natives et relaxées nous utilisons pour chaque gel, les mêmes sources d'ARN. Les structures natives sont obtenues en fin de transcription et sont conservées dans leur configuration native en les stockant au congélateur à -20°C pendant quelques heures voir à -70°C si le stockage devait se prolonger. Les structures relaxées sont obtenues après un vieillissement à 37°C pendant plusieurs heures, ou en dénaturant et en renaturant les molécules d'ARN en les plongeant dans un bain d'eau chauffé à 80°C et en laissant le bain se thermaliser à la température ambiante. Les deux espèces natives et renaturées, sont ensuite chargées en gels de polyacrylamide.

Chapitre 7

Résultats expérimentaux

Tous les ADN clonés sont séquencés par le service de séquençage de l'IBMC.

7.1 Gels de contrôle

Masse d'ARN à temps court

Nous nous sommes intéressés à connaître le temps incompressible le plus court pour transcrire une masse suffisante d'ARN. Durant une transcription dans un volume total de 30μl, nous avons pipeté et mis en gel agarose un volume fixé de 3.5μl à intervalle de temps régulier. La concentration en ADN et en polymérase reste constante tout au long de l'expérience et l'électrophorèse fonctionne en continu pendant la durée de l'expérience. Le champ n'est que brièvement coupé lors de chaque chargement. Nous obtenons une image de migration sur laquelle les bandes représentant des polynucléotides de longueur identique ne sont pas horizontales mais obliques. La figure 7.1 montre la variation de la masse d'ARN transcrite au cours du temps. Les trois lignes obliques visibles sont le substrat ADN (plasmide et insert) linéarisé, l'ARN synthétisé et entre les deux, le bleu de bromophénol. La première piste est changée avec une solution référence d'ADN lambda digérée avec l'enzyme HindIII. Chaque bande est de masse connue. A partir de cette image, un logiciel d'intégration permet de définir la masse

relative des différentes bandes. Ce gel nous permet de conclure qu'au bout d'une demi-heure de transcription, nous obtenons une concentration en ARN proche de 30ng/μl.

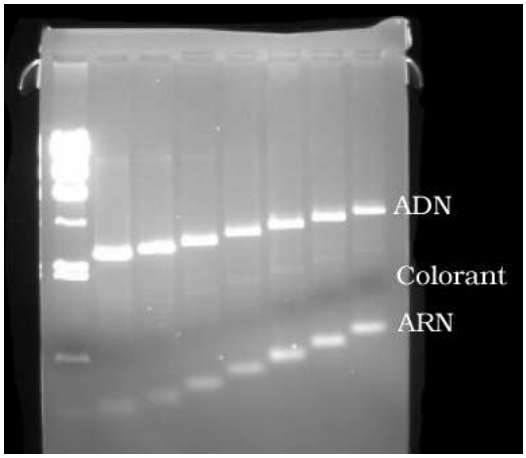


FIG. 7.1 – Variation de la masse d'ARN transcrite au cours du temps. Chaque piste représente une prise, l'intervalle de temps entre les prises est de 7 minutes. Tout à fait à gauche se trouve la piste de référence contenant une masse volumique connue d'ADN Lambda digéré par l'enzyme HindIII.

Masse visible en gel de polyacrylamide

De même que nous connaissons la vitesse caractéristique de transcription de notre séquence d'ADN, nous avons testé différentes masses d'ARN mise en gel de polyacrylamide. Ce test permet de connaître la masse d'ARN minimum qu'il faut déposer en puits pour obtenir des gels lisibles. Nous nous sommes fixé comme limite inférieure, une concentration minimale de 30ng/μl. La figure 7.2 montre un gel obtenu avec une masse de 30ng/μl. Les bandes ARN sont visibles en milieu de gel. L'ADN ne rentre quasiment pas en gel, il est bloqué dans le fond des puits de chargement. Les pores des gels de polyacrylamide sont plus petits, le plasmide avec insert est trop gros pour pouvoir migrer profondément en gel.

Pour éviter que le gel de polyacrylamide ne blanchisse trop, il est préférable de ne pas laisser le gel dans le bain de bromure d'éthidium plus d'une vingtaine de minutes.

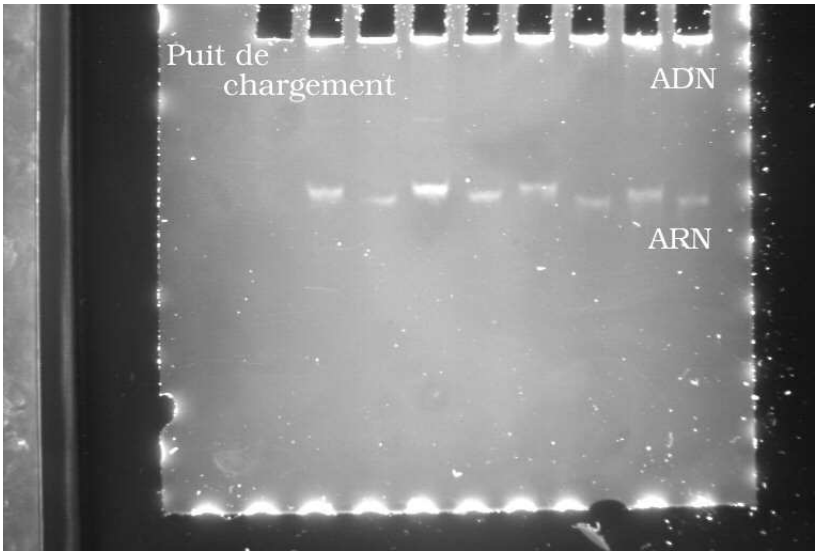


FIG. 7.2 – Gel polyacrylamide chargé avec une masse d'ARN de 30ng/μl par puit. Les gels de polyacrylamide sont plus sélectifs, les ADN (plasmide et insert linéarisé) ne rentrent pas en gel, ils sont bloqués dans les puits de chargement.

Relaxation des molécules

Pour tous les couples étudiés, nous avons vérifié qu'il n'y avait pas de biais dû au mode de relaxation. Que les molécules relaxées soient obtenues soit par dénaturation/renaturation à haute température, soit par relaxation thermique plusieurs jours à 37°C, les intensités des bandes sont identiques.

7.2 Optimisation des gels de polyacrylamide

Nous avons cherché à optimiser les paramètres du gel pour obtenir la meilleure séparation possible. A la lecture de la littérature concernant la séparation de brin hétéroduplexe [56, 77, 51, 74], il apparaît, que les paramètres de concentration totale (T) et relative (C) des deux monomères d'acrylamide et de bisacrylamide jouent un rôle prépondérant dans la séparation de conformation. Mais la présence ou l'absence de glycérol à différentes proportions peut aussi améliorer la séparation des espèces. Nous avons fait varier la concentration totale en monomères

(T) de 6% à 20%, et la concentration relative des espèces dans un ratio de 12 monomères d'acryl pour 1 monomère de bisacryl (12 :1) à (37.5 :1). L'effet majeur est porté par la concentration totale (T) en monomère : au-delà d'une concentration massique supérieure à 18%, les molécules d'ARN ne peuvent quasiment pas migrer en gel et aucune séparation n'est effective. Les gels de 6% s'avèrent délicats à manipuler et présentent des fronts de migration relativement dispersés pour nos molécules, et nous ne pouvons pas exclure la possibilité de fusion entre deux bandes proches. Les gels de 12% et 15% s'avèrent optimum pour la séparation attendue. Mais pour des raisons de dissipation thermique, nous avons décidé d'utiliser des gels de concentration T=12% qui permettent à la fois d'obtenir des bandes fines et une température estimée en cœur de gel raisonnable. La variation de la concentration relative en espèce de monomère n'influe que peu sur la mobilité obtenue, mais permet d'affiner la dispersion des bandes. L'ajout de glycérol nous a permis de consolider l'hypothèse d'une seconde bande pour la molécule réverse, mais s'est avéré inutile avec la modification de la séquence de la molécule comme nous le verrons plus loin dans le chapitre.

7.3 Séquence directe Dsma

Le but de l'expérience est de mettre en évidence l'existence d'une structure unique à temps court, et la relaxation de la molécule à temps long dans les deux configurations d'énergie similaire. Nous espérons pouvoir discriminer les deux conformations par électrophorèse en gel de polyacrylamide. La discrimination sur gel d'électrophorèse joue sur les paramètres de charge, de masse, et de conformation. Comme nous avons à faire à deux conformères, le seul paramètre pertinent est la différence de conformation. Donc, l'apparition de bandes séparées en gel, qui traduit la différence de mobilité électrophorétique, ne peut être attribuée qu'à l'existence de conformations tridimensionnelles différentes.

Le résultat obtenu pour la séparation de possibles conformères de la molécule directe Dsma est montré sur la figure 7.4. Le gel comporte deux séries de transcrit vieilli à $37°C$ pendant des temps variant de 2^{h00} à plus de 90^{h00}.

Protocole expérimental : à la solution initiale de transcription est ajouté un volume d'EDTA en excès pour capter les ions magnésium et arrêter la transcription des ARN. La solution est

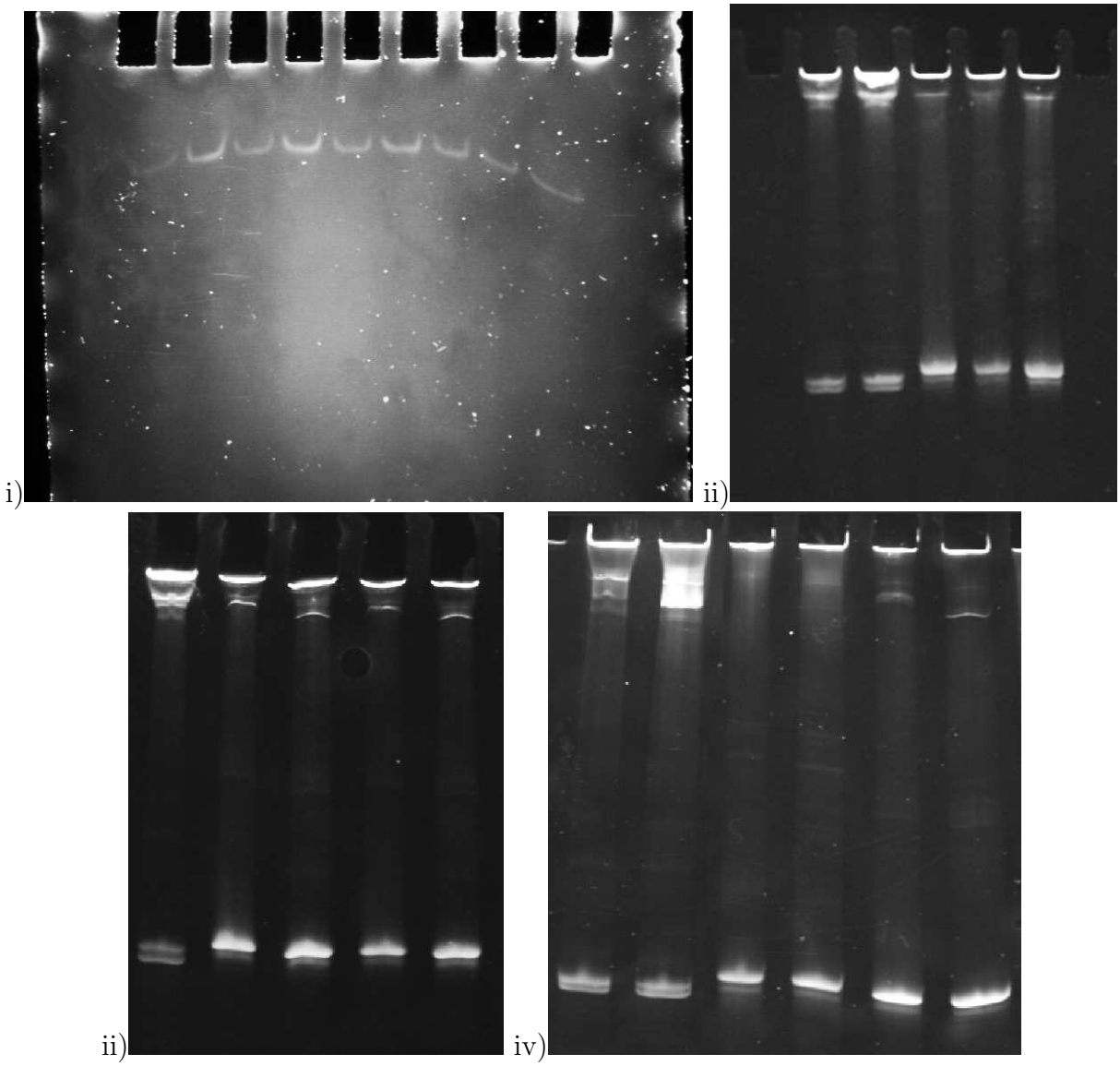


FIG. 7.3 – Migration obtenue pour différents paramètres de gel. i) Gel à T=20%, contenant des molécules directe et réverse natives et renaturées thermiquement. ii) Gel à T=12%, glycérol 5% : les deux premières bandes correspondent à la séquence Dsma renaturée, les trois suivantes sont issues de la transcription et la renaturation thermique de la séquence réverse allongée (cf. texte). iii) Gel à T=12%, glycérol 10% : la première bande est la molécule Dsma renaturée, les autres correspondent à la molécule réverse allongé renaturée. iv) Gel à T=12%, glycérol 15% : les deux premières bandes sont caractéristiques de la séquence Dsma renaturée, les quatre autres sont la signature de la séquence réverse allongée. Lorsque la concentration en monomère est plus élevée, les ARN ne migrent pas assez en gel, de plus toute anisotropie dans la concentration volumique est immédiatement visible (bandes courbées). L'ajout de glycérol réduit la dispersion des bandes et fait apparaître la probable existence d'une seconde bande pour la molécule réverse.

ensuite titrée et aliquotée pour être vieillie durant des temps différents. Les solutions vieillies sont ensuite stockées sur glace et au congélateur, à $-20°C$. Les tampons de chargement sont préparés à l'avance et mixés avec les solutions d'ARN fondantes juste avant le chargement en gel.

