Université de Montréal

PRÉDICTION DE LA STRUCTURE COMMUNE AUX ARN MESSAGERS CODANT POUR LA PROTÉINE STG

Par
Ratiba TERBAOUI

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maître ès sciences en informatique (M.Sc.)

Décembre 2004

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

PREDICTION DE LA STRUCTURE COMMUNE AUX ARN MESSAGERS CODANT POUR LA PROTEINE STG

Présenté par :
Ratiba TERBAOUI

a été évalué par un jury composé des personnes suivantes :

El Mostapha Aboulhamid
Président rapporteur

François Major
Directeur de recherche

Miklós Csũrős
Membre du jury

Mémoire accepté le : 30/03/05

Sommaire

Avec la technique de microdissection au laser, l'équipe du Dr Mauricio Neira a trouvé un nouveau gène dans les papilles gustatives du macaque rhésus. Cette technique permet d'extraire des cellules spécifiques dans les tissus sans détruire la morphologie et la structure de la molécule. La même séquence a été retrouvée dans des régions reliées à la maladie du psoriasis vulgaris.

De nouvelles séquences ont été rajoutées aux bases de données sans que leur structure soit connue. Le but de ce mémoire est de trouver une structure commune à ces nouvelles séquences pour connaître la fonction appropriée et cela en tentant de prédire les interactions avec les autres molécules d'ARN ou entre éléments de la même molécule ou bien avec des protéines. Vu les limitations de la cristallographie et de la résonance nucléaire moléculaire dans la résolution de structure, des outils informatiques seront utilisés pour simuler des effets biologiques réels.

Ce mémoire va décrire le chemin qui mènera à une structure tridimensionnelle possible. De la structure primaire à la tertiaire des programmes dédiés à la prédiction de structure secondaire, à leur visualisation ainsi qu'à leur construction en 3D.

Le processus de recherche est présenté de manière hiérarchique en cinq chapitres comme il suit :

- Le premier couvre l'état-de-l'art du domaine de la structure.
- Le second chapitre entame la recherche de la structure secondaire en exposant les problèmes apparus lors des exécutions et cela pour chacune des hypothèses.
- Le troisième chapitre montre le modèle 3D avec les sites d'interactions possibles avec d'autres molécules ou d'autres protéines. Ce modèle sera exposé à un champ de force pour étudier sa stabilité.
- Le dernier chapitre va suggérer des améliorations à appliquer aux programmes pour tenir compte du polymorphisme observé dans les séquences d'ARN messagers.

MOTS CLES : Prédiction de structure secondaire, polymorphisme, recherche de motifs, structure tertiaire.

Summary

By using the LCM technique, Dr Mauricio Neira's team found a new rhesus monkey gene expressed specifically in taste cells. The laser capture microdissection makes it possible to extract specific cells from tissue by laser without destroying the morphology and the structure of molecules. The same sequence was localized in human and mouse species parts related to psoriasis vulgaris disease.

The new sequences were added to databases without knowing their structures. The goal of this thesis is to find a common structure of these new sequences. In addition we will try to predict their appropriate function. This will be done by trying to predict interactions with other RNA molecules or proteins, including between elements of the same RNA messenger molecule.

Due to X-Ray crystallography and NMR limitations in structure resolution, software tools will be used to simulate real biological effects.

This thesis will describe the path that leads to a possible three dimensional structure. From the primary structure to the tertiary, we will use software dedicated to predict secondary structure, to visualize them and to build them in 3D space

The research's process is presented hierarchically in four different chapters as follows:

- The first chapter covers the state-of-the-art in the structure field.
- The second chapter shows the 2D structure research with the problems encountered by using the programs for each hypothesis
- The third chapter shows a 3D model by showing the interaction sites with other molecules or proteins. This model is finally exposed to a force field to study its stability in the fourth chapter.
- The last chapter will suggest some future improvements to apply to the programs used, to handle the polymorphism observed in the new RNA messengers.

KEYWORDS: Secondary structure prediction, polymorphism, motif's research, tertiary structure.

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

Chapitre 1

Généralités

1.1. Exposition du problème

La découverte de la protéine STG a été faite par Mauricio Neira en 2001 [1]. Cette protéine est localisée au niveau des papilles gustatives du macaque rhésus (Identificateur séquence ARN messager: AF245204), une espèce de singe qui partage un grand pourcentage de son code génétique avec l'espèce humaine. La même séquence se retrouve également chez l'homme (Identificateur séquence ARN messager: AB031481), sur le chromosome 6p21, partie la plus susceptible d'être liée au psoriasis vulgaris et plus précisément les régions entourant le gène HLA-C (dans un intervalle de 111kb allant de 80 à 200kb du télomère du gène [4]. Par homologie partielle, la séquence est également retrouvée chez la souris et plus précisément dans la région du complexe majeur d'histocompatibilité contenant la région Q des gènes de classe I (Identificateur séquence ARN messager: AF111103). Les particularités de ces séquences sont l'occurrence successive d'une sous séquence de 11 ou 13 acides aminés, selon l'espèce, ainsi que leur similarité avec la protéine du prion (voir figure 1-1). Ce polymorphisme observé au niveau des protéines est également observé au niveau de leurs séquences d'ARN messagers respectives.

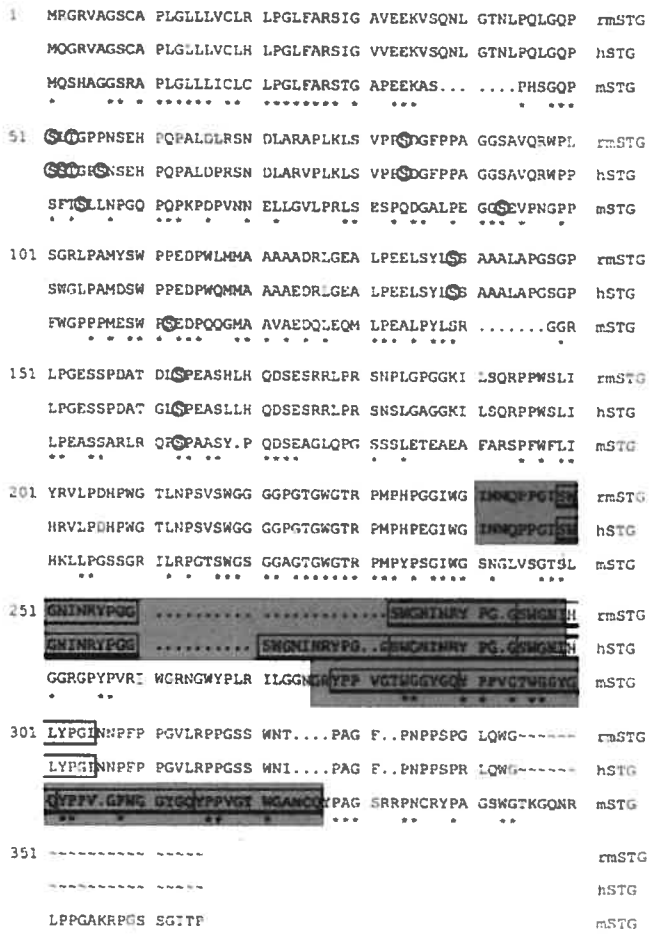


Figure 1- 1: Alignement entre les protéines STG des trois espéces : macaque rhésus (rmSTG), humain (hSTG) et la souris (mSTG). Les acides aminés encerclés Sérine et Thréonine sont les sites de O-glycosilation. Les régions grises montrent la similarité avec le prion. Les rectangles mettent en évidence les sous séquences répétées. Les astérisques indiquent les acides aminés communs aux trois protéines.

En observant l'alignement de la figure 1-1, on voit clairement la séquence SWGNINRYPGG qui se répète deux fois chez le macaque rhésus et trois fois chez l'humain. Au niveau de la souris la séquence YPPVGTWGGYGQ de douze acides aminés se répète quatre fois. Ce qui est aussi perceptible de la figure 1-1 est que pour chacune de ces trois séquences, au niveau du dernier rectangle gris, on retrouve les cinq

acides aminés de la séquence répétée. Autrement dit, on retrouve SWGNI pour le macaque rhésus et l'humain, et les cinq premiers acides aminés YPPVG pour la souris.

Le polymorphisme observé dans notre cas est la répétition d'une même séquence un nombre de fois qui diffère selon l'espèce. Nous faisons l'hypothèse que cette répétition pourrait être fonctionnelle et affecter la régulation du gène, ou de la séquence de la protéine étant donné sa localisation dans la partie codante de l'ARN messager. Cet effet pourrait causer ou contribuer au développement de certaines maladies.

1.2. Les papilles gustatives

Etant donné que le gène trouvé est en relation avec le sens du goût, ce paragraphe va introduire brièvement le mécanisme impliqué dans sa transmission. Longtemps l'idée que des parties de la langue étaient dédiées uniquement à un des quatre stimuli de base: l'acidulé, l'amer, le sucré et le salé. En fait ces différentes saveurs peuvent être senties par toute la surface de la langue, mais par contre chaque cellule gustative possède un seuil de sensibilité à chacune des saveurs et selon son affinité pour la substance, elle transmettra au cerveau un stimulus de différente intensité.

La perception des différentes saveurs est possible en partie grâce aux papilles gustatives. Ces organes sensoriels qu'on retrouve sur la langue, sur le palet de la bouche, la trachée ainsi que la partie supérieure de l'œsophage. Les papilles gustatives sont à leur tour formées de bourgeons gustatifs et chacun possède 100 récepteurs qui sont renouvelés toutes les deux semaines. En moyenne une personne normale possède 10000 bourgeons gustatifs.

Il y a trois sortes de papilles gustatives :

- les papilles caliciformes, au nombre de neuf à douze, formant le V lingual
- les papilles fongiformes, petites et nombreuses, situées en avant du V et le long de ses branches.
- Les papilles foliées sur le côté proche du V lingual.

Les papilles caliciformes, fongiformes et foliées renferment des bourgeons gustatifs en relation avec le nerf du goût. De même il existe deux sortes de papilles tactiles : les papilles filiformes (au centre de la langue) et corolliformes.

Chaque cellule réceptrice a une forme unique, qui répond à un type de signal chimique donné. Les signaux émis par les récepteurs gustatifs sont véhiculés principalement par trois nerfs crâniens. Ils sont transmis au système nerveux central, où des régions du cerveau décodent l'information chimique et la traduisent en sensation gustative (voir figure 1-2).

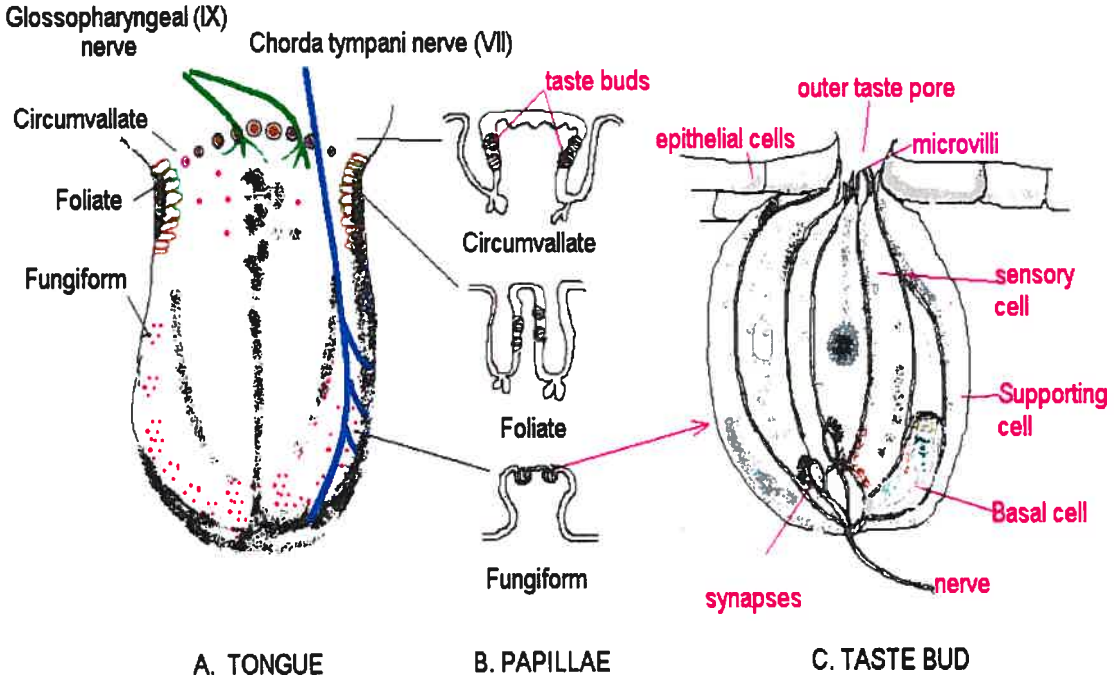


Figure 1-2 : Langue et papilles gustatives : La section A localise les 3 papilles gustatives sur la langue avec les nerfs de transmission du goût: caliciformes sur le V-lingual(cercle marron), foliée sur le cote proche du V-lingual (marron), fongiforme sur le côté (point rouge). La section B représente chaque bourgeon gustatif selon la forme de papille. La section C montre la composition d'un bourgeon gustatif des trois papilles. Droit d'utilisation accordé par l'auteur Tim Jacob.

1.3. Le système HLA

La présence des gènes dans la région reliée au système immunitaire mérite une brève introduction par ce paragraphe. On entend souvent parler de rejet suite à une transplantation d'un organe tel que le rein, le cœur ou autre. Ces rejets impliquent le système HLA [5], la version humaine du CMH, le complexe d'histocompatibilité majeur, qui est un élément fondamental parmi d'autres dans l'initiation et le développement du système immunitaire. La région du Complexe Majeur d'histocompatibilité se retrouve sur le bras court du chromosome 6 (6p21.31), région avec plus de 220 gènes dont environ 40 codent pour des glycoprotéines membranaires et ont une capacité d'induire une forte réponse immunitaire allogénique. Le chromosome 6 de l'humain a été complètement séquencé[6] et représente 6% du génome humain avec une taille de 166, 880, 988 paires de bases.

Les gènes du système sont divisés en deux classes I et II qui sont différents du point de vue structural et fonctionnel. Dans le cadre de ce projet ceux de la classe I uniquement sont concernés vu que la séquence de la protéine STG (humain et souris) est retrouvée dans cette catégorie. Les gènes de la classe I sont exprimés à la surface de toutes les cellules nucléées de l'organisme. Les antigènes majeurs d'histocompatibilité sont codés par trois gènes: HLA-A, HLA-B, HLA-C (voir figure 1-3). L'antigène HLA-Cw6 est associé à la maladie du psoriasis vulgaris, une maladie de la peau qui se manifeste par une irritation accentuée de l'épiderme à n'importe quel endroit du corps. La cause de cette maladie est inconnue, le traitement de la maladie par les dermatologues se fait par prescription de cortisone sous différentes formes, de la vitamine D, une exposition au soleil aide les patients, le but étant de soulager l'irritation de la peau et le contrôle de sa mue.

L'importance du système HLA a été démontrée par les résultats des transplantations réalisées à partir de donneurs vivants apparentés, la survie du greffon

étant significativement supérieure chez les sujets identiques par rapport aux sujets semi identiques. Dans les cas de rejets aigus chez des sujets identiques, il est suggéré le rôle d'autres systèmes antigéniques, en particulier ceux portés par les cellules endothéliales dans les vaisseaux sanguins du rein et par les monocytes qui sont les plus grosses globules blanches du système immunitaire et se trouvent en forte concentration au niveau de la rate.

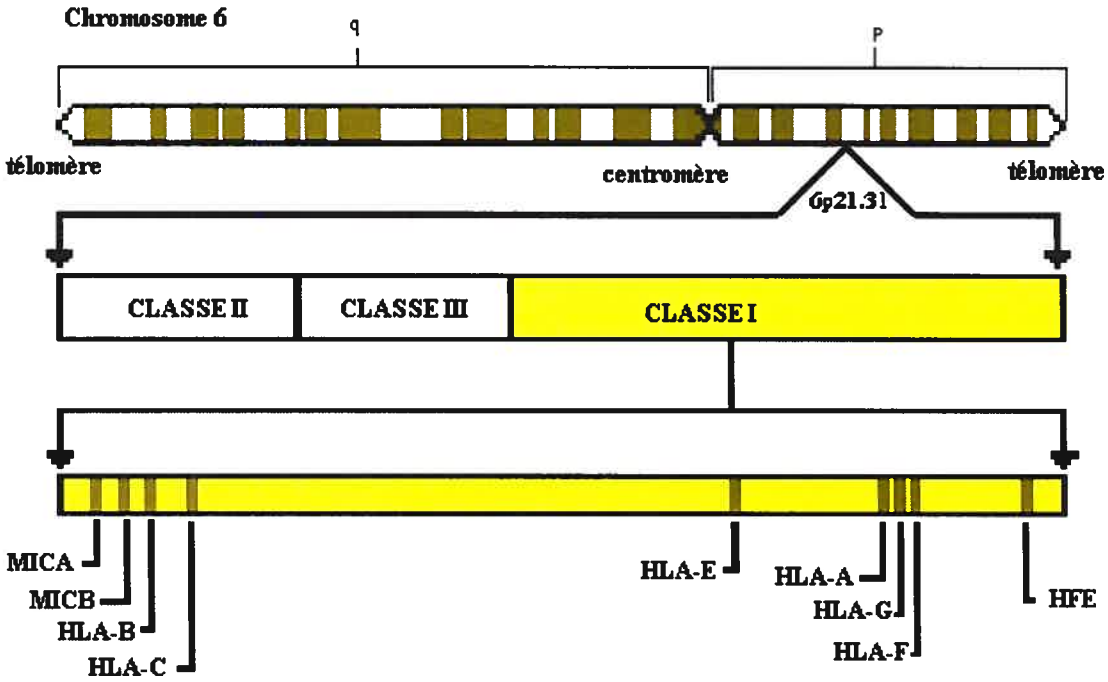


Figure 1-3 : Organisation du système HLA sur le chromosome 6. Le bras long (q) est à gauche du centromère. Le bras court (p) est à droite du centromère. Les extrémités du chromosome 6 sont les deux télomères. Sur le bras court (6p21.31) on retrouve les trois classes I, II et III du système immunitaire. Les gènes de la classe I sont représentés par des bandes vertes avec leur loci.

1.4. La molécule d'ARN et son repliement

La molécule d'ADN est, selon le dogme central, porteuse du patrimoine génétique. Une copie complémentaire d'un de ces brins produit la molécule d'ARN après

une transformation chimique qui substitue l'hydrogène de l'ADN au niveau du sucre par le groupe hydroxyle (OH) et la thymine en uracil. Trois éléments constitutifs de la molécule sont les anneaux de sucre qui se lient aux groupes phosphates, et à chacun des anneaux une des quatre base azotée (A, C, G ou T) y est attachée, de cette façon on parle de chaîne qui a un sens unidirectionnel du 5' au 3', ceci est dû au fait que le carbone 5' du sucre se lie au carbone 3' du suivant, le début de la chaîne se retrouve donc avec un carbone 5' non lié et à la fin le carbone 3' (voir figure 1-4).

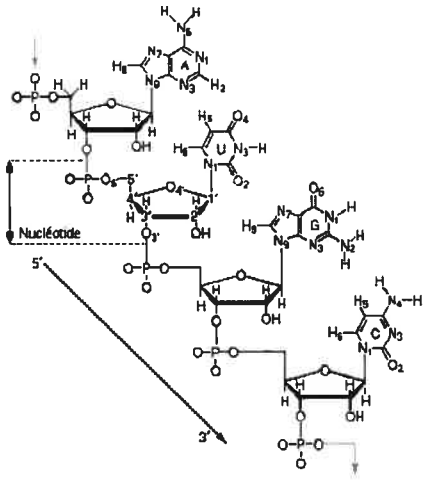


Figure 1-4: Composition chimique de la chaîne AUGC. Un nucléotide est composé d'un groupe phosphate (rouge), d'un ribose (bleu) et de la base (vert). Chaque nucléotide se lie au suivant par un lien covalent entre le ribose et le groupe phosphate (flèche grise en haut). Le lien avec le nucléotide qui précède se fait entre le phosphate et le ribose (flèche grise en bas). Ces liens assurent la continuité de la chaîne et lui donne un sens (la flèche qui va du 5' au 3'). Sur chaque entité sont indiqués les éléments chimiques : H (hydrogène), O (oxygène), N (azote), P (phosphate), OH (groupe hydroxyle). Les jonctions des arêtes sans aucun symbole représentent le Carbone. Les chiffres sur le ribose sont mentionnés avec un prime pour les différencier des autres chiffres.

Il existe plusieurs sortes d'ARN en voici les trois majeures :

ARN ribosomal : représente 80% de l'ARN total d'une cellule. Associé à des protéines, il forme le ribosome qui constitue la tête de lecture de l'information génétique

transcrite par l'ARN messager. C'est dans le ribosome que sont enchaînées les séquences d'acides aminés qui constituent les molécules de protéine.

ARN de transfert : sont des molécules qui se placent sur les sites du ribosome où va être lu l'ARN messager. Leur rôle fondamental est celui d'adaptateurs car ils permettent de reconnaître les acides aminés dans le cytoplasme pour les amener jusqu'au brin d'ARN messager et les positionner de manière à ce que leur enchaînement dans la protéine à synthétiser corresponde aux instructions précisées par la séquence des bases du segment d'ADN codant pour cette protéine.

ARN messager : est transcrit à partir de l'ADN puis est exporté du noyau vers le cytoplasme. Le ribosome, l'unité d'assemblage des protéines, vient se placer sur l'ARN messager et permet ainsi sa traduction en protéine. L'expression de l'ARN messager comporte plusieurs étapes qui sont:

1) la transcription du gène en pré-ARN messager,
2) la maturation du pré-ARN messager, qui comprend l'épissage des introns, l'addition de la coiffe et la polyadénylation,
3) le transport de l'ARN messager mature depuis le noyau jusqu'au cytoplasme, dans lequel sa concentration est contrôlée par des éléments permettant sa stabilisation ou sa dégradation.

L'expression d'un gène dans une cellule résulte donc de la régulation coordonnée de chacune de ces étapes. En particulier, lors du transfert de l'ARN messager depuis le noyau vers sa destination cytoplasmique.

La région nécessaire à la production de la protéine est délimitée par des zones non traduites (UTR) 5'et 3' auxquelles un grand intérêt y est apporté, ceci s'explique par le rôle qu'elles jouent dans la régulation post transcriptionnel. L'ARN messager se retrouve toujours lié à des cations ou à des protéines, ce qui contrôle sa localisation au niveau du cytoplasme, sa stabilité, sa dégradation et sa longueur de vie dans le cytoplasme, ce qui permettra la production de plus de protéines.

Le repliement de la molécule d'ARN à partir de sa structure primaire résultante d'un séquençage qu'on représente comme une suite de caractères pris dans l'alphabet {A, C, G et U}, se fait selon les règles d'appariements des paires Watson Crick A-U, G-C et G-U (voir figure 1-5), ce qui donne naissance à des motifs structuraux tel que les hélices, les bulges, les boucles, les boucles internes (voir figure 1-6).

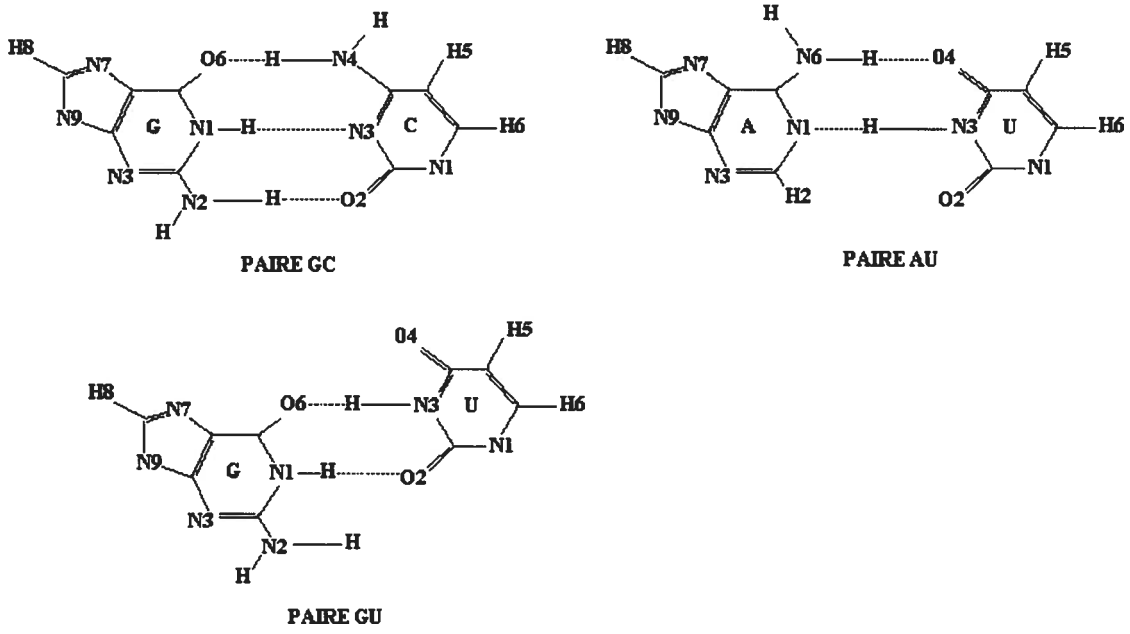


Figure 1-5: Les appariements des bases canoniques AU, CG et non canonique GU. Chaque base est représentée par une lettre au centre de l'hexagone: A (Adénine), G (Guanine), U (Uracil) et C (Cytosine). La lettre N représente l'azote, la lettre O l'oxygène et la H l'hydrogène. Les numéros après chacune des lettres représentent la position de chacun des atomes dans la base. Les intersections des arêtes sans lettres indiquent la présence du carbone. Les liaisons hydrogènes entre les bases sont en pointillés.

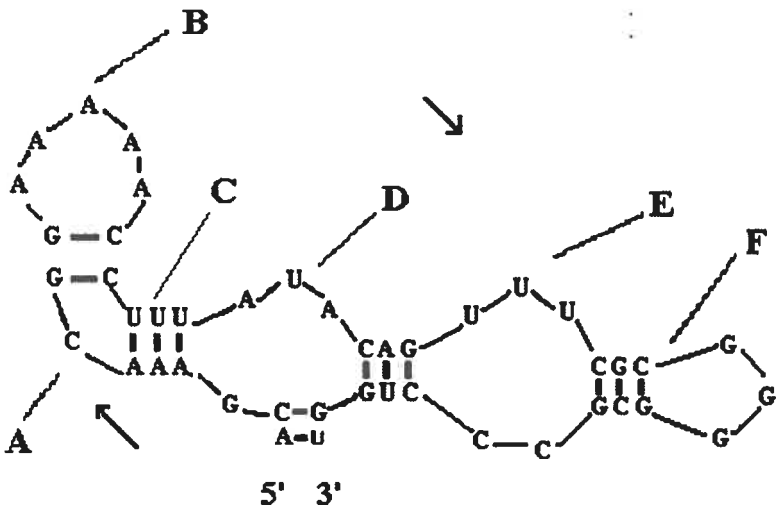


Figure 1-6: Les différents éléments structuraux formant la structure secondaire ; Le tiret rouge indique les trois liaisons hydrogènes entre les bases et le bleu indique les deux. Chacune des lettres, en dehors de la structure, correspond à un élément structural. La lettre A est pour le Bulge (nucléotide libre entre deux hélices). La lettre B est pour la boucle avec les nucléotides non appariés. La lettre C est pour l'hélice avec les trois paires. La lettre D est pour la boucle de multi-branchement (reliant les trois hélices). La lettre E est pour la boucle interne entre deux hélices. La lettre F est pour l'empilement des bases.

Une fois réunis, ces motifs forment à leur tour la structure secondaire (voir figure 1-7), qui est un arrangement de motifs dans un espace à deux dimensions. La structure secondaire est une représentation théorique abstraite permettant aux programmeurs de déduire la structure tertiaire, qui elle est la structure réelle de la molécule dans le contexte cellulaire. Donc du point de vue théorique elle sert comme intermédiaire entre le passage de la molécule de sont état primaire à sont état final de structure tertiaire.

