Université de Montréal

Détection et analyse de motifs structuraux et fonctionnels dans les acides ribonucléiques

par

Patrick Gendron

Département d'informatique et recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maîtrise ès sciences (M.Sc.)

en informatique

Août, 2000

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

Détection et analyse de motifs structuraux et fonctionnels dans les acides ribonucléiques

présenté par:

Patrick Gendron

a été évalué par un jury composé des personnes suivantes:

Nadia El-Mabrouk,	président-rapporteur
François Major,	directeur de recherche
Bernard Gendron,	membre du jury

Mémoire accepté le 17 octobre 2000

Sommaire

L'étude comparative de la structure des acides ribonucléiques (ARN) est un des principaux vecteurs de découverte du rôle joué par ces ARN dans les mécanismes biologiques. En outre, la compréhension du lien entre la structure adoptée par un ARN et la fonction qui y est associée rend possible la conception d'agents thérapeutiques à hauts degrés de spécificité. Le développement de méthodes efficaces et objectives d'analyse de la structure des ARN est donc essentiel. Le présent travail propose trois ajouts importants à ce domaine.

Dans un premier temps, une description géométrique précise de la conformation adoptée par les nucléotides formant l'ARN et des interactions moléculaires agissant entre eux est présentée. Une catégorisation symbolique de ces informations structurales est aussi décrite, de même qu'une validation statistique par rapport à une base de données de conformations et d'interactions définie préalablement.

Ensuite, nous décrivons un cadre d'analyse structurale des régions conservées dans la structure des ARN, aussi appelées motifs. Celui-ci emploie une méthode basée sur l'isomorphisme de graphes et la recherche du graphe sous-jacent commun maximal, afin d'identifier, dans la structure d'ARN équivalents mais d'organismes différents, des régions possédant une composition structurale similaire. Ces régions sont responsables à la fois du repliement spécifique des ARN et du maintien de la fonction biologique de ces derniers. Les méthodes développées ici, en ce qui concerne la recherche de motifs conservées dans la structure des ARN, constituent une approche avant-gardiste à l'étude du lien entre la structure et la fonction des ARN.

Finalement, nous proposons une méthodologie permettant l'étude détaillée de nouveaux motifs structuraux, allant de la séquence à la structure tertiaire de l'ARN. Cette dernière cible des régions à haut potentiel structural dans une séquence par évaluation de

l'énergie libre de repliement. L'utilisation d'algorithmes de prédiction de structures secondaires et de modélisation tridimensionnelles permet de caractériser la structure véritable de ces régions et de mieux en comprendre le rôle.

La pertinence de ces nouvelles approches à l'analyse structurale des ARN est confirmée par de nombreux résultats positifs, allant de l'identification de nouveaux motifs dans les ARN ribosomaux des *Archaea*, *(eu)Bacteria* et *Eukarya*, à la découverte d'une région structurée de l'ARN messager responsable de certains cas de la maladie de Creutzfeldt-Jakob, en passant par la validation de la singularité du site d'interaction de l'ARN ribosomal avec la protéine L11.

Mots clés: ARN, recherche de motifs, isomorphisme de graphe, analyse structurale, détection de régions structurées.

Table des matières

Liste des Tables

Liste des Figures

À ma famille,

Chapitre 1

Introduction

L'avènement de nouvelles technologies de séquençage de l'ADN, initié par la création, en 1990, du projet de séquençage du génome humain (ainsi que par plusieurs projets similaires pour différentes espèces) amena la génération d'une quantité phénoménale d'information sur le code génétique. En date de juin 2000, le nombre de bases séquencées atteint près de 9 milliards, réparties dans 7 077 000 de séquences, et suit une courbe de croissance exponentielle [9]. Parallèlement à ces travaux, de nombreuses méthodes de détermination de structure ont permis d'obtenir un grand nombre de structures candidates de molécules diverses. Parmi celles-ci, on retrouvent des méthodes physiques, telles que la cristallographie par diffraction aux rayons X ou la spectroscopie par résonance magnétique nucléaire (RMN), et théoriques, telles que la prédiction de structures secondaires par analyse comparative de séquences [66, 109, 154] ou la modélisation tridimensionnelle (3D) [104]. À titre indicatif, les structures tertiaires de 11416 protéines, de 857 acides nucléiques et de 558 complexes entre des protéines et des acides nucléiques et les structures secondaires de plus de 300 acides nucléiques sont présentes dans diverses banques de données [12, 13, 68].

Une grande proportion des travaux effectués en bioinformatique et en chimie computationnelle vise l'analyse et le traitement de cette information de séquence et de structure. L'intérêt réside dans la compréhension des mécanismes moléculaires responsables de l'activité cellulaire des organismes vivants. En particulier, il est maintenant généralement accepté que la présence de régions conservées et/ou structurées de manière spécifique dans ces molécules est directement reliée à la conservation d'une fonction biologique précise.

En effet, l'activité des cellules des organismes ayant été sélectionnés au cours de l'évolution est gérée par l'entremise de quantité de molécules diverses. La conservation de

chacune des fonctions biologiques de ces molécules nécessite le maintien de structures précises, et en particulier de régions structurales relativement invariantes.

L'étude de ces régions constitue donc une étape cruciale du développement d'agents thérapeutiques pouvant y interagir de façon spécifique. Déjà, plusieurs laboratoires pharmaceutiques misent sur la conception de molécules conçues en fonction de cibles préalablement identifiées dans la mise au point de nouveaux médicaments antimicrobiens.

Plusieurs méthodes expérimentales sont adoptées afin de localiser de nouvelles régions structurales et fonctionnelles. Parmi celles-ci, on retrouve la mutagenèse dirigée, lors de laquelle on altère la séquence d'une molécule en y incorporant tour à tour des résidus modifiés à chacune des positions. Il est donc possible de vérifier l'influence de chacun des résidus ou de groupes de résidus pour le maintien de la fonction. Ce processus s'avère toutefois très coûteux et laborieux vue la quantité de sites potentiels.

Diverses méthodes de prédictions ont donc été développées afin de restreindre la quantité de régions fonctionnelles potentielles et d'accélérer le processus de création de nouveaux médicaments. Ces méthodes aident par le fait même à l'étude et à la compréhension générale des mécanismes biologiques ayant lieu dans les cellules vivantes.

1.1 Fonction des ARN

La vue traditionnelle du fonctionnement d'une cellule vivante met en jeu trois types de macromolécules. Ces macromolécules sont des polymères linéaires composés d'un nombre variable de sous-unités, ou résidus, distinctes. La quantité et l'ordre de ces sous-unités, ainsi que le repliement du polymère, déterminent l'information qui y est contenue de même que la fonction biologique qui y est associée.

D'abord, l'acide désoxyribonucléique (ADN) formé des nucléotides A, C, G ou T est responsable du stockage et de la propagation de l'information génétique. Tel que montré par Watson & Crick en 1953 [147], l'ADN se présente de manière générale sous la forme d'une double hélice complémentaire dans l'espace tridimensionnel.

Puis, l'acide ribonucléique (ARN) composé des nucléotides A, C, G et U sert à la synthèse des protéines. Il se divise en trois sous-groupes. L'ARN messager est responsable du transport de l'information génétique encodée dans l'ADN du noyau cellulaire vers le cytoplasme où s'effectue la synthèse grâce à d'autres ARN spécialisés, les ARN de transfert et ribosomaux, de concert avec une panoplie de protéines. La synthèse des protéines, ainsi que l'implication des ARN dans ce processus, est détaillée dans [145].

Enfin, les protéines formées de 20 types d'acides aminés sont les acteurs essentiels au maintien et à la régulation de l'activité cellulaire. Elles jouent aussi un rôle structural comme par exemple dans la constitution des muscles. Leur implication dans la plupart des réactions chimiques intracellulaires suppose que les protéines adoptent une grande diversité de formes permettant le positionnement précis des groupements chimiques mis en jeu dans chacune des réactions. Le repliement des protéines s'effectue par le biais de la formation d'hélices α, de feuillets β et de boucles (pour une introduction détaillée à la structure et à la fonction des protéines, voir [15]).

Cette vue traditionnelle, dans laquelle l'ARN joue un rôle de transporteur inerte d'information et où l'importance de son repliement est négligeable, a par contre été révolutionnée au cours des deux dernières décennies par la découverte des propriétés catalytiques de petits ARN, les ribozymes [24, 64]. Les ribozymes jouent, entre autre, un rôle dans la réplication des virus, leur évitant d'avoir recours à la machinerie de traduction des cellules hôtes.

L'ARN s'avère être une molécule des plus versatile et flexible, capable d'interagir, comme les protéines et contrairement à l'ADN, avec d'autres molécules de manières autant spécifiques que diverses. Cette variété d'activité de l'ARN est attribuée à ses capacités particulières de repliement. Plusieurs théories actuelles décrivent les mécanismes de synthèse, de maturation, de traduction et de dégradation de l'ARN messager en faisant appel à des structures spécifiques de l'ARN. De plus, la régulation de certains processus biologiques fait appel à des ARN structurés précis, les ARN antisens. Le chapitre 5 illustre l'importance de la structure des ARN messagers dans la synthèse et le repliement de la protéine impliquée dans la forme familiale de la maladie de Creutzfeldt-Jakob, de même

que la participation de potentiels ARN antisens. Les ARN antisens sont principalement des inhibiteurs du mécanisme de traduction des ARN et régule par conséquent l'expression des gènes dans la cellule.

Les propriétés régulatrices des antisens et des ribozymes font de l'ARN un outil de recherche thérapeutique puissant de par ses possibilités d'inhibition de cibles fonctionnelles d'ARN. Pour une revue des connaissances actuelles sur la structure et la fonction de ces ARN particuliers, voir [70], [132] et [159].

La compréhension de la versatilité fonctionnelle de l'ARN, ainsi que la capacité de manipuler ces fonctions est rendue possible par l'étude des propriétés structurales de l'ARN. Aussi, une compréhension précise de la fonction des ARN passe par l'analyse des possibilités d'interactions des éléments structuraux avec d'autres ARN ou avec des protéines liantes. Le chapitre 3 présente l'analyse du site de contact entre une partie de l'ARN ribosomal et la protéine L11.

1.2 Structure des ARN

Différents niveaux d'organisation interviennent dans le repliement et la structuration de l'ARN. Les liens phosphodiesters entre nucléotides maintenant la chaîne polymérique représentent le premier de ces niveaux, soit la structure primaire ou séquence.