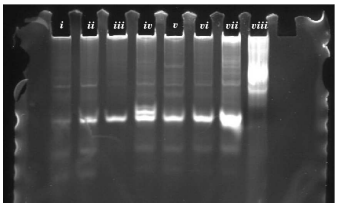


FIG. 7.4 – Etude de la cinétique de passage de la barrière de potentiel pour la molécule directe. Les molécules d'ARN sont vieillies à 37°C pendant des temps différents, stockées sur glace à -20°C puis chargées en gel. De gauche à droite, les bandes obtenues pour un temps de vieillissement croissant : i)2h20, ii)7h00, iii)20h00, et iv)54h00, pour une première série, v)2h11, vi)5h25, vii)97h00, pour une seconde série issue d'une transcription différente, et viii) bande utilisée pour orienter le gel

A temps long, la molécule relaxe et présente deux bandes distinctes caractéristiques de deux conformations différentes. La plus lourde des deux bandes migre sur le même front que la bande unique obtenue à temps court. Cette observation apporte deux remarques. Premièrement, la bande unique obtenue à temps court correspond à une et une seule configuration stable et non à la superposition des signatures des deux conformères. Secondement, la bande lourde correspond à la conformation la plus adaptée à migrer "facilement" en gel, soit à la molécule branchée.

En conclusion, la molécule théorique directe valide nos hypothèses. Elle est bien bistable, et sa cinétique de relaxation est suffisamment lente pour que la molécule n'ait pas le temps de thermaliser lors de l'électrophorèse. La configuration piégée en fin de transcription correspond à la molécule branchée, souple.

7.4 Comparaison avec la séquence réverse Rsma

Les premiers essais ont été effectués sur les séquences dites directes et réverses théoriques. Ces séquences sont obtenues à partir des séquences théoriques en ajoutant un triplet de base

G à l'extrémité 5' et un triplet de base C à l'extrémité 3'. Les premiers gels sont effectués à température ambiante. Le champ et le courant appliqués sont faibles : 6mA fixe, soit environ 50V de tension, et la migration dure entre 12 et 15 heures.

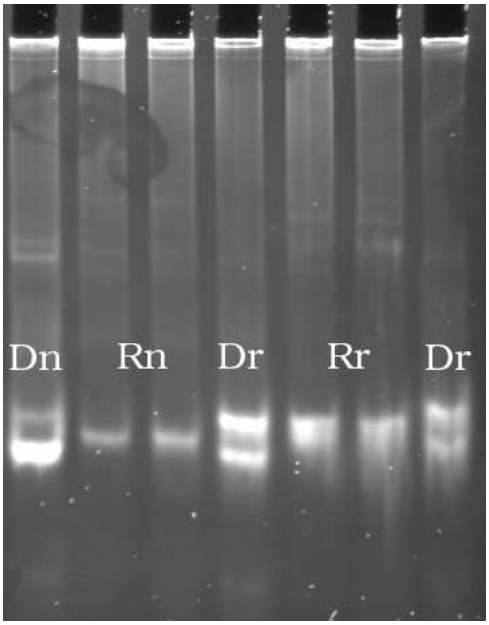


FIG. 7.5 – Migration des structures directes(Dn) et réverses(Rn) natives et renaturées thermiquement (Dr et Rr). La molécule directe présente bien deux configurations stables après relaxation, et essentiellement une seule configuration en fin de synthèse comme attendue. Pour la molécule réverse, les choses sont moins claires, il est possible que nous observions en fait une bande unique qui correspond à la migration moyenne des deux attendues.

La molécule directe en fin de transcription présente une bande intense (en bas) ainsi parfois qu'une bande plus légère au-dessus alors qu'après renaturation, deux bandes équilibrées apparaissent. Les deux bandes obtenues à temps long démontrent que la molécule présente bien deux configurations bistables d'énergie similaire [1]. La bande intense obtenue à temps court laisse à penser que seule l'une des deux configurations est présente en fin de synthèse. Donc, pour la molécule directe, l'hypothèse de piégeage cinétique semble être vérifiée à temps court

[1]Une différence d'intensité intégrée d'un facteur 2 ne correspond qu'à une différence d'énergie de $0.3kT$ sur environ $60kT$ d'énergie libre totale

et le codage des deux configurations stables est vérifié à temps long.

La molécule réverse ne semble présenter en revanche qu'une seule bande à temps court et une bande identique ou proche à temps long. En comparant la position de cette bande avec celles de Dsma, nous voyons que cette bande apparemment unique ne migre ni comme la configuration branchée ni comme la configuration cigare mais semble plutôt migrer à une vitesse moyenne de celles des deux configurations de Dsma. A ce stade, nous avons posé comme hypothèse qu'au cours de l'électrophorèse, la molécule réverse a le temps nécessaire pour changer plusieurs fois de configuration. Cette bande unique reflèterait la vitesse moyenne de migration des deux configurations attendues. En effet, il n'est pas exclu en principe que les deux molécules Dsma et Rsma aient des temps de relaxation assez différents bien que leurs états d'équilibres supposés soient similaires par construction.

Apparatus de refroidissement pour les gels de polyacrylamide

Dans cette hypothèse, pour tenter d'obtenir la dissociation des deux bandes de la molécule réverse nous devons réduire le temps d'électrophorèse donc augmenter le courant traversant le gel. Or l'augmentation du courant induit une augmentation de la puissance dissipée en gel et donc une augmentation de température du gel qui accélère la dynamique de relaxation entre les structures. Pour réduire la température de cœur, nous avons effectué les électrophorèses dans un réfrigérateur à 4°C avec des solutions tampons thermalisées. Une plaque d'aluminium de 0.5 mm d'épaisseur est accolée à l'une des faces de la cellule et plonge dans l'un des bains de tampon salin pour évacuer plus rapidement la chaleur dissipée dans le gel. De plus, dans chaque bain de tampon, nous avons plongé des blocs de refroidissement préalablement stockés à -20°C. Avec ce dispositif, la température en gel est uniforme dans son épaisseur et peu fluctuante dans sa longueur. Avec le montage de refroidissement du gel, nous pouvons obtenir une migration suffisante en 3 heures au lieu de 15. L'intensité est augmentée à une valeur de 45 mA soit une dissipation d'environ 12 Watts. En fin de migration, la température constatée du gel est de 12°C.

Dans les mêmes conditions de préparation des échantillons, la molécule directe native ne présente toujours bien qu'une seule bande comme nous l'avions attendu. Par contre, la molécule réverse présente toujours une bande moyenne.

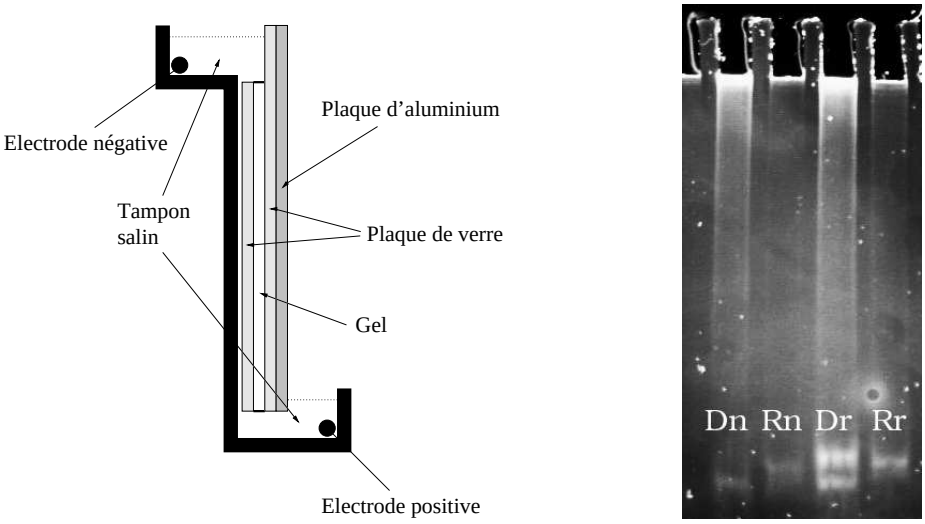


FIG. 7.6 – Schéma de principe de la cuve pour gel de polyacrylamide. A droite, résultat obtenu pour le couple directe/réverse théorique avec le montage de refroidissement.

Identification des structures

Pour la molécule directe Dsma, les gels de polyacrylamide présentent bien deux bandes d'ARN après renaturation. Ces deux bandes sont la signature des deux configurations de la molécule. La structure qui migre le plus rapidement en gel correspond à la structure piégée obtenue en fin de synthèse, c'est-à-dire d'après notre étude théorique, à la structure branchée. Il existe un certain nombre d'exemples de molécules d'ARN naturelles connues qui sont aussi piégées dans des structures de types de types branchées en fin de transcription [6] . En revanche, il n'existe pas à notre connaissance d'exemple connu de molécule piégées en fin de transcription dans des structures de type cigare qui nécessite d'après notre étude théorique un échange de stabilité entre les hélices au cours de la transcription. Restropectivement (voire suite de l'étude) c'est sans doute la raison pour laquelle les séquences réverses étudiées ont généralement un taux d'échec plus important que les molécules directes qui ont toutes fonctionnées.

7.5 Séquences directes et réverses allongées

Nous voulons améliorer le piégeage des structures de plus basse énergie. Pour ce faire, nous devons augmenter la hauteur de la barrière cinétique. Or toute modification de la séquence, entraîne une modification dans les chemins cinétiques de relaxation.

Si nous ne pouvons pas connaître avec précision toutes les relations entre séquence et dynamique de relaxation, nous savons qu'il faut à tout prix garder l'équilibre énergétique entre les structures de plus basse énergie. Donc plutôt que de chercher à améliorer la dynamique de relaxation en mutant quelques bases de la séquence initiale, nous avons préféré ajouter des paires de bases aux hélices définies. Cet ajout stabilise la molécule et maintient la partition énergétique entre les molécules. Nous avons ensuite étudié numériquement les différents couples pour en choisir un qui présente une dynamique de relaxation suffisamment lente pour les deux molécules, pour que l'on puisse discriminer les deux structures sur gel.

Remarque 1 : L'ajout d'une seule paire de bases dans une hélice nécessite ici l'ajout de trois autres pour garder le caractère palindromique des hélices de la configuration et permettre le repliement de la molécule dans les deux configurations. Les énergies de *stacking* ajoutées aux structures sont les mêmes dans les structures #1 et #2 deux à deux semblables, il n'y a donc pas de modification de l'équilibre énergétique entre les configurations.

Remarque 2 : Lors du premier essai avec les séquences directe et réverse théoriques, nous avons choisi de couper la séquence avec l'enzyme SmaI. Cette enzyme laisse un triplet de bases CCC en positions terminales. Si la polymérase commence correctement la transcription après la fin de la séquence promotrice TATA, le triplet de base GGG (inséré entre la fin du promoteur et le début de la séquence à transcrire cf. 6.2.2) doit être transcrit. Dans ce cas, une petite hélice de base GC peut se former entre les deux extrémités de la molécule. Pour palier à cette indéterminée, nous avons choisi de remplacer l'enzyme de restriction SmaI par l'enzyme StuI qui reconnaît le site AGG—CCT et coupe après AGG. Le triplet de C 3' terminal est alors remplacé par un doublet de G (cf. figure 6.3).

Le tableau ci-dessous donne les temps caractéristiques de relaxation numérique des molécules directe et réverse théoriques et allongées, ainsi que le rapport du temps de relaxation de chaque molécule par rapport au temps de relaxation de la molécule directe théorique.