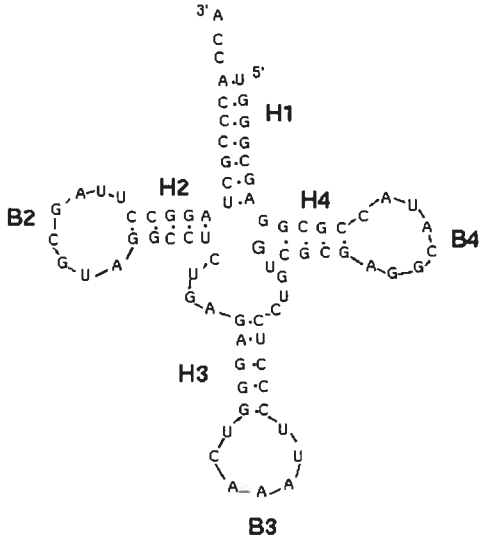


Figure 1-7: Structure secondaire en forme de trêfle à quatre feuilles d'un ARN de transfert, la structure est composée de quatre hélices (H1, H2, H3 et H4) et de trois boucles (B2, B3 et B4). Le 5' indique le début de la chaîne et le 3' sa fin. Les points entres les bases indiquent un appariement entre elles. Les tirets entre les nucléotides indiquent la continuité de la chaîne.

Une hélice d'ARN contenant des paires de bases Watson-Crick adopte la conformation la plus commune de type A (voir figure 1-9), où le sillon mineur est large et non profond, ce qui le rend plus accessible contrairement au sillon majeur qui lui est profond et étroit qui n'est accessible qu'aux molécules d'eau et aux ions de métal. Ces deux sillons sont dûs aux appariements des bases qui présentent une distance de décalage par rapport à l'axe de l'hélice de 4.4 Angström .

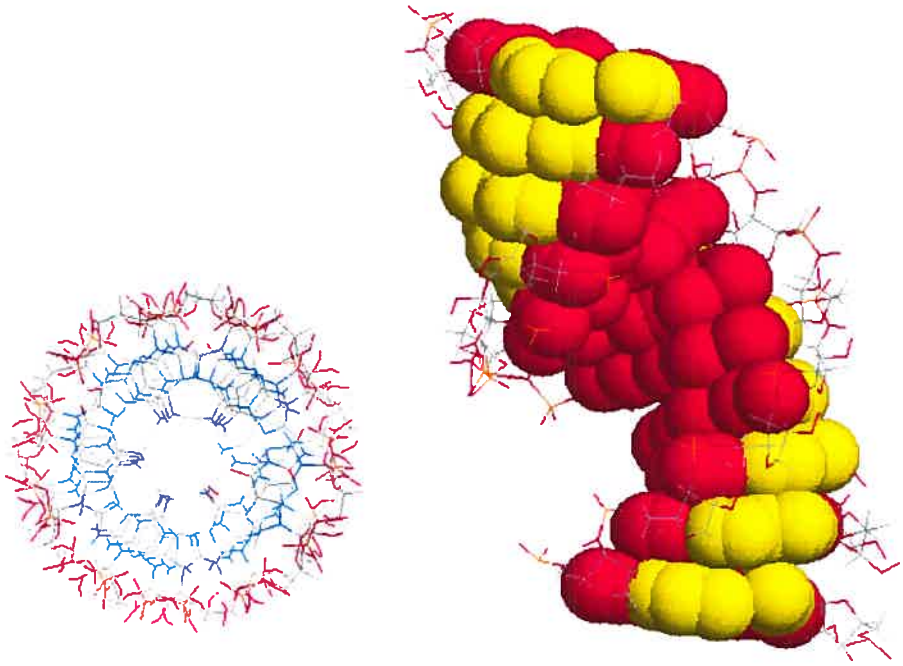


Figure 1-8: La forme A de l'hélice d'un ARN ; à gauche c'est la vue de haut: les bases azotées (en bleu) se retrouvent à l'intérieur de l'hélice ; à droite c'est la vue de face de l'hélice. Chaque paire est composée d'une partie jaune et d'une rouge. Le jaune représente le sillon mineur plus à l'extérieur et le rouge représente le sillon majeur qui lui est plus à l'intérieur de la molécule.

Dans la forme A d'une hélice d'ARN, la conformation anti est adoptée, elle correspond à l'orientation de la base par rapport au lien glycoside C1'-N (N9 pour les purines A ou G et N1 pour les pyrimidines C ou U). En ce qui concerne le sucre en général sa surface n'est pas planaire, de ce fait il y a toujours un ou deux atomes qui sortent du plan, on retrouve en général le C3'-endo de type N (N pour Nord), néanmoins la conformation C2'-endo on la retrouve dans les brins non appariés comme une tétra boucle GAAA, le mot 'endo' signifie au dessus de la surface.

1.5. Prédiction de la structure secondaire

Comme il a été mentionné à la section 1.4., la structure secondaire est vue comme une abstraction entre les structures primaire et finale. Elle montre les appariements des paires de nucléotides mais ne donne aucune information sur les positions des hélices et des boucles, les unes par rapport aux autres, dans l'espace tridimensionnel. Selon le principe de repliement hiérarchique de l'ARN [7], la formation de la structure secondaire se produirait avant la structure tertiaire. Ce principe s'appui sur l'information contenue dans la séquence primaire qui guiderait le repliement. De manière hiérarchique, la structure primaire donnerait naissance à la structure secondaire qui à son tour mènerait à la tertiaire. D'après ce principe, l'énergie d'appariement des paires de bases est plus élevée que celle pour des interactions tertiaires, cela permet de séparer les éléments de la structure secondaire des éléments de la tertiaire. Les règles de ce principe permettent de simplifier le problème du repliement de la molécule d'ARN et de ce fait permettent le développement de programmes de prédiction obéissant à des règles simples.

Cette perception du repliement de l'ARN a permis certes la mise au point d'un bon nombre d'algorithmes mais leurs critères diffèrent. Ces critères peuvent être de nature énergétique, cinétique ou autre [8] selon que l'on possède une ou plusieurs séquences.

La catégorie la plus connue est celle des algorithmes de programmation dynamique, leur critère de sélection est la recherche de la structure avec l'énergie minimale. L'énergie minimale reflète la grande stabilité de la structure et correspond à l'empilement successif des paires de bases. Cet empilement permet la décomposition de la structure de façon récursive en sous structures et l'énergie minimale totale de la structure sera égale à la somme des énergies minimales de ces sous structures. Supposons qu'on veuille calculer l'énergie minimale E entre deux bases i et j, il faudra tenir compte de trois énergies correspondant à trois états d'appariements des bases i et j selon la formule suivante :

E(i,j) = min {

E(i,j-1), *Si i et j-1 sont liées*

E(i+1,j), *Si i+1 et j sont liées*

min E(i+1,j-1), *Energie minimale de la structure entre i+1 et j-1*

}

La règle à respecter est d'avoir un minimum de quatre nucléotides entre deux bases qui s'apparient pour assurer le renversement de la chaîne car du point de vue calcul d'énergie si (j-i < 4) alors E(i,j) = 0.

La construction d'une matrice (N*N), N est le nombre de nucléotides de l'ARN à replier, symétrique supérieure va permettre de regrouper les informations de cette récursivité et l'énergie minimale va se retrouver à la case (1,N) (si (1,N) appariées). Avec un chaînage arrière, les pointeurs sauvegardés à chaque étape de calcul, vont être retracés de la diagonale jusqu'à E(1,N). Cette méthode a été implémenté dans deux programmes *mfold* [9] et *pknots* [10]. Cette implémentation diffère d'un programme à l'autre en deux points, le premier concerne leur complexité en temps qui est de l'ordre de $O(n^3)$ et $O(n^6)$ respectivement ; Le second point est la prise en compte de la structure de pseudo nœud par le programme *pknots* uniquement.

Le programme *mfold* ne considère que les structures récursives et donc les hélices qui ne se croisent pas, par contre *pknots* prend en compte dans sa recherche l'entrelacement des éléments structuraux qui vérifient la condition des indices suivante: (i< i'< j <j'). La condition des indices précise l'appariement des bases i avec j et i' avec j', en donnant l'ordre d'occurrence de chacune des quatre bases i, j, i' et j'.

Pour mieux comprendre, voici des exemples pour les deux types de structures, la représentation est sous forme de matrices.

Le premier exemple pris en compte par *mfold* est illustré par la figure 1-9.

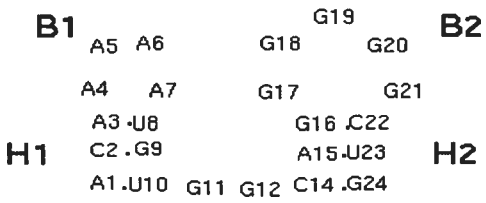


Figure 1-9: Deux hélices successives d'une séquence de 24 nucléotides. H1 et H2 sont 2 tiges formées par trois paires de bases ; B1 et B2 sont deux boucles formées par des nucléotides non appariés. Les 2 tiges boucles, (H1+B1) et (H2+B2), sont reliées par deux nucléotides G11G12 aux positions 11 et 12.

La matrice correspondante à la structure de la figure 1-9 est illustrée par la figure 1-10 où les antidiagonales avec des 1 représentent les paires des hélices H1 et H2 et les 0 dans les cases horizontales représentent les bases non appariées des boucles B1 et B2 :

	A..A	A	A	A	A	U	G	U	G	G	C	A	G	G	G	G	G	G	C	U	G
A								1													
C							1														
A		0	0	0	0	1															
G.. G																					
C																					1
A																				1	
G													0	0	0	0	0	1			

Figure 1- 10: Matrice de construction de la structure avec deux hélices (figure 1 - 9) Le 1 indique un appariement entre les bases de la ligne et colonne correspondantes, le 0 indique des bases non appariées. Les colonnes de 1 à 3 et les lignes de 4 à 11 sont tronquées pour des raisons d'espace d'affichage et sont non informatives.

Le deuxième exemple de structure prise en compte par le programme *pknot* est illustré par la figure 1-11.

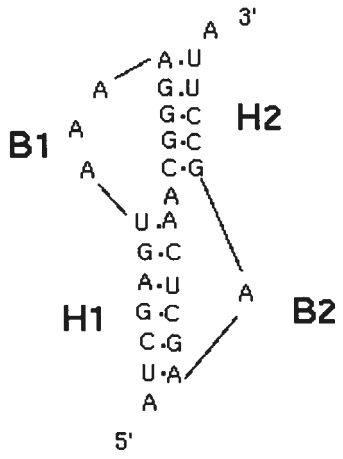


Figure 1-11: Structure de pseudo nœud dans une séquence de 29 nucléotides. Le début de la séquence commence au 5' et finit au 3'. L'hélice H1 est formée de 6 paires de base, la H2 est constituée de 5 paires. Ces deux hélices ont un nucléotide A entre elles et sont fermées respectivement par deux boucles B1 et B2 ne contenant que des adénines (A). Les points entre les bases représentent un appariement. Les traits indiquent la continuité de la chaîne de nucléotides.

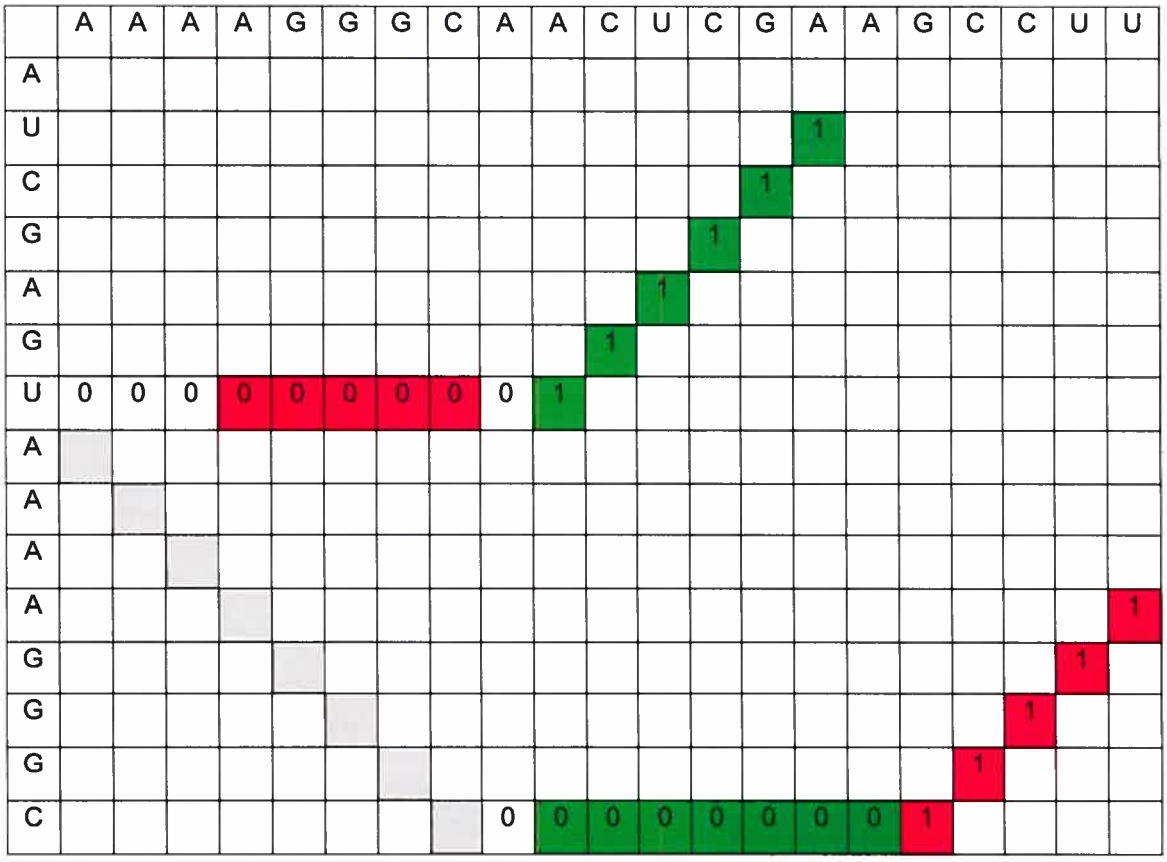


Figure 1-12: Matrice du pseudo noeud de la figure 1-11, les hélices correspondent aux anti-diagonales avec des cases remplies de 1 et les boucles correspondent aux cases horizontales remplies avec des 0. Les couleurs vert et rouge montrent les deux états des bases appariées et non appariées.

De la figure 1-12, on voit que certains nucléotides peuvent être dans deux états à la fois, appariés et non appariés. Dans ce cas, le programme *pknots* construit deux matrices avec deux espaces vides au milieu qui seront remplis par les bases formant le pseudo nœud (voir figure 1-13).

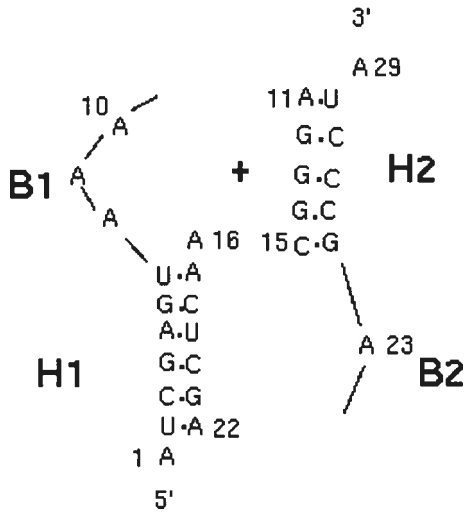


Figure 1-13: Découpage de la structure de pseudo nœud en deux parties discontinues. La boucle B1 de l'hélice H1 n'est pas reliée à l'hélice H2, idem pour la boucle B2 et l'hélice H1. Le (+) indique l'union des deux parties (H1B1 + H2B2). Les points indiquent une paire de base et les traits indiquent la continuité de la chaîne. Les chiffres à côté des bases indiquent leur position dans la chaîne de nucléotides. L'absence des nucléotides entre 10 et 16 ainsi qu'entre le 15 et 23, correspond à deux trous. Au niveau des matrices, ces trous sont des vides qui seront remplis par les bases formant le pseudo nœud.

De cette décomposition, on peut voit l'ordre de complexité en $O(n^4)$ pour le stockage des matrices et de l'ordre de complexité en $O(n^6)$ pour le temps. Le chaînage arrière est appliqué deux fois et trois pointeurs sont nécessaires pour retrouver le chemin.

En plus de ces algorithmes basés sur la programmation dynamique, il existe d'autres comme les algorithmes génétiques qui tentent de simuler une évolution biologique d'une séquence à replier. Pour ces algorithmes, une séquence peut avoir un ensemble de structures possibles qui sera considéré comme une population. Chaque structure possible de l'ensemble sera à son tour considérée comme un individu avec des capacités de reproduction intégrale, de mutation ou de recombinaison. Le processus d'évolution aura pour but de produire et de sélectionner les structures solutions avec une

meilleure fonction 'fitness'. La fonction 'fitness' est le critère de sélection qui pourrait être une fonction d'énergie d'une structure. Les structures avec une basse énergie auront une 'fitness' élevée ce qui augmentera leurs fréquences et chances de reproduction. Durant la simulation, les structures peuvent muter par suppression ou ajout d'une hélice. Les hélices enfants des structures peuvent aussi se recombiner avec d'autres hélices appartenant à des structures de parents différents. Les structures créées lors de ce processus d'évolution vont servir à retracer le chemin de repliement de la séquence d'ARN. Un exemple de programme qui utilise le principe des algorithmes génétiques est celui de *RAGA* (RNA Alignement by Genetic Algorithm). Il considère l'ensemble des alignements possibles entre deux séquences comme un population et chaque individu représente un alignement possible. Parmi les deux séquences, une d'entre elle a une structure connue et sera la structure maîtresse, par contre l'autre séquence de structure inconnue sera l'esclave. La fonction de 'fitness' se base sur l'alignement et non pas sur l'énergie de structure [11].

Une autre méthode pouvant donner des informations sur la structure secondaire est celle de l'analyse par comparaison de séquences de différentes espèces. Leur hypothèse est qu'une séquence ayant la même fonction dans différentes espèces aura de fortes chances d'adopter la même structure dans l'espace. Son but est de trouver à partir d'un alignement entre les séquences des positions covariantes où un changement de base à une position i de la séquence doit être compensé par un changement à la position j de telle sorte que les bases i et j resteraient appariées (soit dans une séquence à une position *i* on a C et dans *j* un G tel que C et G forment une paire, si un changement occure à la position *i* et on retrouve un A, un G ou un U alors du côté *j* on devrait retrouver un U, un C ou un A pour conserver la paire de base). On pourrait donc déduire l'importance de ces paires de bases à i et j. Pour utiliser cette méthode il faut que les différences entre les séquences soient raisonnables pour pouvoir exploiter leur alignement [8].

Dans le cadre de ce projet, les algorithmes de programmation dynamique ont été utilisés, vu qu'aucune structure n'est connue pour aucune des séquences et le fait aussi

qu'on ne possède pas assez de séquences pour faire une analyse par comparaison pour en extraire les co-variations qui maintiennent les appariements de la structure commune.

1.6. Recherche de la structure secondaire à l'aide de motif

Un autre moyen utile pour la recherche de la structure secondaire est l'utilisation de programmes de recherche de motifs. Parmi ces programmes, on retrouve *RNABOB*, *RNAMot* [12] et *RNAMotif* [13] qui se basent sur le même principe. Ces programmes permettent de trouver un motif bien précis dans une (ou plusieurs) séquence (s). Le motif, sous forme de fichier, est une suite d'instructions écrites en utilisant une syntaxe propre au langage de description d'une structure secondaire. Les instructions servent à décrire un arrangement des différents éléments structuraux (voir figure 1-7). Le fichier est à soumettre au programme pour subir les vérifications classiques d'un compilateur (lexicale, syntaxique et sémantique). Après succès des vérifications, une image interne du motif est construite. L'arbre est la structure de donnée choisie pour la représentation interne. En effet, la représentation arborescente est la plus appropriée pour représenter les éléments structuraux occurrents de façon récursive dans la structure secondaire. Les nœuds de l'arbre représentent les hélices et les feuilles les brins de nucléotides non appariés (éléments structuraux : boucle, bulge, boucle interne). Les éléments structuraux qui se retrouvent à gauche d'une hélice (nœud) sont internes à cette hélice, par contre ceux qui sont à sa droite lui sont externes (voir figure 1-14).

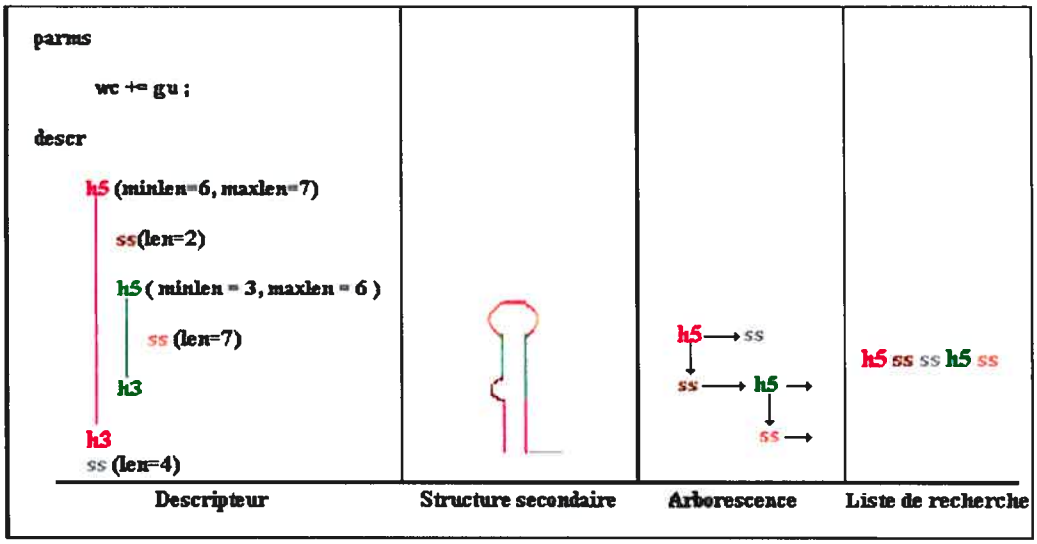


Figure 1-14: Exemple de fichier descripteur, à gauche, selon la syntaxe du langage de structure secondaire. La première section du fichier descripteur, indique au programme qu'aux deux paires standard Watson Crick (wc) il faut rajouter la paire GU (gu). La structure à rechercher est celle du milieu et son arbre correspondant est à la 3eme colonne. Dans le descripteur deux hélices sont décrites, leur imbrication dans la structure se reflète dans les descripteur, la première hélice la plus externe est mentionnée en rouge avec les deux demis brins h5 et h3 pour les deux cotés respectifs 5' et 3'. La deuxième hélice est indiquée en vert. Pour chacune des trois boucles, une couleur est attribuée, la plus externe est en gris et occure après le h3 de la première hélice. Les boucles internes aux hélices se retrouvent, dans le descripteur, entre les deux mots réservés h5 et h3. Les paramètres de longueur minimale et maximale pour chacun des éléments structuraux sont entre parenthèses. Au niveau de l'arbre les nœuds correspondent aux hélices, les flèches horizontales indiquent que l'élément structural qui suit courant, est externe. La flèche verticale orientée vers le bas indique que l'élément structural suivant, est interne à l'élément courant. La liste de recherche, 4eme colonne, est le résultat de la linéarisation de l'arbre.

Ces programmes de recherche de motifs sont utiles quand l'utilisateur a une idée du motif à rechercher, ce qui lui permettra de dresser un fichier descripteur adéquat. Les éléments structuraux seront, certes, précis mais leurs paramètres associés tels que les longueurs minimale et maximale seront flexibles tout en restant dans une fourchette de longueurs cohérente.

Une fois l'arbre construit, il sera linéarisé en se basant sur la hiérarchie des éléments structuraux. Le résultat de cette linéarisation sera une liste de recherche. Le remplissage de la liste se fera en parcourant chacun des sous arbres de l'arbre, de façon à insérer en premier l'élément du nœud, par la suite l'élément interne (flèche verticale) et en dernier l'élément externe (flèche horizontale) (voir figure 1-14). Une fois donc la liste construite, le programme entame sa recherche en parcourant la séquence de gauche à droite pour tenter de retrouver l'élément structural le plus à gauche. Le programme reprend de façon récursive la recherche de l'élément suivant, toujours le plus à gauche, ainsi de suite jusqu'à la fin du motif. Plusieurs candidats seront possibles du plus court au plus long. La solution trouvée sera par la suite soumise à des vérifications pour reconnaître sa validité. Ces vérifications se font à l'aide d'instructions dans une section *score.* Cette dernière, si elle existe, sera dans le fichier décrivant le motif. Les vérifications peuvent, par exemple, soumettre le motif trouvé à un calcul d'énergie, de sorte que ce dernier devrait être négatif pour que le motif soit accepté sinon il sera refusé et ne fera pas partie des résultats possibles. La section *score* n'est valable que dans le programme *RNAMotif* ce qui représente une amélioration par rapport aux programmes antécédents. Dans le cas où cette section serait absente, tous les candidats seraient acceptés sans aucune restriction.

1.7. Le calcul d'énergie

Pour le calcul d'énergie minimale d'une structure secondaire, les programmes des différentes catégories, citées dans les paragraphes précédents, se basent sur des valeurs d'énergies pré-calculées. Certaines de ces valeurs sont d'origine expérimentale et d'autres sont déduites ou extrapolées vu qu'il est impossible de mesurer les paramètres thermodynamiques de toutes les hélices [14].

Les valeurs expérimentales proviennent de deux principales méthodes qui sont : la calorimétrie et la dénaturation thermique au laser [15]. Ces deux méthodes ont chacune leurs avantages et inconvénients, de ce fait elles sont souvent utilisées en parallèle. Les valeurs sont obtenues avec des taux d'erreurs [15] qui proviennent essentiellement des erreurs de calibrage des appareils ainsi qu'aux conditions de préparation des échantillons à étudier. Dans les deux expériences les valeurs sont particulièrement précises autour de 37^{o} qui est la température du corps humain et celle de fusion des nucléotides en duplexe.

Pour les valeurs déduites, elles sont prédites de façon théorique selon un modèle. Ce modèle est celui du plus proche voisin, son utilisation est dû au fait que les résultats mesurés en laboratoire correspondent aux prédits. De ce fait, il est perçu comme étant une bonne approximation du monde réel. Dans ce modèle les paramètres thermodynamiques dépendent des paires de bases présentes dans la séquence. Par exemple, le calcul d'énergie des deux hélices suivantes, composées chacune de six paires GC,

5' - GCCCGGG - 3' 3' - CGGGCCC - 5'	et	5' - CGCGCG - 3' 3' - GCGCGC - 5'

donnera -11.2 et -9.4 kcal/mol respectivement, selon les données de Freier et al [14]. On observe une différence de +1.8, entre les deux énergies, qui est dûe à la différence de position de chacune des six paires. Des variantes du modèle théorique existent tout en gardant le même principe de calcul, leur différence réside dans les paramètres pris en

compte dans leur calcul. Exemple, certains prennent en compte l'initiation de la formation du duplexe selon que l'hélice contiennent au moins une paire GC ou que des paires AU. Cette prise en compte se manifeste par un rajout d'un (ou des) terme(s) d'énergie dans leur calcul selon la composition des paires de l'hélice [14] .

Le calcul d'énergie minimale d'une structure effectué par les programmes reviendra au calcul de l'énergie totale de la structure, en sommant les termes d'énergie attribués à chacune des deux paires successives pour chacune des hélices formées. Pour les éléments structuraux non appariés leur valeur prend en compte d'autres facteurs (longueur, symétrie,..). Par exemple, le stockage des termes d'énergie dans le cas d'un empilement d'une paire GC dans une hélice se retrouvera dans un fichier 'stack.dat'. Ce fichier fera partie de la librairie des valeurs d'énergie. La représentation des données, la plus commune, est sous forme de matrice 4x4 (voir figure 1-14-A). Pour les nucléotides non appariés telle que les boucles, elles sont représentées sous une autre forme ou il est tenu compte de la longueur de la boucle et dans quel type d'élément structural ces nucléotides se retrouvent (voir figure 1-14-B).

A

5' → 3'
GX
CY
3' ← 5'

Y →
X ↓

A	C	G	U
.	.	.	-2.40
.	.	-3.40	.
.	-3.30	.	-1.50
-2.20	.	-2.50	.