La propension des ARN à former un type d'appariement très spécifique entre des nucléotides éloignés dans la séquence est à la base du second niveau d'organisation, soit la structure secondaire. Ces appariements sont dénommés de type Watson-Crick du fait que ce sont les mêmes qui interviennent dans la structure en double hélice de l'ADN. Dans l'ARN, ils sont responsables de la formation d'hélices complémentaires à l'intérieur d'un même brin et par le fait même, de boucles de longueurs diverses et de jonctions entre hélices. La structure secondaire est en fait un artifice conceptuel, une simplification de la structure véritable d'un ARN, dans lequel seuls sont exprimés les appariements ne créant pas la formation de pseudo-nœuds. Un pseudo-nœud apparaît lorsqu'un seul des

nucléotides d'un appariement est situé entre deux autres nucléotides appariés de la chaîne polymérique. Cette représentation prend principalement son origine dans le fait que les premières méthodes de prédiction de structure secondaire ne pouvaient traiter du cas des pseudo-nœuds.

D'autres appariements, dits non-canoniques, peuvent aussi se former entre nucléotides, ainsi que des relations d'empilement permettant de stabiliser la structure tertiaire des ARN, le dernier niveau d'organisation. Bien que se référant à la présence d'interactions entre éléments de la structure secondaire, l'appellation structure tertiaire est généralement employée lorsque les coordonnées spatiales des atomes de la molécule sont connues.

Dans la suite de ce mémoire, il sera plus pratique d'effectuer une distinction entre la *séquence* telle que définie plus haut, le *graphe structural*, dans lequel on retrouve une description symbolique de la structure secondaire et des relations tertiaires, et la *structure tertiaire* pour laquelle les coordonnées et les types de chaque atome sont assignés. Le graphe structural s'avère en fait être une description de la topologie d'une structure pouvant contenir des interactions associées aussi bien à la structure secondaire qu'à la structure tertiaire. Les nœuds dans un tel graphe représentent les nucléotides de la structure, tandis que les arcs représentent les relations spatiales entre nucléotides. Les nœuds et les arcs peuvent au besoin contenir certaines informations sur la conformation des nucléotides et le type des relations.

L'acquisition de la structure d'un ARN est une tâche se divisant en plusieurs étapes de difficulté variable. La composition en nucléotides de la séquence est aisément obtenue expérimentalement par méthodes de séquençage, méthodes qui ont été grandement affinées et automatisées durant les 20 dernières années.

Par contre, la détermination de la structure secondaire représente déjà un défi d'envergure. Nombre de méthodes de prédiction de structure ont été élaborées. Une des méthodes principalement employée de nos jours utilise la programmation dynamique afin de minimiser l'énergie potentielle des interactions prédites [160]. Une amélioration à cet algorithme permet toutefois d'inclure la prédiction de structures possédant des interactions ter-

tiaires [125]. Lorsque plusieurs séquences sont disponibles, représentant la même molécule dans différents organismes, la structure secondaire de l'alignement peut être inférée par analyse comparative de séquence (CSA) [66, 109, 154] ou étude de l'énergie libre et des covariations [78]. Ces méthodes produisent généralement des résultats plus précis que celles n'utilisant qu'une seule séquence. D'autres méthodes utilisant des algorithmes génétiques [27, 65] ou encore des grammaires stochastiques hors-contextes [20] ont vu le jour récemment et leur validité reste à confirmer. Dans tous les cas, il est évidemment nécessaire de valider expérimentalement les prédictions obtenues par ces différents modèles. Des expériences comme la mutagenèse dirigée, la sélection *in vitro* et le clivage par ribonucléase et autres enzymes sensibles à la conformation permettent de confirmer les prédictions [112].

Ces différents modèles de prédiction de structure secondaire ont pour but de s'approcher du processus de détermination de la structure tertiaire des ARN, processus coûteux et complexe mais duquel, lorsqu'il est mis en application, en ressort une caractérisation complète de la géométrie de la molécule. De fait, de l'obtention de la structure 3D découle une connaissance immédiate de la structure secondaire, ou plutôt du graphe structural, moyennant un post-traitement similaire à celui présenté au chapitre 3.

La détermination de la structure tertiaire peut être réalisée à la fois de façon théorique, par modélisation, et de façon expérimentale, par observation directe ou indirecte de la structure. Les méthodes théoriques se basent toutefois sur l'obtention de données expérimentales partielles, telles que l'accessibilité de groupements chimiques, les réactions de pontage ainsi que l'utilisation de nucléotides modifiés (pour une revue exhaustive des méthodes de modélisation de structure tertiaire, voir [25]).

L'inférence théorique de la structure des ARN est couramment réalisée manuellement à l'aide d'outils de visualisation moléculaire comme *InsightII*, *VMD* ou *ERNA-3D* [114]. De cette façon, les connaissances structurales du modélisateur sont mises à l'œuvre ainsi que certaines informations partielles sur la structure, telles que la structure secondaire ou les données expérimentales mentionnées plus haut. La modélisation d'un intron du

groupe I [110] constitue une illustration typique de cette technique. La modélisation peut aussi être réalisée à l'aide d'engin d'exploration exhaustive de l'espace des conformations que peuvent adopter chacune des parties des molécules. Le logiciel *MC-Sym* est sans doute le plus complet, utilisant une base de données de résidus et de relations afin d'échantillonner l'espace des conformations de façon discrète à l'aide d'un algorithme de retour-arrière [104]. Enfin, les méthodes les plus réalistes procèdent à la simulation du repliement des molécules par reproduction, le plus fidèlement possible, des interactions du système physique en question. Utilisant des principes découlant de la mécanique classique, et même de la mécanique quantique (dans le cas de petites molécules), ces procédés tentent de résoudre l'évolution de la structure des molécules dans le temps en tenant compte des différentes forces intra ou extramoléculaires agissant sur chacun des atomes du système [17, 28, 120]. Ces méthodes énergétiques sont toutefois très coûteuses en temps de calcul et tirent donc avantage de la génération d'une solution de départ à l'aide d'une des deux premières méthodes. D'ailleurs, les modèles obtenus par ces dernières sont toujours affinés par minimisation énergétique afin d'en optimiser les propriétés stéréochimiques.

Les principales méthodes expérimentales de détermination de structure tertiaire sont la cristallographie par diffraction aux rayons X et la spectroscopie par RMN, bien que certaines approches novatrices telles que la cryo-microscopie électronique commencent à apparaître. La très grande majorité des structures que l'on retrouve dans les banques de données publiques, telles que PDB [13] et NDB [12], trouvent leur origine dans l'une de ces méthodes. Ces dernières procurent cependant une caractérisation de la structure 3D d'une molécule qui est dépendante des conditions expérimentales, lesquelles ne sont pas nécessairement celles présentes *in vivo* et lors des mécanismes chimiques cellulaires. Il est donc nécessaire d'insister sur le fait qu'une structure d'ARN est un système dynamique dont l'état varie en fonction du milieu, et qu'une structure 3D n'est qu'une représentation de la molécule à un moment et dans un état donné.

1.3 Identification de régions fonctionnelles

Le repliement et la structuration des ARN de manière tout aussi variée que spécifique rend possible la réalisation de fonctions diverses au niveau des mécanismes intracellulaires. De nombreuses méthodes théoriques ont donc été élaborées afin d'identifier des régions de l'ARN ayant une structure particulière ou conservée, supportant ainsi les efforts correspondant au niveau expérimental. Nous dénommerons *motif*, toute région structurale conservée, potentiellement fonctionnelle, qu'il s'agisse d'un motif de séquence ou encore d'un motif de structure secondaire ou tertiaire.

1.3.1 Recherche dans les séquences

La recherche de mots conservés, c'est-à-dire de suite de nucléotides conservées dans la séquence des ARN, constitue un champ de recherche des plus développés en bioinformatique. Parmi les méthodes élaborées, on retrouve nombre d'algorithmes de recherche de mots, avec ou sans insertions, délétions ou erreurs, tels que la recherche exacte de répétitions [11], la recherche approchée de motifs [46], la recherche de palindromes [96] ou de doubles répétitions [10].

Certains de ces algorithmes tirent cependant avantage, lorsque possible, de données phylogénétiques provenant d'alignements de séquences, multiples ou non, et permettant d'intégrer une information implicite de la structure secondaire des séquences par étude des covariations (par exemple, dans [63]). Les alignements sont réalisés à l'aide de méthodes allant de la programmation dynamique [135] jusqu'à des méthodes utilisant des modèles de Markov cachés (HMM) [44, 45], lesquelles s'avèrent être les plus efficaces.

Par ailleurs, certains algorithmes permettent la recherche de motifs structuraux dans les banques de données de séquences. Aucune information préalable sur les appariements présents dans les structures n'est requise. Parmi ces algorithmes, on retrouve les méthodes générales *RNAMOT* [55] et *Palingol* [14] qui permettent la description et la recherche de motifs quelconques (bien que restreints à la structure secondaire) ainsi que des méthodes

adaptées à la recherche de motifs spécifiques, tels que les ARN de transfert [47, 99]. Ces différents programmes évaluent la possibilité de retrouver un ensemble d'appariements potentiels permettant de satisfaire la requête. Ceux-ci nécessitent toutefois une connaissance préalable des motifs recherchés et ne peuvent que permettre d'évaluer la distribution de ces motifs dans les banques de données de séquences [144].

Le chapitre 5 présente une méthode novatrice d'identification de régions structurées, possiblement de nouveaux motifs, dans les séquences d'ARN messagers employant l'évaluation de l'énergie libre de repliement (FFE) de la séquence en comparaison avec l'énergie d'une séquence aléatoire de constitution identique.

1.3.2 Recherche dans les graphes structuraux

Peu de travaux ont été effectués jusqu'à maintenant en ce qui concerne l'identification et l'analyse des motifs présents dans les graphes structuraux d'ARN, qu'ils soient restreints à des appariements de la structure secondaire ou non. La rareté de l'information structurale en est sans doute la cause principale, de même que la complexité algorithmique du traitement des graphes. Néanmoins, étant donné les récentes avancées dans le domaine de la prédiction et de la détermination de structures d'ARN, il devient primordial de se pencher sur ce problème.

D'une part, le logiciel *ESSA* [29] utilise une approche graphique, jumelée à un algorithme de recherche de mots dans la séquence sous-jacente à la structure secondaire, permettant de visualiser les interactions entre ces mots dans la structure. D'autre part, le logiciel *Palingol* cité plus haut [14], permet la recherche de motifs formés d'un ensemble d'hélices et de contraintes de positionnement de ces hélices (le long de la séquence) dans des structures spécifiées de façon similaire. La détection est effectuée par satisfaction des contraintes intra et inter-hélices à l'aide d'un algorithme de retour-arrière.

Ces programmes ne permettent toutefois pas de considérer la forme générale des graphes structuraux d'ARN. En effet, ils sont restreints à l'analyse de motifs dans les gra-

phes de structure secondaire. De plus, la détection de motifs jusque-là inconnus n'est pas possible, la recherche se faisant avec un motif cible.