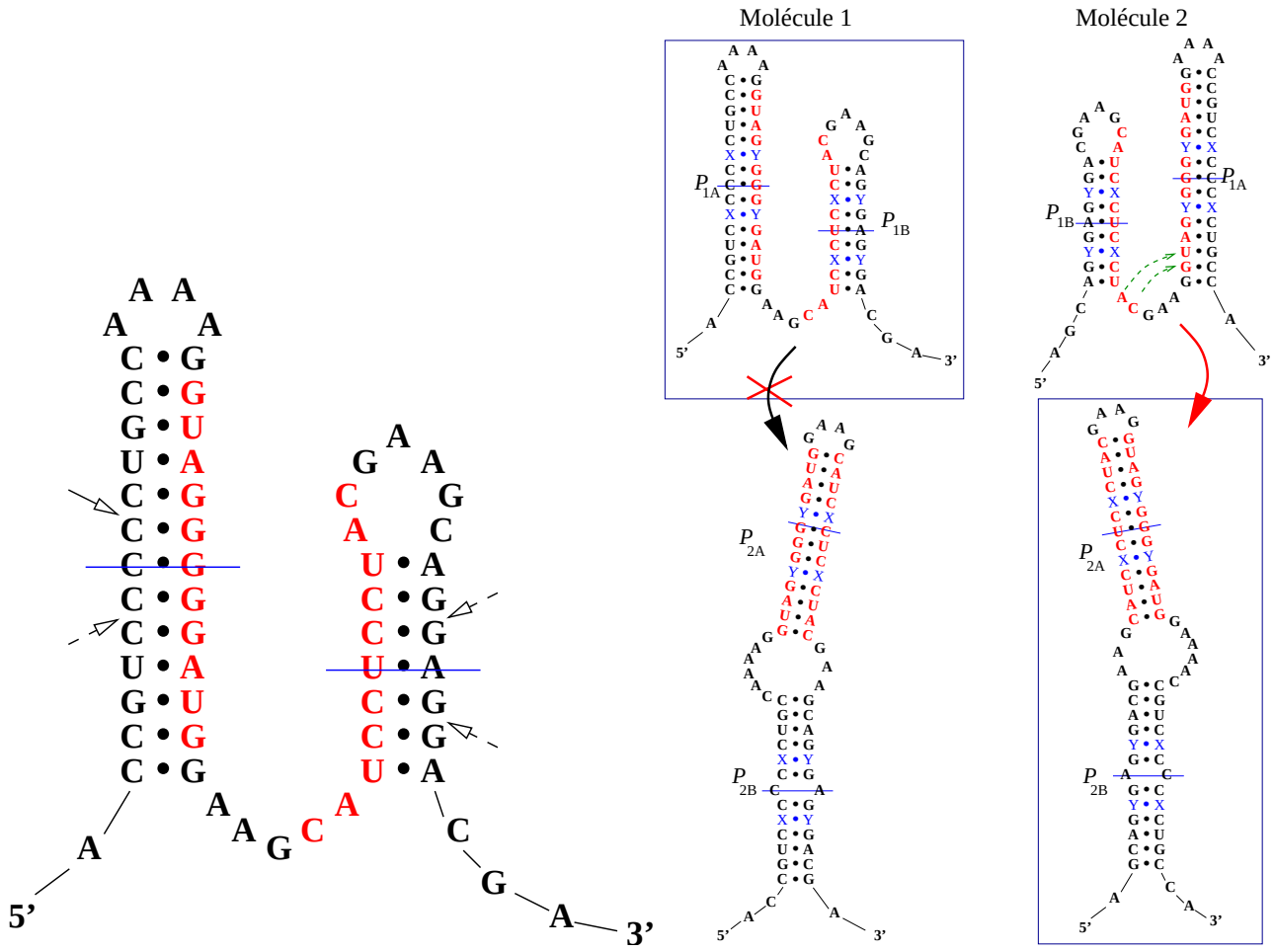


FIG. 7.7 – L'insertion d'une paire de bases supplémentaire dans une hélice induit l'insertion des autres paires de bases dans la configuration pour respecter la symétrie qui nous assure que les énergies des deux structures seront modifiées de la même quantité ΔE. En effet, ces modifications permettent de garder le caractère palindromique de l'hélice modifiée, et assurent que toutes les énergies de *stacking* ajoutées à l'une des structures sont aussi ajoutées à la seconde. La molécule réverse est obtenue en renversant la séquence de la molécule directe.

nom de la séquence	temps moyen d'équilibre en secondes	rapport des temps
directe	6.24E+03	1
reverse	5.79E+02	0.09
directe_alongée	3.08E+04	4.94
reverse_allongée	1.07E+04	1.71

Notons que notre hypothèse sur une relaxation plus rapide de la molécule Rsma est plutôt étayée par ces résultats numériques bien que ces derniers ne sauraient expliquer entièrement nos observations. L'insertion d'une base UA dans les hélices dans les molécules "allongées" permet de ralentir suffisamment la cinétique de relaxation de la molécule réverse pour qu'elle soit a priori visible en gel.

Le résultat numérique est en accord avec le résultat expérimental, comme montré ci-dessous. Les molécules directe et réverse allongées relaxent dans deux structures. A temps court, les

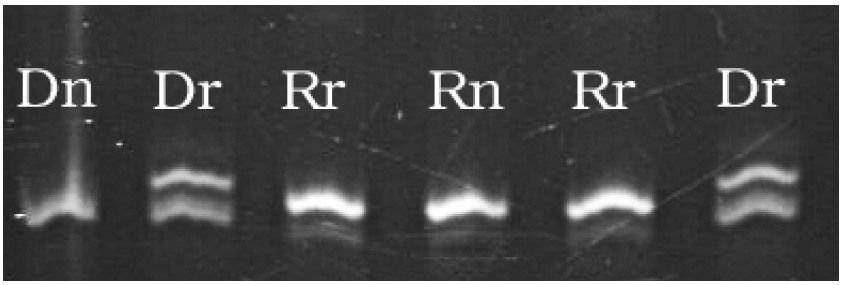


FIG. 7.8 – Résultat pour la migration du couple directe réverse allongée

molécules sont piégées dans une seule configuration. La molécule directe est piégée dans la configuration dont la vitesse de migration est la plus rapide. La molécule réverse est piégée dans la configuration dont la vitesse de migration est la plus lente. Ce gel suggère que les molécules suivent bien notre hypothèse de repliement en cours de synthèse. Malheureusement la molécule réverse ne relaxe pas aussi bien que la molécule directe, les deux bandes sont déséquilibrées.

Le facteur d'intensité entre les deux bandes de la molécule réverse relaxée est de l'ordre de 4. Après avoir relaxé, la différence d'intensité reflète la différence d'énergie des deux configurations de plus basse énergie. La différence d'énergie entre les deux conformations est très petite :

$\Delta E = k_B T ln4$, soit environ $0.6 k_B T$. Comparée à l'énergie moyenne d'un stacking, 1.3 $k_B T$, il semble que cette différence ne soit pas imputable aux hélices (identique par symétrie), mais probablement due aux énergies de formation des boucles. En particulier, cette différence peut être due à l'utilisation de boucles à quatre nucléotides.

De manière générale, les simples brins ont une tendance naturelle à s'empiler pour réduire leurs interactions avec l'eau. Dans le cas des quadriboucles, il est connu que les résidus du simple brin ont tendance à se structurer pour minimiser leur interaction avec le solvant, ce qui apporte une contribution à l'énergie de stabilisation de la boucle. Ceci est particulièrement vrai pour les quadriboules de séquence GNRA, où N est n'importe quel nucléotide et R remplace une base purine. Cette énergie est séquence et solvant dépendant. Il est donc probable que les quadriboucles présentes dans les configurations de la molécule réverse soit la source de la faible différence d'énergie entre les deux configurations.

Actuellement, il n'existe pas de code de repliement de structures secondaires de molécules d'ARN qui prédise la correction énergétique à apporter pour les boucles en général. Par contre la dernière version du code de repliement de Michael Zuker inclut une correction phénoménologique pour certaines quadriboucles particulières. Dans notre cas, il apparaît que la quadriboucle 5'-C-GAAG-G-3' fermée par l'appariement CG, stabilise la structure #1 de la molécule réverse.

La méthode la plus simple pour rééquilibrer les deux configurations est de retirer un nucléotide de la quadriboucle. La boucle 5'-C-GAG-G-3', les boucles à trois nucléotides ne se comportent pas non plus comme des simples brins modèles, la distance curviligne du squelette est de l'ordre du rayon de l'hélice. Il se peut donc, que les bases se structurent, mais la correction à apporter est sans doute plus faible que dans le cas des quadriboucles.

Pour modifier la quadriboucle nous pouvons, soit ajouter une base supplémentaire soit retirer un nucléotide. Dans les deux cas, nous ne contrôlons pas la correction énergétique apportée par la structuration des bases. Néanmoins, dans le cas des boucles à trois nucléotides, les contraintes spatiales font que les résidus sont rejetés vers l'extérieur, au contact du solvant. On s'attend ainsi à ce que les triboucles soient moins stabilisées que les quadriboucles.
Numériquement, les simulations de relaxation effectuées à l'aide de ***KineFold*** nous donnent les temps de relaxation indicatifs suivant :

nom de la séquence	temps moyen d'équilibre en secondes	rapport des temps
directe	6.24E+03	1
reverse	5.79E02	0.09
directe tri-boucle	8.59E+03	1.38
reverse tri-boucle	4.32E+03	0.69

Les temps de relaxation numérique du couple directe/reverse triboucle sont du même ordre de grandeur que celui de la molécule directe théorique de référence. Notons tout de même que le temps de relaxation de la molécule réverse tri-boucle est inférieur au temps de relaxation de la molécule directe théorique.

Parmi quelques autres essais de séquence, nous avons testé ce dernier couple de molécules directes et réverses dans lesquelles la quadriboucle GAAG est remplacée par la triboucle GAG. De plus, nous ajoutons directement 0.5μl de bromure d'éthidium dilué 0.1x dans le tampon de chargement. L'éthidium est ainsi mélangé aux ARNs avant leur chargement en gel. Cette modification dans le protocole nous permet de visualiser le gel directement après électrophorèse. Ci-dessous, le résultat obtenu pour le couple de molécules directe et réverse tri-boucle.

Pour ce couple, les molécules relaxées sont équipartitionnées dans les deux configurations et, à temps court les molécules sont bien piégées dans l'une ou l'autre des configurations. Il est à noter qu'à temps court la molécule réverse présente une faible seconde bande ; lors de la compétition entre les hélices incompatibles, il arrive donc statistiquement que le processus de basculement d'un chemin de repliement vers l'autre n'ait pas lieu.

7.6 Conclusion

Nous avons montré qu'il est possible de concevoir des molécules d'ARN bistables qui soient guidées efficacement vers l'une ou l'autre de leurs configurations au cours du repliement co-transcriptionnel. Les structures ainsi formées peuvent être piégées pendant plusieurs heures voir plusieurs jours à 37°C et relaxées quasiment instantanément à haute température ($\sim$ 80°C). Les exemples des molécules directe et réverse triboucle montrent que la structure piégée en fin de transcription peut en principe avoir n'importe quelles formes souhaités, soit plutôt ramifiée (comme pour la molécule directe) soit plutôt allongée (comme pour la molécule réverse) alors même que leurs hélices sont identiques deux à deux. Ces résultats montrent qu'il suffit finalement

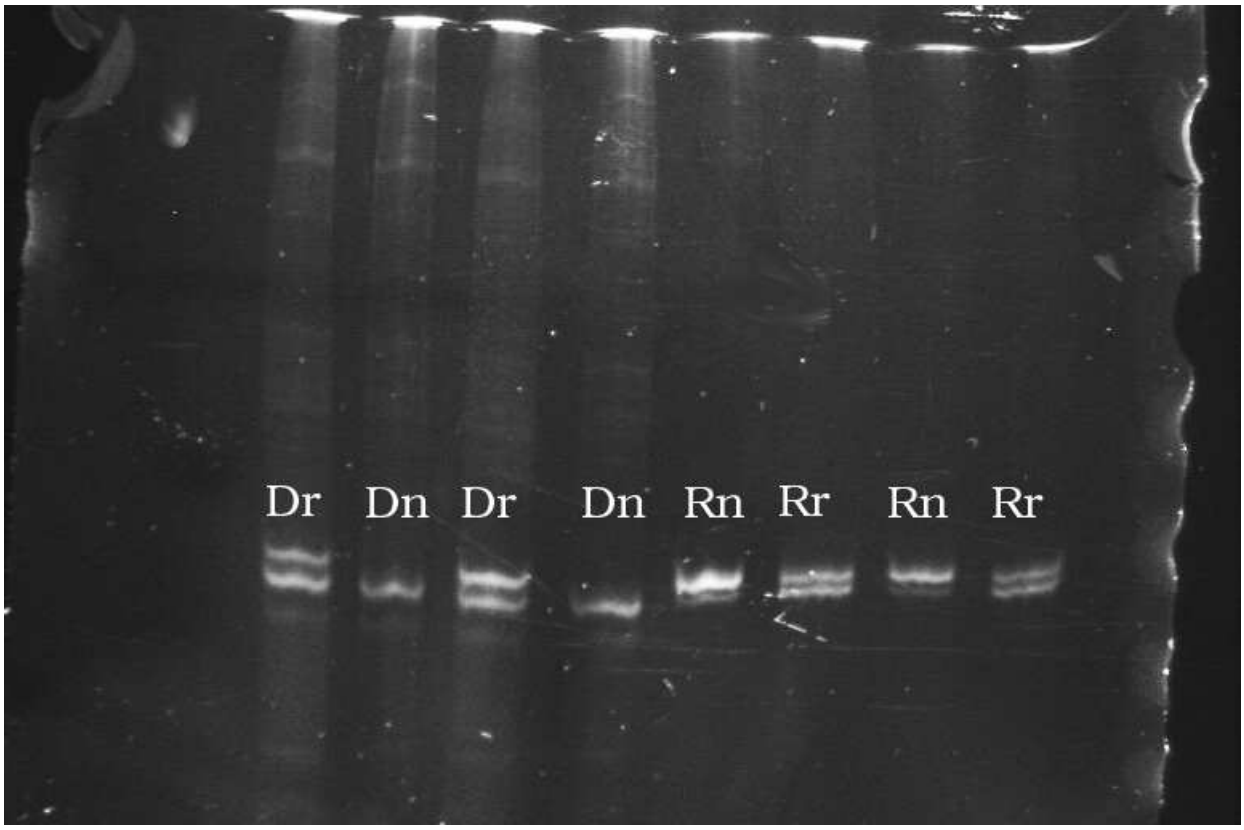


FIG. 7.9 – Résultat pour la migration du couple directe réverse tri-boucle

de coder peu d'informations sur une séquence d'ARN pour obtenir des comportements de **switch** intéressants. Il s'agit essentiellement de coder la longueur relative des hélices successives qui peuvent être formées au cours de la transcription.