B

Taille	Interne	"Bulge"	Tige Boucle
1	.	3.80	.
2	.	2.80	.
3	.	3.20	5.70
4	1.70	3.60	5.60
5	1.80	4.00	5.60
6	2.00	4.40	5.40
7	2.20	4.60	5.90
8	2.30	4.70	5.60
9	2.40	4.80	6.40
10	2.50	4.90	6.50
11	2.60	5.00	6.60
12	2.70	5.10	6.70
13	2.80	5.20	6.80
14	2.90	5.30	6.90
15	3.00	5.40	6.90
16	3.00	5.40	7.00
:	:	:	:

Figure 1-15: Exemple de table des paramètres thermodynamiques en kcal/mol. La figure A correspond aux valeurs d'énergies attribuées aux doublets formés par les paires XY qui suivent la paire GC. X correspond à un des 4 nucléotides [A, G, C, U] pris dans une des 4 lignes de la matrice (flèche verticale), idem pour le Y mais la valeur est prise dans une des 4 colonnes (flèche horizontale). Une paire GC suivie d'une paire AU, aura -2.40 kcal/mol comme énergie, ce qui correspond à la cellule [1,4] de la matrice (1ère ligne et 4ème colonne). La section B donne les énergies des boucles selon leurs longueurs et l'élément structural dans lequel elles se retrouvent. La table se poursuit jusqu'à une longueur de 30, au-delà de cette limite une formule de calcul est appliquée. Une boucle de 16 nucléotides aura une énergie de 7.00 kcal/mol si elle se retrouve dans une tige boucle. Les points noirs indiquent l'absence de valeurs et les rouges pour indiquer la continuité. Le sens du duplexe est donné par les flèches –» allant du 5' au 3'.

Le calcul d'énergie d'une structure secondaire, que ça soit pour un programme de prédiction ou pour un programme de recherche de motifs, dépendra des termes d'énergies disponibles. Mais l'évolution de la technologie des ordinateurs a permis d'élaborer des programmes de déduction de nouveaux termes d'énergies pour de nouveaux motifs tels que les boucles de multi-branchement, où il est tenu compte du nombre d'hélices des nucléotides non appariés des 'bulges' dans les hélices. Les tétra boucles ont pu être traitées à part vu que certaines sont reconnues pour leur stabilité en tenant compte de la

paire de base la fermant. Des valeurs d'énergie sont aussi calculées pour empilements coaxiaux de deux hélices adjacentes qui sont connectées souvent par une paire de base non canonique GA, ce qui leur permettra de former une seule hélice quasi continue. Pour les boucles internes, la symétrie est prise en compte (nombre de nucléotides des deux cotés 5' et 3'). Cette évolution a permis aux programmes de réadapter leur calcul, exemple du programme de *mfold* qui à enrichit ses paramètres thermodynamiques [16][17][18] par contre *pknots* n'a utilisé qu'un seul ensemble de données [16]. Le programme de recherche de motif *RNAMotif* lui utilise les derniers paramètres d'énergie [19].

Chapitre 2

Recherche de la structure secondaire

2.1. Introduction

Dans ce chapitre, il sera question de la recherche de la structure secondaire au niveau des trois séquences d'ARN messagers des protéines STG pour chacune des trois espèces. Cette recherche utilisera les programmes cités au chapitre 1 selon l'hypothèse à prouver ou à proposer. Deux hypothèses seront exposées dans ce chapitre, la première découle de l'observation du chercheur Mauricio Neiro [1] qui se base sur l'alignement entre la séquence de protéine du macaque rhésus avec celle du prion de l'humain. Cette observation nous mènera à rechercher la structure de pseudo nœud en particulier, vu que cette dernière a été proposée par Barrette et al [20] lors de leur étude portant sur la structure du prion. La seconde hypothèse découle de l'observation, au niveau des ARN messagers, de la séquence de quatre nucléotides GAAA ou bien GGAA. Ces deux tétra boucles font partie de la famille des boucles GNRA où le N pourrait être n'importe lequel des quatre nucléotides (A, G, C, U) par contre le R ne peut être qu'une purine, donc un A ou un G. Cette famille de boucles est reconnue pour sa stabilité de structure.

Avant d'exposer les hypothèses, il sera mis en évidence l'existence d'une structure possible autour de la région du polymorphisme. Cette étape va permettre de supposer que le polymorphisme observé pourrait donner naissance à une structure (ou bien juste des parties) bien particulière, pour éventuellement accomplir une tâche bien particulière.

2.2. Existence d'une structure secondaire

La première étape est de savoir si une structure secondaire pourrait exister au niveau des trois ARN messagers. Ceci est faisable avec le programme SPF (Structural Pattern Finder), développé par P.Gendron et al (résultats non publiés, 2003). Le but de ce programme est le calcul des énergies libres des régions composant la séquence biologique en ne sachant au préalable si elles sont structurées ou non. L'énergie la plus faible calculée nous laisserait supposer l'existence d'une structure. Ce programme permet de montrer que la conservation de l'ordre des nucléotides pourrait être utile à la formation d'une structure particulière. Dans le cas des ARN messagers l'ordre pourrait être conservé pour la préservation des codons et la synthèse de la bonne protéine. Il a pour principe de fonctionnement le suivant: il parcourt la séquence sur toute sa longueur avec des fenêtres de 200 nucléotides, et pour chacune d'elle il réordonne les nucléotides 500 fois. Pour chaque réarrangement un calcul d'énergie correspondant à la structure prédite est effectué. On obtient donc 500 valeurs d'énergies incluant la valeur de la structure de la séquence initiale. Ces valeurs sont stockées dans un fichier. Pour chacune des valeurs, un fichier décrivant la structure secondaire est généré. Ce fichier pourra par la suite être visualisé avec un logiciel adéquat tel que *rnaviz* [21][22]. Le passage à la prochaine fenêtre va se faire avec un chevauchement de 50 nucléotides avec la précédente, par exemple si la première fenêtre couvre les 200 premiers nucléotides (de 0 à 199), la seconde s'étalera du $50^{ième}$ au $249^{ième}$ nucléotide, ainsi de suite, jusqu'à ce que la séquence soit parcourue en totalité, en respectant toujours la taille 200 de la fenêtre. Pour chaque fenêtre, une différence d'énergie est calculée entre la moyenne des énergies des séquences aléatoires et l'énergie de la séquence originale. Ce calcul, de différence d'énergie, sera reporté dans un fichier qui servira à tracer un graphe d'évolution d'énergie selon la fenêtre (les abscisses correspondent aux fenêtres et les ordonnés aux différences d'énergie). Les méthodes de prédiction de structure et de calcul d'énergie sont ceux utilisés par le programme *pknots* [10] cité au paragraphe 1-6 du chapitre 1.

L'exécution du programme s'est faite avec les séquences des ARN messagers ayant pour longueur les valeurs du tableau 2-1. Notre intérêt se porte plus particulièrement aux séquences du tableau 2-2 qui correspondent aux régions où sont observées le polymorphisme.

Espèce	Longueur = nombre de nucléotides
Humain	1153
Macaque rhésus	1120
Souris	1144

Tableau 2-1: Longueur des séquences des ARN messagers pour chacune des trois espèces. La longueur est exprimée en nombre de nucléotides.

Le tableau 2-2 résume les séquences étudiées, ainsi que leur position de début et de fin dans le messager avec leurs protéines correspondantes.

Espèce	Début/ Fin	ARN messager	Protéine
Humain	739/877		
Répétition 1		agctggggaaatattaatcggtatccaggaggc	SWGNINRYPGG
Répétition 2		agctggggaaatattaatcggtatccaggaggc	SWGNINRYPGG
Répétition 3		agctggggaatattaatcggtatccaggaggc	SWGNINRYPGG
½ Répétition		agctgggggaatatt	SWGNI
Macaque Rhésus	740/846		
Répétition 1		agctggggaaatattaatcggtatccgggaggc	SWGNINRYPGG
Répétition 2		agctggggaatattaatcggtatccaggaggc	SWGNINRYPGG
½ Répétition		agctgggggaatatt	SWGNI
Souris	806/950		
Répétition 1		taccccccagtagggacctggggcggttatggtcag	YPPVGTWGGYGG
Répétition 2		taccccccagtagggacctgggggggttatggtcag	YPPVGTWGGYGG
Répétition 3		taccccccagtaggaccctggggcggttatggtcag	YPPVGTWGGYGG
Répétition 4		taccccccagtagggacctgggggccaattgccag	YPPVGTWGANCQ

Tableau 2-2: Tableau récapitulatif des régions répétées pour chacune des trois espèces. La première colonne correspondant à la répétition de chacune des trois espèces (humain, macaque rhésus et la souris). La seconde colonne correspond aux positions de début et de fin de la répétition dans chacune des séquences étudiée. La troisième colonne correspond à la portion d'Arn messager répété. La quatrième colonne correspond à la protéine résultante de la traduction de la portion de la colonne numéro 3.

Effectivement pour les trois séquences, les énergies les plus faibles sont observées au niveau des répétitions. Les courbes de la figure 2-1 ci-dessous illustrent les résultats.

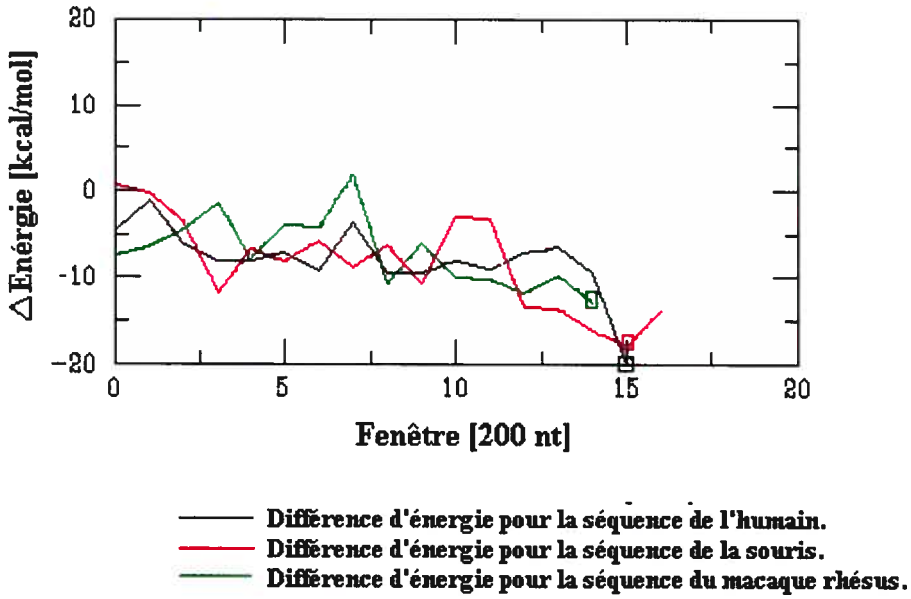


Figure 2-1: Courbes des différences d'énergies pour les trois séquences d'ARN messagers, résultantes de l'exécution du programme SPF (Structural Pattern Finder). Les rectangles indiquent les minimums de différence d'énergies. Ces minimums sont observés au niveau des répétitions pour chacune des trois espèces.

Plusieurs minimums apparaissent sur les trois courbes de la figure 2-1. Ces pics pourraient signifier la présence d'une structure. Pour notre étude la zone d'intérêt est celle autour des répétitions. Effectivement, on retrouve des différences d'énergies minimales dans ces régions qui correspondent aux extrémités finales des courbes de la figure 2-1. Cette observation nous permet de poser l'hypothèse d'une existence possible d'une structure dans cette zone.

A partir des graphes de la figure 2-1 on peut dresser le tableau 2-3 résumant les différences d'énergies calculées pour les fenêtres correspondantes:

Espèce	ΔÉnergie (kcal/mol)	Fenêtre
Humain	-20.0466	15
Macaque rhésus	-12.8122	14
Souris	-17.8174	15

Tableau 2-3: Moyennes d'énergies correspondantes aux fenêtres des prédictions par SPF. La fenêtre 15 de l'humain s'étale du 700^{éme} au 899^{éme} nucléotide. La fenêtre 14 du macaque rhésus s'étale du 650^{éme} au 849^{éme} nucléotide. La fenêtre 15 de la souris s'étale du 700^{éme} au 899^{éme} nucléotide.

Les structures secondaires correspondantes aux différences d'énergie du tableau 2-3 sont celles de la figure 2-2 :

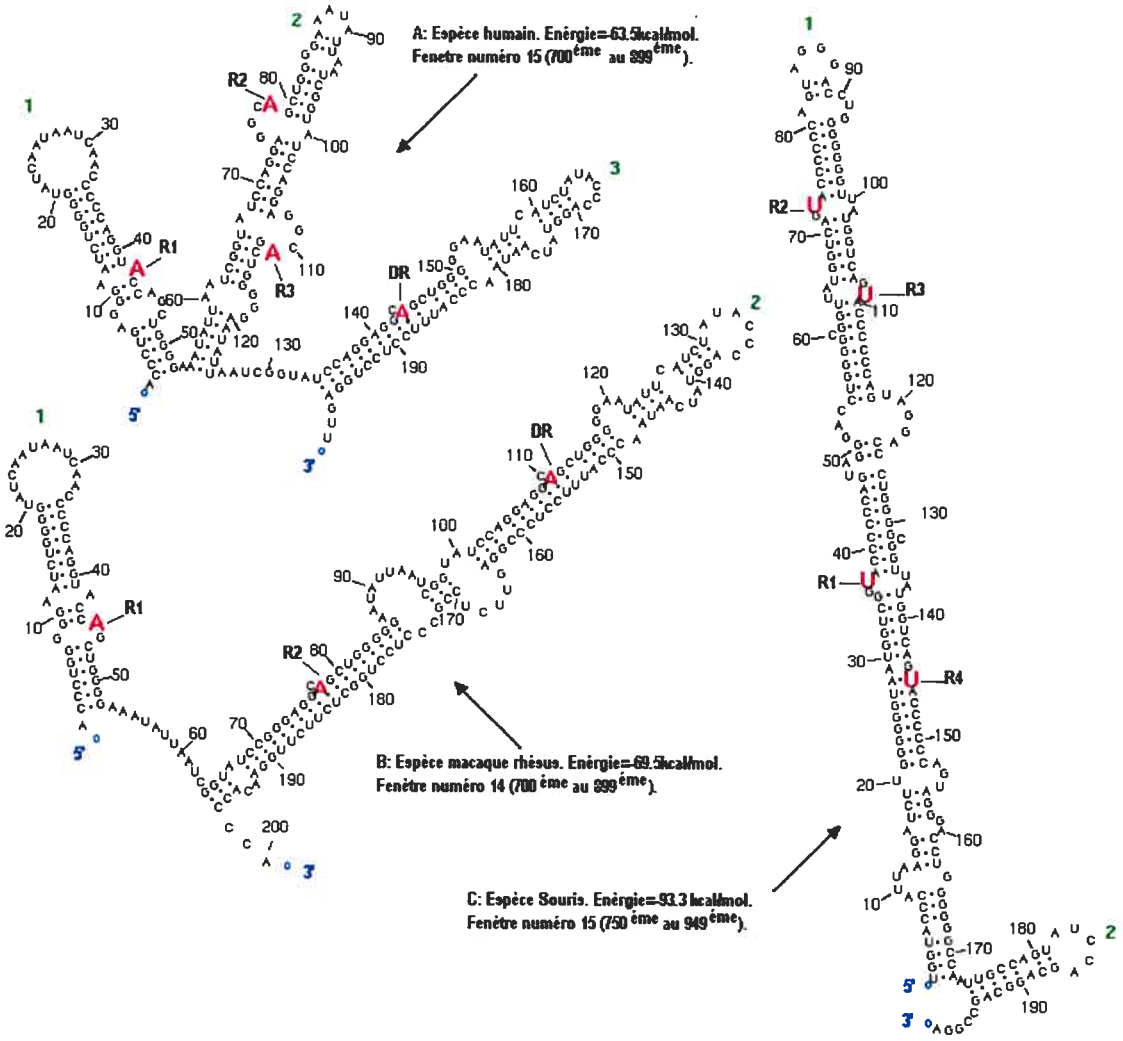


Figure 2-2: Structures secondaires des fenêtres correspondantes au minimum d'énergie. Pour chacune des espèces une lettre est attribuée A, B et C respectivement pour l'humain, le macaque rhésus et la souris. Les flèches mettent en correspondance l'espèce et sa structure. Les débuts de répétitions sont en gras rouge avec leur numéro d'occurrence (lettre R + numéro d'occurrence). Les demie répétitions sont différenciées par le numéro d'occurrence (lettre DR + numéro d'occurrence). Les débuts et fins de structures sont indiqués respectivement par les chiffres 5' et 3'. Les chiffres de part et d'autre de la structure servent à la numérotation relative des nucléotides. Pour chaque structure il est indiqué l'énergie et la fenêtre réelle. Les chiffres en vert indiquent le numéro de l'élément structural.

Pour les trois structures secondaires de la figure 2-2, on voit bien que les régions couvrant les répétitions (débutant par une lettre en rouge) font partie des structures prédites par le programme. Ceci nous permet de supposer l'existence d'une structure au niveau des répétitions. Le programme n'a pas prédit la même structure pour les trois séquences, mais on retrouve quelques ressemblances entre le macaque rhésus et l'humain. Ces deux espèces ont en commun les éléments structuraux suivants: les deux numéros 1 ainsi que le numéro 3 de l'humain qui est totalement inclut dans le numéro 2 du macaque rhésus. L'élément structural numéro 1 des deux espèces couvre un bout des premières répétitions et le numéro 3 couvre les demi répétitions entièrement. On observe également la tétra boucle GGAA au niveau des éléments structuraux 2 et 3 respectivement chez l'humain et le macaque rhésus. Cette boucle, interne et asymétrique, est reconnue pour sa stabilité. La tétra boucle GGCA n'apparaît que chez l'humain en boucle interne à la fin de l'élément structural numéro 2. Il a été montré que cette tétra boucle représente un élément structural stable pour la digestion par l'enzyme ribonucléase T1, quand elle se retrouvait à l'extrémité d'une tige. La raison étant la conformation des nucléotides de cette tétra boucle qui empêchait la ribonucléase T1 de procéder au clivage. Cette enzyme a pour caractéristique de cliver un brin non apparié de l'ARN et spécifiquement après deux résidus G dont deux font partie de la tétra boucle GGCA [23].

Néanmoins, une différence apparaît entre les premiers éléments structuraux des deux espèces du macaque rhésus et de l'humain. Cette différence est au niveau de l'appariement entre le septième nucléotide (A) et le quarante septième (C) de la chaîne. Bien que ce type de paires est observé dans la base de données des structures d'ARN, cet appariement n'a pas été prédit vu que le programme ne permet que les appariements canoniques (AU, CG et GU). La paire AC peut adopter plusieurs conformations, mais la plus prépondérante est retrouvée 245 fois selon la conformation où la face Hoogsteen de A interagit avec la face Watson-Crick du C. Cette interaction se fait à travers deux liens hydrogènes avec le lien glycoside orienté en trans [24]. Cette conformation est illustrée par la figure 2-3.

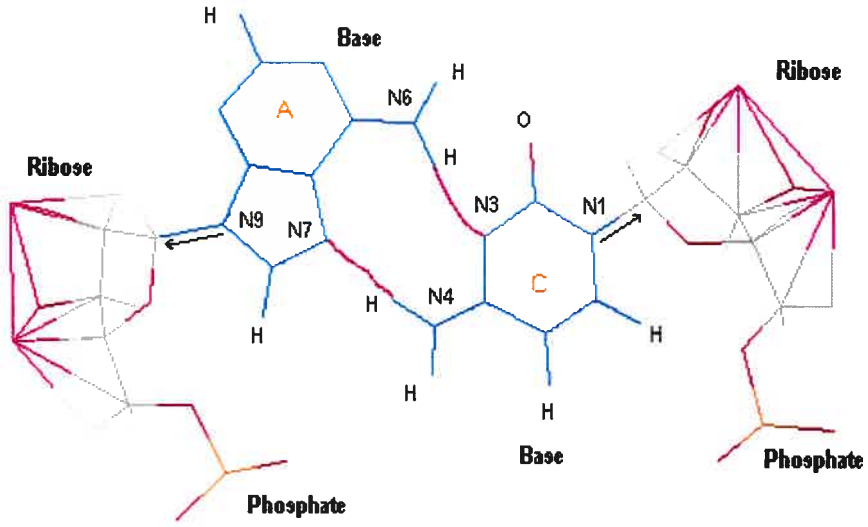


Figure 2-3: La paire AC la plus fréquente avec les deux liens hydrogènes indiqués en fushia entre N7-H et N3-H. Conformation où la face Hoogsteen de la base A se lie à la face Watson-Crick de la base C. Les flèches en noir indiquent l'orientation du lien glycoside (N9-C1') pour la base A et (N1-C1') pour la base C avec le ribose. Ces deux flèches sont parallèles, ce qui correspond à l'orientation trans. Le lien entre le ribose et le phosphate assure la continuité de la chaîne. N : azote, H : Hydrogène, O : oxygène. Les chiffres après chaque atome indiquent sa position dans la base. Les chiffres accompagnés d'un prime (') sont réservés aux atomes du ribose.

Pour la souris, aucun point commun n'est observé avec les deux autres espèces. Mais on retrouve un élément structural, de boucle interne symétrique, au début de chacune des trois répétitions. Cet élément est composé de quatre nucléotides répartis en groupe de deux sur chacun des demi brins. La composition de la boucle interne est résumée dans le tableau 2-4.

Elément structural	Répétition 1	Répétition 2	Répétition 3
Brin 5'→ 3'	5'-GU-3'	5'-GU-3'	5'-UA-3'
Brin 5' ← 3'	3'-AU-5'	3'-AU-5'	3'-UG-5'
Position 5'→ Position 3'	36 – 37	72-73	64-65
Position 5'← Position 3'	137-136	101-100	109-108

Tableau 2-4: Tableau récapitulatif des occurrences de l'élément structural de boucle interne formée de deux nucléotides sur chacun des brins. Dans la 2^{éme} ligne on retrouve les composants de l'élément structural sur chacun des deux brins. Les chiffres 5' et 3' dénotent respectivement le début et la fin de la chaîne des nucléotides. Les positions correspondantes aux trois occurrences sont données en 3^{éme} ligne en tenant compte du sens de la chaîne.

En résumé à cette phase de recherche, on pourrait conclure qu'il serait fort probable qu'une structure pourrait exister dans les régions répétées étant donné l'observation de minimums d'énergies dans ces zones pour chacune des trois espèces. Par contre, le programme *SPF* n'a pas prédit la même structure pour les trois séquences. Il semblerait que le nombre d'occurrence affecte la phase de prédiction du programme *SPF*. Effectivement, le programme a prédit des structures partiellement semblables pour les deux espèces, humaine et macaque rhésus, qui ne différent d'ailleurs que par le nombre d'occurrences de la même répétition. Néanmoins, cette phase a pu mettre en évidence des éléments structuraux qui vont orienter la suite de la recherche en posant différentes hypothèses qui seront présentées dans les paragraphes suivants. Celle des boucles GNRA en est une.

2.3. Hypothèse du pseudo nœud

2.3.1. Le prion et la structure du pseudo nœud

L'hypothèse de la structure du pseudo nœud au niveau de l'ARN messager du macaque rhésus s'est imposée en premier vu l'alignement observé par le chercheur Mauricio Neira avec la protéine du prion (voir figure 2-5). Avant d'entamer la recherche de la structure du pseudo nœud, il serait intéressant de la présenter.

La structure du pseudo nœud au niveau des répétitions de l'ARNm de l'humain (NM_000311) vient de P.R. Wills qui a rapporté sa présence en 1992 [25]. La caractéristique de cette protéine est la présence de quatre copies de l'octapeptide PHGGGWGQ dans la région N-terminal non structurée. La variation du nombre de copies de cette répétition pourrait engendrer la maladie du prion en affectant la conformation de cette protéine. La protéine prion normale est constituée de trois hélices alpha et d'un feuillet bêta. Ce type d'hélice et de feuillet sont deux des motifs structuraux formant la structure tertiaire d'une protéine. En rajoutant deux à neuf répétitions de l'octapeptide, la conformation de la protéine se retrouvera avec plus de feuillets bêta et moins d'hélices [20].

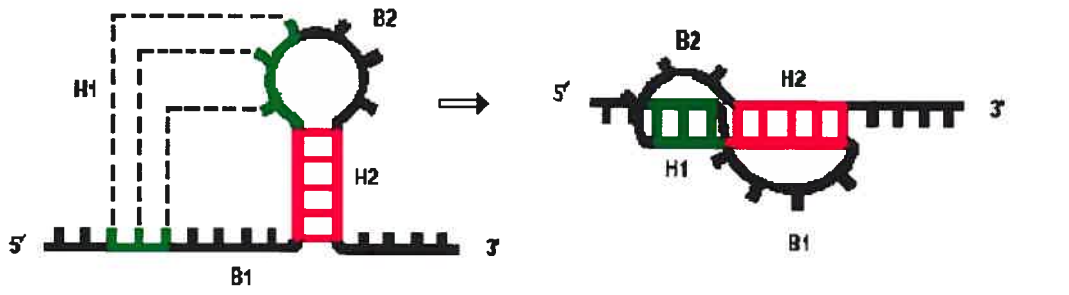
Pour comprendre l'éventuelle implication du pseudo nœud, il faut savoir que dans le bon fonctionnement de la traduction de la protéine du prion normale, où il est supposé y avoir une seule copie du pseudo nœud, le complexe protéine-pseudonœud serait bloqué par un ARN antisens. Ce blocage serait dû à la présence d'un motif structural UNR (U : uridil, N: n'importe quel nucléotide et R : purine A ou G, avec une paire entre U et R). Ce motif est reconnu pour sa fonction de blocage et dans le cas du pseudo nœud il se retrouverait au niveau de la boucle supérieure. Etant donné qu'on a une seule copie de l'antisens pour une seule copie de pseudo-nœud, une augmentation du nombre de copies pourrait donner naissance à plusieurs complexes protéine-pseudonœud et ralentir le processus de traduction. Le blocage des complexes serait éventuellement impossible vu qu'il n'y a pas assez d'ARN antisens pour le faire.

De ce fait, les prions sont des protéines porteuses d'information d'ordre structurelle. Les prions, dérivés de protéines du cerveau, pourraient induire chez des protéines normales un changement de conformation. Ce changement se propagerait graduellement et pourrait être transmis de l'espèce bovine à l'humaine. Les maladies à prions sont des maladies dégénératives du système nerveux central (c.f la maladie de la "vache folle") [25].

Le changement de conformation n'a pas encore été élucidé mais des expériences ont mis en évidence le rôle de la protéine du prion dans le métabolisme du cuivre au niveau du système nerveux central. La portion HGGG de l'octapeptide est suspectée d'être la portion minimale qui contiendrait le site nécessaire pour la liaison avec le cuivre [26].

Une autre hypothèse a été posée par Luck R et al [22] à partir de leur programme de prédiction de structure. Ils constatent une conservation d'une structure de tige boucle dans les 23 séquences d'ARN messagers étudiées. La composition en nucléotides de cette tige boucle en serait peut être la cause. Ils ont effectivement remarqué que les codons du demi brin 5' de l'hélice étaient d'usage courant, contrairement à ceux du demi brin 3'. Cette remarque concerne les codons GGU de glycine (G) qui sont en forte proportion de 70% alors que leur usage en général est en dessous de 30%. Une seconde observation concerne la boucle de la tige dont la première proline est toujours codée par le codon CCA dans les 23 séquences de l'expérience. La boucle est suspectée d'être l'élément responsable de la forme infectieuse du prion.

La structure du pseudo noeud recherchée est celle de forme H la plus répandue. Elle résulte de l'appariement, souvent considéré comme interaction tertiaire, de certains nucléotides qui se trouvent dans la boucle de la première hélice avec des nucléotides qui sont soit en aval ou en amont de la boucle de cette même hélice. Ces nucléotides formeraient une seconde hélice et amèneraient à la création d'une hélice quasi-continue suite à un empilement coaxial entre ces deux hélices (voir la figure 2-4).



A: Structure avant la formation du pseudo noeud **B: Structure après formation du pseudo noeud**

Figure 2-4: Schéma de la structure du pseudo nœud de type H. A droite (A), la structure avant la formation des appariements des nucléotides. A gauche (B), la structure du pseudo-nœud après les appariements. Les pointillés indiquent les futurs appariements des nucléotides. H1 (en vert) et H2 (en rouge) indiquent respectivement les hélices 1 et 2. B1 et B2 indiquent respectivement les boucles 1 et 2. L'empilement coaxial des deux hélices H1 et H2 crée une troisième hélice quasi continue (H1+H2).