L'algorithme décrit au chapitre 2 et employé dans l'analyse de motifs structuraux au chapitre 4, exploite le concept de recherche de sous-graphe maximal commun et d'isomorphisme de graphe dans la recherche de motifs connus mais aussi dans l'analyse comparative de la structure de différents ARN. Cette analyse comparative rend possible l'identification des régions conservées, potentiellement impliquées dans une fonction importante, en établissant une base de données de motifs communs à ces ARN. Les travaux regroupés dans ces chapitres constituent une approche novatrice dans le domaine de la recherche sur les ARN et s'ajoute de façon complémentaire à ceux portant sur l'analyse de séquences.

1.3.3 Recherche dans les structures tertiaires

Quelques travaux concernent l'analyse structurale de motifs tertiaires reconnus. Ceux-ci traitent du rôle de ces motifs dans le maintien de la structure (par exemple, [93, 37]) et de la fonction (par exemple, [50]), ainsi que du comportement de ces motifs sous diverses conditions expérimentales et théoriques. Pour une revue de divers motifs tertiaires reconnus, voir [8] et [113], et pour une description des interactions avec des protéines, voir [41].

La structure tertiaire des ARN renferme, en théorie, toute l'information nécessaire pour identifier les caractéristiques structurales misent en jeu dans ses interactions avec les autres molécules. La description de chacune des composantes responsables du repliement et de la structuration se fait donc à plusieurs niveaux.

D'abord, chaque nucléotide possède une organisation spatiale propre se décrivant par un ensemble de paramètres précis [126]. Une analyse statistique de la conformation des nucléotides montre une forte préférence de l'ARN pour une conformation spécifique présente dans la majorité des hélices, nommée "C3'-endo anti". Les autres conformations se retrouvent dans des régions plus flexibles de l'ARN, souvent impliquées dans des interac-

tions moléculaires.

Puis, les interactions entre nucléotides, les appariements et les relations d'empilement, peuvent être décrites au niveau physique et symbolique, en mettant en évidence les groupements chimiques impliqués [126, 94]. Ainsi, diverses caractérisations géométriques sont proposées, utilisant des paramètres de torsion [126] ou de pseudo-torsion [43] le long de la chaîne polymérique et des paramètres de translation et de rotation [4, 87, 89] entre les nucléotides éloignés dans la chaîne. Le chapitre 3 illustre l'utilisation des matrices de transformations homogènes dans la description uniforme de tout type de relations entre nucléotides. La déviation d'une relation particulière d'un ARN, par rapport à la moyenne des relations observées, donne une indication de l'apport à la spécificité structurale de cette région de l'ARN. D'ailleurs, certaines analyses phylogénétiques mettent en évidence le rôle important des appariements non-canoniques de la structure tertiaire dans les interactions intermoléculaires [54, 95].

Enfin, l'identification de motifs, c'est-à-dire de sous-ensembles récurrents de nucléotides et de relations dans la structure tertiaire des ARN, peut être réalisée par réduction de la structure tertiaire au graphe structural. Le chapitre 3 illustre une telle transformation. Les méthodes mentionnées à la section précédente et en particulier celles utilisant des algorithmes d'isomorphisme de graphe s'appliquent immédiatement aux graphes résultants.

D'autres algorithmes d'alignement optimal partiel de structures tertiaires permettent d'identifier des sous-structures communes. Ces algorithmes se basent, par exemple, sur l'alignement d'éléments de structures secondaires similaires. Le logiciel *VAST* [60] implante ce type d'algorithme uniquement pour des structures de protéines, bien que la méthode semble s'appliquer tout aussi bien à des ARN. La recherche de sous-structures maximales utilisant un graphe complet, dans lequel les nœuds représentent les atomes de la structure et les arcs symbolisent la distance entre chaque atome, a aussi été décrite [16]. Ce type de graphe ne permet cependant l'application de la méthode qu'à de petits composés chimiques étant donné le grand nombre de nœuds et d'arcs.

La majorité des études sur les régions fonctionnelles sont effectuées expérimentalement

sur l'ensemble des structures d'ARN en se basant sur des motifs reconnus. Des méthodes permettant de restreindre l'espace de recherche en identifiant de nouveaux motifs, telles que celles décrites dans ce mémoire, sont donc essentielles.

1.4 Présentation du mémoire

Au chapitre 2 (publié dans le livre *High Performance Computing Systems and Applications* [57]), nous présentons une approche basée sur l'isomorphisme de graphe pour identifier et compiler des motifs (ou sous-graphes récurrents) de structures secondaires d'ARN. Bien que l'approche soit employée dans le cadre de la recherches dans des acides nucléiques, celle-ci se généralise immédiatement à n'importe quel type de macromolécules pour lesquelles un graphe structural, ou graphe de relations, peut être défini. L'application de cette approche à la détection de motifs dans la structure secondaire de l'ARN ribosomal de *E. coli* est présentée. Dans cet article, j'ai contribué à l'ensemble du développement des méthodes, de l'analyse des résultats et de la rédaction.

Le chapitre 3 (soumis à la revue *Journal of Molecular Biology* [58]) consiste en une description détaillée d'une méthode d'annotation des acides nucléiques, dans laquelle nous proposons de représenter symboliquement, sous forme d'un graphe structural, les informations contenues dans la structure tertiaire. Les méthodes d'extraction des conformations géométriques des nucléotides (composantes unaires) et des interactions chimiques binaires y sont présentées. Puisque les interactions binaires dictent en grande partie le repliement final des structures [103], cette analyse permet de mettre en évidence les régions de la structure tertiaire qui dévient de la norme et sont donc propres à un repliement particulier. Les informations extraites sont évaluées en regard des connaissances préalables sur l'ARN, contenues dans une base de données de conformations et de relations. Cette base de données, améliorée ici grâce à une méthode d'annotation plus précise, est à la base de l'engin de modélisation 3D d'ARN *MC-Sym*. L'article présente une description du site de contact de l'ARN ribosomal avec la protéine L11 dans laquelle les relations non standard

se révèlent être situées dans la région d'interaction, et maintenues par les contacts avec la protéine. Dans ce travail, j'ai participé à l'ensemble du développement des méthodes et du programme final. J'ai aussi réalisé l'analyse du site de contact avec la protéine L11, en plus de rédiger une grande partie de l'article.

À un plus haut niveau, au chapitre 4 (soumis à la revue *Nucleic Acids Research* [59]), nous discutons de l'étude des motifs structuraux, ce qui constitue l'étape subséquente à l'analyse de la structure des ARN après l'étude des composantes unaires et des relations binaires. Cette étude utilise, entre autre, les algorithmes décrits au chapitre 2 en ce qui concerne l'isomorphisme de graphe afin de déterminer les régions conservées dans un ensemble de structures d'ARN provenant de différents domaines phylogénétiques.

D'une part, la présence de régions préservées, au cours de l'évolution naturelle de divers organismes, indique la conservation d'une fonction particulière importante pour ces organismes. D'autre part, l'existence de régions propres à un sous-ensemble d'organismes permet à ces derniers d'exercer une spécificité d'interaction face à des agents extérieurs. Cette spécificité provient d'une divergence des espèces durant l'évolution causée par l'adaptation à des milieux différents. C'est cette divergence de la structure des ARN qui pourrait rendre possible le développement d'agents thérapeutiques antimicrobiens efficaces et non-toxiques pour l'humain.

Des motifs extraits de la structure secondaire d'ARN ribosomaux, ainsi que de la structure tertiaire, démontrent la validité de la méthode en permettant d'identifier des sites fonctionnels déjà confirmés expérimentalement. Aussi, nous présentons certains résultats appuyant une des théories actuelles sur la division évolutionniste des trois domaines phylogénétiques (les *Archaea*, *(eu)Bacteria* et *Eucarya*) à l'aide d'une évaluation de leurs distances structurales relatives. J'ai contribué au développement de l'ensemble des méthodes décrites dans cet article. Quant à l'analyse des résultats et la rédaction du manuscrit, ceux-ci sont le résultat d'une collaboration entre les autres auteurs et moi-même.

Finalement, au chapitre 5 (soumis à la revue *Nucleic Acids Research* [7]), nous présentons une nouvelle méthode d'analyse de séquences par recherche de motifs structuraux.

Cette méthode permet l'identification de régions structurées dans des séquences d'ARN, c'est-à-dire de sections de séquences à l'intérieur desquelles des motifs structuraux ont été conservés durant l'évolution des organismes en question. L'étude d'une région à haut potentiel structural permet l'identification de motifs spécifiques qui seront plus amplement étudiés.

Ces régions sont déterminées par une combinaison de calculs de l'énergie de repliement potentiel et d'alignements avec des séquences similaires. Par la suite, la création de motifs consensus et la génération de modèles 3D théoriques conformes permettent de valider les prédictions en plus de rendre possible l'analyse de l'influence de la structure sur la fonction.

L'application de cette méthode à la protéine responsable de la forme familiale de la maladie de Creutzfeldt-Jakob montre la présence d'une répétition d'un motif de pseudo-nœuds dans l'ARN messager suggérant un effet sur la vitesse de traduction de l'ARN et de la structure de la protéine résultante. Dans ce projet, j'ai mené à terme l'automatisation de certaines des méthodes développées par mes collaborateurs en plus de participer, de façon superficielle, à la rédaction du manuscrit.

Chapitre 2

Structural ribonucleic acid motifs identification and classification

P. Gendron[1], D. Gautheret[2] and F. Major[1]

Abstract

A structural ribonucleic acid (RNA) motif is a recurrent subset of nucleotide arrangements in secondary or tertiary structure. RNA motifs are represented by structural graphs where the nodes represent the nucleotides and the edges represent structural relations. In order to identify all RNA motifs from all known secondary and tertiary structures, an incremental search algorithm was developed. Given a list of motifs of size n, the algorithm builds the motifs of size $n + 1$. The building of a RNA motif database, the development of RNA secondary structure prediction energy potentials, and the integration of RNA motifs in 3-D modeling are discussed.

Keywords: RNA motif, database, graph isomorphism, molecular modeling.

[1]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7

[2]Département de Biologie, Université Aix-Marseille II and CNRS, 13 402 Marseille Cedex 20, France

2.1 Introduction

Since the structure of a macromolecule dictates its function, researchers invest considerable efforts in the determination of their 3-D structures. Determining the secondary structure of a ribonucleic acid (RNA) from sequence data is the first step towards the 3-D structure. The secondary structure of a RNA encodes the majority of its nitrogen base interactions (see Fig. 2.1 and 2.2). The second step consists in determining the tertiary structure which includes all inter-nucleotide interactions found in the 3-D structure. The last step consists in building the 3-D structure from the secondary and tertiary structures. All steps involve combinatorial enumerations.

Structural recurrence suggests functional conservation. Knowledge about RNA motifs help us to identify functional sites, to improve secondary structure prediction, and to simplify three-dimensional modeling. In this paper, we present a method for the detection of unknown RNA motifs. Section 2.2 describes the molecular structure of RNAs, and a relational graph representation that will be used in the construction of a RNA motif database. The following sections are respectively dedicated to the motif search algorithm and its application to the yeast $tRNA^{Phe}$, and 16S and 23S ribosomal RNA (rRNA) subunits.