De plus, nous avons mis en évidence la possibilité pour les ARN d'utiliser un processus de **switch** contrôlé, activé par la présence d'hélices transitoires (exemple de notre molécule réverse et de l'hélice P_{1B}), qui permet de guider la molécule vers sa conformation native. L'étude théorique menée à partir de ***KineFold*** sur les molécules directe et réverse se vérifie donc bien en pratique et permet de valider les prédictions de chemin de repliement basé sur ce processus. En particulier, l'analyse théorique [34] menée sur le repliement du ribozyme du virus de l'hépatite delta est confortée.

Quatrième partie

Applications publiques

Chapitre 8

Visualisation de structures secondaires avec pseudonœuds.
RNAMovies *with pseudoknots*

8.1 Introduction

La représentation de la structure secondaire d'une molécule d'ARN sans pseudonœud peut se faire sous forme d'une séquence orientée et linéaire de caractères. En particulier, sous forme d'une séquence de points " $\cdot$ " et de parenthèses "(" ")". Dans cette représentation, une paire de bases (i, j) est indiquée par une parenthèse ouvrante "(", à la $i$ème position et une parenthèse fermante ")" à la $j$ème position. Les bases non appariées sont représentées par un point " $\cdot$ ". A partir de cette séquence, plusieurs représentations bidimensionnelles sont possibles. La moins intuitive est la représentation sous forme de montagne (cf. figure 8.1). Elle a été conçue pour la comparaison de structures secondaires de molécules d'ARN [32, 37]. Les trois symboles "(", ")" et " $\cdot$ " sont associés aux trois directions haut, bas, et horizontal dans le tracé. Le graphique obtenu permet de visualiser les éléments de la structure. Un pic correspond à une tige-boucle. Un plateau à une petite région non appariée. Une vallée indique une région reliant

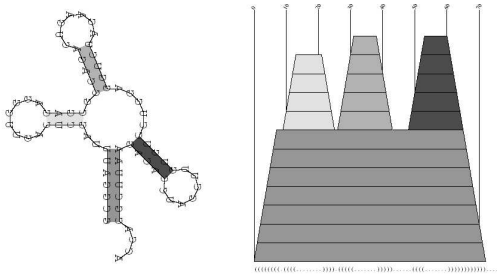


FIG. 8.1 – Représentation polygonale à gauche et montagne à droite de la structure secondaire de l'ARNt^{phe}.

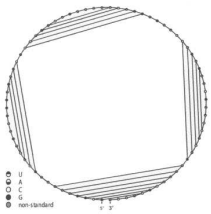


FIG. 8.2 – Représentation ciculaire de la structure secondaire de l'ARNt^{phe}.

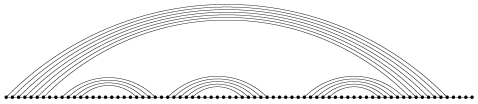


FIG. 8.3 – Représentation en graphe de la structure secondaire de l'ARNt^{phe}.

deux branches d'une multi boucle, ou lorsque la hauteur est nulle, le lien entre des régions séparées de la structure. Une manière plus générale et particulièrement simple de représenter une structure secondaire a été proposée par Ruth Nussinov [54]. Les bases de la séquence sont placées sur un cercle, et chaque paire de bases est représentée par une corde reliant les deux bases appariées 8.1. Sur le même principe, la représentation en graphe place la séquence le long d'une ligne droite et représente les appariements par des arcs de cercles qui relient les parties complémentaires 8.2.

Ces méthodes très simples à implémenter, et très rapides à exécuter, ne sont pas très adaptées pour se faire rapidement une idée sur la structure secondaire de la molécule repliée. Pour ce faire, plusieurs approches proposent à partir du format parenthèse-point de dessiner une structure secondaire comme un circuit complexe qui schématise les motifs de la structure, appelée structure polygonale [73, 8, 25]. Les hélices sont représentées sous forme d'échelles et

les parties simples brins sous forme de cordes à nœuds où chaque nœud représente une base. La représentation polygonale mime l'image obtenue en étalant la structure secondaire d'un ARN dans un plan. La structure polygonale peut s'obtenir à partir de la représentation de Nussinov. En remplaçant chaque corde par un ressort rigide, et en traitant la chaîne de polynucléotides comme infiniment déformable, la structure qui relaxe les contraintes de tension est la représentation polygonale. Les programmes de représentation polygonale s'efforcent de minimiser le nombre d'hélices qui se superposent, en déformant la chaîne ou en modifiant les orientations des hélices dans le plan. L'image obtenue est plus directement instructive sur la structure globale. Parmi l'ensemble des programmes de représentation polygonale, **RNAMovies** [17, 16] propose de plus de chaîner les images produites en une animation très utile pour visualiser la cinétique de repliement de molécules d'ARN. **RNAMovies** est un logiciel sous licence GNU c'est-à-dire, libre de droit d'accès et de modifications annotées du code source. Nous avons donc pu effectuer un certain nombre de modifications pour pouvoir notamment représenter les pseudonœuds. Nous avons aussi développé une version en ligne pour la génération automatique de fichiers images. Ce code est actuellement utilisé par notre serveur de repliement en ligne de molécule d'ARN avec pseudonœuds ***Kinefold***.

8.2 RNAMovies

Initialement **RNAMovies** a été conçu comme un programme de visualisation animé d'espace de structures secondaires de molécules d'ARN sans pseudonœuds. En entrée, il accepte un script contenant les informations sur la structure primaire et les différentes structures secondaires à visualiser au format parenthèse-point "(·)". **RNAMovies** propose toutes les fonctionnalités d'un visualisateur d'animations ; l'animation peut être mise en pause, il est possible d'avancer ou de reculer image par image, une barre de défilement permet d'accéder rapidement aux structures d'intérêt, et la vitesse de défilement est réglable. L'affichage se fait dans une fenêtre gérée par la librairie graphique **OpenGl**. A la souris ou aux menus contextuels, il est possible de déplacer, tourner et zoomer dans la structure courante. Par défaut, le facteur de zoom et de rotation est défini tel que la structure la plus étendue remplisse au mieux la fenêtre courante. **RNAMovies** propose plusieurs options pour l'affichage des structures. Avec ou sans

le nom des résidus. Avec un squelette plus ou moins lissé, d'épaisseur défini, de couleur unie ou peint selon la couleur de l'acide nucléique. Avec ou sans l'affichage des liaisons hydrogènes. Avec ou sans affichage particulier de l'appariement $G = U$. Avec ou sans les indices $5'$ et $3'$ qui donnent l'orientation de la chaîne. Enfin **RNAMovies** permet de sauvegarder les structures secondaires sous différents formats images, soit la structure courante, soit toutes celles qui sont contenues dans un intervalle définit par l'utilisateur.

Le cœur de **RNAMovies**, le calcul de la représentation graphique de la structure, est basé sur l'algorithme **Naview** développé par Bruccoleri RE. et Heinrich G. [8]. **Naview** est un algorithme récursif basé sur une représentation radiale de la structure secondaire. Il conserve les directions des hélices, mais déforme les boucles et le squelette pour éviter que les motifs ne se superposent. **Naview** commence par analyser la structure pour déterminer les connections entre les hélices et les boucles internes et multi branchées. La structure polygonale est dessinée itérativement, de la multiboucle la plus profondément enfuit vers les tiges-boucles terminales, en suivant la connectivité de la structure. Les hélices reliées entre elles au travers d'une boucle multiple sont placées de manière équi angulaire le long d'un cercle représentant la boucle. Puis, la boucle est distendue et déformée pour que les hélices des différentes branches ne se superposent pas.

Entre chaque structure définie dans le script, **RNAMovies** calcule des structures intermédiaires mimant un fondu enchaîné. La figure 8.4 montre un exemple de fondu obtenu sur deux images. Le nombre de structures intermédiaires est réglable. Cette option permet de suivre facilement les changements entre deux structures. Ces intermédiaires ne sont qu'un support visuel, ils ne miment la réalité qu'à condition que les structures visualisées représentent un chemin exhaustif de la cinétique de repliement de la molécule. De tels chemins sont uniquement prédit par la version classique de ***KineFold*** .

Néanmoins, **RNAMovies** ne peut pas afficher correctement des structures avec pseudo-nœuds. Le changement de séquence en cours de script est interdit, que se soit suite à une délétion aux extrémités de la séquence courante ou un changement de séquence.

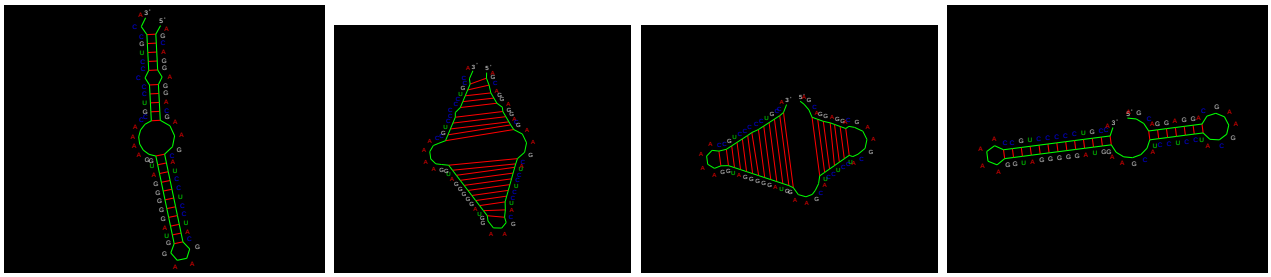


FIG. 8.4 – Exemple de fondu enchaîné sur deux images calculées par **RNAMovies**. Les configurations physiques enregistrées sont présentées aux deux extrémités.

8.3 Affichage des pseudonœuds

Les pseudonœuds forment un motif particulier de la structure secondaire. Il fait intervenir au moins deux hélices dont les séquences complémentaires sont alternées. Schématiquement, il représente l'interaction d'une boucle d'une hélice avec une autre partie simple brin. La figure 8.5 montre les deux principaux types de pseudonœuds rencontrés.