Comme il a été mentionné au début du paragraphe, l'élément précurseur de l'investigation est l'alignement de la figure 2-5.

Pour revenir à la structure de pseudo nœud [20][28], à partir de l'alignement de la figure 2-5, la structure débute à partir de la seconde répétition du prion de l'humain et exactement de l'acide aminé H, mais on voit un décalage entre les répétitions de la séquence de la protéine du macaque rhésus et celle de la protéine du prion humain.

```
                CD-Length = 207 residues, only  35.7% aligned
                Score = 30.5 bits (68), Expect = 0.007

Query:  230  RPMPHPGGIWGINNQPPGTSWGNINRYP--GGSWGNINRYPGGSWGNIHLYPGINNPFPP  287
Sbjct:  1    KKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGWGQPH---GGGWGQPH-GGGWGQPHGG  56

Query:  288  GVLRPPGSSWNTPAGFPN  305
Sbjct:  57   GWGQPHGGGWGQGGGTHN  74
```

Figure 2-5: Alignement avec RPS-BLAST entre la protéine du prion chez l'humain (deuxième ligne : NM_000311) et la protéine STG du macaque rhésus (première ligne). Les deux premières lignes indiquent les résultats statistiques. La valeur du CD-Length (207) correspond au nombre de résidus du domaine conservé qui correspond à la longueur de la séquence consensus de l'alignement. Le pourcentage d'identité entre les deux séquences n'est que de 35.7%. Ce chiffre représente le pourcentage de résidus alignés. La valeur du score brut de 68 entre parenthèse provient de la sommation de scores d'une matrice pré-calculée Blosum80. Ce score est normalisé et a pour unité les bits. La E-value indique qu'on a 0.7% de chances de retrouver la séquence dans la base de données de recherche avec un score égal ou supérieur à 30.5 bits. La troisième ligne correspond à la requête soumise. Les chiffres au début et la fin des séquences correspondent aux positions de début et fin. Les acides aminés en rouge indiquent une similarité totale. En couleur bleu, ils le sont toujours moins avec une similarité moindre. Les traits indiquent les insertions ou suppressions dans les séquences.

L'alignement de la figure 2-5 est le résultat de la soumission de la séquence de la protéine STG du macaque rhésus au programme *RPS-BLAST*. L'algorithme de recherche est le même que celui de *BLAST* [29], il commence à chercher un ou deux mots de longueur par défaut de 3 résidus à partir desquels il étend sa recherche de part et d'autre avec un minimum de trous. Il s'arrête quand l'extension sera maximale. Ce programme recherche les séquences similaires à la requête dans une base de données de modèles. Ces modèles ou profiles sont nommés PSSM (Position Specific Score Matrix) car ils sont des matrices de scores où chaque case contient la probabilité d'avoir un acide aminé donné à une position donnée dans une séquence spécifique[30]. Le but de l'alignement est de retrouver une (ou des) séquence(s) similaire(s) à celle soumise. Un pourcentage d'alignement entre les résidus sera calculé. Cette valeur devrait être d'au moins 75% ce qui permettrait de poser une hypothèse de structure et donc de fonction. Dans le cas de l'alignement de la figure 2-5, la seule et la plus similaire séquence retrouvée est celle du

prion de l'humain mais le pourcentage de résidus alignés n'est que de 35% ce qui n'est pas suffisant pour supposer l'existence de la structure du pseudo nœud supposée être celle de la protéine du prion humain. Une autre observation qui ressort du résultat de la figure 2-5 est la séquence PHGGGWGQ répètée quatre fois qui se retrouve plus vers la fin du polymorphisme observée dans le macaque rhésus. Une bonne valeur de la E-value devrait être de 10^{-5} ou moins mais une plus grande valeur comme dans notre cas qui est de 0.7 % ne nous permet pas d'ignorer notre hypothèse ce qui explique notre poursuite de recherche du pseudo nœud en utilisant d'autres programmes.

2.3.2. Recherche du pseudo nœud avec pknots

Pour investiguer dans cette hypothèse, le programme *pknots* [10] spécialement dédié à la recherche du pseudo nœud a été appliqué. Il a été appliqué en premier sur la séquence d'ARN messager codant pour la protéine du prion humain. La structure secondaire prédite par le programme est une succession de tiges boucles. La figure 2-6 illustre ce résultat.

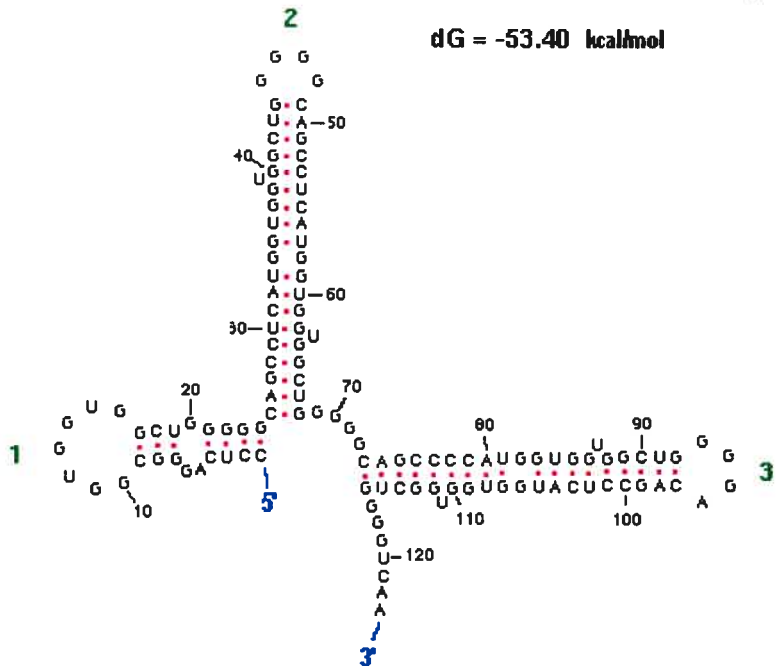


Figure 2-6: Résultat obtenu avec *pknots* appliqué sur la séquence ARN messager de la séquence du prion de l'humain NM_000311 codant pour les cinq répétitions : PQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQ. Les points rouges indiquent les appariements entre les bases. Les chiffres en vert indiquent le numéro de l'élément structural. Les chiffres 5' et 3' en bleu indiquent respectivement le début et la fin de la chaîne de nucléotides. L'énergie de la structure est indiquée par la valeur de dG en kcal/mol.

La structure de la figure 2-6 est principalement composée de trois tiges boucles. La première, la plus courte, est formée d'une hélice de sept paires qui sont fermées par une boucle également de sept nucléotides. Dans cette hélice, il y a une boucle interne asymétrique et les paires qui la constituent sont les cinq CG et les deux GU. Les deux autres tiges boucles sont formées d'une hélice, de dix neuf paires, qui est fermée par une boucle de quatre nucléotides. Les deux hélices sont composées de quatre paires AU, de quatre GU, de 10 GC, d'une boucle symétrique (1x1) et de deux 'bulges' U répartis de façon symétrique de part et d'autre du duplexe. Les deux tétra boucles fermantes sont essentiellement formées de G et en particulier la troisième (GGGA) qui fait partie de la famille des boucles GNRA. La séquence qui relie les deux hélices de dix neuf paires est aussi formée de quatre nucléotides G. Dans cette séquence on retrouve vingt et une glycine (G) dont onze proviennent du codon GGU ce qui correspond à plus de 50% de la

totalité des glycines. La structure prédite par *pknots* n'est certes pas celle du pseudo nœud mais néanmoins elle reflète bien la répétition des séquences surtout pour les quatre séquences des huit acides aminés PHGGGWG. Effectivement ces quatre séquences se retrouvent impliquées dans les deux hélices 2 et 3 qui sont identiques excepté pour leurs boucles fermantes respectives GGGG et GGGA. Ce dernier résultat nous laisse supposer qu'on devrait avoir le même résultat reflétant la répétition de nucléotides sur les trois séquences d'ARN messagers de ce projet. Mais la figure 2-7 illustre un tout autre résultat.

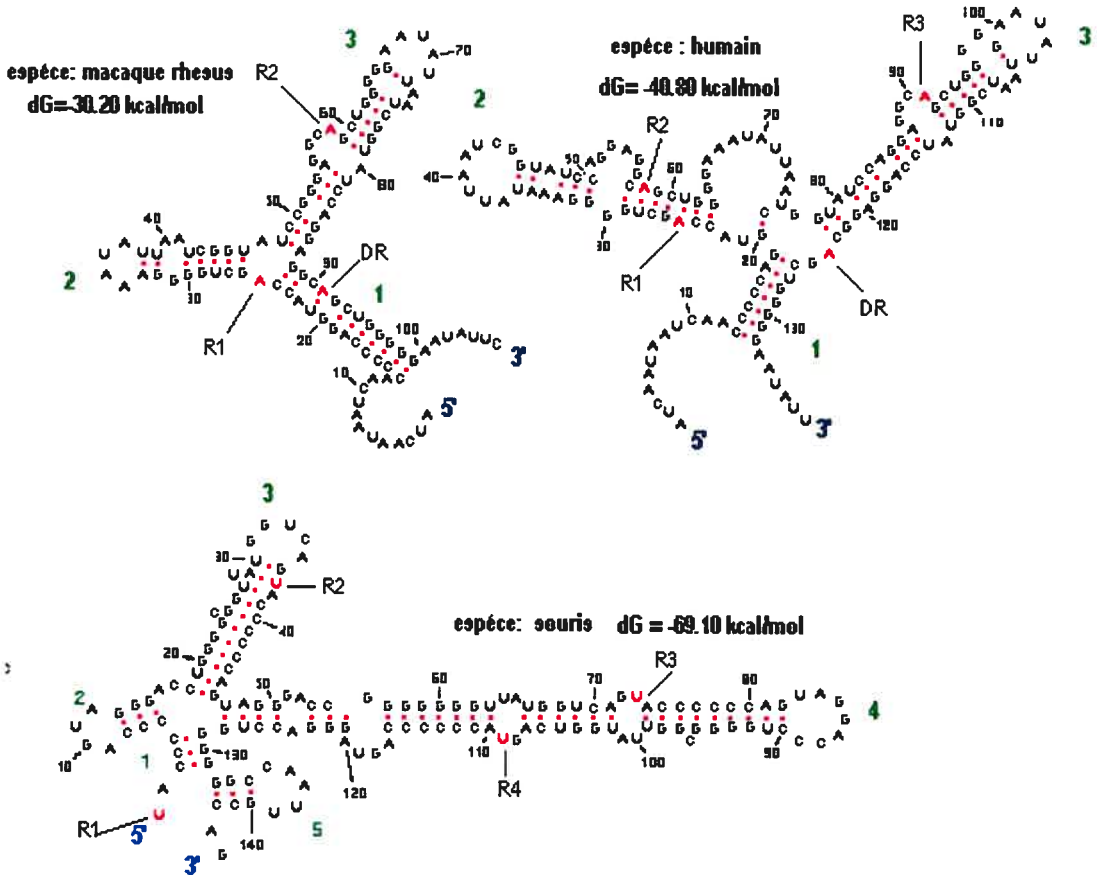


Figure 2-7: Résultats des structures secondaires prédites par *pknots* pour les ARN messagers des trois espèces. Les nucléotides en rouge indiquent le début d'une répétition. Leur numéro d'occurrence est indiqué par une lettre R + le numéro. Les demi répétitions sont indiquées par les lettres DR. Les points rouges indiquent les appariements entre les nucléotides. Les chiffres 5' et 3' indiquent respectivement les débuts et fins de chacune des séquences. Les numéros en vert indiquent le numéro de l'élément structural. Pour chacune des espèces, l'énergie est affichée en haut de la structure secondaire de chacune d'elles.

De la figure 2-7, on constate que dans aucune des trois structures prédites, le polymorphisme de séquences n'est reflété comme dans le cas de l'ARN messager du prion de la figure 2-7. Par contre pour les deux espèces, l'humaine et le macaque rhésus, on remarque que l'élément structural numéro 3 de l'humain est inclus dans le numéro 3 du macaque rhésus.

Une autre tentative a été d'appliquer *pknots* sur une seule séquence répétée de l'ARN messager de la souris composée de 36 nucléotides. Effectivement la structure de pseudo-nœud a été prédite. Cette structure est formée de deux hélices H1 et H2 et de deux boucles B1 et B2 dont une partie de B1 forme H2 comme le montre la figure 2-8.

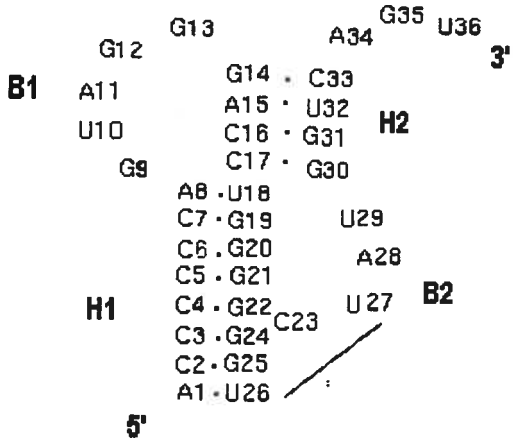


Figure 2-8: Structure de Pseudo-nœud trouvé par le programme *pknots* dans une répétition de 36 nucléotides de la souris. Les points rouges indiquent un appariement entre les nucléotides. Chaque nucléotide de la chaîne est indiqué par la base plus sa position dans la chaîne. Une hélice H1 de 8 nucléotides et une boucle B1 dont les nucléotides de la position 14 à 17 forment une hélice H2 suite aux appariements avec les nucléotides allant de la position 30 à 33. Les nucléotides de 27 à 29 créent la seconde boucle B2 de la structure de pseudo-nœud. L'énergie calculée de la structure est de -12.20 kcal/mol.

La même tentative a été testée sur une répétition de séquence des deux espèces, humaine et macaque rhésus, mais la structure prédite n'était pas celle du pseudo nœud.

2.3.3. Recherche du pseudo nœud avec RNAMot

A partir du résultat de la figure 2-8 de *pknots* sur les 36 nucléotides de la séquence de la souris, il serait intéressant de la rechercher dans toute la séquence. Pour cela, l'outil de recherche de motifs *RNAMot* a été utilisé. Pour ce programme la structure secondaire du pseudo nœud a été décrite selon la syntaxe appropriée au langage. Ces descriptions sont stockées dans un fichier nommé descripteur (voir figure 2-9). Ce dernier va être soumis au programme ainsi que le fichier contenant la séquence en entier sur laquelle la recherche sera effectuée.

```
parms
chk_both_strs = 0;
descr

h5 (tag="1", minlen=2, maxlen=3)
  h5 (tag="2", minlen=2, maxlen=5)
    ss (tag="1s", minlen=3, maxlen=5)
      h5 (tag="3", minlen=3, maxlen=4)
  h3(tag="2")
    ss (tag="2s", minlen=0, maxlen=1)
h3 (tag="1")
  ss (tag="3s", minlen=3, maxlen=6)
    h3 (tag="3")
ss (tag="4s", minlen=2, maxlen=3)
```

Figure 2-9: Descripteur avec *RNAMot* pour la souris. Dans la section débutant par *parms* il est précisé au programme de faire la recherche dans un seul sens (du 5' au 3'). Dans la section débutant par le mot clé *descr*, les informations concernant la structure y sont écrites. Chaque hélice est décomposée en deux parties h5 et h3. *h5* indique la demi hélice du côte 5' et *h3* pour le cote 3'. Les nucléotides non appariés sont indiqués par *ss* (single strand). Les couleurs ont été rajouté pour des fins explicatives visuelles, Les couleurs servent á mettre en les parties qui s'apparient ainsi que leur imbrication l'une par rapport à l'autre. Les paramètres entre parenthèses sont attachés à chacun des éléments structuraux. Le mot clé «*tag*» permet au programme de reconnaître les parties qui s'apparient en leur attachant leurs paramètres associés. Ces paramètres ne sont mentionnés qu'une seule fois pour par hélice du côte 5'.

La recherche de ce motif dans la séquence de la souris donne trois structures uniquement bien qu'on ait quatre répétitions. Le résultat est résumé dans le tableau 2-5.

h5 ("1")	*h5 ("2")*	*ss("1s")*	*h5 ("3")*	*h3("2")*	*ss("2s")*	*h3("1")*	*ss("3s")*	*h3("3")*	*ss ("4s")*
acc	cccca	gtagg	gacc	tgggg	c	ggt	tat	ggtc	agt
acc	cccca	gtagg	gacc	tgggg	g	ggt	tat	ggtc	ag
acc	cccca	gtag	gacc	*tgggg	c*	ggt	tat	ggtc	agt

Tableau 2-5: Résultat avec *RNAMOT* sur la séquence de la souris. La première ligne du tableau contient les éléments structuraux du descripteur de la structure. Les autres lignes contiennent le résultat du descripteur. Les colonnes avec les couleurs rouge, vert et bleu correspondent aux hélices. Le brin coté 5' est indiqué par le mot clé *h5* et celui du coté 3' par *h3*. La mise en correspondance entre les deux demis brins se fait grâce à des étiquettes (*"1", "2" et "3"*). Les colonnes avec des caractères en noir correspondent aux segments de nucléotides non appariés. Ces segments peuvent être des boucles (*ss("1s")* et *ss("3s")*) ou inter hélice *ss("2s")* ou bien après une hélice *ss ("4s")*. Les astérixes indique la suppression d'un nucléotide par le programme. La dernière ligne vide du tableau indique que le programme n'a pas trouvé la 4^{ième} structure dans la 4^{ième} répétition de la séquence de la souris.

Effectivement, le motif de *pknots* est retrouvé par *RNAMOT*, mais sa mauvaise manipulation de la séquence a mené à des résultats incomplets. Cette manipulation est observable à travers le changement que *RNAMot* a apporté à la séquence en supprimant des nucléotides. Ce changement débute à la cinquième colonne (*h3("2")*) de la troisième ligne, ce qui affecté le reste en omettant le motif de la quatrième répétition ce qui explique la quatrième ligne vide du tableau 2-5.

La seconde observation concerne la longueur de la boucle supérieure *ss("3s")* qui n'est que de trois nucléotides. Cette observation soulève la question suivante: « Est-ce cette conformation serait possible du point de vue stérique? » sachant que cette boucle devra couper le sillon mineur de l'hélice [*h5 ("1")+h5 ("2")*] qui est formé de huit paires plus un 'bulge'.

Pour montrer le résultat escompté par *RNAMot*, le tableau 2-6 donne une décomposition manuelle selon le motif du descripteur. En comparant ce tableau avec le tableau 2-5 on voit bien la différence à partir de la troisième ligne.

h5 ("1")	*h5 ("2")*	*ss("1s")*	*h5 ("3")*	*h3("2")*	*ss("2s")*	*H3("1")*	*ss("3s")*	*h3("3")*	*ss ("4s")*
acc	cccca	gtagg	gacc	tgggg	c	ggt	tat	ggtc	agt
acc	cccca	gtagg	gacc	tgggg	g	ggt	tat	ggtc	agt
acc	cccca	gtag	gacc	ctggg	g	cgg	tta	tggt	cagt
acc	cccca	gtag	ggac	ctggg	g	ggc	caa	ttgcc	ag

Tableau 2-6: Décomposition manuelle de la séquence de la souris selon le motif observé (souris). Ce tableau contient le résultat que *RNAMot* aurait du trouvé. La première ligne du tableau contient les éléments structuraux du descripteur de la structure. Les autres lignes contiennent le résultat du descripteur. Les colonnes avec les couleurs rouge, vert et bleu correspondent aux hélices. Le brin coté 5' est indiqué par le mot clé *h5* et celui du coté 3' par *h3*. La mise en correspondance entre les deux demis brins se fait grâce à des étiquettes (*"1", "2" et "3"*). Les colonnes avec des caractères en noir correspondent aux segments de nucléotides non appariés. Ces segments peuvent être des boucles (*ss("1s")* et *ss("3s")*) ou inter hélice *ss("2s")* ou bien après une hélice *ss ("4s")*.

Pour retrouver tous les motifs et résoudre les erreurs du tableau 2-5, un autre descripteur contenant les quatre motifs a été construit. Dans ce dernier, les quatre motifs ont été écrit pour chacune des quatre répétitions en s'assurant de la cohérence des étiquettes. Ces étiquettes sont attribuées à chacun des éléments structuraux qui se répètent dans chaque motif. Les mêmes erreurs se sont reproduites ce qui pourrait mettre en cause la représentation interne que le programme se fait du motif.

La même expérience a été menée sur les séquences de l'humain et le macaque rhésus. Le descripteur du motif de la souris a été repris et adapté pour tenir compte de la nouvelle longueur de répétitions de 32 nucléotides et de la composition de la séquence.

La séquence de la souris est plus riche en résidus G et C ce qui donne une structure plus riche en paire GC. Les pourcentages en GC sont mentionnés dans le tableau 2-7.

ARN messager Espèce	%CG
Macaque rhésus	50.62
Humain	48. 25
Souris	63.89

Tableau 2-7: Pourcentage de paires CG (colonne droite) dans chacune des trois espèces (colonne gauche). Ce pourcentage est calculé à partir des longueurs des séquences du tableau 2-2.

Pour tenir compte de cette différence, on permet a *RNAMot* d'inclure les paires non canoniques GA lors de la sa recherche. Pour les longueurs, les intervalles entre les longueurs minimales et maximales ont été réduits à 33 au lieu de 36 nucléotides. Le descripteur est celui de la figure 2-10.

```
parms
      chk_both_strs = 0;
      wc+=gu;
      ga={"g:a", "a:g" };
descr
h5 (tag="1", minlen=1, maxlen=5, pair+=ga)
  h5 (tag="2", minlen=2, maxlen=5, pair+=ga)
    ss (tag="1s", minlen=4, maxlen=5)
      h5 (tag="3", minlen=2, maxlen=5, pair+=ga)
  h3(tag="2")
     ss (tag="2s", minlen=0, maxlen=1)
h3 (tag="1")
  ss(tag="3s", minlen=4, maxlen=5)
     h3 (tag="3")
  ss (tag="4s", minlen=0, maxlen=1)
```

Figure 2-10: Descripteur avec *RNAMot* pour les deux espèces humaine et macaque rhésus. Chaque hélice est décomposée en deux parties h5 et h3. *h5* indique la demi hélice du brin 5' et *h3* pour le brin 3'. Les nucléotides non appariés sont indiqués par *ss* (single strand). Les couleurs ont été rajouté pour des fins explicatives visuelles, Les couleurs servent à mettre en évidence les parties qui s'apparient ainsi que leur imbrication l'une par rapport à l'autre. Les paramètres entre parenthèses sont attachés à chacun des éléments structuraux. Le mot clé « *tag* » permet au programme de reconnaître les parties qui s'apparient en leur attachant leurs paramètres associés. Ces paramètres ne sont mentionnés qu'une seule fois pour par hélice du côte 5'.

Le résultat du descripteur de la figure 2-10 est résumé par le tableau 2-8.

h5("1")	*h5 ("2")*	*ss("1s")*	*h5 ("3")*	*h3("2")*	*ss("2s")*	*H3("1")*	*ss("3s")*	*h3("3")*	*ss ("4s")*
ag	ctggg	Gaaat	atta	atcgg	t	at	ccag	gagg	c
ag	ctggg	Gaaat	atta	atcgg	t	at	ccag	gagg	c
ag	ctggg	Ggaat	atta	atcgg	t	at	ccag	gagg	c

Tableau 2-8: Résultat du descripteur de *RNAMot* appliqué sur la séquence de l'humain. La première ligne du tableau contient les éléments structuraux du descripteur de la structure. Les autres lignes contiennent le résultat du descripteur. Les colonnes avec les couleurs rouge, vert et bleu correspondent aux hélices. Le brin coté 5' est indiqué par le mot clé *h5* et celui du coté 3' par *h3*. La mise en correspondance entre les deux demis brins se fait grâce à des étiquettes (*"1", "2" et "3"*). Les colonnes avec des caractères en noir correspondent aux segments de nucléotides non appariés. Ces segments peuvent être des boucles (*ss("1s")* et *ss("3s")*) ou inter hélice *ss("2s")* ou bien après une hélice *ss ("4s")*.

Le résultat escompté est celui du tableau 2-9.

h5("1")	*h5 ("2")*	*ss("1s")*	*h5 ("3")*	*h3("2")*	*ss("2s")*	*H3("1")*	*ss("3s")*	*h3("3")*	*ss ("4s")*
ag	ctggg	Gaaat	atta	atcgg	t	at	ccag	gagg	c
ag	ctggg	Gaaat	atta	atcgg	t	at	ccag	gagg	c
ag	ctggg	Ggaat	atta	atcgg	t	at	ccag	gagg	c

Tableau 2-9: Décomposition manuelle de la séquence de l'humain selon le motif observé. Ce tableau contient le résultat que *RNAMot* aurait du trouver. La première ligne du tableau contient les éléments structuraux de la structure observable. Les autres lignes contiennent le résultat de la décomposition. Les colonnes avec les couleurs rouge, vert et bleu correspondent aux hélices. Le brin 5' est indiqué par le mot clé *h5* et celui du 3' par *h3*. La mise en correspondance entre les deux demis brins se fait grâce à des étiquettes (*"1", "2" et "3"*). Les colonnes avec des caractères en noir correspondent aux segments de nucléotides non appariés. Ces segments peuvent être des boucles (*ss("1s")* et *ss("3s")*) ou inter hélice *ss("2s")* ou bien après une hélice *ss ("4s")*.

En comparant les deux tableaux 2-8 et 2-9, on constate qu'ils sont identiques. Cette similarité signifie que *RNAMot* a effectivement trouvé le pseudo-nœud escompté et observé. Le résultat du tableau 2-7 ne présente aucune erreur de manipulation de la séquence. Les trois lignes, correspondant au nombre de répétitions de la séquence, sont correctement remplies. La première particularité de cette structure est la présence de paires GA dans les deux hélices H1 et H2. Comparativement aux longueurs des hélices ces paires forment presque la moitié des hélices. Pour chacune des hélices H1 et H2, on retrouve respectivement 3 sur 7 et 2 sur 4 paires de GA sur la totalité des paires. La seconde particularité est la taille de la boucle inférieure `ss("3s")` de 4 nucléotides comparée à celle de l'hélice H1 formée elle de 7 nucléotides avec un "bulge".

Du point de vue graphique la structure des tableaux 2-17 et 2-18 est illustrée par la figure 2-11 :

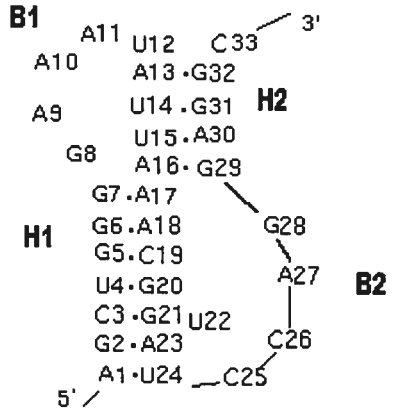


Figure 2-11: Pseudo Nœud trouvé au niveau de l'humain et le macaque rhésus par RNAMOT. Les nucléotides de la séquence sont accompagnés de leur position (nucléotide+position). Les appariements sont indiqués par des points rouges). La structure est formée de deux hélices H1 et H2 avec deux boucles B1 et B2 dont 5 nucléotides de B1 forment un demi brin de l'hélice H2. Le nucléotide U à la position 22 se retrouve en « bulge » dans l'hélice H1. Les traits noirs indiquent la continuité de la chaîne. Les débuts et fins de la chaîne sont indiqués respectivement par les chiffres 5' et 3'.

2.3.4. Pseudo nœud à l'œil nu

Une autre structure de pseudo nœud est observable à l'œil nu pour les trois séquences. Pour celui des deux espèces, humaine et macaque rhésus, il y a une tétra boucle GNRA (GAAA) comme boucle supérieure B1. Deux 'bulges' sont de part et d'autre de l'hélice H1. L'un est le G à la position 5 et l'autre qui est un U est à son opposé à la position 20. La paire GU ferme l'hélice H1, le G étant à la position 9 et le U à la 16 de la chaîne. Ce dernier pourrait s'apparier au G de la position 30, ce qui donnerait naissance au triplet G9-U16-G30. La même situation serait possible entre la paire GU, dont les résidus G et U sont aux positions respectives 8 et 17, et le nucléotide A à la position 29. Ce triplet serait donc G8-U16-A29. Également dans cette structure la boucle inférieure B2 de 6 nucléotides serait plus longue que celle prédite par *RNAMot*. Cette structure est illustrée par la figure 2-12.