2.2 Representation

The primary structure of a RNA is a single chain polymer formed of nucleotide units linked together by phosphodiester bonds which connect the O3'-end of one nucleotide to the O5'-end of the next one. Each nucleotide consists of a nitrogen base, a ribose sugar, and a phosphate group (see Fig. 2.1). There are four types of nitrogen bases, namely the purines adenine (A) and guanine (G), and the pyrimidines cytosine (C) and uracyl (U). The nitrogen bases contain hydrogen bond (H-bond) donors and acceptors. A RNA three-dimensional (3-D) structure is stabilized by the formation of H-bonds between pairs of nitrogen bases. Most of these interactions are Watson-Crick base pairs, such as in deoxyribonucleic acid (DNA), that is the $C \bullet G$ and $A \bullet U$ base pairs. The determination by x-ray crystallogra-

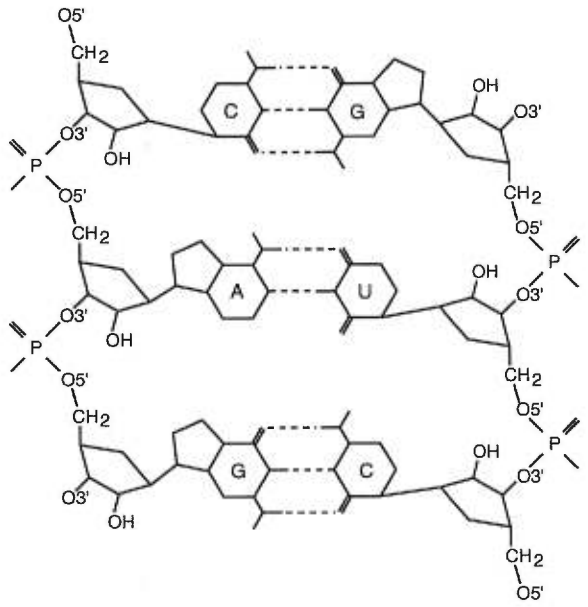


Figure 2.1. *Molecular structure of the largest recurrent subgraph in yeast $tRNA^{Phe}$. The dashed lines indicate H-bonds.*

phy and NMR spectroscopy of new RNA structures has revealed many non-canonical base pairs, such as $G \bullet U$, $G \bullet A$ and $U \bullet U$. RNA 3-D structures are composed of double-helical regions (similar to the DNA double-helix), hairpin loops, multi-branched loops, and bulges that result, in part, from the nitrogen base interactions [102] (see Fig. 2.2).

Secondary structure information is encoded in a structural graph, $G = (V, E)$, where V is the set of vertices, labeled by one of {A,C,G,U} representing each type of nucleotides, and E, is the set of edges representing the structural relations between two nucleotides. In the case of adjacent nucleotides in the sequence, the direction of the edges follows the phosphodiester linkage from 5'→3', and undirected edges are used for non-adjacent nucleotides, such as for H-bonding relations (see Fig. 2.3). Given the particular nature of nucleotide interactions, structural graphs typically possess a number of edges inferior to twice that of the number of vertices.

A linear representation was developed to simplify the computer storage of relational

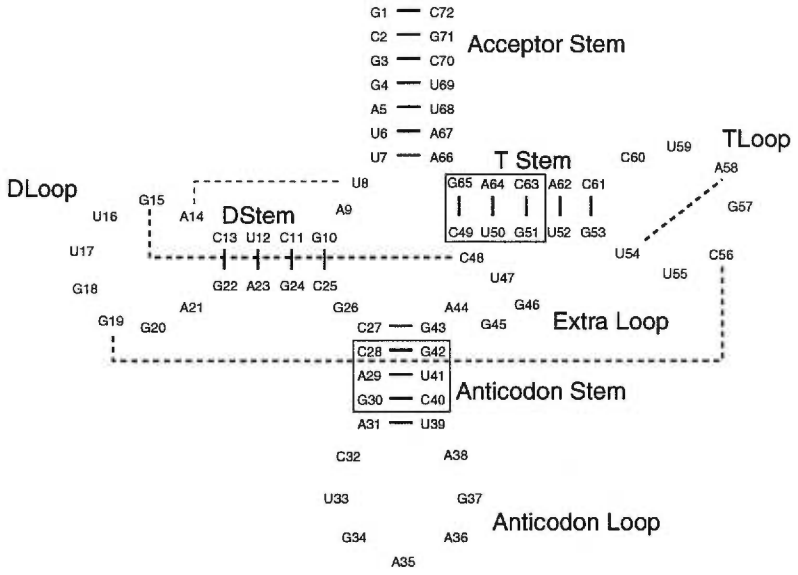


Figure 2.2. *The clover leaf secondary structure of yeast tRNA^{Phe}. The base pairs of the secondary structure are indicated by bold lines linking both nitrogen bases. The tertiary interactions between two nitrogen bases are indicated by dashed lines. The longest subgraph that appears at least twice is shown in boxes.*

graphs in which the *helical* relations are indicated by vertical bars, (|), the adjacent *non-helical* relations are indicated by slashes, (/), the *Watson-Crick* base-pairing relations are indicated by dashes, (−), the *other than Watson-Crick* base pairing relations are indicated by dots, (.), and all other *tertiary* relations are indicated by tildes, (∼). Fig. 2.3 shows the largest recurrent subgraph in yeast tRNA^{Phe} using the graph and linear representations.

2.3 Searching for motifs

The problem of finding structural motifs in RNA secondary structures can be divided into two distinct tasks. First is an enumeration of all possible subgraphs and, second, is their isomorphic classification.

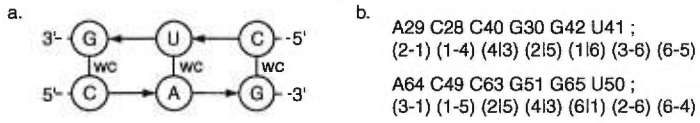


Figure 2.3. *(a) Graph of relations for the largest recurrent subgraph in yeast $tRNA^{Phe}$. (b) A linear representation of the motif shown in (a) as used in the motif database. Note that the linear representation is not unique for each motif since the nucleotides can be ordered in many different ways.*

2.3.1 Enumeration

The size of interesting RNA sequences ranges from several tens to more than three thousand nucleotides. The number of subgraphs grows rapidly as its size increases, and therefore a straightforward enumeration is often impracticable. Different heuristics have been developed to cope with this overwhelming data, and must be used in the identification of RNA motifs. One of them is to limit the motif size to no more than 15 nucleotides. Another one consists in retaining only the most *significant* motifs. One possible definition of significance is the number of occurrences, let us say for instance that we have a motif if the corresponding subgraph occurs at least p times in all considered secondary structures. For this demonstration, we decided to eliminate all the subgraphs that appear only once by fixing the value of p to 2 (see Fig. 2.4). However, in the building of the actual RNA motif database, a more precise evaluation of the significance will be made, based on the relevance of a motif's consistuents.

Central to the motif identification process is the notion of incremental enumeration. In order to find the subgraphs of size n, we consider the subgraphs of size $n-1$. The subgraphs of size $n-1$ are extended by connecting the nodes that are connected to it from the secondary structure. This approach allows us to find all motifs since any subgraph of size n occurring q times contains at least one occurrence of a subgraph of size $n-1$ that occurs at least q times.

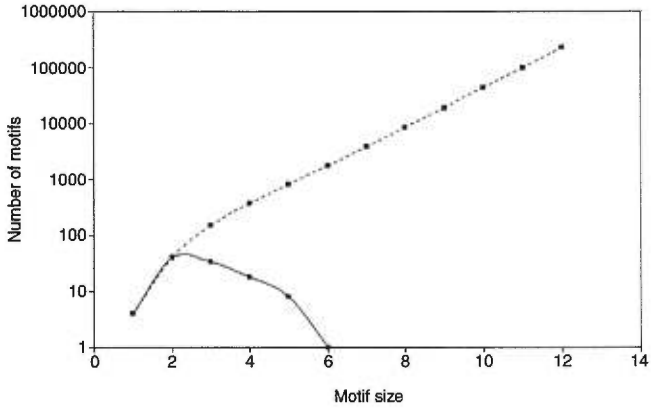


Figure 2.4. *The reduction in the number of motifs by using different definition of significance in yeast tRNA^{Phe}. The dashed line shows the number of motifs occuring at least once, $p = 1$. The plain line shows the number of motifs occuring at least twice, $p = 2$.*

2.3.2 Graph Isomorphism

The classification of subgraphs requires an efficient graph isomorphism algorithm. However, the graph isomorphism problem is one of the long-standing intractable problems in discrete mathematics. Given two graphs, $G_A = (V_A, E_A)$ and $G_B = (V_B, E_B)$, the problem is to establish the existence of a one-to-one onto mapping function, f, from V_A to V_B such that $(i, j) \in E_A$ if and only if $(f(i), f(j)) \in E_B$.

This problem belongs to NP, the class of problems that can only be decided by a *nondeterministic polynomial time algorithm.* Moreover, it is NP-*complete*, and hence, no sub-exponential running time algorithm is known to solve it. Nevertheless, specific RNA secondary structure information allows us to split the isomorphism determination in three stages of increasing complexity. First, a comparison is made between two subgraph vertices, based on their respective type and number of relations, or *degree*. Then, if the subgraphs contain the same nucleotides, their edges are compared, and if they are equal,

a depth-first search is finally applied to verify their isomorphism. The algorithm for this depth-first search was adapted from [143] and can be described as follows.

Let M_A and M_B be the adjacency matrices of each subgraphs, A and B, defined similarly as:

$$M_k[i][j] = \begin{cases} t_{ij} & \text{if edge } (i,j) \in E_k \\ 0 & otherwise \end{cases}$$

where $k \in \{A, B\}$, and $t_{ij} \in \{-, |, /, ., \sim\}$ is the type of relation between vertices i and j. Let M_0 be the matrix of possible equivalences between subgraphs, A and B, where $M_0[i][j] = 1$ if node i in A and node j in B are the same nucleotides and have the same degree. The algorithm generates all permutations of equivalent vertices in the subgraphs based on M_0, stores them in the form of a vector of equivalences H, and tests for isomorphism for each permutation. The two subgraphs are isomorphic if and only if $M_A[i][j] = M_B[[H[i]][H[j]]$ for all i and j. The details of the third stage of the algorithm are shown in Fig. 2.5.

A simple analysis of the algorithm allows us to evaluate the time required during the first two stages, that is, $O(max(|E|, |V|) \log(max(|E|, |V|)))$, due to the sorting of the vertices and edges. The third stage takes a time in $O(|V|^2)$ for each permutation. Section 2.4.2 gives an estimate of the number of permutations based on the results of applying the algorithm on three different secondary structures.