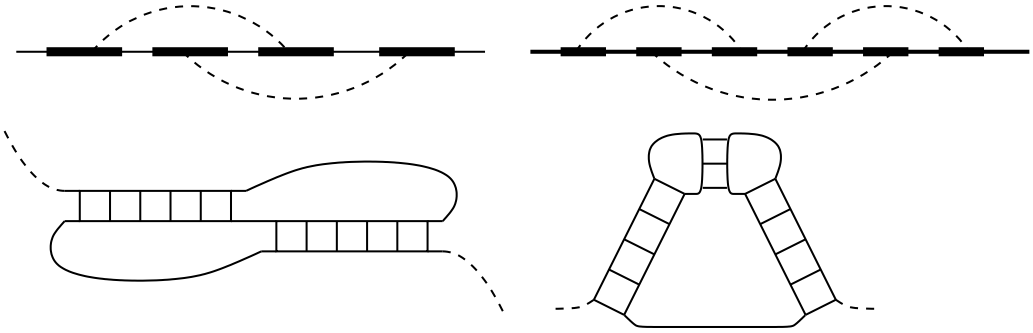


FIG. 8.5 – Les deux types les plus courants de pseudonœuds : les pseudonœuds de type H et I

L'hypothèse sous-jacente de la représentation polygonale est que la structure secondaire suit les règles d'appariement de Waterman définies page 20. La structure obtenue est topologiquement identique à un arbre : les interactions entre branches distinctes sont interdites. Cette définition permet de dessiner des structures qui réalisent les deux contraintes d'esthétisme : les bases appariées doivent se positionner en vis-à-vis et à distance fixe l'une de l'autre, et la

structure ne doit pas présenter de chevauchements d'hélices. En général l'ajout de motifs de pseudonœuds viole la règle de non chevauchement. Pour représenter les structures avec pseudonœuds, la contrainte de distance entre les bases doit être assouplie, pour garantir une image visuelle de la conformation. La manière la plus simple d'opérer est de décider arbitrairement de ne pas contraindre en distance l'ensemble des hélices que l'on associe arbitrairement aux pseudonœuds. Ceci revient à chercher un arbre sous-jacent à la conformation nouée, puis ajouter les hélices qui ferment les pseudonœuds. Le choix de l'hélice qui ferme le motif n'est pas unique, nous avons donc décidé de maximiser l'arbre avec un algorithme de programmation dynamique inclus dans ***KineFold*** . **Naview** calcule ainsi la structure simplifiée puis **RNAMovies** ajoute les appariements complémentaires qui ferment les motifs pseudonœuds. Nous obtenons une image basée sur la représentation polygonale de l'arbre maximum. Les hélices de fermeture du motif pseudoneuds sont schématisées par la coloration des parties complémentaires du squelette reliées par des lignes de rappel. Nous avons modifié **RNAMovies**, pour qu'il reconnaisse deux types de format de codage d'information de structures secondaires avec pseudonœuds. Le premier, limité aux pseudonœuds les plus simples, reprend le format parenthèse-point en y ajoutant le couple de caractère "[", "]", pour définir les appariements des hélices de fermeture de pseudonœuds. Le second, plus général, s'inspire du format d'entrée du programme de visualisation polygonal **RNAviz**. Dans ce format, un numéro unique est associé à chaque hélice. Les deux parties complémentaires d'une hélice sont encadrées par une paire de crochets ouvrant fermant "[", "]" indicée par le numéro de l'hélice correspondante. Nous avons ajouté de l'information supplémentaire délimitée par les caractères spéciaux "#" et "|". Les numéros des hélices de fermeture sont stockés après le caractère spécial "#". Contrairement aux hélices de la structure en arbre, nous avons décidé de ne pas afficher tous les appariements mais seulement les extrémités des hélices pour ne pas surcharger l'image. De plus, pour rendre la lecture de la structure plus claire, pour chaque hélice délocalisée, les rappels sont de couleurs différentes. La figure 8.6 illustre ce cas sur l'exemple d'un intron de groupe I.

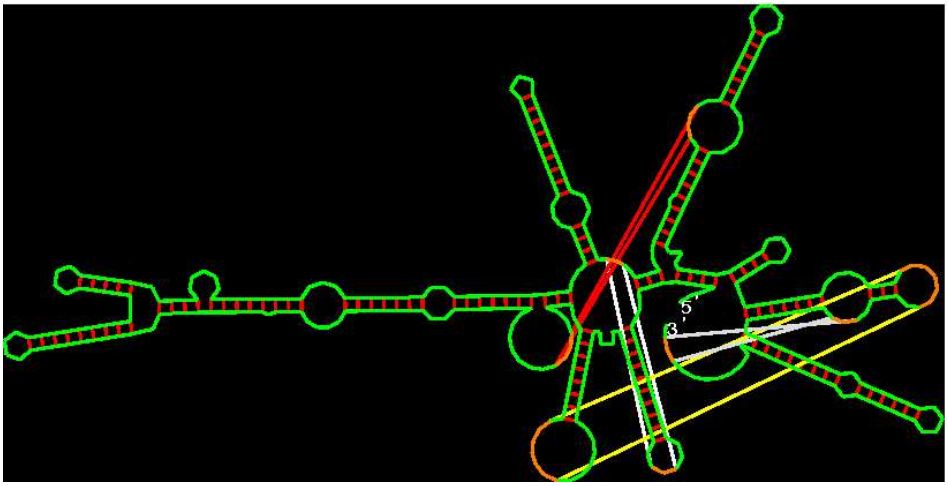


FIG. 8.6 – Affichage avec pseudonœuds sous **RNAMovies**. Chaque hélice délocalisée est symbolisée par les renvois de couleurs délimitant les extrémités de l'hélice. A chaque hélice est associée une couleur différente pour faciliter la compréhension globale de l'image.

8.4 Autres modifications d'intérêt

Nous avons apporté d'autres modifications au code **RNAMovies**. Les plus pertinentes sont la possibilité d'enchaîner plusieurs animations, la possibilité de visualiser des structures dont le nombre de bases décroît, et la possibilité de sauvegarder les images d'intérêts directement en ligne de commande.

Dans sa version initiale, **RNAMovies** ne permet pas de changer de séquences en cours d'animation. Le calcul des structures intermédiaires n'a pas de sens dans ce cas de figure. Pour chaîner plusieurs séquences au cours d'un même script, nous avons remplacé le calcul des structures intermédiaires par la copie de la structure courante. Lors du changement de séquence, la dernière structure courante et affichée autant de fois que nécessaire. Le calcul des structures intermédiaires reprend dès que la séquence est modifiée.

Sur le même principe, il est possible d'afficher des structures dont le nombre de bases décroît au cours du temps. En particulier, il est possible d'afficher des animations de la cinétique de dénaturation mécanique de molécule d'ARN. Au cours de la dénaturation, les bases situées aux extrémités 5' et 3' sont peu à peu débobinées de la conformation repliée. Ces bases ne

présentent aucun intérêt du point de vu structural. En les retirant de la structure visualisée, l'image se focalise sur la partie pertinente de la structure. Nous avons simplement dérivé le calcul du facteur de grossissement pour que l'affichage soit toujours optimum au cours de la visualisation.

Enfin, nous avons permis d'accéder en ligne de commande à toutes les options de sauvegardes accessibles uniquement par menus contextuels. Cette modification est essentielle pour pouvoir piloter **RNAMovies** à l'aide d'un programme annexe. Le gestionnaire de tâches du serveur de repliement ***KineFold*** peut ainsi générer toutes les images des structures secondaires de plus basse énergie. L'utilisation de **RNAMovies** en ligne de commande est explicitée dans le chapitre "serveur ***KineFold*** ".

Chapitre 9

Serveur *KineFold*

Avant-propos

Je remercie Thomas Bucher pour sa participation à la mise en place du serveur ***KineFold*** lors de son stage d'Iut d'informatique. Son aide et ses connaissances pointues ont été précieuses pour aboutir à la fonctionnalité et à la stabilité du serveur actuel. ***KineFold*** est accessible à l'adresse suivante : **http ://kinefold.u-strasbg.fr/**.

9.1 Introduction

La communauté scientifique a toujours eu à cœur de partager avec le plus grand nombre les outils que chacun développe. L'accessibilité et la vitesse de diffusion de l'information sur l'Internet en fait le vecteur principal de la communication actuelle. La facilité avec laquelle tout un chacun peut s'afficher sur le WEB permet à beaucoup de proposer des accès à des bases de données et des outils actifs. La bioinformatique, de fait, n'échappe pas à la règle et il est de mise qu'un travail achevé se concrétise par un site proposant l'accès aux nouveaux services développés. Dans le monde de la prédiction des structures secondaires de molécules d'ARN, quelques minutes sur Internet permettent d'accéder à au moins un exemplaire de code de chaque type d'approche : appariement maximum (*maximum matching*), minimisation d'énergie libre (*mfe*), ou phylogénétique. Le site *mFold* de Michael Zuker [92], avec près de deux cents visites journalières, est sans aucun doute le site le plus connu et le plus utilisé d'entre tous.

Lorsque nous avons décidé d'interfacer ***KineFold*** sur le WEB, deux approches s'offraient à nous : proposer le code source au téléchargement, ou proposer un service totalement en ligne. Dans le premier cas, des connaissances rudimentaires en HTML, ou un programme de conception assisté de page WEB suffisent pour réaliser le site. Mais, dans un service de téléchargement, le concepteur se doit de s'assurer de la compatibilité de son code source avec la majeure partie des compilateurs répandus sur la diversité des systèmes existants. De plus, l'utilisateur final doit connaître un minimum d'administration système pour compiler et installer correctement l'application sur son propre ordinateur. Le service en ligne évite ces complications tout en protégeant la propriété intellectuelle des codes sources. Le programme est installé sur une machine reliée au serveur de page WEB. Toute l'administration et l'interfaçage WEB reste sous le contrôle du concepteur. Pour l'utilisateur, l'outil est immédiatement fonctionnel, et un simple formulaire permet l'interactivité.

La réussite d'un site repose sur deux points essentiels ; en premier, la requête doit être suffisamment simple à formuler et en second les résultats doivent être donnés rapidement tout en étant les plus complets possibles. Du côté de la requête, nous nous sommes inspirés du formulaire du site *mFold* de Michael Zuker. Néanmoins, le temps de calcul de la simulation peut être plus ou moins long selon la longueur et la composition en bases de la séquence de la molécule d'ARN à replier. Nous proposons donc deux types de fonctionnement. Un mode immédiat : un résultat intermédiaire ou définitif est renvoyé au plus tard dix secondes après le début de la simulation. Un mode différé : en fin de simulation, un courriel est automatiquement envoyé à l'utilisateur, avec le lien vers la page de résultat. Du côté des résultats, ***KineFold*** renvoit en page principale, la structure la plus stable rencontrée jusqu'ici au cours du repliement, annotée de son énergie libre, et permet d'accéder de même aux vingt-cinq structures de plus basses énergies visitées. De plus, l'accent est mis sur l'approche cinétique du repliement. Un applet Java permet la visualisation en ligne du film du chemin de repliement, c'est-à-dire la succession des structures intermédiaires visitées. Enfin un outil graphique relate l'évolution de l'énergie libre de la structure la plus stable au cours du temps, ainsi que le taux de formation d'hélices d'intérêts au cours du temps.

Du côté du serveur, les données de chaque utilisateur sont à codifier et sauvegarder pour permettre un accès différé aux résultats. L'interface entre le formulaire et l'application ***Kine-***

Fold se fait au travers d'une succession de petites applications écrites en Perl, en C, et en shell. Le tout est architecturé autour d'un programme de gestion qui orchestre les flux de données et qui gère les files d'attentes.

9.2 Eléments fondateurs du serveur

Le but de l'ensemble des modules du serveur est de mimer la chaîne de commande qui relie la séquence à replier aux différents résultats renvoyés. Notre programme ***KineFold*** prédit la cinétique de repliement des molécules d'ARN et il sauvegarde les structures et les informations d'intérêt dans des fichiers textes. Le fichier de structure, d'extension *rnm*, contient toutes les informations utiles pour l'application **RNAMovies** adaptée pour l'affichage des structures secondaires avec pseudonœuds. A partir de l'application **RNAMovies**, nous pouvons sauvegarder les différentes images des structures au format EPS ou au format de coordonnées de points. Quant au fichier de sortie principale, il contient toutes les informations nécessaires pour tracer les courbes temporelles d'évolution du minimum d'énergie libre de la structure la plus stable courante et des taux de formation des hélices d'intérêt particulier. Les différents modules du serveur, sont installés sur une machine architecturée autour d'un biprocesseur Athlon MP cadencé à 1.2 Ghz et doté de 1 giga-octet de mémoire vive. **Apache** sert de serveur Web et le système d'exploitation choisi est Linux Suse version 7.2, régulièrement mis à jour.

KineFold étant une application gourmande en ressources, il est impératif de gérer efficacement les différentes requêtes soumises au serveur. L'élément central du serveur est le gestionnaire de tâches, il relie l'ensemble des composants entre eux et s'assure de ne pas surcharger les ressources physiques du système.

Afin de faire le lien entre le gestionnaire de tâches, et le site Internet, nous avons développé des petits programmes d'interfaces au format **CGI** -*Common Gateway Interface*-. En général, ces petits programmes font l'interface entre les données récoltées des sites Web et les programmes résidants du serveur. Les **CGI** s'occupent de récupérer les données entrées par les utilisateurs pour les fournir aux applications résidantes sur le serveur et réciproquement. Dans notre cas, un **CGI** écrit en Perl vérifie, et transmet les données entrées dans le formulaire de lancement de ***KineFold*** . Il sécurise aussi l'accès au serveur. A la fin d'une simulation, ces **CGI**

afficheront les résultats sur le navigateur WEB du client et permettront leur téléchargement.

Enfin, pour permettre aux utilisateurs de visualiser à partir d'une page Web la dynamique de repliement, nous avons développé un applet **JAVA** qui clone la fonction d'animation de **RNAMovies**. A l'aide de la sortie *rnm* de ***KineFold*** , **RNAMovies** peut générer deux fichiers de sorties graphiques directement exploitables pour l'affichage : un au format EPS, et l'autre au format de coordonnées de points. L'applet télécharge le fichier de coordonnées et utilise les ressources matérielles du client pour recalculer chaque image et l'afficher.