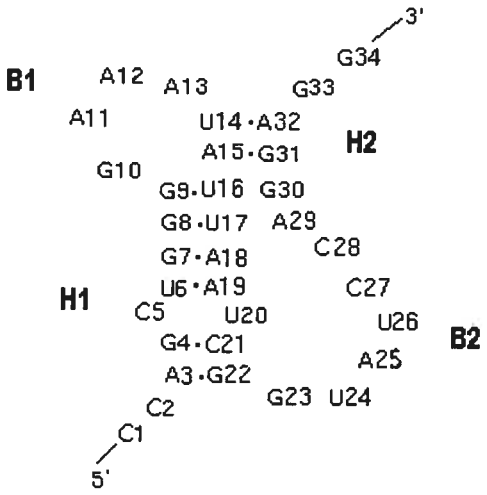


Figure 2-12: Structure avec le descripteur Pseudo Nœud et boucle GNRA de l'humain et le macaque rhésus. Les nucléotides de la séquence sont accompagnés de leur position (nucléotide+position). Les appariements sont indiqués par des points rouges). La structure est formée de deux hélices H1 et H2 avec deux boucles B1 et B2 dont 5 nucléotides de B1 forment un demi brin de l'hélice H2. Les nucléotides C à la position 5 et le U à la position 20 se retrouvent en « bulge » dans l'hélice H1. Les débuts et fins de la chaîne sont indiqués respectivement par les chiffres 5' et 3'.

En ce qui concerne la souris, le pseudo nœud est lui aussi observable mais sa boucle supérieure est la GGGA qui fait partie de la famille des tétra boucles GNRA. Au niveau de l'hélice H1 on ne retrouverait qu'un seul 'bulge' sur le brin 5'. Mais si le nucléotide à la position 18 interagirait avec le nucléotide à la position 28 de la chaîne, on aurait un triplet avec la paire AU fermant H1. On pourrait également faire la même supposition entre la paire UG, dont les nucléotides sont aux positions respectives 10 et 19, et le nucléotide U à la position 27. Si ces deux triplets se formeraient on obtiendrait un 'bulge' à la position 29. La figure 2-13 illustre la structure du pseudo nœud au niveau de la souris.

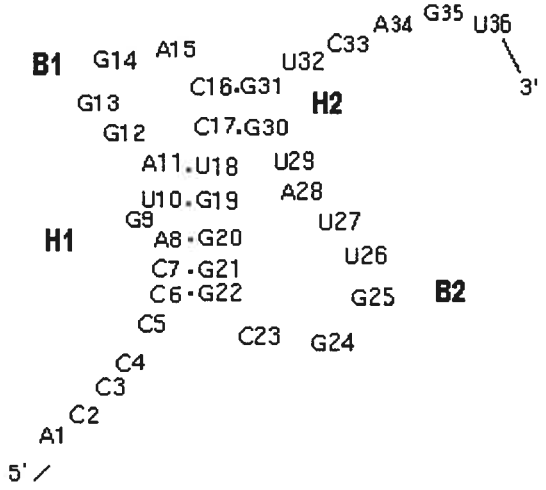


Figure 2-13: Structure avec le descripteur Pseudo Nœud et boucle GNRA de la souris. Les nucléotides de la séquence sont accompagnés de leur position (nucléotide+position). Les appariements sont indiqués par des points rouges). La structure est formée de deux hélices H1 et H2 avec deux boucles B1 et B2 dont 5 nucléotides de B1 forment un demi brin de l'hélice H2. Le nucléotide G à la position 9 se retrouve en 'bulge' dans l'hélice H1. Les débuts et fins de la chaîne sont indiqués respectivement par les chiffres 5' et 3'.

2.3.5. Conclusion de l'hypothèse du pseudo nœud

En conclusion à l'hypothèse du pseudo nœud, les résultats de l'alignement de la protéine rmSTG avec celle du prion ne nous ont pas permis de confirmer ou infirmer l'hypothèse. Cette situation nous a poussé à investiguer plus en utilisant d'autres programmes. En premier, on s'est aperçu que le programme *pknots* dédié à la recherche de ce type de structure ne l'a pas prédite dans aucune des trois séquences d'ARN messagers. Néanmoins ça n'exclut pas totalement la possibilité d'existence de la structure vu qu'un pseudo nœud a été retrouvé dans une séquence répétée de la souris de 36 nucléotides. Dans cette dernière prédiction, le programme a retrouvé une boucle inférieure B2 courte comparée à l'hélice H1. Cette observation a soulevé un questionnement concernant la possibilité de construire la structure dans un espace 3D. Le résultat de *pknots* sur la souris nous a aidé à faire la recherche dans les deux autres espèces avec l'outil de recherche de motif *RNAMot*. Ce changement de programmes reflète la combinaison des différents outils informatiques qui est souvent nécessaire pour mener une recherche de structure. Les résultats du paragraphe 2.2.4 montrent des structures observables uniquement à l'œil nu et que le programme *pknots* aurait pu trouver s'il avait été possible de rajouter certaines contraintes à travers un fichier. Il aurait été intéressant si *pknots* pouvait faire des vérifications sur les longueurs des boucles par rapport à celles des hélices afin de maintenir une proportionnalité. Cette proportionnalité serait à déterminer en fonction des limites stériques de la structure.

2.4. Hypothèse des boucles GNRA

Cette seconde hypothèse ressort après observation de la suite des quatre nucléotides GAAA et GGAA au niveau des séquences des ARN messagers de l'humain et du macaque rhésus. Pour la séquence de la souris on observe plutôt la GGGA. Ces observations ont permis d'orienter la recherche vers une structure avec des boucles formées de quatre nucléotides observées ci-dessous. Le programme *RNAMot* a été utilisé pour écrire le descripteur. Ce descripteur va servir à reconnaître une structure de deux hélices qui s'empilent coaxialement l'une sur l'autre avec une tétra boucle fermant l'hélice gauche (H1). Ces boucles font partie de la famille des boucles GNRA et dans notre cas ça serait soit une GAAA, une GGAA ou une GGGA selon l'espèce. On aurait également un 'bulge' dans l'hélice H1. La position de ce 'bulge' dépendra de la paire fermante de H1. Le motif est illustré par la figure 2-14.

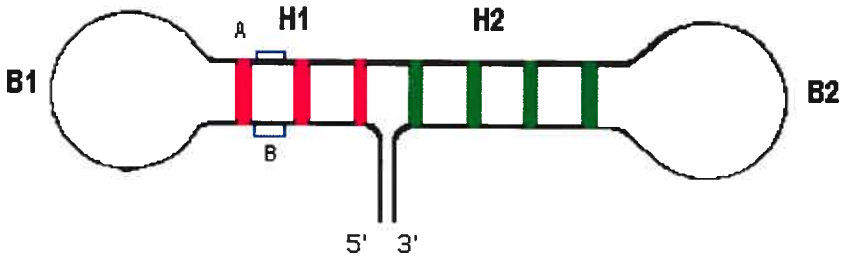


Figure 2-14: Motif recherché avec RNAMotif avec un bulge. Deux hélices H1 (en rouge) et H2 (en vert) sont coaxialement empilées et sont fermées respectivement par les boucles B1 et B2. Les deux ainsi empilées forment une hélice quasi continue. Le bulge est en bleu. Il peut être soit à la position A ou B selon la paire qui le précède.

Les résultats de la recherche pour chacune des trois espèces sont illustrés par les figures 2-15, 2-16 et 2-17. La stabilité de ces trois structures pourrait être dûe à l'empilement coaxial en créant une hélice quasi continue.

Les structures de l'humain et du macaque rhésus sont semblables. Elles ne différent que par leur nombre de copies du motif qui est le même que celui des séquences répétées. Le point commun entre ces deux structures est les boucles GNRA. L'avant-dernière et dernière boucle sont toujours des GGAA. De la première jusqu'à l'avant-dernière c'est une GAAA. Le 'bulge' qui est une adénine devance la paire fermante GU tout en étant sur le brin 3'. Les deux hélices H1 et H2 sont reliées par une paire canonique GA. Au niveau des séquences demi répétées (SWGNI et YPPVGTW) on retrouve la moitié gauche du motif de la figure 2-15 avec la boucle GNRA. On retrouve également trois paires GU successives du côté de la tétra boucle.

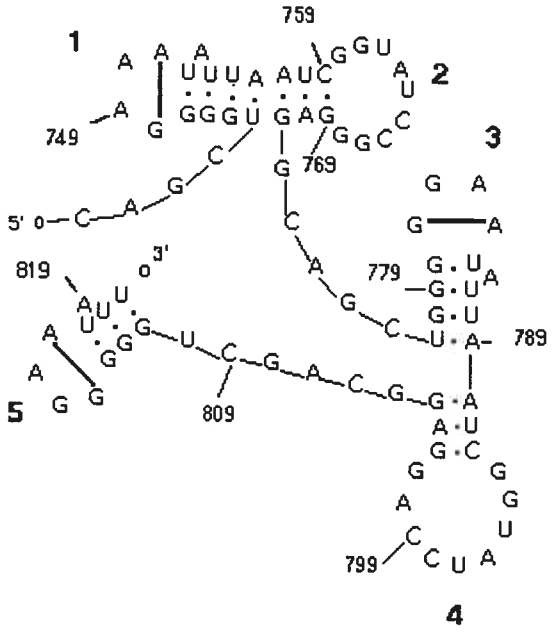


Figure 2-15: Structure du macaque rhésus selon le descripteur du 740^{éme} au 821^{éme} nucléotide. Les positions des nucléotides sont données avec des pas de 10 (749,759..819). Les appariements entre les nucléotides sont indiqués par des points rouges. Les traits noirs indiquent la connexion entre les nucléotides. Les numéros en gras (1, 2, 3, 4 et 5) indiquent le numéro de l'élément structural. Les traits noirs et gras entre les nucléotides G et A des boucles 1, 3 et 5 indiquent la formation de la paire de la boucle G(A/G)AA. Les débuts et fins de la chaîne sont indiqués respectivement par les lettres 5' et 3'.

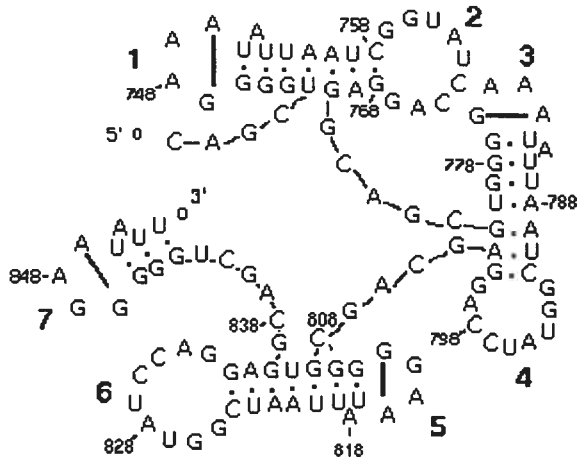


Figure 2-16: Structure de l'humain selon le descripteur du 739^{éme} au 753^{éme} nucléotide. Les positions des nucléotides sont donnés avec des pas de 10 (748, 758, .., 848). Les appariements entre les nucléotides sont indiqués par des points rouges. Les traits noirs indiquent la connexion entre les nucléotides. Les numéros en gras (1, 2, 3, 4, 5, 6 et 7) indiquent le numéro de l'élément structural. Les traits noirs et gras entre les nucléotides G et A des boucles 1, 3, 5 et 7 indiquent la formation de la paire de la boucle G (A/G)AA. Les débuts et fins de la chaîne sont indiqués respectivement par les lettres 5' et 3'.

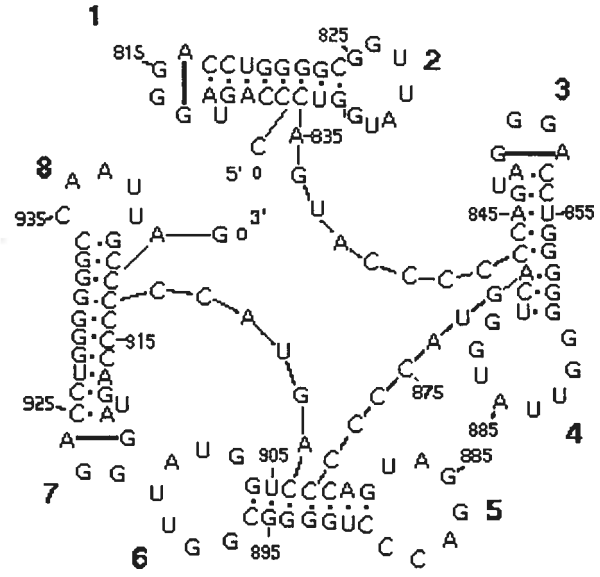


Figure 2-17: Structure de la souris selon le descripteur du 806^{éme} au 944^{éme} nucléotide. Les positions des nucléotides sont donnés avec des pas de 10 (815, 825,...,935). Les appariements entre les nucléotides sont indiqués par des points rouges. Les traits noirs indiquent la connexion entre les nucléotides. Les numéros en gras (1, 2, 3, 4, 5, 6, 7 et 8) indiquent le numéro de l'élément structural. Les traits noirs et gras entre les nucléotides G et A des boucles 1, 3, et 7 indiquent la formation de la paire de la boucle GGGA. Les débuts et fins de la chaîne sont indiqués respectivement par les lettres 5' et 3'.

En ce qui concerne la structure de la souris (voir figure 2-18), elle est différente des autres (voir figures 2-16 et 2-17) du point de vue de la composition des paires. Elle ressemble aux deux autres structures par la présence des tétra boucles GNRA mais le 'bulge' est un U qui est sur le brin 5' et devance la paire non canonique AC.

Plusieurs éléments structuraux sont apparus des structures retrouvées. Pour tenter d'expliquer la structure il va falloir exposer les caractéristiques de ces composants entre autre la tétra boucle GAAA/GGGA qui ressort dans toutes les boucles impaires des trois structures ainsi que la paire GU et sa paire isostérique AC. Les paragraphes qui vont suivre sont dédiés à ces éléments.

2.4.1. Motif de la tétra boucle GNRA

La boucle GNRA est un motif structural dans lequel le N à la deuxième position peut être n'importe quel nucléotide (A, G, C, U) mais le R à la troisième position ne peut être qu'une purine donc un A ou un G. Le nucléotide G à la première position et le A à la quatrième position forment une paire dans laquelle deux liens existent entre les deux hydrogènes de chacun des atomes. Ces liens sont illustrés par la figure 2-19.

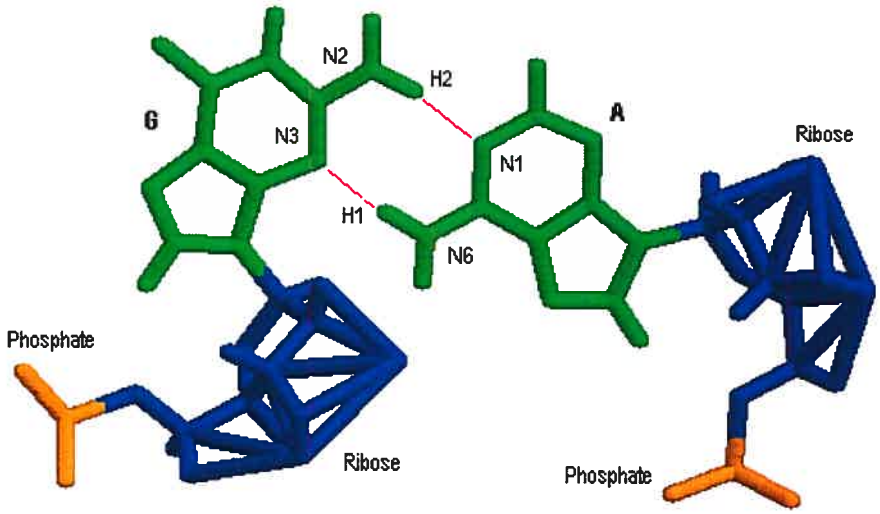


Figure 2-18: Les liens de la paire GA entre le sillon mineur du G et le majeur de A. Le phosphate est en orange, le ribose en bleu et la base en vert. G est le symbole de la base guanine et le A pour l'adénine. Les atomes sont indiqués par leur symbole chimique approprié ainsi que leur position dans la base. N indique l'azote et le H l'hydrogène. Les liens hydrogènes sont en rouge entre les atomes concernés. Le troisième et le deuxième hydrogène du deuxième azote du G sont liés respectivement au premier hydrogène du sixième azote et au premier azote du A.

La stabilité de cette structure a été trouvé par des expériences de RMN. Elles mettent en évidence un réseau de liaisons hydrogènes illustré par la figure 2-19 en incluant celui de la figure 2-18 et un empilement entre les trois dernières bases qui lui est illustré par la figure 2-20.

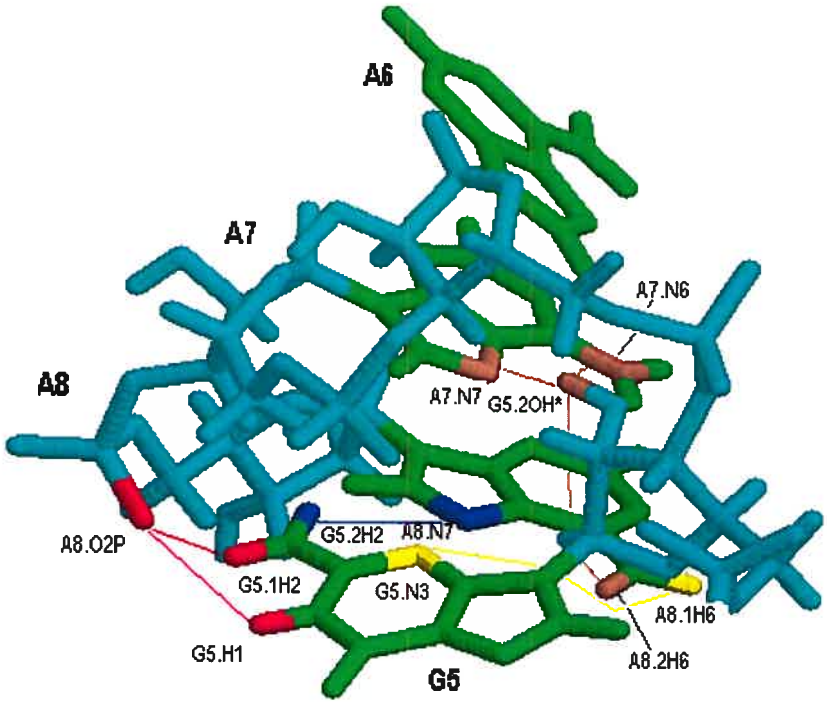


Figure 2-19: Réseau des 7 liens hydrogènes dans une GAAA (numérotée de 5 à 8) en 3D. Les phosphates et les riboses sont en cyan. Les bases sont en vert. Les atomes concernés par les liens sont mis en rouge, jaune, marron ou bleu. Des traits de même couleur entre ces atomes indiquent la liaison hydrogène. Les bases sont indiqués par un symbole, G pour Guanine ou A pour adénine, ainsi qu'un numéro pour indiquer leur position dans la boucle. Les éléments en rouge indiquent les deux liens de l'oxygène numéro 2 du phosphate de l'adénine à la position 8 avec l'hydrogène numéro 1 et le numéro 2 de l'azote numéro 1 du G à la position 5. Les éléments en bleu indiquent la liaison de l'hydrogène numéro 2 de l'azote numéro 2 du G avec l'azote numéro 7 à la position numéro 8. Les éléments en jaune indiquent la liaison entre l'azote numéro 3 du G avec l'hydrogène numéro 1 de l'azote numéro 6 du A. Les éléments en marron indiquent la liaison du groupe hydroxyle du ribose du G avec les trois éléments suivants : 1) l'hydrogène numéro 2 de l'azote numéro 6 du nucléotide A à la position 8 de la boucle. 2) L'azote numéro 6 du A à la position 7. 3) L'azote numéro 7 du nucléotide à la position 7.

La boucle GAAA a pour caractéristique d'être stable de part la liaison GA et l'empilement des trois A. Une autre caractéristique de cette tétra boucle est le changement de direction du phosphate du squelette entre le premier et le deuxième nucléotide. Les nucléotides qui suivent le phosphate tournant adoptent une conformation

dans l'espace qui les prédisposent à former des interactions tertiaires en exposant les faces Watson-Crick vers le solvant (voir figure 2-20).

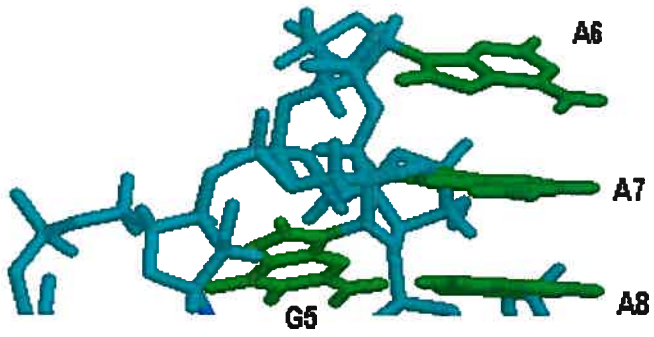


Figure 2-20: Conformation des nucléotides d'une boucle GAAA (source: fichier 1ZIF.pdb/ GAAA). Les phosphates et les riboses sont en cyan. Les bases sont en vert. Les bases sont indiqués par des symboles, G pour Guanine et A pour adénine. Elles sont accompagnées d'un chiffre indiquant leur position dans la chaîne. Les trois derniers A de la boucle sont bien empilées l'une sur l'autre et leurs phosphates de l'autre côté créent une courbure.

Le motif GNRA est très répandu dans les longs ARN tel que le ribozyme groupe intron I, il a pour rôle principal d'interagir avec des éléments d'une hélice tel que les paires GC en tandem ou bien des structures en tiges boucles d'une autre molécule. La figure 2-21 illustre une interaction entre la GAAA et son récepteur.

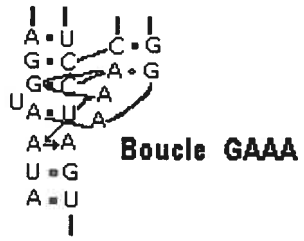


Figure 2-21: Interaction entre la GAAA et son récepteur dans P5B du l'intron du groupe I (1ajf.pdb). Les points rouges pleins indiquent des appariements canoniques et les creux les non canoniques tel que le « wobble » et la paire AG. Les traits en bleu indiquent les interactions entre les nucléotides de la boucle GAAA à droite avec un bout de l'hélice à gauche.

Le second mode d'interaction tétra boucle-ARN que fait une boucle GNRA avec le sillon mineur de deux paires Watson-crick GC en tandem a été observé à partir de la structure de cristal du ribozyme en tête de marteau, où une GAAA se lie à deux paires de GC successives [31].

Le motif GNRA peut se retrouver sous d'autres formes [32], en boucle interne, ou d'un seul côté d'une boucle interne ou bien en tétra boucle. La dernière forme citée est la plus connue en général.

2.4.2. Les paires GU et le 'bulge' A

La plupart des contacts d'une structure d'ARN impliquent le sillon mineur de la forme A de l'hélice, vu son accessibilité. Cependant, les paires non canoniques, les boucles et les 'bulges' peuvent perturber la conformation de l'hélice ainsi que son potentiel électrostatique. Ces perturbations peuvent par contre être des endroits cibles pour permettre à des ions de métaux de s'y lier pour neutraliser les charges négatives. Avec l'intron du groupe I (domaine P4-P6), il a été mis en évidence la présence de ce genre de sites dans le sillon majeur. Ces sites sont proches des jonctions d'hélices. Un

tandem de paires GU fermant la tétra boucle est également présent. La figure 2-22 illustre cette interaction.

Figure 2-22: Site d'interaction avec le métal formé par les paires GU en tandem. Les deux appariements GU sont indiqués par des cercles noirs creux entre les nucléotides. A coté de chacun des nucléotides, un numéro indique leur position dans la chaîne dont ce bout a été extrait. A la fin de l'hélice on retrouve une tétra boucle indiquée par deux traits noirs creux. Le site d'interaction de l'osmium, entre les deux paires GU, est indiqué par des traits noirs. L'osmium (III) hexamine est utilisé pour la détermination de la structure de l'ARN.

L'ion d'osmium (III) hexamine est un métal utilisé dans la détermination de la structure de l'ARN vu sa ressemblance géométrique avec l'ion de magnésium hydraté. Dans la figure 2-22, il se lie directement aux bases du tandem GU par des liens hydrogènes. Il se retrouve entre les oxygènes de phosphate d'un côté et de l'autre les hydrogènes donneurs des sillons majeurs de G et U [33].

La paire GU se retrouve dans la plupart des ARN fonctionnels et il a été prouver qu'elle joue plusieurs rôles vu ces propriétés chimiques et structurales. L'ARN messager codant pour les protéines S15 en est un exemple. Ceci est dû à la présence des paires GU en tandem qui est un signal de reconnaissance pour l'auto régulation de la synthèse des protéines.

Dans certains contextes, les paires GU stabiliseraient les courbures raides du squelette de l'ARN, comme par exemple dans l'ARN de transfert où la paire GU se retrouve au niveau de la jonction de la boucle V simple brin et l'hélice T avec une

courbure raide de la chaîne [34]. Ce qui pourrait expliquer dans le cadre de cette recherche leur présence proche de la tétra boucle GAAA, vu le changement de courbure raide du squelette de phosphate (voir figure 2-21).

En ce qui concerne le 'bulge' A sa conformation dans l'hélice peut être de deux types: à l'intérieur ou à l'extérieur de la chaîne. Ces deux conformations sont illustrées par la figure 2-23.

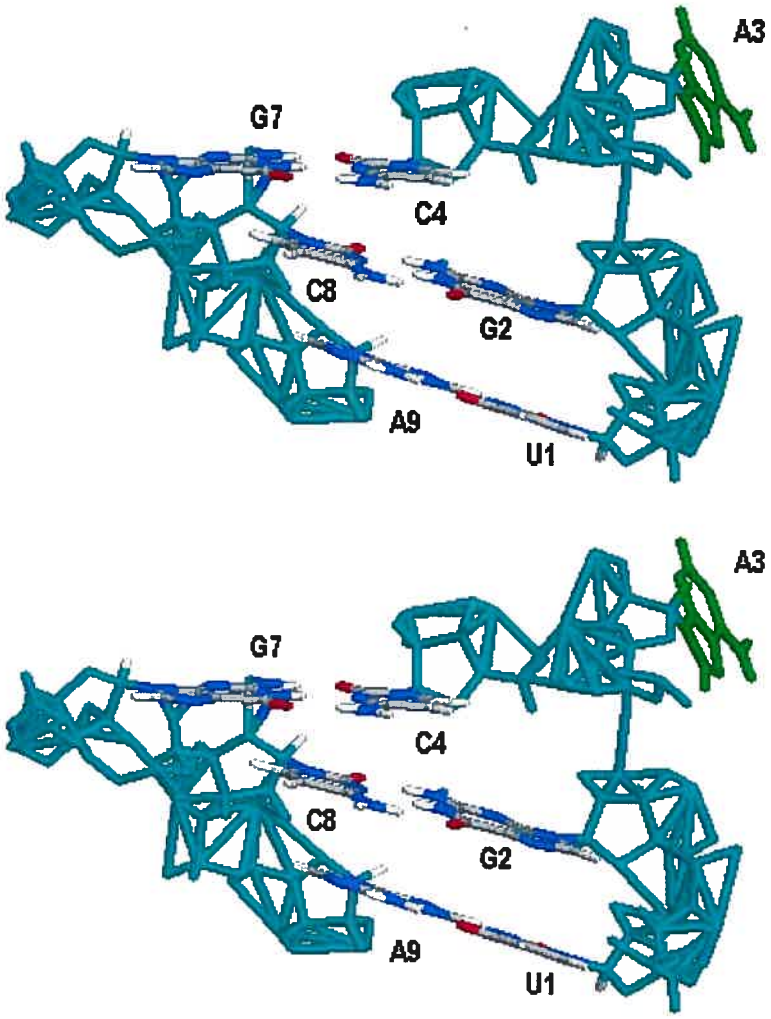


Figure 2-23: Conformations pour l'hélice formée de 3 paires {[U1-A9], [G2-C8], [C4-G7]}. Chacune des bases est accompagnée de sa position dans l'hélice (nucléotide+position). Le nucléotide A3 est un 'bulge'. Pour la même hélice on a deux conformations possibles du 'bulge'. En haut le 'bulge' est à l'intérieur de l'hélice. En bas, il est à l'extérieur. Les phosphates et les sucres sont en cyan. La base du 'bulge' A3 est en vert. Le bleu est pour l'azote. Le gris pour le carbone et le rouge pour l'oxygène. Ces trois atomes composent la base.