Several alternative approaches to the depth-first search algorithm were considered. All required modifications of the graph structure in order to meet the characteristics of the concerned graphs. For instance, Hopcroft and Wong [74] showed that isomorphism of planar graph can theoretically be tested in linear time. If tertiary interactions are not considered, a secondary structure becomes a planar graph which would allow the use of the approach proposed by Hopcroft and Wong.

```
GraphIsomorphism (depth):
  finished := false;
  i := 1;
  while (!finished & i ≤ graphSize) do
    if (M0[depth][i] = 1 and G[i] = 0)
      H[depth] := i;
      G[i] := depth;
      if (depth < graphSize)
        finished := GraphIsomorphism (depth + 1);
      else
        return (TestIsomorphism (MA, MB, H));
      H[depth] := 0;
      G[i] := 0;
    i := i + 1;
  return finished;

TestIsomorphism (MA, MB, H):
  for j := 1 to graphSize do
    for k := 1 to graphSize do
      if (MA[j][k] ≠ MB[[H[j]][H[k]])
        return false;
  return true;
```

Figure 2.5. *Pseudo-code for the third stage in the isomorphism evaluation. The function is initially called with depth=1.*

2.4 Applications

Prior to the construction of a complete RNA motif database, our algorithms were tested on a reduced number of secondary structures, namely, the yeast $tRNA^{Phe}$ [82], and *Escherichia coli* 16s and 23s rRNAs [118].

2.4.1 Secondary structure motifs

For the considered structures, a minimum occurrence of $p = 2$ was used to define the significance of a motif. In fact, without this distinction, the number of motifs grows ex-

a.

	16S	23S
b	3	4
c	2	8
d	54	118
e	1	1
f	4	0
g	0	2

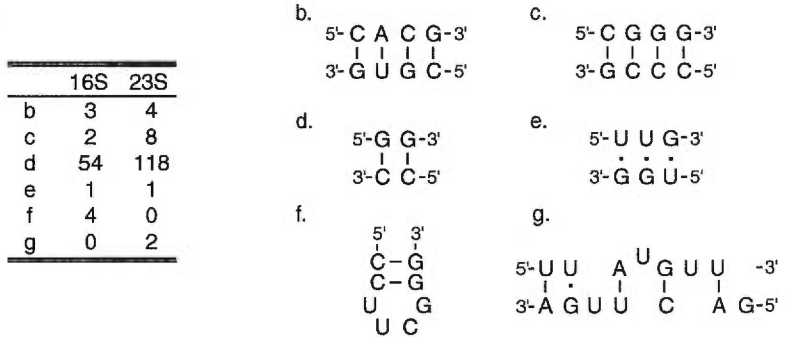


Figure 2.6. *Interesting motifs found and (a) their respective occurrences in* E. coli *16S and 23S rRNAs: (b) and (c) stems, (d) double G•C pairing, (e) triple G•U mismatches, (f) hairpin loop, and (g) one of the longest motifs found.*

ponentially and becomes intractable even for the small secondary structure of the yeast $tRNA^{Phe}$ (see Fig. 2.4).

The largest motif identified in the yeast $tRNA^{Phe}$ is composed of three base pairs and appears twice, as shown in Fig. 2.3. The rRNA secondary structures of *Escherichia coli* contains 1542 (16S) and 2904 (23S) nucleotides. These secondary structures were first treated separately, and then together. A large number of motifs (see Fig. 2.6) was found, but no motifs of size larger than 14 nucleotides were found.

Several stems, such as those shown in Fig. 2.6b and 2.6c, as well as more complex motifs, such as the hairpin loop of Fig. 2.6f which appears four times in the 16S rRNA, were among the motifs found by our algorithm. The double guanine-cytosine Watson-Crick base-pairing (see Fig. 2.6d) appears twice as many times as any other motif of the same size. This suggests that this particular tandem of base pairs might have been selected for a peculiar stability or function. Also of interest is the occurrence in both rRNAs of a triple $G•U$ mismatches (see Fig. 2.6e), which embed the double $G•U$ mismatches that was identified by [54].

Many more motifs were found in these preliminary studies, and their significance will

be determined precisely when the RNA motif database will be constructed.

2.4.2 Algorithm efficiency

In order to determine the efficiency of the algorithm, a simple benchmark was used. First, the time required by the construction of a given set of motifs was evaluated. Second, the mean number of evaluated permutations for each subgraph pair was determined. These evaluations were made using the *E. coli* rRNAs, as above, on a Silicon Graphics Origin 2000 equipped with R10000 CPUs.

The 34350 identified motifs that range from size 1 to 14 were found in less than seven minutes. This is rather fast considering that more than one million subgraphs were tested for isomorphism. For all motif sizes, the last stage of the isomorphism algorithm, involving a depth-first search of all possible permutations of vertices, took in O($|V|$) trials in practice (results not shown). It is worth noting that more computation is required on average to compare two non-isomorphic subgraphs than to compare two isomorphic subgraphs since in the former case all permutations need to be evaluated, whereas the evaluation of only a subset is required in the latter case. The complexity of the entire algorithm would then be in the O($max(|E|, |V|) \log(max(|E|, |V|))$) for the first two stages, and in the O($|V|^3$), otherwise. This satisfies Corneil and Gotlieb's criterion, that is, a graph isomorphism algorithm is efficient if the observed running time is polynomial [34]. Over 5000 secondary structures will be considered for the construction of the RNA motif database. Since larger motifs are expected to be found, we predict the running time to be bounded by no more than a higher degree polynomial given the particular class of structural graph that encodes RNA secondary structure.

2.5 Perspectives

The described algorithm defines the basic component for the creation of a RNA motif database. This algorithm allowed us to identify and classify the structural motifs of three

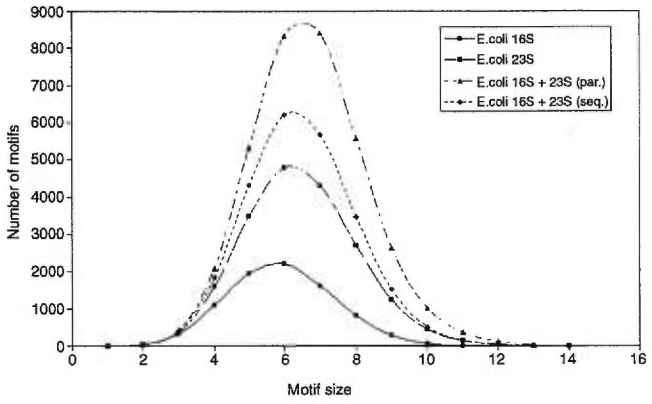


Figure 2.7. *Variation of the number of motifs found in Escherichia coli rRNAs.*

secondary structures.

Two approaches can be adopted for the building of a motif database. First is sequentially, that is by finding all significant motifs for each RNA one after the other. This approach has the downside that even when the most relevant motifs are present in a majority of secondary structures, one motif may be missed if it is not found at least p times in one secondary structure. Second is in parallel, that is by considering all secondary structures at once. In this case, all motifs would be identified but a very efficient external memory would be required since the subgraphs that must be considered for each motif size is so large that they cannot be kept in central memory. Both approaches can be parallelized since the subgraph enumeration can be done independently for each structure and motif size. The curves in Fig. 2.7 illustrate the number of motifs that would have not been identified by treating the structures independently. These results suggest that the first approach is not appropriate for the study of RNA motifs.

Several improvements will be necessary before we can apply our algorithm to a large number of secondary structures. The parallelization of the enumeration and isomorphism

evaluation procedures will be considered. The only other option would be to neglect tertiary information and use a planar graph representation.

Knowledge of RNA motifs will allow us to develop secondary structure energy potentials, such as those used in the Zuker's algorithm [161]. The basic assumption being that a secondary structure that contain motifs would be favored. We would even be in a position to evaluate how much motifs a secondary structure must contain to be valid. Finally, motif three-dimensional information would provide a library of building blocks for three-dimensional modeling.

Acknowledgments

We thank Elie Hanna for his early work on this project and useful discussions. This work is funded by the Medical Research Council (MRC) of Canada. Patrick Gendron holds a FCAR scholarship. François Major is a MRC fellow.

Chapitre 3

Quantitative Analysis of Nucleic Acid Three-Dimensional Structures

P. Gendron[1], S. Lemieux[1], and F. Major[1]

Abstract

We present *MC-Annotate*, a computer program for analyzing nucleic acids three-dimensional structures was developed. *MC-Annotate* uses a general computer representation based on a structural graph that we developed for the encoding of RNA or DNA structure information at all abstraction levels, from sequence to quaternary structure. *MC-Annotate* analyzes the conformations and spatial relations of nucleotides in a three-dimensional structure, and reports geometrical information such as atomic distances and torsion angles, as well as higher-order structural information such as sugar puckering modes, nitrogen base orientations around the glycosyl bond, hydrogen bonding patterns and stacking interactions. We developed distance metrics to compare nucleotide conformations and spatial relations, such as those involved in base pairing and stacking. *MC-Annotate* was used to build a structural database of nucleotide conformations and spatial relations employed by the *MC-Sym* program to construct RNA three-dimensional models. New structural features, as well as potential errors, in a given structure are detected by mapping the graph of relations produced by *MC-Annotate* to the structures in the database. The analysis of the ribosomal RNA fragment that binds to protein L11 was made, and showed peculiar nucleotide conformations and spatial relations in the regions where the

[1]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7

RNA interacts with the protein. The question whether the current database of nucleic acid three-dimensional structures is complete was addressed.

Keywords: Nitrogen base interactions, three-dimensional structure and modeling, quantitative analysis, structure comparison, RNA structure database, RNA-protein complexes, computer algorithm.

3.1 Introduction

The function of ribonucleic acid molecules (RNA) goes far beyond the roles of genetic information repository and carrier. The structural flexibility of RNAs confers a large diversity of three-dimensional (3-D) shapes and functions [134]. The properties of RNA to interact with other macromolecules, and in particular to perform catalytic activities, have considerably increased the scientific interest for RNAs and, consequently, the number of individuals and industries involved in RNA research.

For over fiftheen years, three transfer RNA, the yeast $tRNA^{Phe}$ and $tRNA^{Asp}$, and the E.coli $tRNA^{Gln}$, were the only available x-ray crystal structures of biologically active RNA. As a consequence, the reliability of most RNA structure prediction was evaluated on the capacity in reproducing the tRNA structures, which is a very restrictive learning set. Recently, however, several new RNAs of biological interests were discovered, and experimental techniques that yield medium- and high-resolution structural information, such as x-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, are now commonly applied to RNA.

The newly determined structures, as made available in the Protein DataBank (PDB) [13], range from a few nucleotides, such as the lead-activated [73, 148] (PDB codes: 1LDZ and 429D) and the hammerhead [123] (PDB code: 1HMH) ribozymes, to several hundreds of nucleotides, such as the P4-P6 domain of the *Tetrahymena thermophila* group I intron [61] (PDB code: 1GID) and the 23S rRNA [6] (PDB code: 1FFK). More RNA 3-D structures are expected to be released in a near future, as research groups have been able to obtain low- and medium-precision x-ray density maps of the complete ribosomal assembly [22, 122].