En fin de chaîne, les résultats de chaque requête sont sauvegardés sur le serveur. Tous les fichiers de résultats sont compressés au format zip et sont proposés au téléchargement. De plus, chaque utilisateur peut accéder directement en ligne aux dernières simulations qu'il a effectué durant une période d'un mois. Son identification se fait automatiquement par la détection du numéro IP de la machine à partir de laquelle la requête est lancée.

9.3 Mise en œuvre

9.3.1 Partie WEB

Afin de répondre aux besoins de simplicité des utilisateurs, nous nous sommes limités à une seule page pour le formulaire de requête. Elle est aussi concise que possible et contient entre autres : un champ de nom de la séquence, un champ de séquence des bases et le choix d'autoriser ou non les pseudonœuds. Chaque champ du formulaire est relié par hyperlien à une aide détaillée.

Les scripts Perl ne font pas seulement office de liens entre le formulaire et le gestionnaire de tâche. Le script traitant le formulaire vérifie aussi l'intégrité et la validité des données entrées par l'utilisateur. Si un ou plusieurs champ ne sont pas renseignés, il prévient l'utilisateur et lui permet de compléter le formulaire. Pour bloquer les requêtes contenant des caractères non autorisés qui peuvent provoquer des failles de sécurité, nous avons décidé d'être relativement restrictif par rapport aux possibilités de saisie dans les champ disponibles. Dans le champ *Nom de la séquence* les caractères autorisés sont limités aux seuls caractères alphanumériques et à l'underscore. En cas de manquement à la règle, l'utilisateur est invité à modifier sa saisie. De

même, la taille des champ a été limitée pour éviter les manœuvres connues d'exécution de codes pirates par dépassement de mémoire. Cette technique consiste à entrer dans un champ une chaîne plus grande que sa place allouée en mémoire. Une partie de la chaîne est alors écrite dans l'espace mémoire du code du programme en cours et permet à l'utilisateur indélicat d'y substituer les commandes préprogrammées par ces propres requêtes.

9.3.2 Le gestionnaire de tâches

Le cœur du serveur est constitué du gestionnaire de tâches. Ce programme écrit en C, attend les requêtes provenant des scripts Perl, lance la chaîne de traitement et renvoie les résultats à l'utilisateur. D'un autre côté, il s'occupe aussi de la gestion des ressources matérielle et gère les files d'attentes des requêtes. A chaque requête, l'ensemble de la chaîne doit être déroulé pour libérer le programme de gestion. Grossièrement, le temps d'immobilisation du processus et de l'ordre du temps de calcul de la simulation. Pour éviter de bloquer le serveur lors des différents traitements, nous avons utilisé la méthode classique de dédoublement de processus. Le gestionnaire principal de tâches, donne naissance à un processus jumeau appelé processus fils, qui hérite de toutes les informations du processus père et prend en charge la fin de l'exécution de la chaîne. En fin de traitement, le processus fils renvoie un signal de fin d'exécution au processus père, qui l'acquitte. Le père se décharge ainsi d'une partie de ses tâches sur ses fils. Le temps d'immobilisation de celui-ci est réduit au strict minimum, quelques dixièmes de secondes. Le serveur apparaît toujours disponible à tous les utilisateurs. Chaque processus fils assure le suivi d'une simulation particulière et informe le processus père de l'avancée du traitement. Finalement, le schéma du gestionnaire est le suivant : un processus père qui s'occupe de recevoir les requêtes envoyées par les scripts Perl et de gérer ses fils (simulations en mode immédiat et différé), des processus fils qui gèrent indépendamment chaque requête, traitent les résultats et les envoient à l'utilisateur. Selon le mode choisi : immédiat ou différé, la gestion des requêtes et légèrement différente. Le schéma général de fonctionnement est donné figure 9.1.

Le repliement peut être effectué selon deux modes différents : le mode immédiat idéal pour replier des séquences courtes ou pour avoir rapidement un résultat intermédiaire, et le mode différé conçu pour la simulation de repliement de molécules complexes. En mode immédiat, la requête est prioritaire et un résultat est renvoyé au plus dix secondes après le début de la

simulation de repliement. La page de résultat est automatiquement générée et affichée dans la fenêtre courante du navigateur WEB du client. Si le temps physique atteint lors de la simulation n'est pas suffisant comparé au temps physique estimé nécessaire pour replier la molécule, l'utilisateur en est averti, et peut choisir de continuer la simulation en mode différé. En mode différé, les requêtes sont enfilées dans une liste d'attente et ne sont lancées que lorsque les ressources du système sont suffisantes. La simulation de la cinétique de repliement est arrêtée après un temps physique déterminé par l'utilisateur ou estimé par notre heuristique. En mode différé, un courriel de notification contenant l'adresse WEB d'un script Perl et le nom de la simulation est envoyé dès la fin de la simulation. Le script récupère le nom de la simulation, génère la page des résultats qui lui correspond et l'affiche dans la fenêtre courante du navigateur WEB du client.

Quel que soit le mode, au maximum 5 simulations sont lancées en parallèle, les autres sont enfilées sur leur liste respective. S'il advient qu'une requête en mode immédiat ne puisse être lancée dans la seconde, l'utilisateur est averti que sa requête à caractère prioritaire est mise en attente.

9.3.3 Visualisateur des résultats

Pour afficher à la fois les structures et le film de la cinétique de repliement de la molécule d'ARN, nous utilisons une version modifiée de **RNAMovies** (cf. "chapitre **RNAMovies**"). Malheureusement, il n'est pas possible d'interfacer directement **RNAMovies** sur le web. Nous avons donc développé une interface écrite en **JAVA** pour simuler les fonctionnalités de **RNAMovies**. Cet applet ne calcule aucune structure, il se sert de la sauvergarde des images des structures calculées par **RNAMovies** au format coordonné de points.

Au cours de la simulation de la cinétique de repliement, ***KineFold*** stocke chaque nouvelle structure d'énergie minimale dans un fichier texte. A partir de ces données brutes, **RNAMovies** réalise une image par structure. Ces images peuvent être sauvegardées sous deux formats : un format graphique, le format EPS, et un format texte contenant les coordonnées de chaque point de chaque image. Dans la version modifiée de **RNAMovies**, nous pouvons effectuer les tâches de sauvegarde directement en ligne de commande.

Nous utilisons la sauvegarde au format EPS pour afficher les images des structures de plus basses énergies. Le gestionnaire de tâches invoque **RNAMovies** puis convertit les images du

volumineux format EPS, vers le format d'images compressées JPEG. Les images obtenues sont insérées par des scripts Perl dans les différentes pages de résultats.

Pour la fonction d'animation, nous utilisons la sauvegarde au format de coordonnées de points. En effet, même compressés, les fichiers graphiques sont très volumineux, or comme ils doivent transiter par le réseau pour arriver chez l'utilisateur, ils utiliseraient une grande partie de notre bande passante. De plus leurs traitements ralentiraient fortement les performances de l'application. Nous avons écrit un applet **JAVA** qui mime les fonctionnalités de **RNAMovies** : il permet de régler la vitesse de défilement de l'animation, d'avancer, de reculer pas à pas, et de modifier le facteur de grossissement. L'applet télécharge le fichier texte, peu volumineux, contenant les coordonnées de chaque élément de chaque structure. A partir de ces informations relatives au format d'affichage de **RNAMovies**, l'applet adapte ces coordonnées en des coordonnées XY relatives à sa fenêtre d'affichage, calcule, puis affiche les différentes structures à la fréquence réglée par l'utilisateur. L'applet apporte plusieurs intérêts par rapport à un programme résidant. Dès lors qu'il est chargé, et qu'il a téléchargé le fichier de coordonnées, son fonctionnement est autonome. La connection avec le serveur peut être coupée, et toutes les ressources CPU et mémoire sont délocalisées sur la machine du client. Les ressources du serveur ne sont donc pas mises à contribution. Deux versions de l'applet ont été développées par Thomas Bucher. La plus classique et écrite en **JAVA** 1.1 et est accessible à partir de n'importe quel navigateur. La plus élaborée utilise la version 1.2 et une bibliothèque spécifique pour l'affichage : la bibliothèque d'objets graphiques pré déclarés **SWING**. Cette seconde version est plus ergonomique, plus rapide mais seuls les navigateurs récents la supportent. Actuellement, nous utilisons la version classique.

9.3.4 Interaction des différents éléments

Après avoir énuméré et expliqué chacun des éléments qui composent le serveur, il est utile de montrer la manière dont ils sont chaînés. Le schéma récapitulatif général (9.1) retrace la chronologie des différentes étapes majeures du fonctionnement du serveur.

– Dans un premier temps, l'utilisateur accède au site ***KineFold*** [91]. Il remplit son formulaire et le valide. Un premier script Perl vérifie et traite les données saisies afin de créer les deux fichiers nécessaires au lancement de la simulation : les fichiers .req et .dat (flèche

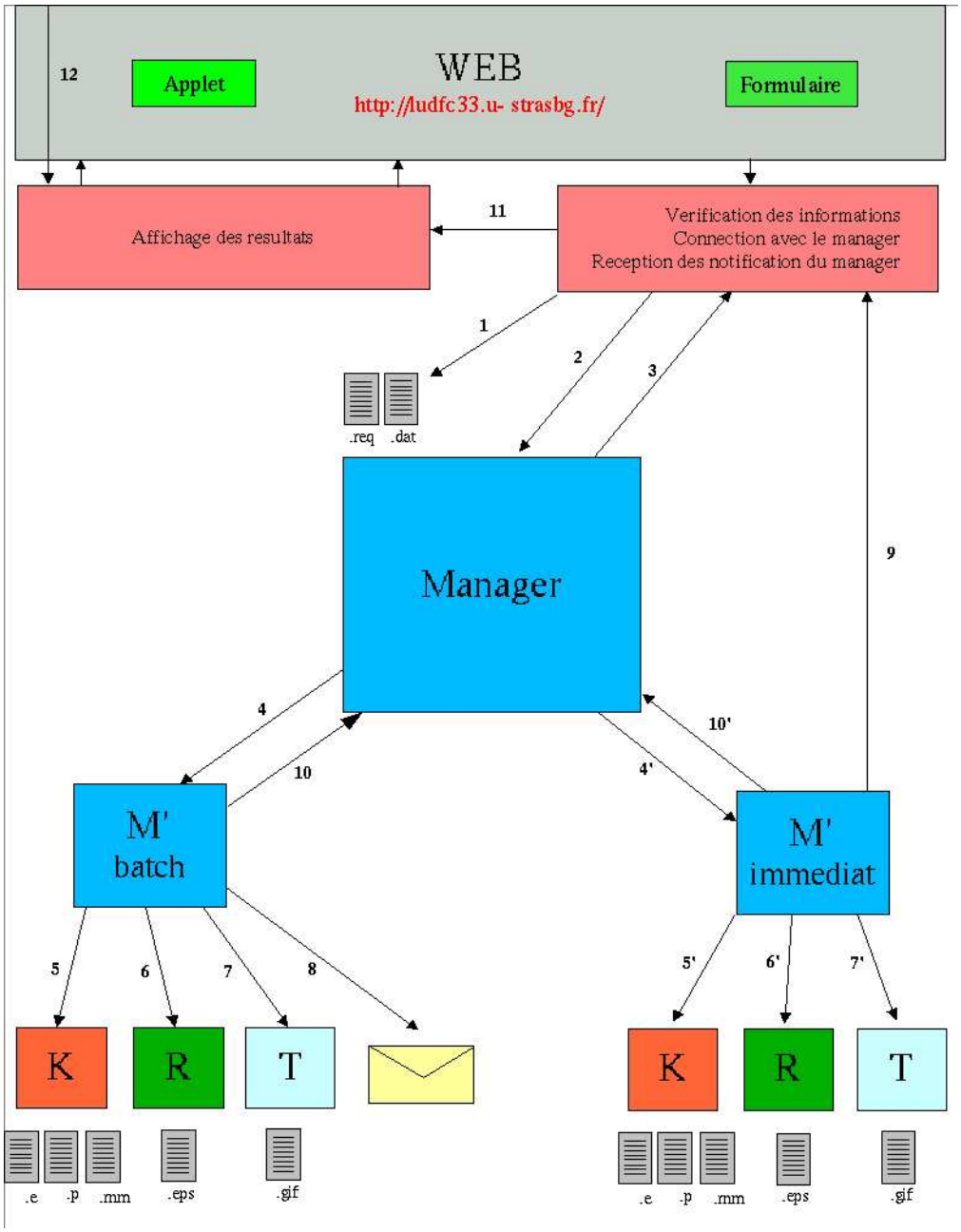


FIG. 9.1 – Schéma général du server. Explication des flèches du schéma dans le texte

1 sur le schéma). En cas d'informations non valides, le script génère une fenêtre reprenant l'ensemble champ à ressaisir. Une fois les fichiers créés, la requête (flèche 2) est envoyée au gestionnaire, appelé *Manager* sur le schéma.