Dans les deux conformations, on observera des déformations de l'hélice. Les phosphates chargés négativement auront tendance à se rapprocher. Ce rapprochement va favoriser les liens avec des ions de métal chargés positivement et donc atténuer la densité de charge négative. Dans le cas du 'bulge' A des figures 2-15 et 2-16, il pourrait élargir la surface exposée au solvant vu qu'il est proche de la GNRA qui expose déjà ces deux derniers nucléotides au solvant. Si le 'bulge' serait au contraire orienté vers l'intérieur, il pourrait créer des liens avec chacun des G des deux paires entre lesquelles il se retrouve. Chacun des G aurait un lien hydrogène libre qui serait disponible pour une liaison avec le « bulge » A.

Pour la structure de la souris de la figure 2-17, la paire la plus proche de la tétra boucle GNRA est la paire AC. Du point de vue structural, cette paire serait la plus proche de la paire GU. Cette isostérie permettrait de remplacer la paire GU par la AC [35]. La figure 2-24 illustre cette isostérie.

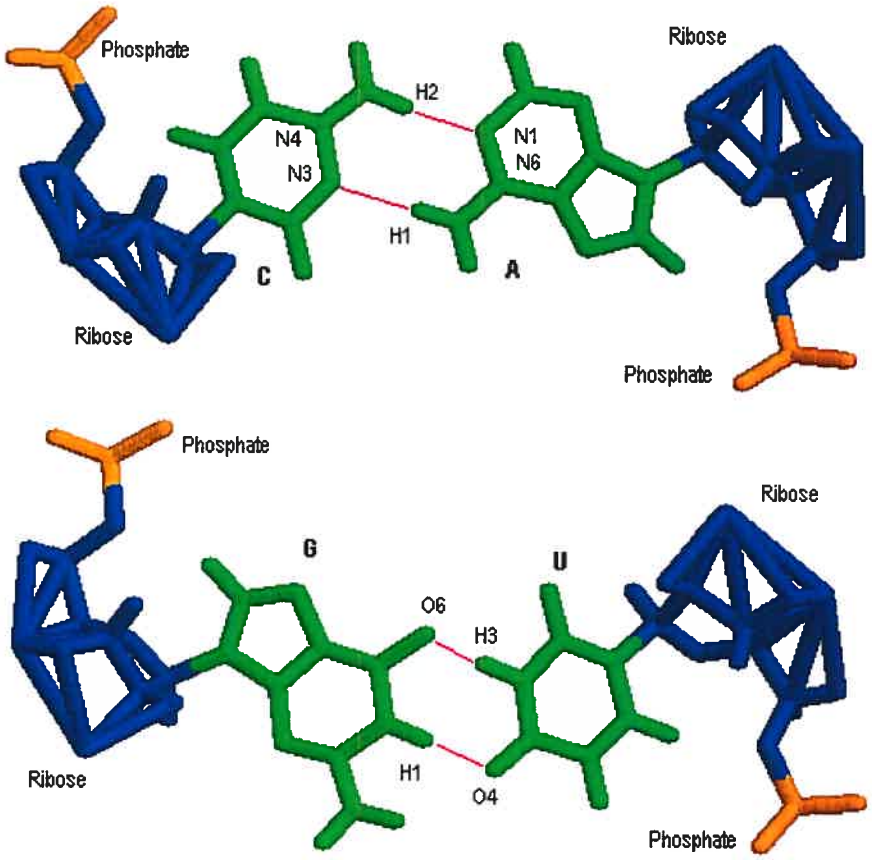


Figure 2-24: Comparaison d'une des conformations de la paire AC avec une de la paire GU iso stérique qui lui est similaire et pourrait la remplacer. Les phosphates sont en orange, les riboses en bleu et les bases en vert. Chacune des bases est indiquée par un symbole, G pour guanine, A pour adénine, C pour cytidine et U pour uracile. L'azote est indiqué par la lettre N, l'hydrogène par H et l'oxygène par O. Chacun des atomes est accompagné de sa position dans la base ou bien par rapport à l'azote auquel il est attaché en ce qui concerne l'hydrogène, s'il y a lieu. Les liens hydrogènes sont en rouge. Pour la paire AC (en haut de l'image), on a 2 liens : l'hydrogène numéro 1 de l'azote à la position 6 de A avec l'azote à la position 3 du C et l'azote à la position 1 du A avec l'hydrogène numéro 2 de l'azote à la position 4 de C. Pour la paire GU (en bas de l'image), on a également deux liens: l'oxygène à la position 4 du U se lien à l'hydrogène à la position 1 du G et l'hydrogène à la position 3 du U se lien à l'oxygène a la position 6 du G.

2.4.3. Conclusion de l'hypothèse des boucles GNRA

Pour cette hypothèse les structures trouvées reflètent bien le polymorphisme en répétant les motifs. Effectivement *RNAMotif* a donné des résultats cohérents. Son seul inconvénient, dans ce cas, est son incapacité de calculer les énergies libres. Les raisons sont les absences des valeurs d'énergie de la tétra boucle GAAA fermée par une boucle GU et celles des hélices qui s'empilent coaxialement.

Le second point est la présence du motif GNRA qui pourrait être utile á la thérapie antisens [36]. Cette supposition est due au fait que la GNRA est similaire au motif UNR qui lui est connu d'être une cible pour des antisens [37]. Dans cette thérapie des fragments courts d'ADN pourraient se lier au motif selon les règles de complémentarité Watson-Crick qui inhiberait la production de la protéine. Le processus est illustré par la figure 2-25.

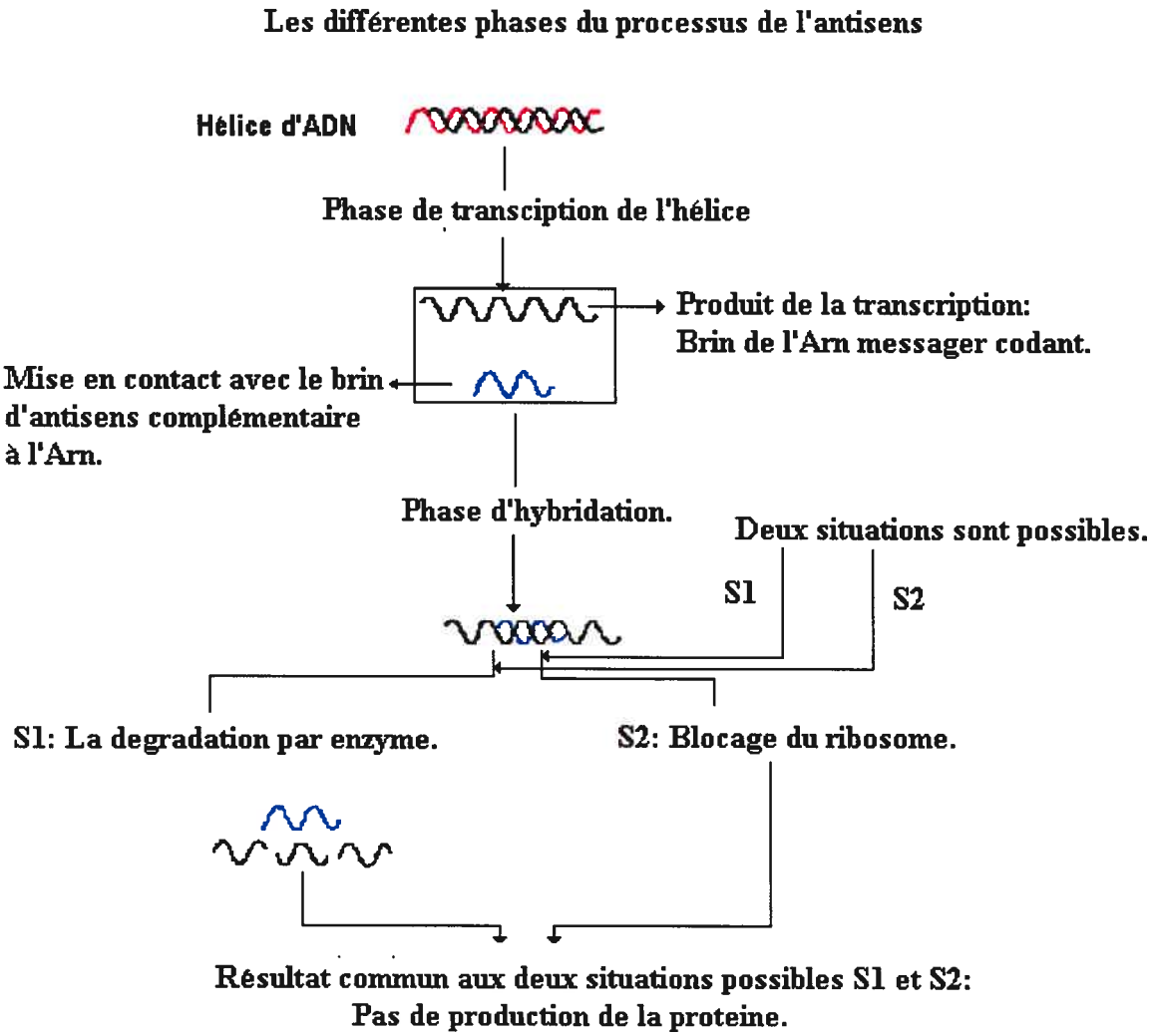


Figure 2-25: Inhibition de la production d'une protéine par le technique de l'antisens. Les deux brins de l'hélice sont en rouge et noir. Le brin antisens est en bleu. Avec double hélice d'ADN et suite à une étape de transcription, un ARN messager codant est produit. Ce dernier est mis en contact avec un brin antisens auquel il va s'apparier. Après hybridation, la traduction sera impossible par le ribosome et il y aura dégradation de ce double brin par un enzyme. Ce processus empêchera la formation de la protéine.

Le processus débute dans une zone non structurée tel que les boucles des tiges, qui sont d'ailleurs les plus communes pour débuter une nucléation et de plus doivent avoir au minimum quatre nucléotides, par la suite l'hélice adjacente sera défaite. Ce

processus s'arrête quand il se retrouve à la fin de l'hélice ou bien il rencontre une courbure raide [38].

Dans le cas des structures de cette recherche, le processus pourrait débuter au niveau des boucles paires ou bien sur le brin reliant les deux répétitions et il s'arrêterait dés qu'une GNRA serait rencontrée. La difficulté majeure de cette technique est de trouver justement ces sites de début de nucléation. Cette hypothèse de structure pourrait éventuellement donner des indices et aiderait à accélérer la recherche de structure de manière expérimentale.

Chapitre 3

Recherche de la structure tridimensionnelle

3.1. Introduction

Dans ce dernier volet de la recherche on s'intéressera à la construction de la molécule en 3D à partir de la structure secondaire. Une fois la construction terminée certains traitements seront appliqués aux fichiers représentants le modèle brut. Ces traitements seront sous forme de réadaptions syntaxique et de rééquilibrage de charges. Par la suite les processus de minimisation et de dynamique moléculaire qui sont la dernière phase de la recherche du modèle théorique avant de passer aux expériences en laboratoire. Cette étape de modélisation nous informera sur l'agencement des résidus dans l'espace tridimensionnel ce qui nous permettra de poser une hypothèse. Le modèle produit ne sera pas totalement correct mais pourrait mettre en évidence des éléments structuraux qui permettraient de mettre en œuvre les expériences existantes déjà ou bien donner naissance à de nouvelles.

3.2. Construction des modèles avec MC-SYM

Pour la construction de la molécule et sa modélisation, le programme *MC-SYM* (Macromolecular Conformations by SYMbolic programming) [39] [39] a été utilisé. Ce programme repose sur la technique du chaînage arrière pour la construction de la molécule en respectant des contraintes. La construction se fait à partir d'informations émanant de la structure secondaire. Les informations ainsi que d'autres contraintes sont décrites dans un fichier ASCII. Ce dernier est décomposé en sept sections et chacune

d'elles est dédiée à la description d'un aspect nécessaire à la construction d'une molécule. Ces aspects sont introduits dans le fichier script dans l'ordre de citation qui suit: la séquence biologique en elle-même, informations sur les résidus, informations sur les paires à construire, informations sur l'empilement des résidus successifs, la description de la construction de la molécule en chaînage arrière, la description des contraintes d'acceptation de la molécule et les informations sur l'exploration des différentes possibilités de construction de la molécule en précisant le lieu de stockage du résultat. Ces différentes sections sont illustrées par la figure 3-1.

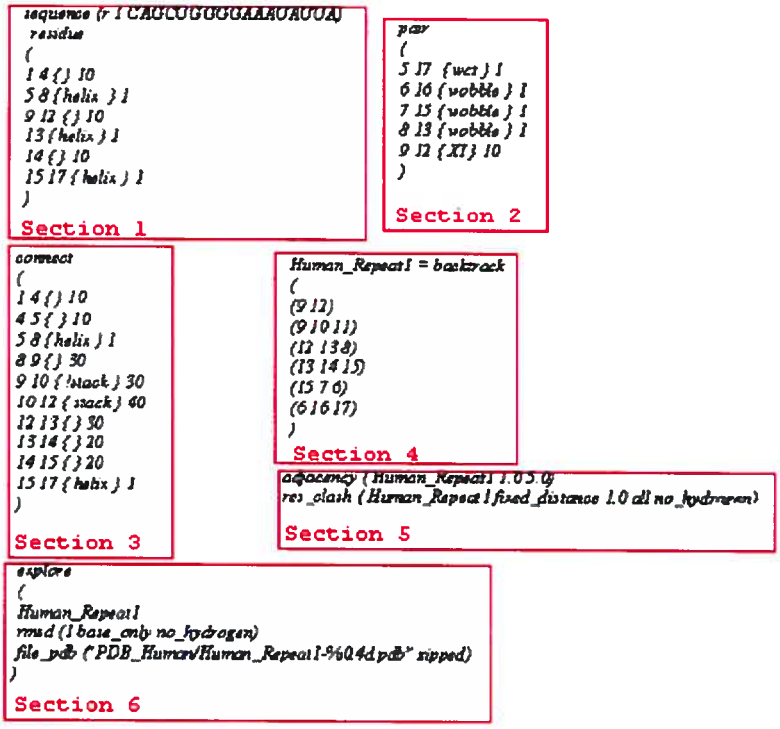


Figure 3-1: Exemple des différentes sections d'un script *MC-SYM*. Ce script a été utilisé pour construire une partie de la répétition de la séquence de l'humain. La section 1 (rassemble deux sections) décrit la séquence de nucléotides à construire et la conformation des résidus. La séquence est entre parenthèse et précédée par le numéro 1 qui représente le début de la numérotation de la séquence et le r avant ce numéro précise que c'est un ARN. Par la suite la description de la conformation adoptée par des résidus, pour les résidus libres les parenthèses et le nombre de conformations possibles suit, ce chiffre est donné par l'utilisateur et le programme se limite à ce nombre lors de sa recherche. Pour les résidus qui font partie d'une hélice, on retrouve entre parenthèses le mot réservé *helix* et le chiffre qui suit est toujours 1. La section 2 décrit les résidus qui forment des paires dont le type est mentionné entre { }. La section 3 informe le programme sur la relation entre les résidus successifs. La section 4 informe le programme sur l'ordre de construction de la structure en indiquant l'ordre de placement des résidus. La section 5 indique les contraintes d'acceptation des structures, la première concerne la contrainte de distance entre les atomes O3' (Oxygène du ribose) et P (Phosphate) de résidus adjacents qui doit être comprise entre 1.0 Å et 5.0 Å et la seconde contrainte permet de refuser une structure contenant des atomes plus proches que 1.0 Å. La section 6 est la section ou est entamée la recherche des conformations possibles et où est mentionné le chemin de stockage des structures trouvées et répondant aux contraintes.

La génération des structures se fait selon les contraintes géométriques telles que celles décrites dans le script de la figure 3-1. L'agencement des paires et des atomes est extrait d'une base de donnée (espace de recherche de conformations). Cette base de données accompagne *MC-SYM* et englobe les géométries des paires provenant d'expériences de laboratoire. Dans le cas des scripts de ce travail, les contraintes sont larges, excepté pour les boucles GNRA ou l'empilement des bases NRA a été précisé. Pour les nucléotides qui sont au niveau de la jonction des hélices, dans la section 3 « *connect* » de figure 3-1, l'interaction base-base est du type hélice. Les bornes d'adjacence entre deux bases sont entre 1.0 et 5.0 Angström. Au niveau des paires, quand une n'est pas canonique comme la paire AC au niveau de la souris, aucune contrainte n'a été fixé et il a juste été mentionné l'appariement.

Les scripts *MC-SYM* ont été exécutés sur des machines d'un cluster et qui ont chacune 2 micros processeurs AMD Athlon de 1.2Ghz avec 1 GB de mémoire. L'exécution des scripts a demandé des temps de calcul d'environ 3 à 4 semaines. Les constructions des molécules dans leur totalité se sont faites par assemblages partiels. Chacun de ces derniers correspond à une séquence répétée du polymorphisme. Un ensemble de molécules en 3D a été généré selon différentes conformations possibles des résidus. La flexibilité des scripts a donné un grand nombre de solutions possibles. Le tableau 3-1 donne le nombre de structures générées et il faut savoir aussi que les programmes ont été arrêtés délibérément vue leur nombre très élevé.

Espèce	Nombre de structures
Humaine	2765
macaque rhésus	1833
Souris	3565

Tableau 3-1 : Le nombre de structures trouvées par *MC-SYM* dans chacune des rois espèces.

3.3. Minimisation des modèles avec AMBER

Le but de cette étape est de corriger les faux contacts des modèles bruts générés par *MC-SYM*, ils ont été minimisés par *Sander* un des modules du programme AMBER4.1 (Assisted Model Building with Energy Refinement) [41] de champ de force. Ce programme est composé de modules et chacun accomplit une tâche. La recherche consiste à faire une minimisation de l'énergie intra moléculaire, elle ne tient pas compte des interactions avec le solvant mais juste des interactions entre les atomes constituants les résidus de la molécule. La minimisation d'énergie en mécanique moléculaire revient à optimiser la géométrie initiale de la molécule mais avant d'entamer ce processus il a fallu transformer et réadapter les structures produites par *MC-SYM*.

3.3.1. Étapes de Traitement

La première étape est la transformation des structures selon le format PDB, en le soumettant au programme *toleap-3.2.5* qui a pour rôle de changer la syntaxe des fichiers et de rééquilibrer les charges de la molécule aux deux extrémités 5' et 3'. Les changements de syntaxe sont effectués sur chacun des atomes des résidus en les réarrangeant dans un certain ordre. Le premier résidu ainsi que le dernier subissent des équilibrages de charge en supprimant certains atomes sur le sucre et le phosphate qui sont aptes à faire des liaisons entre eux. Les transformations sur ces résidus sont illustrées respectivement par les figures 3-2 et 3-3. Les transformations que subissent les autres résidus compris entre le premier et le dernier sont illustrées par la figure 3-4.

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z
ATOM	1	C1*	C	1	-1.866	-12.894	1.804
ATOM	2	C2*	C	1	-2.482	-12.478	3.143
ATOM	3	C3*	C	1	-3.979	-12.587	2.819
ATOM	4	C4*	C	1	-4.021	-13.020	1.916
ATOM	5	C5*	C	1	-5.207	-13.911	0.977
ATOM	6	H1*	C	1	-0.930	-13.416	2.005
ATOM	7	H2*	C	1	-2.195	-11.500	3.529
ATOM	8	H3*	C	1	-4.380	-11.684	2.358
ATOM	9	H4*	C	1	-4.111	-14.710	2.539
ATOM	10	O1P	C	1	-6.449	-11.230	-1.496
ATOM	11	O2*	C	1	-2.099	-13.429	4.100
ATOM	12	O2P	C	1	-7.795	-12.957	-0.265
ATOM	13	O3*	C	1	-4.788	-12.751	3.975
ATOM	14	O4*	C	1	-2.801	-13.719	1.126
ATOM	15	O5*	C	1	-5.342	-12.685	0.234
ATOM	16	P	C	1	-6.498	-12.568	-0.865
ATOM	17	1H5*	C	1	-6.115	-14.093	1.551
ATOM	18	2H5*	C	1	-5.065	-14.743	0.287
ATOM	19	HO2*	C	1	-2.499	-13.176	5.015
ATOM	20	HO3*	C	1	-5.234	-13.679	3.954
ATOM	21	C2	C	1	-0.406	-11.028	1.207
ATOM	22	C4	C	1	-0.950	-9.512	-0.657
ATOM	23	C5	C	1	-2.142	-10.302	-0.862
ATOM	24	C6	C	1	-2.408	-11.359	-0.076
ATOM	25	H5	C	1	-2.827	-10.040	-1.655
ATOM	26	H6	C	1	-3.306	-11.934	-0.248
ATOM	27	N1	C	1	-1.563	-11.721	0.939
ATOM	28	N3	C	1	-0.158	-9.947	0.385
ATOM	29	N4	C	1	-0.556	-8.451	-1.345
ATOM	30	O2	C	1	0.357	-11.336	2.106
ATOM	31	1H4	C	1	-1.111	-8.114	-2.106
ATOM	32	2H4	C	1	0.296	-7.987	-1.104

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z
ATOM	1	C3'	RC5	1	-3.979	-12.587	2.819
ATOM	2	C4	RC5	1	-0.950	-9.512	-0.657
ATOM	3	C2	RC5	1	-0.406	-11.028	1.207
ATOM	4	C2'	RC5	1	-2.482	-12.478	3.143
ATOM	5	C1'	RC5	1	-1.866	-12.894	1.804
ATOM	6	H6	RC5	1	-3.306	-11.934	-0.248
ATOM	7	C4'	RC5	1	-4.021	-13.820	1.916
ATOM	8	C5	RC5	1	-2.142	-10.302	-0.862
ATOM	9	C5'	RC5	1	-5.207	-13.911	0.977
ATOM	10	C6	RC5	1	-2.408	-11.359	-0.076
ATOM	11	2H4	RC5	1	0.296	-7.987	-1.104
ATOM	12	H5	RC5	1	-2.827	-10.040	-1.655
ATOM	13	N1	RC5	1	-1.563	-11.721	0.939
ATOM	14	N3	RC5	1	-0.158	-9.947	0.385
ATOM	15	N4	RC5	1	-0.556	-8.451	-1.345
ATOM	16	O2	RC5	1	0.357	-11.336	2.106
ATOM	17	O2'	RC5	1	-2.099	-13.429	4.100
ATOM	18	O3'	RC5	1	-4.788	-12.751	3.975
ATOM	19	O4'	RC5	1	-2.801	-13.719	1.126
ATOM	20	O5'	RC5	1	-5.342	-12.685	0.234
ATOM	21	1H4	RC5	1	-1.111	-8.114	-2.106

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z

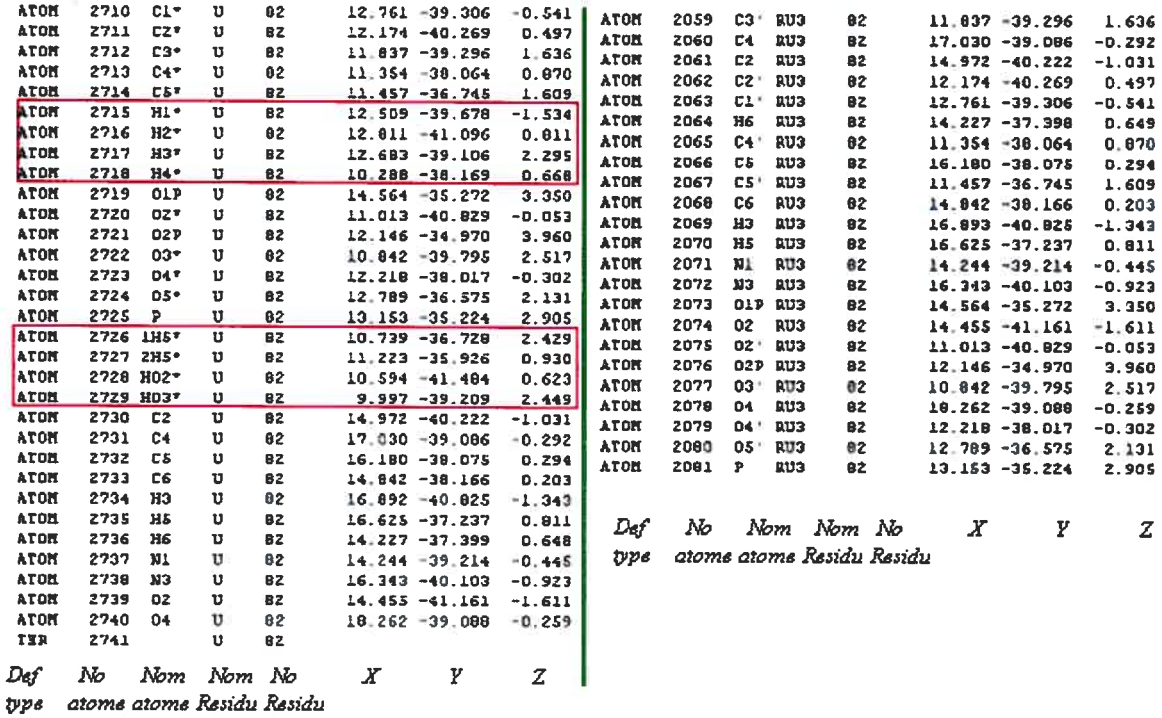
Figure 3-2: Transformations apportés au premier résidu de la molécule d'ARN. A droite on retrouve les atomes du résidu avant l'application du programme *toleap-3.2.5* et à gauche le résidu après transformation. Les 11 atomes dans le rectangle rouge sont supprimés par le programme. Ces atomes sont les 2 oxygènes (O1P et O2P) sur le phosphate et les 4 autres sur l'oxygène (O2*, O3*, O4* et O5*), les 2 groupes hydroxyles sur le ribose (HO2* et HO3*) et les deux hydrogènes sur le carbone 5 du ribose (1H5* et 2H5*). L'astérixe (*) accompagnant les atomes du ribose est remplacée par une apostrophe ('). La lettre R est rajoutée avec caque nom de résidu pour préciser que c'est un résidu d'une molécule d'ARN. Les coordonnées des atomes sont mentionnées dans les colonnes X, Y et Z.