It is clear that the understanding and establishment of the variety of RNA structures and activities, as well as our ability to manipulate RNA function, depend upon the acquisition and analysis of RNA structures. For instance, the success of rational development of pharmaceutical products based on RNA relies on a better understanding of RNA structure-

function relationships, as well as on the localization of essential RNAs in the living cell.

Geometrical or quantitative analysis of RNA 3-D structures is employed in the validation of new RNA 3-D structures, the comparison and identification of RNA structural features, the development of empirical modeling systems, and, more generally, the studying and learning of RNA structure-function relationships. However, a very limited number of structure analysis methods that can be applied specifically to RNA structures are available, and most of them are based on interactive visualization, or atomic distance and torsion angle calculations. Interactive visualization is limited to subjective analysis of only a limited number of small RNA domains. The quantitative methods that have been employed in the past to compare two or more structures were mainly based on the computation of bond and torsion angles values [4, 87, 89]. Because different torsion angle patterns can result in similar conformations [121], comparative analysis based on torsion angle evaluations are not always informative [56]. We present here new methods and a computer program, *MC-Annotate*, to quantitatively and objectively analyze the nucleotide conformations and base-base interactions in RNA and DNA 3-D structures.

Our work focused on nucleotide conformations and base-base interactions since they are determinant in RNA 3-D structure folding and stabilization, both locally and globally. In an attempt to exhaustively and quantitatively analyze the nucleotide conformations and base-base interactions of all available RNA 3-D structures, as well as to update the parameters and conformational sampling of the *MC-Sym* 3-D modeling program [104] as new RNA and DNA structures are made available, we developed the new computer program *MC-Annotate*. Annotating a RNA or DNA 3-D structure consists in transforming its 3-D atomic coordinates into a set of symbols and mathematical parameters which makes the analysis and comparison of 3-D structures simpler and free of human intervention. The validity of nucleotide conformations and base-base interactions is confirmed using specific distance metrics. Each conformation or interaction instance is evaluated relatively to each other in its given conformational space, and its peculiarity, or adversely conformity, assessed using a peculiarity value. *MC-Annotate* was also used for analyzing the struc-

tural features in the ribosomal RNA fragment that binds to protein L11, and to assess the completeness of the current set of available 3-D structure.

3.2 Methods

The methodology used in the analysis and annotation of RNA and DNA structures consists in the conversion of geometrical data into the corresponding structural graph and the evaluation of this graph in regard to the current knowledge on nucleic acid structure. Here, a *structural graph* is a symbolic representation of a RNA or DNA three-dimensional structure formed of a set of features, that is vertices connected by edges. Vertices represent nucleotides that possess certain attributes based on the geometrical disposition of their atoms. Edges are representative of the relative spatial relations characterizing binary nucleotide interactions (see Figure 3.1). Since many nucleic acid structures in the Protein Data Bank (PDB) [13] or the Nucleic Acid Database (NDB) [12] lack hydrogen atoms, those need to be added prior to the analysis from bond lengths and angles described in the Cornell *et al.* force field [35]. Precise distance metrics allow us to relate structural features to each other in their relative conformational space and evaluate the peculiarity of each feature.

3.2.1 Nucleotide conformations

Based on the definition by Saenger [126], the symbolic characterization of nucleotide conformations in a structural graph takes place on two levels. The first one is the position of the furanose ring atoms relative to the general plane of the ring, which determines the sugar puckering mode. The second is the orientation of the nitrogen base relative to the sugar, which can be determined by the angle around the glycosyl bond, χ, defined by the atoms O4', C1', N9 and C4 for purines and the atoms O4', C1', N1 and C2 for pyrimidines. Since the other parts of a nucleotide are mostly rigid, the two above properties represent a fair qualitative description of nucleotide conformations.

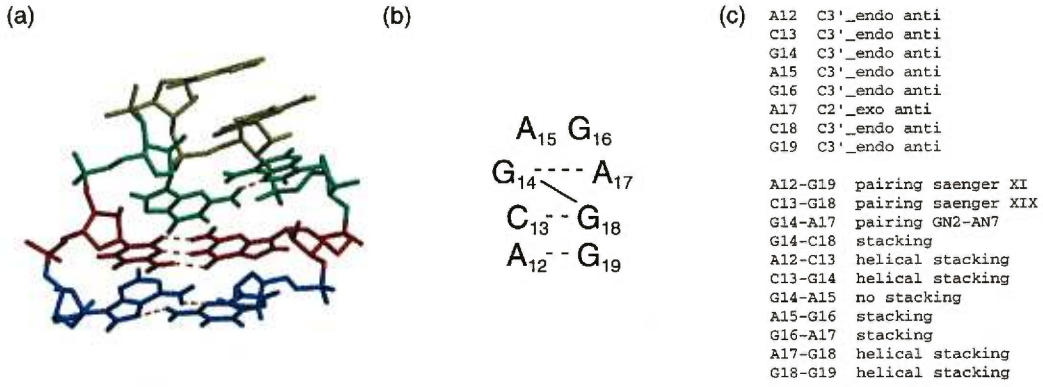


Figure 3.1. *Part of the crystal structure of Sarcin/Ricin loop from rat 28S ribosomal RNA [36] (PDB code: 430D). a) Three-dimensional structure showing the GAGA tetraloop. b) Secondary structure in which dashed lines represent base pairings and solid lines represent non-adjacent stacking interactions. c) Syntactic description of the corresponding structural graph showing attributes of nucleotide conformations and chemical interactions.*

3.2.1.1 Distance metric

The comparison of nucleotide conformations is based on a superimposition of the nitrogen bases local referentials as defined below. The distance metric, $d(\mathbf{b}_1, \mathbf{b}_2)$, calculates the root mean square deviation (RMSD) between the heavy atoms of the backbone of the two nucleotides $\mathbf{b}_1$ and $\mathbf{b}_2$ after the superimposition [56]. Compared to a standard RMSD distance metric calculated on all atoms after optimal alignment, performed using the analytical method described in [79] and [80], our distance metric allows one to avoid the consideration of variations among nitrogen base atom positions while showing a good correlation with this standard, all-atom RMSD (see Figure 3.2). It emphasizes the distinction in the backbone conformations, the nitrogen bases being mostly rigid as shown in Figure 3.3. Figure 3.4 shows that the proposed metric represents more accurately the variations of the backbone relative to the nitrogen base than does the standard RMSD.

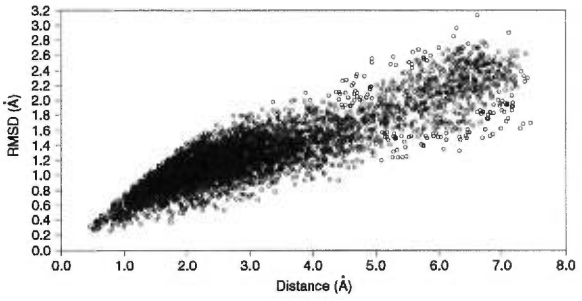


Figure 3.2. *Correlation between standard RMSD and the proposed distance metric for nucleotide conformations. Each dot represents a randomly selected pair of residues from PDB or NDB nucleic acid structures.*

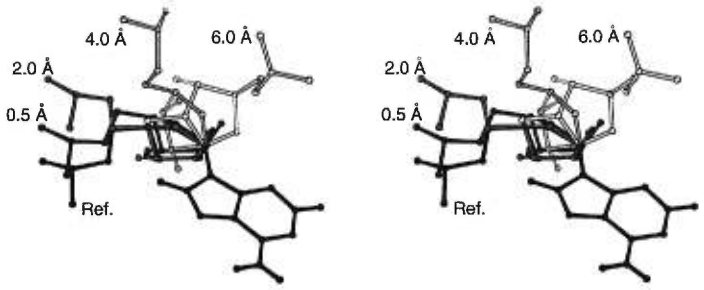


Figure 3.3. *Stereo view of the variation of the nucleotide backbone conformation with increasing distance from the reference nucleotide shown in black.*

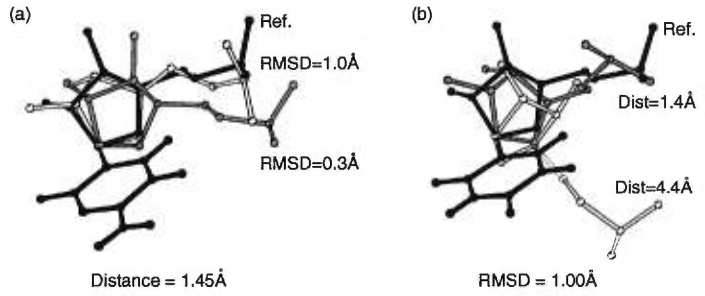


Figure 3.4. *Disagreement between standard RMSD and the proposed distance metric favoring the latter. a) Two nearly identical conformations (in light and medium gray) having the same distance to a reference (in black) but showing a variation in RMSD. b) Two rather different conformations (in light and medium gray) with a large conformational distance but that yield the same RMSD value.*

3.2.2 Spatial relations

Spatial relations in a structural graph represent nucleotide interactions that not only stabilize the local conformation of a nucleic acid structure, but are believed to determine the general arrangement of the whole structure. For example, in $tRNA^{Phe}$ (PDB accession number 6TNA [139]), knowledge of the tertiary interaction involving the base pairing between U8 and A14 (see Figure 2.2) was crucial for the modeling of the $tRNA^{Phe}$ into its characteristic L-shape [103].

In classifying spatial relations, we consider two types of nucleotide interactions: adjacent and non-adjacent. Nitrogen base spatial relations are thus of five distinct types: adjacent-stacked, adjacent-paired, adjacent-unstacked, non-adjacent-stacked and non-adjacent-paired (see Figure 3.5). Adjacency of nucleotides is determined either using the PDB numbering (when valid) or the O3'-P bond distance.

Traditional encodings of adjacent spatial relations use the six backbone torsion angles α, β, γ, δ, ϵ and ζ [126] or, more recently, two pseudotorsion angles η and θ [43]. These parameters accurately describe the relative placement of nucleotides linked by phosphodi-

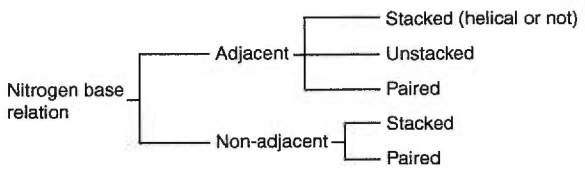


Figure 3.5. *Hierarchical classification of the nitrogen base spatial relations.*

ester bonds. However, it has already been observed that distinct torsion angle combinations produce similar backbone and nitrogen base orientation; this phenomenon is known as the "crankshaft effect" [121]. Also, non-adjacent long-range relations, like base pairings stabilized by hydrogen bonds or non-adjacent stackings, cannot be represented in this way. Rather, a plethora of rotation parameters have been used to describe this type of interaction [4, 87, 89]. Here, we introduce a simplified and unified encoding for any type of nucleotide relations that uses homogeneous transformation matrices.