– Le gestionnaire effectue d'autres traitements de vérification, la plus importante étant la vérification des limites de requêtes par utilisateurs. Chaque utilisateur est identifié par son numéro IP, et chaque numéro IP ne peut faire qu'un nombre limité de requêtes en mode immédiat et en mode différé. Cette limite évite la saturation des ressources par un même utilisateur. Le lancement de la simulation dépend de la place disponible. Si le nombre maximum de simulations effectuées en parallèle est atteint, la requête est enfilée. Dans le cas contraire, le processus fils approprié au mode est créé et la file est mise à jour. Dans tous les cas, une notification est renvoyée au script Perl (3).
– Dès qu'une place se libère, la chaîne de traitement est alors prise en charge par le processus fils, qui lance ***KineFold*** (5 et 5'). Ensuite, le processus génère les fichiers nécessaires aux représentations graphiques à l'aide de **RNAMovies** (6 et 6'). Puis effectue des traitements sur les fichiers résultats (7 et 7'). Ces traitements consistent par exemple à insérer des commentaires dans les images proposées aux utilisateurs ou encore à compresser les résultats en archive au format zip.
– En fin de chaîne, selon le type de mode, les processus fils se terminent de manière différente. En mode différé (8), il envoie un courriel de notification à l'utilisateur, contenant le lien vers la page de résultat. En mode immédiat, la notification se fait directement au script Perl, qui redirige automatiquement l'utilisateur vers la page des résultats (9+11). Dans les deux cas, la fin du processus est notifié au père (10 et 10')
– Pour les utilisateurs ayant reçus le courriel de notification (12), le lien écrit dans le corps du message active un script Perl spécifique qui génère la page des résultats. C'est le même script qui est invoqué lorsqu'un utilisateur se connecte au travers de la page des résultats.

9.4 Gestion automatique du serveur

9.4.1 Gestion des erreurs

Lorsqu'une erreur intervient au niveau du processus père, il n'y a pas d'autre choix que de l'arrêter manuellement. Un courriel notifiant cette erreur nous est envoyé, afin que nous puissions intervenir. Etant conscients de cette possibilité, nous avons mis en place un système de sauvegarde automatique des simulations en cours qu'elles soient en attente ou en cours de traitement. Ainsi, lorsque le gestionnaire sera relancé après un "crash", il commencera par relancer en priorité toutes les requêtes précédemment sauvegardées. Malheureusement, lors d'un dysfonctionnement du serveur, nous ne pouvons rien faire pour les requêtes en immédiat, les connections avec les **CGI** étant perdues.

Lorsque ***KineFold*** sort en erreur, le processus fils qui gère cette occurrence en notifie le processus père. Celui-ci tient à jour un compteur d'essais aboutis en erreur pour chaque simulation. Tant que le nombre d'essais ayant abouti en erreur ne dépasse pas un seuil critique (fixé par l'administrateur), la simulation est relancée avec un nombre aléatoire générateur différent. Sinon, l'utilisateur et nous-même sommes avertis que la simulation n'a pas pu aboutir.

9.4.2 Mise à jour des fichiers de résultats sauvegardés

Les fichiers de chaque requête et les fichiers résultats peuvent prendre une place énorme selon leurs nombres et leurs contenus. C'est pourquoi nous avons mis en place un nettoyage automatique du disque. L'opération est simple et ne nécessite pas beaucoup de code. Nous utilisons le petit programme de lancement automatisé de tâche *cron*. A intervalle régulier et défini, il lance la recherche de tous les fichiers définis avec une permission particulière et datant d'un nombre de jours déterminés, et les efface. Toute cette opération est effectuée sous Linux en une seule ligne de commande : "*find /directory/ -mtime 5 -perm 644 -exec rm ;*" écrite dans le fichier *crontab* lu par l'application *cron*. Dans cet exemple, tous les fichiers de permission 644 situés dans le répertoire *directory* et âgés de cinq jours sont automatiquement effacés.

Cinquième partie

Conclusion

Conclusion

La découverte d'ARN catalytiques au début des années 1980 indépendamment par T. Cech et S. Altmann, a ouvert la voie à l'étude de la relation entre structure et fonction pour les molécules d'ARN. Aujourd'hui, on connaît un vaste champ d'application de ces ribozymes intervenant aussi bien dans la maturation d'ARN messagers que dans la régulation phénotypique ou le transport nucléaire.

La fonction d'un ribozyme est intimement liée à sa structure tridimensionnelle. Contrairement aux protéines, les interactions mises en jeu dans le repliement de chaînes d'acides nucléiques sont séparables en deux groupes distincts : les interactions secondaires apportées par la structuration à courte portée des résidus, et les interactions tertiaires essentiellement électrostatiques et médiées par les molécules d'eau ou les ions de la solution. Ce découplage énergétique permet de définir de manière stricte la structure secondaire d'une molécule d'ARN [84]. Cette définition a donné naissance à un ensemble d'algorithme de prédiction *ab initio* de structure secondaire d'ARN [94, 48, 54, 83, 30, 47, 31]. La programmation dynamique appliquée au problème de repliement permet de trouver de manière certaine la meilleure configuration pour les paramètres donnés, et s'avère extrêmement efficace en temps de calcul. Néanmoins, pour des ribozymes présentant des motifs pseudonœuds, ces codes numériques ne prédisent au mieux qu'une proportion variable de bases communes. L'approche de certains motifs pseudonœuds par programmation dynamique s'avère ardue mathématiquement et numériquement peu efficace [68], ou limité au problème d'appariement maximum qui néglige les propriétés de la chaîne structurée [2]. L'approche par la simulation de la cinétique de repliement de la

chaîne polynucléique permet à la fois d'inclure sans heuristique les motifs pseudonœuds et d'étudier les relations entre séquence et structure, ainsi que la dynamique du repliement elle-même [34, 90, 26].

L'approche de la cinétique de relaxation par les méthodes de Monte-Carlo, s'avère souvent peu efficace du fait du piégeage du système dans des cycles d'échange extrêmement rapide entre les différentes structures [90]. La faible probabilité de s'échapper augmente de manière prépondérante le temps de visite de nouveaux états du système. Nous avons montré qu'il était à la fois possible d'intégrer de manière *exacte* cette dynamique sur un ensemble d'états connectés mais aussi d'obtenir les valeurs moyennées de toutes les observables le long du chemin moyen emprunté par le système [90]. La méthode permet de visiter à coup sûr un nouvel état du système à chaque itération. Le nouvel état est ensuite intégré dans l'ensemble des états connus du système et toutes les propriétés du nouveau cluster sont mises à jour. L'utilisation de la relation de *Shermann-Morrison*, nous permet d'obtenir un algorithme d'ordre $O(n^2)$. Idéalement, le facteur d'accélération obtenu, c'est-à-dire le rapport entre le nombre de mouvements effectués sur le nombre de mouvements intégrés, grossit exponentiellement avec la taille du cluster des états visités. Malheureusement, les limitations physiques des ordinateurs actuels ne nous permettent pas d'augmenter indéfiniment la taille de notre ensemble, d'une part, par manque de place mémoire et d'autre part, par le ralentissement observé dû aux calculs de la mise à jour des matrices des propriétés du système. Le facteur d'accélération optimum est séquence dépendant et machine dépendant. Néanmoins, sur un Athlon Mp 1.2Ghz doté de 1Go de mémoire vive et 256 ko de cache second niveau, une taille de cluster de 33 unités est relativement universelle, et permet d'obtenir des facteurs d'accélération moyens de l'ordre de 15.
L'application de ***KineFold*** à l'étude de proportion de pseudonœuds dans des séquences aléatoires courtes, nous a permis de mettre en évidence une proportion non négligeable d'appariements fermant des motifs pseudonœuds. Ce résultat est contraire à l'idée souvent évoquée que les pseudonœuds sont marginaux dans les structures de molécules d'ARN. La comparaison entre les résultats obtenus avec nos ensembles de séquences aléatoires et des séquences génomiques issues des organismes d'*Escherichia Coli* et de *Saccharomyces Cerevisiae* ne fait pas apparaître de différences évidentes, ce qui nous conforte à penser que ce résultat reste applicable au cas

d'ARN messagers. Au-delà de cette constatation, nous nous sommes intéressés à quantifier l'erreur commise en n'intégrant pas les motifs pseudonœuds dans les codes classiques. Nous avons relaxé un ensemble de structures en autorisant ou non la formation de pseudonœuds, à l'aide de ***KineFold*** , et de quantifié la similitude topologique entre les deux séries de repliements, séquence par séquence. Le résultat montre clairement qu'il n'y a aucune relation entre les structures obtenues. Autrement dit, la connaissance de l'une n'indique rien sur la conformation de l'autre. Il nous semble ainsi difficile de découpler les motifs pseudonœuds des interactions secondaires.

Dans une seconde partie de ce travail, nous nous sommes intéressés plus spécifiquement à la relation entre séquence et structure. En particulier, nous avons étudié la possibilité de coder de l'information de type cinétique le long de la séquence des bases, en utilisant un processus hypothétique de **switch** contrôlé par le repliement co-transcriptionnel. La mise en évidence de réseaux neutres de mutation qui permettent à une séquence d'évoluer tout en conservant la faculté d'exprimer une fonction particulière est à la fois un outil d'étude de l'évolution génétique des séquences et une preuve patente de la plasticité des organismes face à une sollicitation extérieure. Mais le maintien de la fonction chez un grand nombre de séquences différentes induit aussi que cette information essentielle n'utilise pas tous les degrés de liberté offerts par la succession déterminée des nucléotides de la séquence. Il y a donc de la place pour coder de l'information d'un autre type en particulier cinétique. Les travaux de Tao Pan [60] ont directement prouvé cette hypothèse sur le ribozyme de la RNAseP. A l'aide de nos observations faites sur l'importance des structures intermédiaires dans le repliement de molécules d'ARN [26] et les résultats sur les vitesses de repliements de mutants [57, 61, 69], nous avons construit une méthode de guidage de la cinétique tirant partie de la compétition cinétique entre des hélices intermédiaires. Pour prouver notre hypothèse, nous avons dessiné deux molécules bistables composées deux à deux des mêmes hélices palindromiques. Au cours de la synthèse, la compétition entre les hélices incompatibles guide les molécules vers l'une ou l'autre des deux configurations. Cette hypothèse a été testée préalablement numériquement à l'aide de ***KineFold*** . Expérimentalement, nous avons dû modifier légèrement les séquences initiales pour pouvoir discriminer sur gel d'électrophorèse l'existence des deux configurations distinctes obte-

nues après transcription ainsi que le caractère bistable des molécules après relaxation. A l'aide de cette étude expérimentale, nous avons pu prouver qu'il était possible de coder de l'information de type cinétique le long d'une séquence. De plus, nous avons vérifié expérimentalement l'hypothèse de repliement co-transcriptionnel et nous avons mis en évidence la probable existence d'un processus général de **switch** contrôlé activé par l'existence d'hélices transitoires. Cette étude ouvre la voie à la stylistique de molécules cinétiquement controlées.

Pour faire ce travail, nous avons dû modifier un programme libre de visualisation de structures secondaires d'ARN, **RNAMovies**, pour lui faire entre autre afficher des structures peudonouées. Enfin ***KineFold*** et **RNAMovies** ont été adaptés sur le web pour permettre à l'ensemble de la communauté scientifique de pouvoir tester notre approche en complément des autres outils mis à leur disposition.

Bibliographie

[1] Enzime ressource guide. Technical report, promega.

[2] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Dis. Appl. Math.*, 104 :45–62, 2000.

[3] R.C. Allen and B. Budowle. *Gel electrophoresis of proteins and nucleic acids*. W. de gruyter edition, 1994.

[4] AR. Banerjee, JA. Jaeger, and DH. Turner. Thermal unfolding of a group I ribozyme : the low-temperature transition is primarily disruption of tertiary structure. *Biochemistry.*, 32 :153–163, 1993.

[5] BA. Berg. Introduction to multicanonical monte carlo simulations. *cond-mat/9909236*, 1999.

[6] C.K. Biebricher and Luce R. In vitro recombination and terminal elongation of rna by q beta replicase. *EMBO J.*, 11 :5129–35, 1992.

[7] I. Brierley, AJ. Jenner, and SC. Inglis. Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J Mol Biol.*, 220 :889–902, 1991.

[8] RE. Bruccoleri and G. Heinrich. An improved algorithm for nucleic acid secondary structure display. *Comput Appl Biosci.*, 4 :167–173, 1988.

[9] S. Buchner and A. Heuer. Potential energy landscape of a model glass former : thermodynamics, anharmonicities, and finite size effects. *Phys Rev E*, 60 :6507–6518, 1999.

[10] M. Chamorro, N. Parkin, and HE. Varmus. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc Natl Acad Sci.*, 89 :713–717, 1992.