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z
ATOM	33	C1*	A	2	-0.319	-9.627	5.901
ATOM	34	C2*	A	2	-0.004	-8.940	7.141
ATOM	35	C3*	A	2	-2.245	-9.623	7.267
ATOM	36	C4*	A	2	-1.925	-11.055	6.040
ATOM	37	C5*	A	2	-3.094	-11.864	6.313
ATOM	38	H1*	A	2	0.760	-9.698	6.039
ATOM	39	H2*	A	2	-0.952	-7.953	7.105
ATOM	40	H3*	A	2	-3.022	-9.145	6.670
ATOM	41	H4*	A	2	-1.589	-11.616	7.712
ATOM	42	O1P	A	2	-5.549	-10.885	3.511
ATOM	43	O2*	A	2	-0.045	-9.286	8.229
ATOM	44	O2P	A	2	-5.939	-12.336	5.522
ATOM	45	O3*	A	2	-2.752	-9.578	8.603
ATOM	46	O4*	A	2	-0.947	-10.893	5.773
ATOM	47	O5*	A	2	-3.733	-11.159	5.232
ATOM	48	P	A	2	-4.992	-11.828	4.505
ATOM	49	1H5*	A	2	-3.813	-12.033	7.115
ATOM	50	2H5*	A	2	-2.739	-12.832	5.959
ATOM	51	HO2*	A	2	-0.395	-8.839	9.089
ATOM	52	HO3*	A	2	-2.790	-10.532	8.988
ATOM	53	C2	A	2	1.644	-6.180	4.110
ATOM	54	C4	A	2	0.156	-7.788	4.200
ATOM	55	C5	A	2	-0.404	-7.422	3.011
ATOM	56	C6	A	2	0.160	-6.317	2.352
ATOM	57	C8	A	2	-1.509	-9.096	3.684
ATOM	58	H2	A	2	2.483	-5.651	4.536
ATOM	59	H8	A	2	-2.233	-9.895	3.749
ATOM	60	N1	A	2	1.205	-5.710	2.944
ATOM	61	N3	A	2	1.201	-7.197	4.816
ATOM	62	N6	A	2	-0.287	-5.848	1.177
ATOM	63	N7	A	2	-1.463	-8.262	2.692
ATOM	64	N9	A	2	-0.559	-8.872	4.638
ATOM	65	1H6	A	2	0.155	-5.057	0.754
ATOM	66	2H6	A	2	-1.062	-6.289	0.725

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z
ATOM	22	H2	RA	2	2.483	-5.651	4.536
ATOM	23	C3'	RA	2	-2.245	-9.623	7.267
ATOM	24	C4	RA	2	0.156	-7.788	4.200
ATOM	25	C2	RA	2	1.644	-6.180	4.110
ATOM	26	C2'	RA	2	-0.884	-8.940	7.141
ATOM	27	C1'	RA	2	-0.319	-9.627	5.901
ATOM	28	H8	RA	2	-2.233	-9.895	3.749
ATOM	29	C4'	RA	2	-1.925	-11.055	6.040
ATOM	30	C5	RA	2	-0.404	-7.422	3.011
ATOM	31	C5'	RA	2	-3.094	-11.864	6.313
ATOM	32	C6	RA	2	0.160	-6.317	2.352
ATOM	33	C8	RA	2	-1.509	-9.096	3.684
ATOM	34	2H6	RA	2	-1.062	-6.289	0.725
ATOM	35	N1	RA	2	1.205	-5.710	2.944
ATOM	36	N3	RA	2	1.201	-7.197	4.816
ATOM	37	N6	RA	2	-0.287	-5.848	1.177
ATOM	38	N7	RA	2	-1.463	-8.262	2.692
ATOM	39	N9	RA	2	-0.559	-8.872	4.638
ATOM	40	O1P	RA	2	-5.549	-10.885	3.511
ATOM	41	O2'	RA	2	-0.045	-9.286	8.229
ATOM	42	O2P	RA	2	-5.939	-12.336	5.522
ATOM	43	O3'	RA	2	-2.752	-9.578	8.603
ATOM	44	O4'	RA	2	-0.947	-10.893	5.773
ATOM	45	O5'	RA	2	-3.733	-11.159	5.232
ATOM	46	P	RA	2	-4.992	-11.828	4.505
ATOM	47	1H6	RA	2	0.155	-5.057	0.754

Def type	No atome	Nom atome	Nom Residu	No Residu	X	Y	Z

Figure 3-3: Transformations apportés au second résidu de la molécule d'ARN. A droite on retrouve les atomes du résidu avant l'application du programme *toleap-3.2.5* et à gauche le résidu après transformation. Les 8 atomes dans le rectangle rouge sont supprimés par le programme. Ces atomes sont les 4 hydrogènes (H1*, H2*, H3* et H4*) sur le ribose, les 2 groupes hydroxyles sur le ribose (HO2* et HO3*) et les deux hydrogènes sur le carbone 5 du ribose (1H5* et 2H5*). L'astérixe (*) accompagnant les atomes du ribose est remplacée par une apostrophe (').La lettre R est rajoutée avec caque nom de résidu pour préciser que c'est un résidu d'une molécule d'ARN. Les coordonnées des atomes sont mentionnées dans les colonnes X, Y et Z.



ATOM	2710	C1*	U	82	12.761	-39.306	-0.541
ATOM	2711	C2*	U	82	12.174	-40.269	0.497
ATOM	2712	C3*	U	82	11.837	-39.296	1.636
ATOM	2713	C4*	U	82	11.354	-38.064	0.870
ATOM	2714	C5*	U	82	11.457	-36.745	1.609
ATOM	2715	H1*	U	82	12.509	-39.678	-1.534
ATOM	2716	H2*	U	82	12.811	-41.096	0.811
ATOM	2717	H3*	U	82	12.683	-39.106	2.295
ATOM	2718	H4*	U	82	10.288	-38.169	0.668
ATOM	2719	O1P	U	82	14.564	-35.272	3.350
ATOM	2720	O2*	U	82	11.013	-40.829	-0.053
ATOM	2721	O2P	U	82	12.146	-34.970	3.960
ATOM	2722	O3*	U	82	10.842	-39.795	2.517
ATOM	2723	O4*	U	82	12.218	-38.017	-0.302
ATOM	2724	O5*	U	82	12.789	-36.575	2.131
ATOM	2725	P	U	82	13.153	-35.224	2.905
ATOM	2726	1H5*	U	82	10.739	-36.728	2.429
ATOM	2727	2H5*	U	82	11.223	-35.926	0.930
ATOM	2728	HO2*	U	82	10.594	-41.484	0.623
ATOM	2729	HO3*	U	82	9.997	-39.209	2.449
ATOM	2730	C2	U	82	14.972	-40.222	-1.031
ATOM	2731	C4	U	82	17.030	-39.086	-0.292
ATOM	2732	C5	U	82	16.180	-38.075	0.294
ATOM	2733	C6	U	82	14.842	-38.166	0.203
ATOM	2734	H3	U	82	16.892	-40.825	-1.343
ATOM	2735	H5	U	82	16.625	-37.237	0.811
ATOM	2736	H6	U	82	14.227	-37.399	0.648
ATOM	2737	N1	U	82	14.244	-39.214	-0.445
ATOM	2738	N3	U	82	16.343	-40.103	-0.923
ATOM	2739	O2	U	82	14.455	-41.161	-1.611
ATOM	2740	O4	U	82	18.262	-39.088	-0.259
TER	2741		U	82			

Def type	*No atome*	*Nom atome*	*Nom Residu*	*No Residu*	*X*	*Y*	*Z*

ATOM	2059	C3'	RU3	82	11.837	-39.296	1.636
ATOM	2060	C4	RU3	82	17.030	-39.086	-0.292
ATOM	2061	C2	RU3	82	14.972	-40.222	-1.031
ATOM	2062	C2'	RU3	82	12.174	-40.269	0.497
ATOM	2063	C1'	RU3	82	12.761	-39.306	-0.541
ATOM	2064	H6	RU3	82	14.227	-37.398	0.649
ATOM	2065	C4'	RU3	82	11.354	-38.064	0.870
ATOM	2066	C5	RU3	82	16.180	-38.075	0.294
ATOM	2067	C5'	RU3	82	11.457	-36.745	1.609
ATOM	2068	C6	RU3	82	14.842	-38.166	0.203
ATOM	2069	H3	RU3	82	16.893	-40.825	-1.343
ATOM	2070	H5	RU3	82	16.625	-37.237	0.811
ATOM	2071	N1	RU3	82	14.244	-39.214	-0.445
ATOM	2072	N3	RU3	82	16.343	-40.103	-0.923
ATOM	2073	O1P	RU3	82	14.564	-35.272	3.350
ATOM	2074	O2	RU3	82	14.455	-41.161	-1.611
ATOM	2075	O2'	RU3	82	11.013	-40.829	-0.053
ATOM	2076	O2P	RU3	82	12.146	-34.970	3.960
ATOM	2077	O3'	RU3	82	10.842	-39.795	2.517
ATOM	2078	O4	RU3	82	18.262	-39.088	-0.259
ATOM	2079	O4'	RU3	82	12.218	-38.017	-0.302
ATOM	2080	O5'	RU3	82	12.789	-36.575	2.131
ATOM	2081	P	RU3	82	13.153	-35.224	2.905

Def type	*No atome*	*Nom atome*	*Nom Residu*	*No Residu*	*X*	*Y*	*Z*

Figure 3-4: Transformations apportés au dernier résidu (3') de la molécule d'ARN. A droite on retrouve les atomes du résidu avant l'application du programme *toleap-3.2.5* et à gauche le résidu après transformation. Les 8 atomes dans le rectangle rouge sont supprimés par le programme. Ces atomes sont les 4 hydrogènes (H1*, H2*, H3* et H4*) sur le ribose, les 2 groupes hydroxyles sur le ribose (HO2* et HO3*) et les deux hydrogènes sur le carbone 5 du ribose (1H5* et 2H5*). L'astérixe (*) accompagnant les atomes du ribose est remplacée par une apostrophe (').La lettre R est rajoutée avec caque nom de résidu pour préciser que c'est un résidu d'une molécule d'ARN. Les coordonnées des atomes sont mentionnées dans les colonnes X, Y et Z. La fin de la chaîne est indique par le mot clé réservé TER.

La seconde étape sera de soumettre les fichiers transformés au programme *tleap* (un des modules d'*AMBER*) qui aura pour tâche de générer deux fichiers, un contenant la topologie de la structure et l'autre les coordonnées cartésiennes des atomes constituant le modèle. Les coordonnées cartésiennes en général proviennent des expériences de cristallographie aux rayons X, ou bien de la résonance magnétique nucléaire ou bien de modèles théoriques construits comme dans le cas des structures de ce projet. Les fichiers de topologie contiennent les informations concernant la connectivité, les noms des

atomes, le nom des résidus ainsi que leur charge. Les topologies proviennent de la base des données accompagnant le programme *AMBER*.

3.3.2. Soumission au champ de force

Après traitements, la molécule sera soumise à un champ de force qui est un ensemble de fonctions d'énergies associé à une série de paramètres numériques obtenus expérimentalement ou évalués théoriquement. Le but de cette soumission est de rechercher la conformation de la structure la plus stable en déterminant les minima d'énergie globale d'interaction intramoléculaire. Le terme variable de cette énergie dépend de la construction de la molécule et de l'arrangement de ses atomes. Cette énergie potentielle est fractionnée en un certain nombre de termes additifs indépendants.

3.3.2.1. Énergies composant le champ de force

En mécanique moléculaire, la structure moléculaire est considérée comme étant composée de billes et de ressorts (forces harmoniques) associée à une série de fonctions de potentiel. La somme de ces fonctions est exprimée sous la forme d'un champ de force moléculaire entre atomes liés et non liés selon l'équation 1.

$$E_{total} = E_{liés} + E_{non\ liés} \quad \textit{(Equation 1)}$$

a) Énergie d'interactions entre atomes liés

L'énergie d'interaction des atomes liés revient à l'énergie de déformation du squelette décrite à l'aide des élongations des liaisons et des distorsions des angles de valence. Les termes de déformations sont formulés par les équations 2 et 3.

$$E_{liaison} = \tfrac{1}{2} \sum K_B(b-b_o)^2 \quad \textit{(Equation 2)} \text{ et}$$

$$E_{angle\ de\ valence} = \tfrac{1}{2} \sum K_\theta(\theta-\theta_o)^2 \quad \textit{(Equation 3)}$$

K_B et K_θ sont des constantes de force dérivées de l'analyse vibrationnelle des molécules modèles. $K_B = 400$ kJ.mol^{-1}.Å^{-2} et $K_\theta = 40$ kJ.mol^{-1}.deg^{-2}. b_o et θ_o sont respectivement les élongations en Å et rotations en degrés de la molécule à l'état initial.

En plus de ces deux termes, il y a la déformation des angles de dièdres impropres, qui décrit la sortie de plans de certains atomes par rapport à une conformation donnée. Un angle impropre est défini par trois atomes (A1, A2 et A3) qui sont liés à un même atome (A4) qui sera le quatrième. Cette déformation est formulée par l'équation 4.

$$E_{\text{dièdre impropre}} = \tfrac{1}{2} \sum K_\zeta(\zeta-\zeta_o)^2 \quad \textit{(Equation 4).}$$

La valeur de $\zeta_o = 0^o$ si les quatre atomes sont dans une configuration plane sinon sa valeur est de 35.26°.

Une autre énergie associée à la rotation autour d'une liaison BC définie par quatre atomes consécutifs ABCD. Cette énergie de torsion estt formulée par l'équation 5.

$$E_{\text{angle de torsion}} = \tfrac{1}{2} \sum K\phi[1+\cos(n\phi-\delta)] \quad \textit{(Equation 5).}$$

La constante de torsion $K\phi$ est de l'ordre de 40 à 70 kJ. Mol^{-1} et δ angle de phase et ϕ angle de torsion, n : périodicité de la rotation (2 ou 3).

b) Énergie d'interaction entre atomes non liés

La première énergie tient compte des effets de répulsion et d'attraction entre deux atomes non liés. Cette énergie est répulsive à très courte distance et attractive jusqu'à l'infini. Elle est formulée par l'équation 6.

$$E_{VdW} = \tfrac{1}{2} \sum_{\text{paire}(i,j)}(D_{ij}/r_{ij}^{12} - C_{ij}/r_{ij}^{6}) \quad \textit{(Equation 6).}$$

D_{ij}/r_{ij}^{12} : terme répulsif de Lennard-Jones (Van der Waals)

C_{ij}/r_{ij}^{6}: terme attractif dispersion de London.

La seconde énergie est électrostatique. Elle est dûe aux charges q_i et q_j des atomes i et j séparés par une distance r_{ij} . Elle est formulée par l'équation 7.

$$E_{\text{électrostatique}} = \sum_{ij} (1/4\pi\epsilon) (q_i q_j / r_{ij}) \textit{ (Equation 7)}$$

Le paramètre ϵ représente la permittivité du milieu où se trouvent les atomes. Dans le vide sa valeur est égale à 1. Pour ces deux énergies le potentiel est pris en compte pour une distance comprise entre 8 et 10 Å.

La troisième énergie tient compte des liaisons hydrogènes entre deux atomes dont un est déficient en électrons.

Dans le cadre de ce projet le champ de force utilisé est celui d'amber41 dont les paramètres sont une librairie *param94.dat* qui lui ne tient pas compte de l'énergie des liaisons hydrogènes. L'équation 1 reviendrait donc à la somme des énergies formulées par l'équation 2 jusqu'à la 7. C'est donc cette formule empirique qui va tenter de simuler au mieux le comportement des molécules pour optimiser leurs géométries et minimiser leurs énergies potentielles.

3.3.2.2. Minimisation de l'énergie potentielle de la structure

En soumettant la structure au champ de force décrit au paragraphe précédant on tentera de minimiser l'énergie potentielle. Le module principal d'*AMBER* chargé de la minimisation est *Sander*. Le but de cette minimisation est de rechercher un minimum local. Cette recherche va se faire en modifiant la configuration géométrique initiale de la structure. La modification se fait de façon itérative le long d'axes particuliers du repère cartésien. Pour cette phase de minimisation deux algorithmes sont utilisés, le premier étant l'algorithme de 'recherche de ligne' (steepest descsent) et le second est celui du gradient conjugué (premier ordre). Le premier cité sera appliqué lors des premiers cycles par la suite le second sera utilisé.

La méthode de recherche de ligne (de pente) consiste à calculer l'énergie de la géométrie initiale en premier et de déplacer chaque atome selon ses trois axes par la suite et un calcul d'énergie est effectué après chaque déplacement. Ensuite tous les atomes sont déplacés sur une distance qui dépend de dE/dr_i (dr_i étant le pas) et ainsi de suite. Cet algorithme suivra la direction imposée par les forces inter atomiques dominantes. Il est très efficace pour supprimer les principaux problèmes stéréochimiques qui existent dans les coordonnées brutes de la molécule modélisée et celà tout en perturbant très peu cette dernière. Cette méthode est généralement longue vers la fin de chaque cycle de minimisation et la convergence devient très lente (phénomènes oscillatoire et remontée d'énergie). Pour cela la méthode du gradient conjugué prend le relais qui adopte le même principe mais le pas est ajusté à chaque cycle pour obtenir de meilleures diminutions d'énergie, ce qui évitera un comportement oscillatoire et accélérera la convergence.

3.4. Résultats de la modélisation

3.4.1. Scripts de minimisation d'énergie

L'étape de minimisation demande la connaissance de certains paramètres qui sont mentionnés dans un script. La figure 3-5 illustre un exemple de script qui a été utilisé pour la minimisation de l'énergie potentielle du macaque rhésus.

```
minimization, bases frozen
 &cntrl
    imin=1,
    ntmin=1,
    idiel=0, dielc=4,
    ichdna=1,
    ncyc=1000, maxcyc=50000, ntpr=100, scee=1.2,
    drms=1e-2,
    cut=15.0,
 &end

Group input for restrained atoms
1.0

RES 1 82
END
END
LISTIN=POUT
LISTOUT=POUT
```

Explication des parametres du script

imin = 1 : faire juste une minimisation

ntmin = 1 : combiner la méthode de recherche de ligne avec la celle du gradient conjugué.

ncyc = 1000 : pour les ncyc utiliser la méthode de recherche de ligne et pour maxcyc-ncyc utiliser celle du gradient conjugué

ntptr = 100 : afficher le résultat de la minimisation à chaque $100^{ième}$ cycle.

scee = 1.2 : les interaction électrostatiques sont divisées pas la valeur de scee.

drms = 1^{e}-2 : critère de convergence pour le gradient d'énergie, la minimisation s'arrête quand la racine carrée des coordonnées des éléments est inférieure à drms

cut = 15.0 : le seuil de distance pour les interactions des atomes des résidus, il faut que la distance les séparant soit inférieur à la valeur de cut pour pouvoir les considérer assez proche.

ichdna=1 : modification des charges des hydrogènes au deux bouts de la chaîne, ce qui empechera la création des liens entre le 5' et le 3'.

idiel=0 : type de fonction diélectrique pour simuler la présence d'eau.

dielc=4 : Facteur multiplicatif pour les interactions électrostatiques, *idiel* et *dielc* combinés dans mon cas vont donner une constante diélectrique de $4r_{ij}$

Figure 3-5: Un exemple de script de minimisation. A droite on retrouve le script utilisé pour minimiser la structure du macaque rhésus. A gauche les différents paramètres du script sont expliqués.

Le lancement du script de la figure 3-5 va se faire avec le module *Sander* d'*AMBER* selon la commande suivante :

Sander –i mdinX –o min[n].out –p RM_Structure-[n].pdb.out.top –c RM_Structure-[n].pdb.out.crd –x RM_Structure-[n].pdb.out.xyz

La lettre n entre [] représente le numéro de la structure à minimiser. Les fichiers nécessaires pour entamer cette étape sont les fichiers de topologies et des coordonnées de la structure initiale. Le fichier de topologie définit la connectivité et les paramètres pour chacun des modèles crées par *MC-SYM*. Ces deux fichiers ne subiront aucun changement lors de la minimisation ou de la dynamique moléculaire. Les résultats de chaque cycle de minimisation vont être sauvegarder dans le fichier min[n].out. Les nouvelles coordonnées de la structure minimisée vont être sauvegarder dans le fichier avec l'extension *.xyz* .A partir de ce dernier fichier on pourra générer le fichier correspondant selon le format *.PDB*. Pour se faire il faudra exécuter la commande suivante:

Ambpdb –p RM_Structure-[n].pdb.out.top RM_Structure-[n].pdb.out.xyz RM_Structure-0001.pdb.out.min.pdb

3.4.2. Courbes d'énergie potentielle minimisée

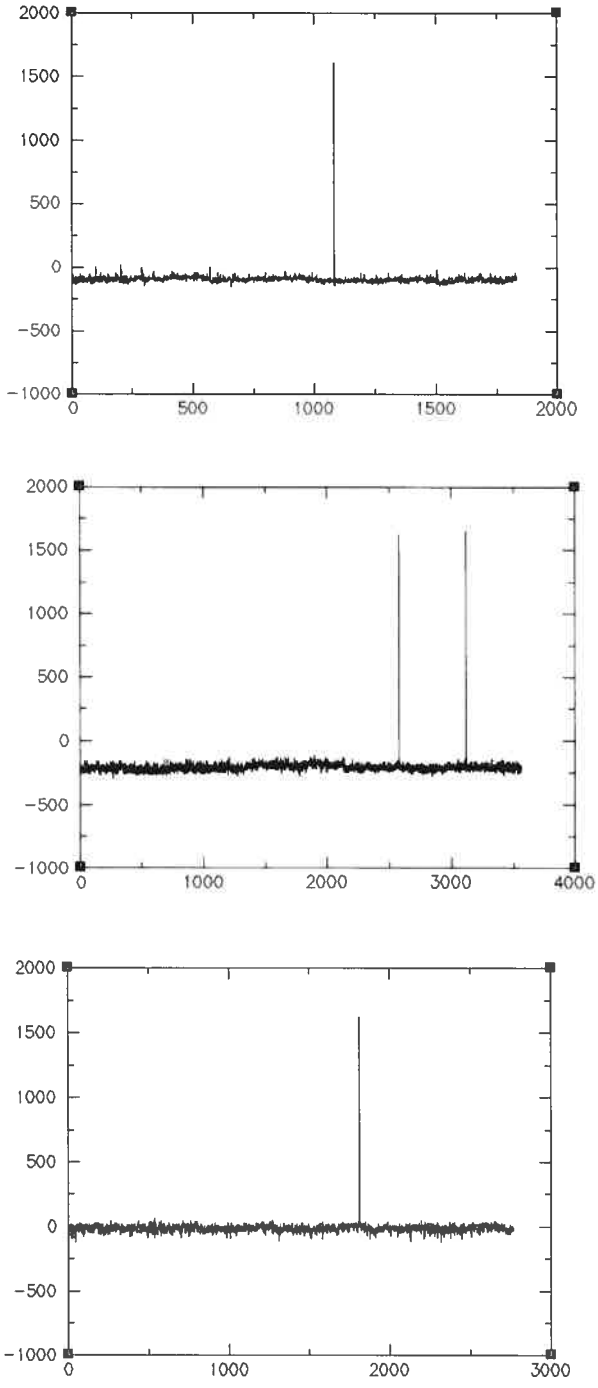


Figure 3-6: Courbes d'énergies de minimisation pour les trois espèces. En haut c'est la courbe du macaque rhésus. Au centre celle de la souris et en bas celle de l'humain. En abscisse on retrouve le numéro de structure et en ordonnée l'énergie en kcal/mol.

En général, les courbes de la figure 3-6 oscillent autour d'une même valeur d'énergie négative. Cette observation est cohérente vu que le but de la minimisation est la recherche du minimum local le plus proche et la suppression de tout mauvais contact introduit par l'hydrogénisation. Néanmoins on observe quelques exceptions pour quelques unes des structures, selon les espèces. Ces observations se manifestent sur la figure 3-1 par des pics positifs qui sont résumés par le tableau 3-2.

Espèce	Numéro de Structure	Énergie (kcal/mol)
Macaque Rhésus	1082	1.6082^{E}+03
Souris	2580	1.6203^{E}+03
	3121	1.6484^{E}+03
Humain	1810	1.6254^{E}+03

Tableau 3-2: Valeurs d'énergie présentant des pics positifs pour chacune des trois espèces.

La présence de ces pics est causée par l'énergie des atomes liés. Ces pics reflètent la déformation du squelette de la molécule. Dans le cas de ces structures, ils sont supérieurs à 100 kcal/mol alors que normalement ils devraient être inférieurs comme dans le cas du reste des structures. Ces pics signifient que les distances entre les atomes sont trop longues. Les valeurs d'énergie causant ces pics et correspondant à chacune des structures du tableau 3-2 sont résumées dans les tableaux qui suivent tableau 3-3.

Structure	1082 du	Macaque	Rhésus		
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
20831	**1.6082E+03**	9.8626E-03	8.6608E-02	C3'	2132
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
1168.8561	454.7650	1479.2589	-1312.6906	-73.9524	0.0000
1-4VDW	1-4 EEL				
596.44421	-704.4385				
Structure	1810 de	l'humain			
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
16233	**1.6254E+03**	9.7016E-03	2.0118E-01	P	3207
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
1269.0603	583.7654	2091.6110	-2164.7648	-37.5855	0.0000
1-4VDW	1-4 EEL				
848.0624	-964.7240				
Structure	3121 de la	souris			
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
21406	**1.6484E+03**	9.9300E-03	1.7296E-01	H3	734
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
1295.7439	656.7406	2312.8409	-2216.7992	-13.3915	0.0000
1-4VDW	1-4 EEL				
926.2920	-1312.9837				
Structure	2580 de la	souris			
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
21432	**1.6203E+03**	9.5053E-03	1.5410E-01	N9	3301
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
1291.8043	644.3712	2315.5234	-2217.2992	-30.8361	0.0000
1-4VDW	1-4 EEL				
927.6821	-1310.9518				

Tableau 3-3: Résultats des structures avec des énergies positives après minimisation. ENERGY est > 0 car BOND est > 100. Pour chacune des structures 1082 du macaque rhésus, la 1810 de l'humain, la 3121 et la 2580 de la souris, on a le cycle de minimisation (NSTEP), l'énergie totale (ENERGY), la RMS (Root Mean Square), La pente de la surface du potentiel calculé (GMAX), le nom du résidu (NAME) et son numéro (NUMBER), l'énergie due aux atomes liés (BOND) supérieure à 100, l'énergie de flexion (ANGLE), l'énergie de torsion (DIHED). Pour les atomes non liés séparés par plus de trois liaisons chimiques on a l'énergie d'interaction de Van Der Waals (VDWAALS) et l'énergie électrostatique (EEL). Pour les atomes séparés exactement par trois liens on a les valeurs dans 1-4 VDW et 1-4 EEL. Ces dernières valeurs sont multiplié par un facteur de 2 dans le premier cas et divisées par 1.2 dans le second.

Dans le tableau 3-4, on retrouve des valeurs d'énergies normales avec une énergie des liaisons inférieures à 100 kcal/mol

Structure	1081 du	Macaque	Rhésus		
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
17331	**-9.9306E+01**	9.7970E-03	9.9047E-02	H4'	2211
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
49.5221	402.5585	1483.6431	-1804.7554	-126.3042	0.0000
1-4VDW	1-4 EEL				
601.0893	-705.0591				
Structure	1811 de	l'humain			
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
11507	**-2.7407E+01**	9.9054E-03	1.1212E-01	P	2235
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
69.9576	574.7948	2094.4432	-2488.5851	-105.3918	0.0000
1-4VDW	1-4 EEL				
847.1967	-965.0086				
Structure	2581 de la	souris			
NSTEP	ENERGY	RMS	GMAX	NAME	NUMBER
21106	**-2.1321E+02**	9.9553E-03	1.3335E-01	HO'2	2543
BOND	ANGLE	DIHED	VDWAALS	EEL	HBOND
72.8791	577.5524	2290.1162	-2764.0711	-8.4589	0.0000
1-4VDW	1-4 EEL				
929.5544	-1310.7814				

Tableau 3-4: Résultats des structures avec des énergies négatives après minimisation. Pour chacune des structures 1081 du macaque rhésus, la 1811 de l'humain et la 2581 de la souris, on a le cycle de minimisation (NSTEP), l'énergie totale (ENERGY), la RMS (Root Mean Square), La pente de la surface du potentiel calculé (GMAX), le nom du résidu (NAME) et son numéro (NUMBER), l'énergie due aux atomes liés (BOND) inférieure à 100, l'énergie de flexion (ANGLE), l'énergie de torsion (DIHED). Pour les atomes non liés séparés par plus de trois liaisons chimiques on a l'énergie d'interaction de Van Der Waals (VDWAALS) et l'énergie électrostatique (EEL). Pour les atomes séparés exactement par trois liens on a les valeurs dans 1-4 VDW et 1-4 EEL. Ces dernières valeurs sont multiplié par un facteur de 2 dans le premier cas et divisées par 1.2 dans le second.

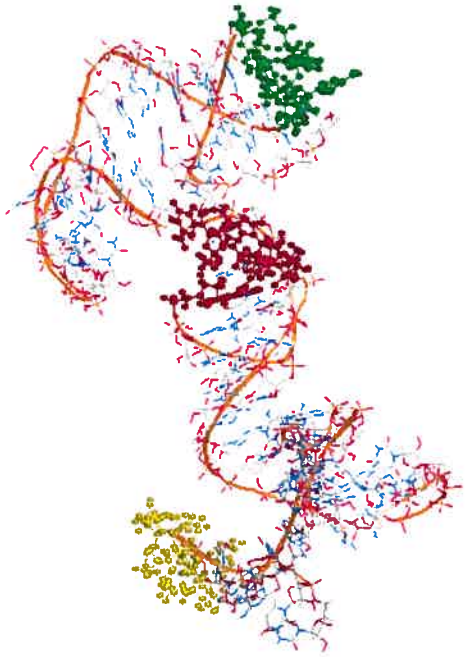


Figure 3-7: Vue de la structure du Macaque Rhésus numéro 1081. Les nucléotides en vert représentent la boucle GNRA du nucléotide 9 au 12, en rouge la GNRA du nucléotide 42 au 45 et en jaune la GNRA du nucléotide 75 au 78 au niveau de la demi répétition. En orange c'est la chaîne du phosphate de la molécule. En bleu les autres résidus avec leur atomes.