3.2.2.1 Homogeneous transformation matrices

Homogeneous transformation matrices, or HTMs, were first developed in the field of geometry [106] and later extensively used in computer graphics and robotics. They encode, in the form of a 4x4 matrix, the geometric operations needed to transform objects in three-dimensional space from one local referential to another. In the present context, a HTM describes a spatial relationship as a composition of translation and rotation between the two local referential of the nitrogen bases involved in the relation.

The local referential of a nitrogen base can be viewed as a Cartesian coordinate system whose position relative to the base is computed from the atomic coordinates (see Figure 3.6). The position of the local referential is arbitrary but should be identically set for each type of nucleotides. We propose the following construct in the determination of local referentials. Let $\mathbf{u}$ be the normalized vector between coordinates of atom N1 and C2 for

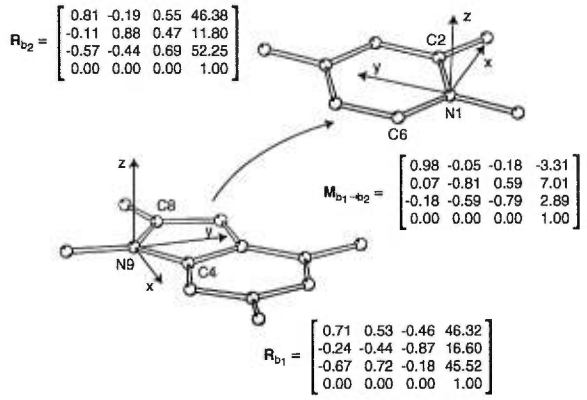


Figure 3.6. *Homogeneous transformation matrices used to represent spatial relations.* $\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}$ *encodes the position of* $\mathbf{R}_{\mathbf{b}_2}$ *relative to* $\mathbf{R}_{\mathbf{b}_1}$.

pyrimidines and N9 and C4 for purines. Also, let $\mathbf{v}$ be the normalized vector between coordinates of atom N1 and C6 for pyrimidines and N9 and C8 for purines. Then, the unit vector $\mathbf{y}$ of the Cartesian coordinate system lie in the direction given by the sum $\mathbf{u} + \mathbf{v}$, the unit vector $\mathbf{z}$ is oriented along the cross product $\mathbf{u} \times \mathbf{v}$ whereas the unit vector $\mathbf{x}$, following the right hand rule for Cartesian coordinate system, is given by $\mathbf{y} \times \mathbf{z}$.

Let the HTMs $\mathbf{R}_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_2}$ be the local referentials of two nucleotides $\mathbf{b}_1$ and $\mathbf{b}_2$ as expressed relative to the global referential centered at $(0, 0, 0)$. The spatial relation between $\mathbf{R}_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_2}$ is then given by the HTM $\mathbf{M}_{\mathbf{b_1} \to \mathbf{b_2}} = \mathbf{R}_{\mathbf{b}_1}^{-1}\mathbf{R}_{\mathbf{b}_2}$. In a molecular modeling context, given two arbitrary nucleotides $\mathbf{b}'_1$ and $\mathbf{b}'_2$, the spatial relation can be reapplied to position atoms of $\mathbf{b}'_2$ relative to $\mathbf{b}'_1$ using the transformation $\mathbf{R}_{\mathbf{b}'_1}\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}\mathbf{R}_{\mathbf{b}'_2}^{-1}$. In a similar way, atoms of $\mathbf{b}'_1$ can be positioned relative to $\mathbf{b}'_2$ using the inverse transformation $\mathbf{R}_{\mathbf{b}'_2}\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}^{-1}\mathbf{R}_{\mathbf{b}'_1}^{-1}$. It is to be noted here that $\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}^{-1} = \mathbf{M}_{\mathbf{b}_2 \to \mathbf{b}_1}$, that is the inverse of the transformation extracted between $\mathbf{R}_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_2}$ is equivalent to the one that would be extracted between $\mathbf{R}_{\mathbf{b}_2}$ and $\mathbf{R}_{\mathbf{b}_1}$.

3.2.2.2 Distance metric

In order to effectively compare spatial relations, a distance metric between two transformation $\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}$ and $\mathbf{N}_{\mathbf{b}'_1 \to \mathbf{b}'_2}$ should possess the following characteristics (see Figure 3.7 for a two-dimensional vectorial analogy):

$$d(\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}, \mathbf{N}_{\mathbf{b}'_1 \to \mathbf{b}'_2}) = d(\mathbf{N}_{\mathbf{b}'_1 \to \mathbf{b}'_2}, \mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}) \tag{3.1}$$

$$d(\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}, \mathbf{M}^{-1}_{\mathbf{b}_1 \to \mathbf{b}_2}) = 0 \iff \mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2} = \mathbf{M}^{-1}_{\mathbf{b}_1 \to \mathbf{b}_2} \tag{3.2}$$

$$d(\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2}, \mathbf{N}_{\mathbf{b}'_1 \to \mathbf{b}'_2}) = d(\mathbf{M}^{-1}_{\mathbf{b}_1 \to \mathbf{b}_2}, \mathbf{N}^{-1}_{\mathbf{b}'_1 \to \mathbf{b}'_2}) \tag{3.3}$$

Equation 3.1 states that the distance metric should obviously be commutative. Equation 3.2 states that a relation should have a null distance with itself but not with it's inverse (unless they are equal). Equation 3.3, however, states that the distance metric should not depend on the direction of application that is implicit in the HTM representation since we want the metric to discriminate non-directional nucleotide relations.

Simple Euclidean distance, in the 16 dimensional space of HTMs does not satisfy these characteristics since HTMs embed a combination of translation and rotation terms, which cannot be mixed together, and describe oriented relations. Thus, a more appropriate definition for the distance metric follows.

A HTM can be decomposed in the product of two HTMs as $\mathbf{M} = \mathbf{TR}$, where $\mathbf{T}$ contains only the translation part of the original HTM and $\mathbf{R}$ contains the rotation part. Paul [119] showed how to extract the length l of the translation as well as the angle θ and axis of rotation k from the two matrices.

We then define $S(\mathbf{M})$, the *strength* of a transformation, regardless of the axis of rotation, as:

$$S(\mathbf{M}) = \sqrt{l^2 + (\frac{\theta}{\alpha})^2} \tag{3.4}$$

where α represents a scaling factor between the translation and rotation contribution. A

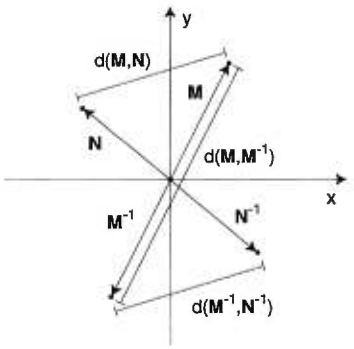


Figure 3.7. *Two-dimensional vectorial representation of the distance metric characteristics. If* $\mathbf{M}$ *and* $\mathbf{N}$ *are two vectors representing spatial relations, one can see that the distance between the extremities of these vectors is independent of the order in which the distance is computed (Equation 3.1) and is equal to the distance between the extremities of their inverses (Equation 3.3). Also, the distance between the extremity of a vector and it's inverse will be zero only if they are equal (Equation 3.2).*

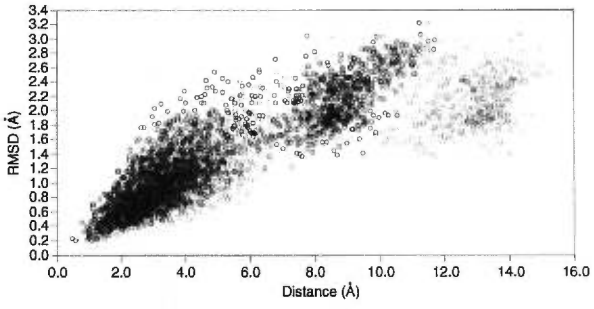


Figure 3.8. *Correlation between standard RMSD and the proposed distance metric for spatial relations (shown are results for the stacking relation). Gray dots are results with* $\alpha = 15°/Å$ *whereas black dots are results with* $\alpha = 30°/Å$.

scaling factor of 30°/Å has been found to yield good results (see Figure 3.8). This implies that a rotation of 30° around any axis is equivalent to a displacement of 1 Å between two nucleotides' local referentials. Using this definition, we propose the following expression for the spatial relation distance metric:

$$d(\mathbf{M},\mathbf{N}) = \frac{[S(\mathbf{M}\mathbf{N}^{-1}) + S(\mathbf{M}^{-1}\mathbf{N})]}{2} \tag{3.5}$$

which satisfies the requirements of equations 3.1 through 3.3. In this expression, the composition of transformation $\mathbf{M}\mathbf{N}^{-1}$ can be seen as the necessary transformation needed to align the local referential $\mathbf{R}'_{\mathbf{b}_2}$ with $\mathbf{R}_{\mathbf{b}_2}$ when $\mathbf{R}'_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_1}$ are aligned with the global referential. Similarly, $\mathbf{M}^{-1}\mathbf{N}$ can be viewed as the transformation required to align $\mathbf{R}'_{\mathbf{b}_1}$ with $\mathbf{R}_{\mathbf{b}_1}$ when $\mathbf{R}'_{\mathbf{b}_2}$ and $\mathbf{R}_{\mathbf{b}_2}$ are aligned with the global referential.

Figure 3.8 shows that this distance metric is roughly equivalent to a standard RMSD calculated on the alignment of the two pairs of nucleotides using the analytical method described in [79] and [80]. Furthermore, our distance metric can better discriminate between two relations that differ by rotations or flip of the nitrogen bases (see Figure 3.9).

Although HTMs are perfectly suited to uniformly encode spatial relationships between nitrogen bases, the information they contain is in a too compact form to identify the type of relations they encode without reproducing the relation in 3-D space and evaluating other parameters. For this reason, we base our symbolic evaluation of spatial relations on an all-atom representation, allowing easier identification of interacting chemical groups.