[11] SN. Cohen, AC. Chang, and L. Hsu. Nonchromosomal antibiotic resistance in bacteria : genetic transformation of *escherichia coli* by r-factor DNA. *Proc Natl Acad Sci USA.*, 69 :2110–2114, 1972.

[12] PS. Doyle, J. Bibette, A. Bancaud, and Viovy JL. Self-assembled magnetic matrices for DNA separation chips. *Science.*, 295 :2237, 2002.

[13] DE. Draper. Strategies for RNA folding. *Trends Biochem Sci.*, 21 :145–149, 1996.

[14] VL. Emerick and SA. Woodson. Fingerprinting the folding of a group I precursor RNA. *Proc Natl Acad Sci.*, 91 :9675–9679, 1994.

[15] Ausubel et al. *Current Protocols in Molecular Biology*, volume 1. John willey & sons edition, 1997.

[16] D. Evers and R. Giegerich. http ://bibiserv.techfak.uni-bielefeld.de/rnamovies/.

[17] D. Evers and R. Giegerich. RNA movies : visualizing RNA secondary structure spaces. *Bioinformatics.*, 15 :32–37, 1999.

[18] AR. Ferre-D'Amare, K. Zhou, and JA. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature.*, 395 :567–574, 1998.

[19] DR. Gallie, JN. Feder, RT. Schimke, and V. Walbot. Functional analysis of the tobacco mosaic virus tRNA-like structure in cytoplasmic gene regulation. *Nucleic Acids Res.*, 19 :5031–5036, 1991.

[20] J. Gelas, C. Rijk, and B. Neal. Opteron : Pushing x86 to the limit. Technical report, www.aceshardware.com/read.jsp?id=55000262, 2003.

[21] U. Gerland, R. Bundschuh, and T. Hwa. Mechanically probing the folding pathway of single RNA molecules. *Biophys J.*, 84 :2831–2840, 2003.

[22] Atkins JF. Cech TR Gesteland, RF. *The RNA World*. Cold spring harbor laboratory press edition, 1999.

[23] RJ. Grainger, AI. Murchie, DG. Norman, and DM. Lilley. Severe axial bending of RNA induced by the U1A binding element present in the 3' untranslated region of the U1A mRNA. *J Mol Biol.*, 273 :84–92, 1997.

[24] ES. Haas, DP. Morse, JW. Brown, FJ. Schmidt, and NR. Pace. Long-range structure in ribonuclease P RNA. *Science*, 254 :853–856, 1991.

[25] K. Han, D. Kim, and HJ. Kim. A vector-based method for drawing RNA secondary structure. *Bioinformatics.*, 15 :286–297, 1999.

[26] S. Harlepp, T. Marchal, J. Robert, JF. Leger, A. Xayaphoummine, H. Isambert, and D. Chatenay. Probing complex RNA structures by mechanical force. *Eur Phys J E*, 12 :605–615, 2003.

[27] C. Haslinger. *Prediction algorithms for restricted RNA pseudoknots*. PhD thesis, universität Wien, 2001.

[28] C. Haslinger and PF. Stadler. RNA structures with pseudo-knots : Graph-theoretical and combinatorial properties. *Bull. Math. Biol.*, 61 :437–467, 1999.

[29] PG. Higgs. Overlaps between RNA secondary structures. *Phys. Rev. Lett.*, 76 :704–707, 1996.

[30] PG. Higgs. RNA secondary structure : physical and computational aspects. *Q Rev Biophys*, 33 :199–253, 2000.

[31] IL. Hofacker, W. Fontana, PF. Stadler, LS. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125 :167–188, 1994.

[32] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic Acids Res*, 12 :67–74, 1984.

[33] S. Hu and N. J. ; Dovichi. Capillary electrophoresis for the analysis of biopolymers. *Anal. Chem.*, 74 :2833–2850, 2002.

[34] H. Isambert and ED. Siggia. Modeling RNA folding paths with pseudoknots : application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci.*, 97 :6515–6520, 2000.

[35] BD. James, GJ. Olsen, and NR. Pace. Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol.*, 180 :227–239, 1989.

[36] V. Juan and C. Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J Mol Biol.*, 289 :935–947, 1999.

[37] DA. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure similarity and consensus of minimal-energy folding. *J Mol Biol.*, 207 :597–614, 1989.

[38] LG. Laing and DE. Draper. Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J Mol Biol.*, 237 :560–576, 1994.

[39] LS. Lerman and HL. Frisch. Why does the electrophoretic mobility of DNA in gels vary with the length of the molecule ? *Biopolymers.*, 21 :995–997, 1982.

[40] F. Lisacek, Y. Diaz, and F. Michel. Automatic identification of group I intron cores in genomic DNA sequences. *J Mol Biol.*, 235 :1206–1217, 1994.

[41] A. Loria and T. Pan. Recognition of the T stem-loop of a pre-tRNA substrate by the ribozyme from Bacillus subtilis ribonuclease P. *Biochemistry.*, 36 :6317–25, 1997.

[42] DK. Lubensky and DR. Nelson. Single molecule statistics and the polynucleotide unzipping transition. *Phys Rev E*, 65 :031917(25pages), 2002.

[43] OJ. Lumpkin. Mobility of DNA in gel electrophoresis. *Biopolymers.*, 21 :2315–2316, 1982.

[44] M. Mandel and A. Higa. Calcium-dependent bacteriophage DNA infection. *J Mol Biol.*, 53 :154–159, 1970. reprint in Biotechnology(1992)24 :198-201.

[45] RM. Mans, CW. Pleij, and L. Bosch. tRNA-like structures. structure, function and evolutionary significance. *Eur J Biochem*, 201 :303–24, 1991.

[46] HM. Martinez. An RNA folding rule. *Nucleic Acids Res.*, 12 :323–34, 1984.

[47] DH. Mathews, J. Sabina, M. Zuker, and DH. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol.*, 288 :911–940, 1999.

[48] JS. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29 :1105–19, 1990.

[49] MB. Michael. *Enzymes of molecular biology*, volume 16 of *methods in molecular biology*. Humana press edition, 1993.

[50] A. Montanari and M. Mezard. Hairpin formation and elongation of biomolecules. *Phys Rev Lett*, 86 :2178–2181, 2001.

[51] AJ. Nataraj, I. Olivos-Glander, N. Kusukawa, and WE. Highsmith, Jr. Single-strand conformation polymorphism and heteroduplex analysis for gel-based mutation detection. *Electrophoresis.*, 20 :1177–1185, 1999.

[52] J. Nowakowski and I. Tinoco, Jr. RNA structure and stability. *Seminars in Virology.*, 8 :153–165, 1997.

[53] R. Nussinov and AB. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci.*, 77 :6309–13, 1980.

[54] R. Nussinov, G. Pieczenik, JR. Griggs, and DJ. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35 :68–82, 1978.

[55] AG. Ogston. The spaces in a uniform random suspension of fibres. *Trans. Faraday Soc.*, 54 :1754–1757, 1958.

[56] M. Orita, H. Iwahana, H. Kanazawa, K. Hayashi, and T. Sekiya. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci USA.*, 86 :2766–2770, 1989.

[57] J. Pan, ML. Deras, and SA. Woodson. Fast folding of a ribozyme by stabilizing core interactions : evidence for multiple folding pathways in RNA. *J Mol Biol.*, 296 :133–144, 2000. et références incluses.

[58] J. Pan and SA. Woodson. Folding intermediates of a self-splicing RNA : mispairing of the catalytic core. *J Mol Biol.*, 280 :597–609, 1998.

[59] T. Pan. How RNA function requires structure. *Nat. Struc. biol.*, 5 :540–541, 1998.

[60] T. Pan, X. Fang, and T. Sosnick. Pathway modulation, circular permutation and rapid RNA folding under kinetic control. *J Mol Biol.*, 286 :721–731, 1999.

[61] T. Pan and TR. Sosnick. Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and uv absorbance spectroscopies and catalytic activity. *Nat Struct Biol.*, 4 :931–938, 1997.

[62] Thirumalai D. Pan, J. and and SA. Woodson. Folding of RNA involves parallel pathways. *J Mol Biol.*, 273 :7–13, 1997.

[63] JM. Pipas and JE. McMahon. Method for predicting RNA secondary structure. *Proc Natl Acad Sci.*, 72 :2017–21, 1975.

[64] CW. Pleij, K. Rietveld, and L. Bosch. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Res.*, 13 :1717–1731, 1985.

[65] A. Pluen. *Mécanisme de l'électrophorèse de l'ADN en simple brin sur gel de polyacrylamide et sur gels à matrice mixte.* PhD thesis, ULP.

[66] AM. Pyle, JA. McSwiggen, and TR. Cech. Direct measurement of oligonucleotide substrate binding to wild-type and mutant ribozymes from tetrahymena. *Proc Natl Acad Sci USA.*, 87 :8187–8191, 1990.

[67] C. Reidys, PF. Stadler, and P. Schuster. Generic properties of combinatory maps : neutral networks of RNA secondary structures. *Bull Math Biol.*, 59 :339–397, 1997.

[68] E. Rivas and SR. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol.*, 285 :2053–2068, 1999.

[69] MS. Rook, DK. Treiber, and JR. Williamson. Fast folding mutants of the tetrahymena group I ribozyme reveal a rugged folding energy landscape. *J Mol Biol.*, 281 :609–620, 1998.

[70] W. Saenger. *Principles of nucleic acid structure.* Springer-verlag edition, 1984.

[71] J. Sambrook, EF. Fritsch, and T. Maniatis. *Molecular Cloning*, volume 1. Cold spring harbor laboratory press edition, 1989.

[72] EA. Schultes and DP. Bartel. One sequence, two ribozymes : implications for the emergence of new ribozyme folds. *Science.*, 289 :448–452, 2000.

[73] BA. Shapiro, J. Maizel, LE. Lipkin, K. Currey, and C. Whitney. Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic Acids Res.*, 12 :75–88, 1984.

[74] VC. Sheffield, JS. Beck, AE. Kwitek, DW. Sandstrom, and EM. Stone. The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. *Genomics.*, 16 :325–332, 1993.

[75] G.W. Slater, S. Guillouzic, Gauthier M.G., J.F. Mercier, Kenward M., L.C. McCormick, and Tessier F. Theory of DNA electrophoresis. *Electrophoresis*, 23 :3791–3816, 2002.

[76] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Q Rev Biophys.*, 4 :213–253, 1971.

[77] P. Strippoli, S. Sarchielli, R. Santucci, GP. Bagnara, G. Brandi, and G. Biasco. Cold single-strand conformation polymorphism analysis : optimization for detection of APC gene mutations in patients with familial adenomatous polyposis. *Int J Mol Med*, 8 :567–572, 2001.

[78] M. Tacker, PF. Stadler, EG. Bornberg-Baeur, IL. Hofacker, and P. Schuster. Algorithm independant properties of RNA secondary structure predictions. *Eur. Biophys. J.*, 25 :115–30, 1996.

[79] SA. Teukolsky, WT. Veterling, and BP. Flannery. *Numerical recipes*, volume 2nd Ed. University press, cambridge edition, 1992.

[80] I. Tinoco, Jr. and C. Bustamante. How RNA folds. *J Mol Biol.*, 293 :271–281, 1999.

[81] DK. Treiber, MS. Rook, PP. Zarrinkar, and JR. Williamson. Kinetic intermediates trapped by native interactions in RNA folding. *Science.*, 279 :1943–1946, 1998.

[82] JL. Viovy. Electrophoresis of DNA and other polyelectrolytes : Physical mechanisms. *Rev. Mod. Phys.*, 72 :813–872, 2000.

[83] MS. Waterman. Secondary structure of single stranded nucleic acids. *Adv. Math. Suppl. Stud.*, 1 :167–212, 1978.

[84] MS. Waterman and TF. Smith. RNA secondary structure : A complete mathematical analysis. *Math. Biosci.*, 42 :257–266, 1978.

[85] CJ. Weitzmann, PR. Cunningham, and J. Ofengand. Cloning in vitro transcription and biological activity of *escherichia coli* 23S ribosomal RNA. *Nucleic Acids Res.*, 18 :3515–3520, 1990.

[86] E. Westhof and V. Fritsch. RNA folding : beyond watson-crick pairs. *Structure Fold Des.*, 8 :55–65, 2000.

[87] NM. Wills, RF. Gesteland, and JF. Atkins. Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus gag stop codon. *Proc Natl Acad Sci.*, 88 :6991–6995, 1991.

[88] S. Wuchty, W. Fontana, IL. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers.*, 49 :145–165, 1999.

[89] S. Wutchy. Suboptimal secondary structures of RNA. Master's thesis, universität Wien, 1998.

[90] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc Natl Acad Sci.*, 100 :15310–15315, 2003.

[91] A. Xayaphoummine, T. Buchert, and H. Isambert. http ://kinefold.u-strabg.fr.

[92] M. Zuker. http ://www.bioinfo.rpi.edu/applications/mfold/.

[93] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science.*, 244 :48–52, 1989.

[94] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46 :591–621, 1984.

[95] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9 :133–48, 1981.