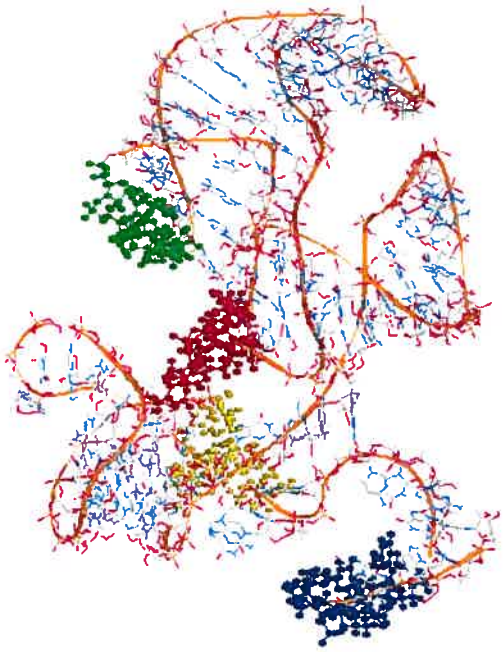


Figure 3-8: Vue de la structure de l'humain numéro 1811. Les nucléotides en vert représentent la boucle GNRA du nucléotide 9 au 12, en rouge la GNRA du nucléotide 42 au 45, en jaune la GNRA du nucléotide 75 au 78 et en bleu la GNRA du nucléotide 108 au 111 au niveau de la demi répétition. En orange c'est la chaîne du phosphate de la molécule. En bleu les autres résidus avec leur atomes.

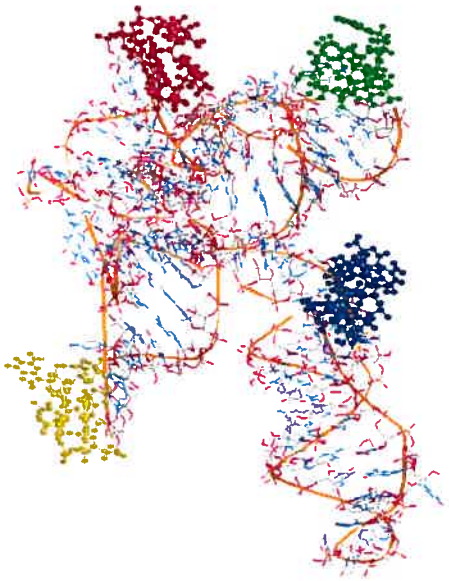


Figure 3-9: Vue de la structure de la souris numéro 2581. Les nucléotides en vert représentent la boucle GNRA du nucléotide 8 au 11, en rouge la GNRA du nucléotide 44 au 47, en jaune la GGAC du nucléotide 81 au 84 et en bleu la GNRA du nucléotide 116 au 119. En orange c'est la chaîne du phosphate de la molécule. En bleu les autres résidus avec leur atomes.

Des figures 3-7, 3-8 et 3-9, on voit bien l'orientation des tétra boucles vers l'extérieur de la molécule ce qui les exposent vers le solvant et les prédisposent à établir des connexions avec d'autres molécules d'ARN ou bien des protéines.

Les distances au niveau des boucles des structures ont été comparées aux distances entre l'atome O3' (oxygène 3 du sucre) et le P (phosphate) de deux résidus successifs. Ces valeurs sont extraites de la GAAA du fichier PDB 1HR2, du domaine P4-P6 de l'intron du groupe I. Elles sont résumées dans le tableau 3-5.

Distances	De Résidu-Position.Atome	A Résidu Position.Atome	Valeurs (Å)
	G-149.O3'	A-150.P	1.61Å
	A-150.O3'	A-151.P	1.62 Å
	A-151.O3'	A.152.P	1.61 Å

Tableau 3-5: Tableau des distances dans une tétra boucle GNRA entre les atomes d'oxygènes et phosphates successifs. Ces distances sont extraites du fichier PDB 1HR2 du domaine P4-P6 de l'intron du groupe I.

Le tableau 3-5 va servir de référence pour comparer les distances d'une GNRA expérimentale avec celles des structures minimisées. Les tableaux 3-6, 3-7 et 3-8 ont été dressés pour l'espèce du macaque rhésus.

Boucle 1 GAAA :

Distances	De Résidu-Position.Atome	A Résidu Position.Atome	Valeurs (Å)
	G-9.O3'	A10.P	1.618
	A10.O3'	A11.P	1.615
	A11.O3'	A12.P	1.628

Tableau 3- 6: Résultats des distances de la première boucle GAAA de la structure minimisée du macaque rhésus entre les atomes d'oxygènes O3' sur le ribose qui sont directement liés au phosphate P.

Distances	De Résidu-Position.Atome	A Résidu Position.Atome	Valeurs (Å)
	G-42.O3'	A-43.P	1.620
	A-43.O3'	A-44.P	1.618
	A-44.O3'	A-45.P	1.639

Tableau 3-7: Résultats des distances de la deuxième boucle GAAA de la structure minimisée du macaque rhésus entre les atomes d'oxygènes O3' sur le ribose qui sont directement liés au phosphate P.

Distances	De Résidu-Position.Atome	A Résidu Position.Atome	Valeurs (Å)
	G75.O3'	A76.P	1.613
	A76.O3'	A77.P	1.636
	A77.O3'	A78.P	1.632

Tableau 3-8: Résultats des distances de la troisième boucle GGAA de la structure minimisée du macaque rhésus entre les atomes d'oxygènes O3' sur le ribose qui sont directement liés au phosphate P.

Les résultats des tableaux 3-6, 3-7 et 3-8 montrent une différence de distance de 1% avec celles provenant du cristal ce qui rapproche cette partie du modèle théorique du réel, après minimisation d'énergie. Pour les deux autres espèces, on retrouve les valeurs dans le même ordre de distances.

La boucle GNRA a la réputation de se lier à des doubles brins contenant les deux paires AU et CG et qui se succèdent. Ce double brin appartiendrait à une autre molécule [26]. Pour cette raison, un script a été dressé pour tenter de construire ce genre

d'interaction mais dans une même molécule. Dans notre cas le test a été fait sur la molécule du macaque rhésus en essayant de créer une interaction entre la première boucle GAAA du 9^{iéme} au 12^{iéme} nucléotide avec le demi brin de la seconde répétition. Cette construction s'est avérée impossible vu qu'un nucléotide de la chaîne n'a pas pu être placé.

3.5. Dynamique moléculaire

La dernière étape à faire est celle de la dynamique moléculaire, où l'évolution en fonction du temps d'une molécule est décrite par les principes de la mécanique classique Newtonienne (F=ma/F; F:force de Newton; a: accélération; m: masse de l'atome).

Le but est de simuler les mouvements intramoléculaires au cours du temps correspondants à des vibrations autour d'un minimum qui sera celui de la structure minimisée et non l'initiale. En dynamique moléculaire le concept de température est fondamental puisque la vitesse est proportionnelle à la racine carrée de la température, à l'équilibre l'énergie potentielle d'un système qui est égale à son énergie potentielle ($3/2KbT = ½ \sum m_i v_i^2$). La méthode utilisée par *AMBER* pour la dynamique moléculaire est celle du recuit simulé ('anneal') qui cherchera à aller au-delà du minimum local trouvé lors de la minimisation. Un exemple de script utilisé pour la dynamique moléculaire est illustré par la figure 3-10.

```
molecular dynamic
 &cntrl
    imin=0, ntb=0,
    ntpr = 100, ntwx = 100,
    ntt=1,scee=1.2,tempi=300.0,temp0=300.0
    nstlim=100000, dt=0.001,
    cut=15.0,
 &end
```

Expliction des parametres du script de dynamique moleculaire
imin = 0 pas de minimisation mais juste une dynamique moleculaire.
ntb = 0 la periodicite n'est pas appliquee
ntpr = 100 a chaque 100 cycles les resultats sont inscrits dans un fichier
ntwx = 100 a chaque 100 cycles les coordonnees sont inscrits dans le fichier des coordonnees "mdcrd"
ntt = 1 tous les atomes du systeme sont maintenus au meme thermostat.
scee = 1.2 facteur par lequel les interactions 1-4 electrostatique sont divisees
tempi = 300.0 temperature initiale en Kelvin
temp0 = 300.o temperature a maintenir.
nstlim=100000,dt=0,001 avec 100000 pas de 0.001ps, donc une simulation durera 100 ps.
cut = 15.0 limite de distance pour les interactions entre atomes non lies.

Figure 3-10 : Exemple de script appliqué sur les structures minimisées du macaque rhésus. A droite on a le script et à gauche l'explication de ses paramètres.

L'exécution du script de la figure 3-10 se fait avec le module *Sander* selon la commande qui suit :

Sander –i dynIn.md –o md[n].out –p RM_Structure-[n].pdb.out.top –c RM_Structure-[n].pdb.out.xyz –x RM_Structure-[n].pdb.out.XYZ –r RM_Structure-[n].pdb.out.mdcrd

Le fichier de dynamique moléculaire de la figure 3-10 correspond à *dynIn.md*. Les résultats seront sauvegardés dans le fichier *md[n].out* pour la n^{ième} structure et pour chacune des étapes du cycle. Le fichier de topologie de la structure initiale et celui des coordonnées résultants de la phase de minimisation sont nécessaires pour cette étape. Le résultat étant la création d'un nouveau fichier de coordonnées avec l'extension *.XYZ* et la trajectoire de la molécule en fonction du temps qui sera sauvegardée dans le fichier ayant l'extension *.mdcrd*.

A partir du fichier *md[n].out* on peut extraire les valeurs d'énergies calculées à chaque étape de la simulation qui dure 100 ps. Un exemple de simulation est illustré par la figure 3-11 pour la structure 1759 du macaque rhésus.

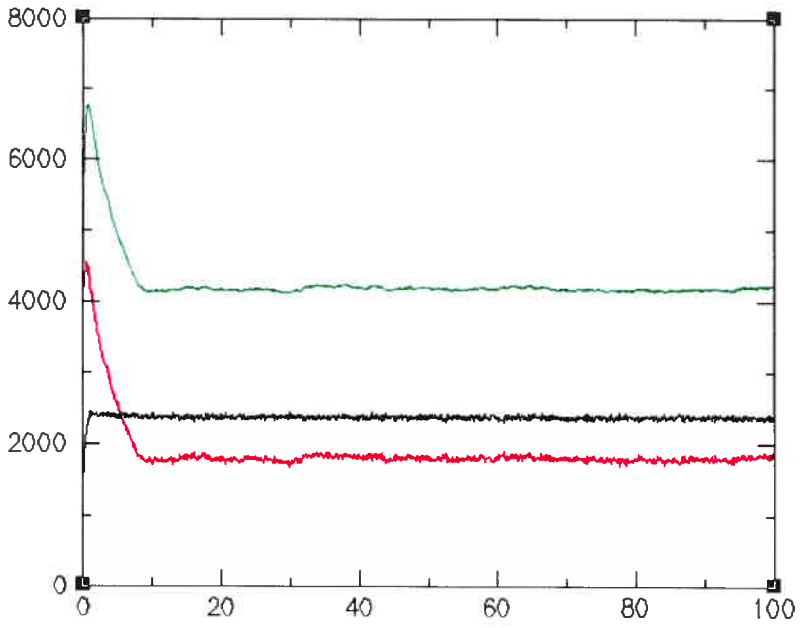


Figure 3-12 : Graphe des énergies cinétique et potentielle de la Structure du macaque rhésus numéro 1759 après la dynamique moléculaire sans Ions de sodium (Na+). La courbe en noir représente l'énergie cinétique, celle en rouge la potentielle et la verte représente la somme des deux. En abscisse on a le temps de simulation en picosecondes (ps) et en ordonnée les énergies en calories par mole (cal/mol).

La visualisation de la simulation avec le logiciel *vmd (Visual Molecular Dynamics)* [42], en entrant les deux fichiers de topologie ainsi que le fichier décrivant la trajectoire *(.mdcrd)*, a montré une molécule d'ARN qui finit par se défaire complètement, ce qui est normal vu que les charges négatives n'ont pas été neutralisées par l'ajout d'ions porteurs de charges positives et la molécule n'a pas été mise dans un solvant tel que l'eau, de ce fait on est loin des conditions d'expérience. Ce qui explique sur le graphe, des énergies très élevées, même après une décroissance et une stabilisation, elles le restent toujours. Une autre expérience a été tentée pour contre carrer l'effet des ions négatifs de la molécule. La façon de réaliser ceci est de rajout des ions de sodium Na+. L'ajout de ces ions à la molécule se fait par une commande qui va permettre de les placer aux endroits où sont placés les charges négatives. Une fois les ions rajoutés, il faudra avoir un autre fichier de topologie qui reflétera la nouvelle information structurale de la

molécule ainsi que le fichier des coordonnées. L'effet de l'ajout des ions de sodium Na+ sur l'énergie potentielle de la molécule est illustré par le graphe de la figure 3-13.

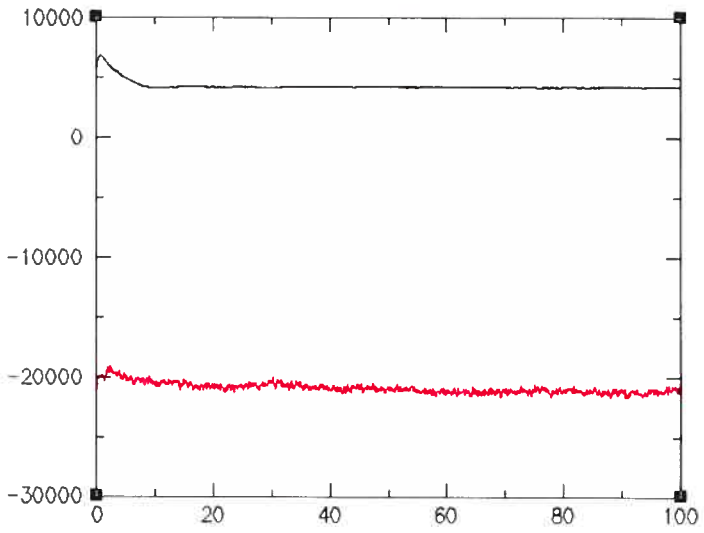


Figure 3-13 : Graphe comparative de l'énergie totale en fonction du temps entre la structure de rhésus macaque numéro 1759 suite à une dynamique moléculaire 'in-vacuo' en noir et la courbe qui reflète l'effet des ions Na+ est en rouge. En abscisse on a le temps en picosecondes (ps) et en ordonnée l'énergie calculée en calories par mole (cal/mol).

L'ajout des ions pour neutraliser à améliorer les valeurs de l'énergie potentielle, ceci s'observe sur la figure 3-13. Effectivement le rajout des ions Na+ a rabaissé l'énergie totale de 5kcal/mol à -20kcal/mol. La dernière expérience a été de mettre la molécule avec des ions dans de l'eau, ceci a été impossible pour la molécule du macaque ainsi que pour les deux espèces vu la longueur des chaînes de nucléotides qui est trop grande et rend les molécules très volumineuses.

Chapitre 4

Conclusion

A travers ce travail de recherche, l'obtention des résultats a été possible grâce aux outils et matériels informatiques mis à la disposition du chercheur. Ces deux moyens de recherche sont constamment en évolution et en développement en vue de rendre les résultats de plus en plus précis afin de se rapprocher au maximum des évènements biologiques se produisant dans la réalité. Néanmoins, certaines prédictions ont été impossibles vu la non adaptation des outils au problème de recherche de structure dans des séquences contenant des régions répétées.

Pour pallier à ces imperfections, certaines améliorations algorithmiques pourraient être apportées à celles déjà existantes. La première amélioration algorithmique concerne la formule de récursivité utilisée par les programmes de prédiction qui se basent sur la programmation dynamique. Pour pouvoir tenir donc compte des séquences avec un polymorphisme, il faudrait pouvoir changer les bornes de recherche de structure pour accentuer la prédiction dans les régions répétées. Pour se faire, il faudrait changer les bornes i et j de la récurrence par les indices i et $i+TM$ avec TM la taille de la séquence répétée. Dans le cas d'une répétition imparfaite, il faudrait tenir compte du fait qu'à chaque i+3, qui est la troisième position de chaque triplet de nucléotides codant pour un acide aminé, pourrait prendre une des quatre valeurs de l'alphabet des nucléotides {A,C,G,U}. Pour tenir compte des répétitions dans une séquence, il serait aussi possible d'intégrer un algorithme dédié à cette tâche tel que celui des arbres de suffixes, par la suite la recherche de structure proprement dite pourrait être entamée en tenant compte des résultats de la phase précédente.

En utilisant le programme *pknots* une structure de pseudo nœud a été effectivement retrouvée sur une séquence de 36 nucléotides uniquement avec une boucle

inférieure très courte comparativement à l'hélice qui la suit ce qui a suscité un questionnement concernant la construction de la structure retrouvée en 3D. Ce genre de structure est considérée par *pknots* étant donné qu'il n'y a aucune vérification entre la longueur de la boucle et celle de l'hélice qui la précède. Du point de vue matricielle cette vérification se manifesterait par le calcul du nombre de colonnes séparent les deux hélices de la structure du pseudo nœud. Le rapport entre la longueur de l'hélice à croiser et celle de la boucle inférieure qui la croise devrait être défini de façon à permettre ce croisement et de là la construction de la structure en 3D.

Lors de la prédiction et recherche de structure, le calcul énergétique a été impossible vu l'absence de données expérimentales. Il serait utile de pouvoir afficher un message explicatif expliquant l'impossibilité du calcul énergétique.

Lors de la modélisation, la dynamique moléculaire de toute la molécule dans un milieu aqueux s'est avérée impossible vu la longueur de la structure ce qui a causé un dépassement de mémoire. Il serait donc préférable de refaire la dynamique moléculaire avec des parties de la structure et plus précisément celles des hélices fermées par les tétra boucles GNRA.

En plus des hypothèses proposées dans ce travail de recherche, une autre serait encore possible vu l'observation de G successifs dans les séquences répétées. Cette observation est illustrée par le tableau 4-1.

Séquence	Espèce	Nombre de fois
UGGGGA	Humain	2 (Dans répétitions 1 et 2)
	Macaque Rhésus	1 (Dans répétition 1)
UGGGGC	Souris	2 (Dans répétitions 1 et 3)
UGGGGA	NM_000311 (PRION)	1
UGGGGGC	NM_000311 (PRION)	3
UGGGGAC	NM_000311 (PRION)	1
UGGGGGA	Humain	2 (Dans répétition 3 et la ½ répétition)
UGGGGGGGU	Macaque Rhésus	2 (Dans répétition 2 et la ½ répétition)
UGGGGGGC	Souris	1 (Dans répétition 4)
UGGGGGGGU	Souris	1 (Dans répétitions 2)

Tableau 4- 1: Tableau résumant les séquences se localisant dans les régions répétées. La première colonne contienne la séquence avec les G. La seconde indique l'espèce dans laquelle la séquence est retrouvée. La troisième colonne indique le nombre de fois que la séquence est observée et dans quelle répétition dans le cas des séquences de ce projet.

Le tableau 4-1 montre les séquences observées dans les régions répétées, le nombre de G successif pourrait indiquer qu'au niveau de la structure tridimensionnelle, il pourrait former un motif connu comme étant le G-quartet [43], où les G créeraient des liens de sorte à former un canal pour laisser passer certains ions de métaux et pas d'autres.

Du point de vue biologique, la vision et l'odorat participent indirectement au sens du goût. La fonction principale du goût est certes transmise par la protéine STG, mais il

serait aussi intéressant de se pencher sur celles qui seraient responsables de la vision et de l'odorat afin d'étudier leurs structures respectives et de les regrouper avec celle de la STG. Il serait possible que les trois structures soient dépendantes l'une de l'autre.

Ce travail a donc montré les différentes étapes de recherche de structure pour des séquences avec un polymorphisme. De ce projet on voit l'implication de l'informatique dans la résolution des problèmes biologiques. Des améliorations restent comme même à apporter au niveau des programmes pour pouvoir résoudre des problèmes biologiques particuliers. Il serait donc intéressant qu'un laboratoire de recherche se penche sur l'étude de ces séquences en se basant sur les hypothèses de ce travail ce qui permettrait de les confirmer ou bien au contraire de les rejeter pour en proposer d'autres avec le but de rechercher des médicaments qui soulageraient les malades vu qu'une des trois séquences du projet se retrouvait sur le chromosome 6 et plus précisément sur le bras soupçonné d'être impliqué dans la maladie du psoriasis vulgaris. Cette maladie de la peau que les dermatologues n'en connaissent pas l'origine.

Bibliographie

[1] Mauricio Neira, Viktoria Danilova, Göran Hellekant, Edwin A. Azen : A new gene (rmSTG) specific for taste buds is found by laser capture microdissection, *Mammalian Genome 12, 60-66 (2001).*

[2] Bonner RF, Emmert-Buck M, Cole K, Pohida T, Chuaqui R et al Laser capture microdissection : molecular analysis of tissue. *Sciences 278, 1481-1483 (1997).*

[3] Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z et al: Laser capture microdissection. *Science 274, 998-1001 (1996).*

[4] Akira Oka, Gen Tamiya, Maiko Tomizawa, Masao et al: Asssociation analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111 kb segment telomeric to the HLA-C gene. *Human Molecular Genetics, Vol. 8, No. 12 2165-2170 (1999).*

[5] Jan Klein, Ph D. : The HLA system (First and Second Part). *The New England Journal of Medicine No. 10 Vol. 343:702-709, No. 11 Vol.343:782-786 (2000).*

[6] A.J. Mungall, J. Rogers & S. Beck et al: The DNA sequence and analysis of human chromosome 6. *Nature 425, 805-811 (2003).*

[7] Ignacio Tinoco Jr, Carlos Bustamente. How RNA folds. *J. Mol. Biol 293, 271-281 (1999).*

[8] Paul G. Higgs. RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics 33, pp. 199-253 (2000).*

[9] A. M. Zucker, B. D.H. Mathews & C. D.H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction. *In RNA biochemistry and biotechnology , NATO ASI Series, Kluwer Academic Publishers (1999).*

[10] E. Rivas, S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol 285:2053-2068 (1999).*

[11] Notredame C, O'Brien EA, Higgins DG. RAGA: RNA sequence alignement by genetic algorithm. *Nucleic Acid Research Vol. 25, No. 22, 4570-4580, (1997).*

[12] Gautheret, D., Major, F. and Cedergen, R.: Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comp. Appl. Biosci., 6, 325-331 (1990).*

[13] Thomas J. Macke, David j. Ecker, Robin R. Gutell, Daniel Gautheret, David A. Case and Rangarajan Sampath: RNAMotif, an RNA secondary structure definition and search algorithm. *NAR, Vol. 29, No.22, 4724-4735 (2001).*

[14] Freier, S., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. & Turner, D. H. improved free-energy parameters for predictings of RNA duplex stability. *Proc. Natl. Acad. Sci. USA, 83, 9373-7*

[15] John SantaLucia, Jr. Douglas H. Turner: Measuring the thermodynamics of RNA secondary structure formation. *1998 John Wiley & Sons, Inc. Biopoly 44: 309-319, (1997).*

[16] D.H. Turner , N.Sugimoto, J.A. Jaeger, C.E. Longfellow, S.M. Freier, R. Kierzek. Improved parameters for prediction of RNA structure. *Cold Spring Harb. Symp. Quant. Biol, 52, 123-133 (1987).*

[17] A.E. Walter, D.H. Turner, J. Kim, M.H. Lyttle, P. Muller, D.H. Mathews, M. Zucker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad Sci USA, 95, 1460-1465 (1998).*

[18] N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, T. Ohmichi, M. Yoneyama, and M. Sasaki. Thermodynamic parameters to predict stabilikty of RNA/DNA hybrid duplexes. *Biochemistry, 34, 11211-11216 (1995).*

[19] D. H. Mathews, J. Sabina, M. Zucker, D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol. 288, 911-940 (1999).*

[20] Isabelle Barrette, Guylaine Poisson, Patrick Gendron, Francois Major. Pseudoknots in prion protein mRNA confirmed by comparative sequence analysis and pattern searching. *Nucleic Acid Research Vol. 29, No. 3 753-758 (2001).*

[21] Peter De Rijk and Rupert De Wachter. RnaViz, a program for the visualization of RNA secondary structure. *NAR. 25(22): 4679-4684 (1997).*

[22] Peter De Rijk, Jan Wuyts and Rupert De Wachter. RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics 19(2): 299-300 (2003).*

[23] Thomas Greiner-Stoffele, Hans-Heinrichtal et al. RNASE Stable RNA conformational parameters of the nucleic acid backbone for binding to RNASE T1. *Biol. Chem. Vol. 382, 1007-1017 (2001).*

[24] Sébastien Lemieux, Francois Major. Recognition of base pairing types in RNA three dimensional structures. ??????

[25] Wills, P. R. Potentiel pseudoknots in the PrP-encoding mRNA, *J. Theor. Biol., 159, 523-527 (1992*).

[26] Eliah Aronoff-Spencer, Colin S. Burns et al. Identification of the Cu^{2+} Binding sites in N-Terminal domain of the prion protein by EPR and CD spectroscopy. *Biochemistry, 39, 13760-13771,(2000).*

[27] Luck R, Steger G, Riesner D. Secondary structure of prion mRNA. *J. Mol. Biol, 258: 813-26 (1996).*

[28] Jan Pieter Abrahams, Miriajm van den Berg, Eke van Batenburg, Cornelis Pleij. Prediction of RNA secondary structure, including pseudoknotting by computer simulation. *Nucleic Acid Research, Vol. 18, No, 10, 3035 (1990).*

[29] Alschtul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology 215: 403-410 (1990*).

[30] Aron Marchler-Bauer, Anna R. Panchenko, Benjamin A. Shoemaker, Paul A. Thiessen, Lewis Y. Geer and Stephen H. Bryant. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *NAR, Vol. 30, No. 1 281-283 (2002).*

[31] Robert T. Batey, Robert P. Rambo, and Jennifer A. doudna. Tertiary Motifs in RNA Structure and Folding. *Reviews. Angew. Chem..Int. Ed. 38,2326-2343 (1999).*

[32] Dana L. Abramovitz and Anna Marie Pyle. Remarkable morphological variability of a common RNA folding motif: The GNRA tetraloop-receptor interaction. *J.Mol.Biol 266, 493-506 (1997).*

[33] Jamie H Cate and Jennifer Doudna. Metal-Binding sites in the major groove of a large ribozyme domain.*Research Article, Structure 4: 1221-1229 (1996).*

[34] Thomas Hermann and Dinshaw J Patel. RNA bulges as architectural and recognition motifs. *Minireview Structure 8: R47-R54, (2000).*

[35] Gabriele Varani and William H. McClain. The G-U wobble base pair. *EMBO Reports Vol.1, No.1, 18-23 (2000).*

[36] Guang-chou Tu, Qing-na Cao, Feng Zhou and Yedy Israel. Tetranucleotide GGGA motif in primary RNA transcripts Novel target site for antisense design. *J.Biol. Chem. Vol.273, Issue 39, 25125-25131 (1998).*

[37] Franch, T., Peterson, M., et al. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-Turn loop structure. *J. Mol. Biol. 294, 1115-1125 (1999).*

[38] Ye Ding, Rational Statistical design of antisense oligonucleotides for high throughput functional genomics and drug target validation. *Statistica Sinica 12, 273-296 (2002).*

[39] F. Major, D.Gautheret, and R. Cedergen. Reproducing the three-dimensionnal structure of a tRNA molecule from structural constraints. *Proc. Natl. Acad. Sci USA, 90(20):9408-9412, (1993).*

[40] MC-SYM User Manual *Version 3.3.2* (Laboratoire de Biologie Informatique et Théorique. Université de Montréal) *(2002).*

[41] D. A. Pearlman, D. A. Case, P. Kollman and al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to stimulate the structural and energetic properties of molecules. *Comp. Phys. Commun. 91,1-41 (1995).*

[42] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33-38, 1996

[43] Stefan Weiss, Daniela Proske et al. RNA Aptamers specifically interact with the prion protein PrP. *Journal of Virology, pp. 8790-879, (1997).*