3.2.2.3 Base pairing

Hydrogen bonds are weak electrostatic interactions involving hydrogen atoms located between two atoms of higher electronegativity. Being weaker than covalent bonds, they are nevertheless the most significant interaction responsible for the folding and stabilization of complex RNA and DNA molecules. Since these H-bonds are somewhat directional due to the orbital shape of the electron density distribution involved, they favor planar nucleotide

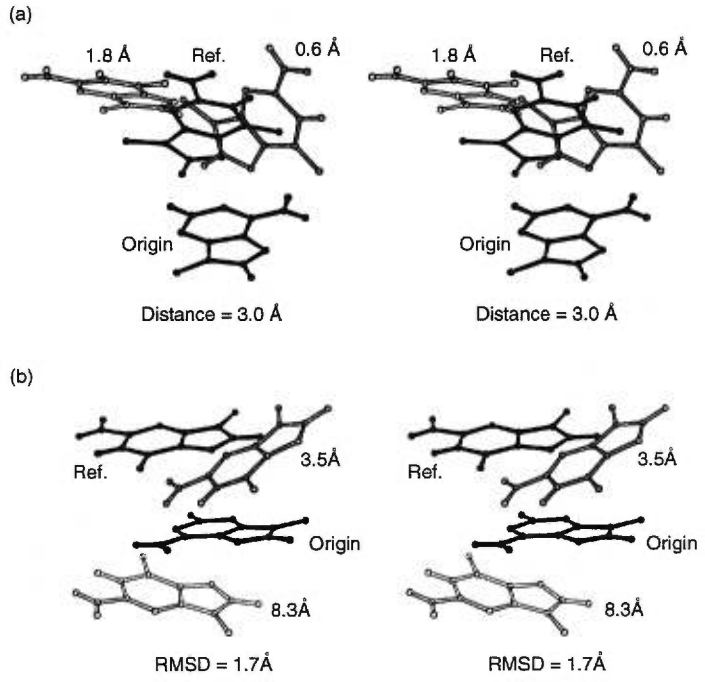


Figure 3.9. *Differences between the RMSD metric and the proposed distance metric. In a), the relations are at the same distance from the reference shown in black but exhibit a significant difference in RMSD. In b), standard RMSD cannot discriminate between a standard stacking and a reverse stacking of the bases. Note: RMSD is computed on and after alignment of the nitrogen bases only; backbone atoms are not considered.*

pairings that comprise at least one hydrogen bond but as much as three, as in the case of the Watson-Crick pairs. Base pairing between two nucleotides is determined using the probabilistic method described in [91] which yields a symbolic classification of the possible hydrogen bonding patterns. For a given pair of residues the list of possible H-bonding patterns is considered, $\mathcal{H} = \{H_1, H_2, \ldots, H_n\}$. This list contains all pairing patterns involving two or three H-bonds defined by Saenger [126] and all the theoretical pairing patterns involving only one H-bond [53]. The base pair satisfies the pairing H_k if

$$p(H_k) > p(H_i), \quad H_i \in \mathcal{H} \text{ and } i \neq k$$

and $p(H_k) \geq c$, where c is a cutoff empirically fixed to 0.3 (see [91]). The probability of a possible H-bonding pattern, $p(H)$, is obtained by considering H to be the set of donor/acceptor pairs that should form H-bonds in a given base pairing pattern and $\overline{H}$ the set of H-bond that should not be formed.

$$p(H) = \Pi_{h \in H} p(h) \cdot \Pi_{h \in \overline{H}} (1 - p(h)) \tag{3.6}$$

In this equation, $p(h)$ represents the probability of a specific H-bond, h, defined by the distance between the donor and acceptor, the angle between the acceptor, the donor and the hydrogen and the angle between the donor, the acceptor and the lone electron pair direction on the acceptor. We obtain the probability of this H-bond by multiplying the probability associated to each parameter. The probability of a given parameter is evaluated by the following function:

$$p(x) = \begin{cases} 1 & if\, x < \mu \\ e^{-\left(\frac{\sqrt{-\log(0.5)}(x-\mu)}{\sigma}\right)^2} & otherwise \end{cases} \tag{3.7}$$

where the constants μ and σ are obtained empirically for each parameter (see [91]).

3.2.2.4 Base stacking

Vertical base stacking is a most significant stabilizing interaction between nucleotides in RNA and DNA three-dimensional structure. Base stacking is believed to play a significant role in the folding and complexation of RNA and DNA structure. It occurs both between adjacent and non-adjacent nucleotides mostly in helical stranded regions. Stabilization of base stacking is believed to involve both London dispersion forces [69] and interactions between partial charges within adjacent rings [128]. Evidences for hydrophobic forces between bases in solution [141] as well as a contradictory nonclassical hydrophobic effect [115] have been observed. However, all these interactions are not characterized well enough for us to use energy parameters in base stacking detection [137]. For this reason, a geometrical approach to stacking identification has been chosen based on the method proposed by Gabb *et al.* [52]. Relaxed values for the different parameters described in this method are used to allow for the large deviation from ideal atom positions found in the various structures of the databases. Indeed, it has been shown that there are many inconsistencies in atomic coordinates and base interactions of NMR and X-ray diffraction structures, due to variations in the refinement protocols, force field differences and artifacts resulting from the determination process [40, 149]. Two nucleotides are part of a stacking relation if the distance between one of their rings is less than 5.5 Å, if the angle between the two normals to the base planes is inferior to 30° and if the angle between the normal of one base plane and the vector between the center of the rings from the two bases is less than 40°.

3.2.3 Structural database

Nucleotide conformations and spatial relations discovered from the annotation process are stored in a database. Nucleotide conformations or spatial relations originating from new structures can then be compared to those of the database in order to evaluate their degree of peculiarity or to scan for close similarity.

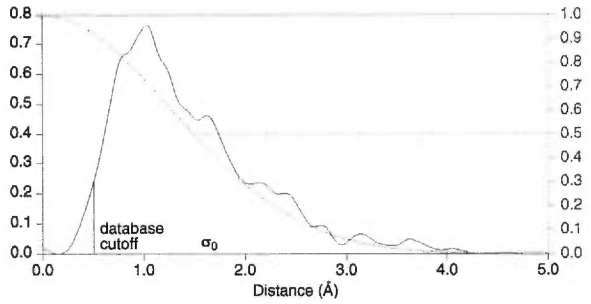


Figure 3.10. *Superimposition of the Gaussian distribution used to determine the peculiarity onto the distribution of distances between an idealized planar C•G Watson-Crick pairing and all other occurrences of that pairing in the database.*

The *degree of peculiarity* of a feature $\mathbf{V}$ is a measure, in the conformational space, of how scarce the space surrounding that particular feature is. A Gaussian distribution centered on each feature is used to evaluate the contribution of surrounding features to the peculiarity factor. This factor is given by:

$$P(\mathbf{V}) = 1 - \frac{1}{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{d(\mathbf{V},\mathbf{U}_i)}{\sigma}\right)^2} \tag{3.8}$$

where the standard deviation $\sigma = \frac{\sigma_0}{\sqrt{-2\log(0.5)}}$ determines the "size" of the Gaussian distribution, that is, the extent to which a distant feature $\mathbf{U}_i$ in the space of conformations contributes to the peculiarity at point $\mathbf{V}$. Figure 3.10 shows a one dimensional representation of the peculiarity factor calculation. Using this definition, we can identify new and original features as well as potentially faulty ones. This tool should allow for a faster way to probe novel RNA structures and determine regions of interest as well as to identify known structures containing similar features.

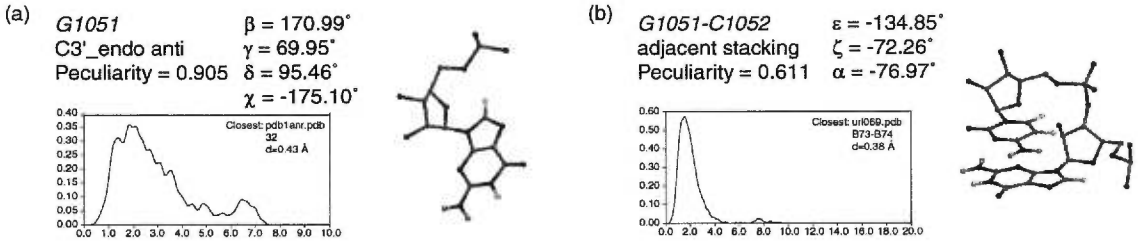


Figure 3.11. *Output of the annotation program. For each feature in the structural graph, the program generates: a symbolic description of the feature, a quantification of the relevant torsion angles, a distance distribution to the other features of the database, a peculiarity value and an identification of the closest feature. a) Characterization of nucleotide conformations. b) Characterization of spatial relations.*

3.3 Results

Our program *MC-Annotate* generates the structural graph of a RNA or DNA structure submitted as a PDB coordinate file. Figure 3.11 illustrates the various informations generated for each nucleotide and relation in the structural graph.

The database of conformational features used for comparison is built from 2434 structures currently originating either from the PDB [13], the NDB [12], or a limited number of theoretical and NMR structures obtained from personal contributions[2]. A distance cutoff of 0.5 Å is used for nucleotide conformations as well as spatial relations to avoid redundancy. The resulting database contains 12 239 nucleotides and 52 673 relations.

3.3.1 Analysis of ribosomal binding site of protein L11

As a test case for the proposed structure analysis procedure, we have selected the recently published crystal structure of the 58-nucleotide domain of the large subunit ribosomal RNA

[2]Some files are not used since they do not conform to the PDB file format specifications. Examples of faulty PDB files include files containing multiple models without the ENDMDL tag and files with misidentified or incomplete nucleotides.

(rRNA) of *E. coli* that forms a complex with ribosomal protein L11 (PDB accession number 1QA6) [33]. This highly conserved domain of rRNA having been identified as an important functional site makes it a preferred target for antibiotics. It is believed that part of the role of protein L11 is to maintain the unusual fold of the domain. The thorough description of the geometrical structure of this domain is therefore essential to the understanding of the interactions with protein L11 and other antibiotics. We propose here to validate our analysis method and provide further evidence to the unusual structural features reported by Conn *et al.* [33].

The results of the annotation procedure are shown in table I. This table shows the symbolic description of each nucleotide and spatial relation as well as the associated degrees of peculiarity; high values indicating high degrees of peculiarity. One can see that all the relationships originally predicted by Gutell [68] using comparative sequence analysis (CSA) on the large subunit rRNA are present in the crystal structure and were correctly identified by the program.

Figure 3.12a illustrates, in a graphical fashion, the degree of peculiarity of the different regions of the domain. The structure exhibits a large portion in which most of the relations have a high degree of peculiarity. This portion is located near the site of interaction with protein L11 which would indicate non ambiguously that there is a contribution of this protein to the stabilization of the rRNA domain conformational structure.

The RNA fold is centered around a four way junction loop between stem-loops A and C, helix B and loop B (see Figure 3.12b). In three-dimensional space, stem-loops A and C lie side by side whereas helix B and Loop B are extending in the opposite direction. This particular orientation of the different regions introduces strain in the junction loop as revealed by the relatively large deviation of the Watson-Crick base pairs A1057•U1081 and G1087•C1102 from their ideal conformations.

Many tertiary interactions between these four regions lock the domain in place to generate a large interior core of stacked and paired bases and to expose a surface of conserved bases available for molecular interactions. Among those is the unusual stacking relation

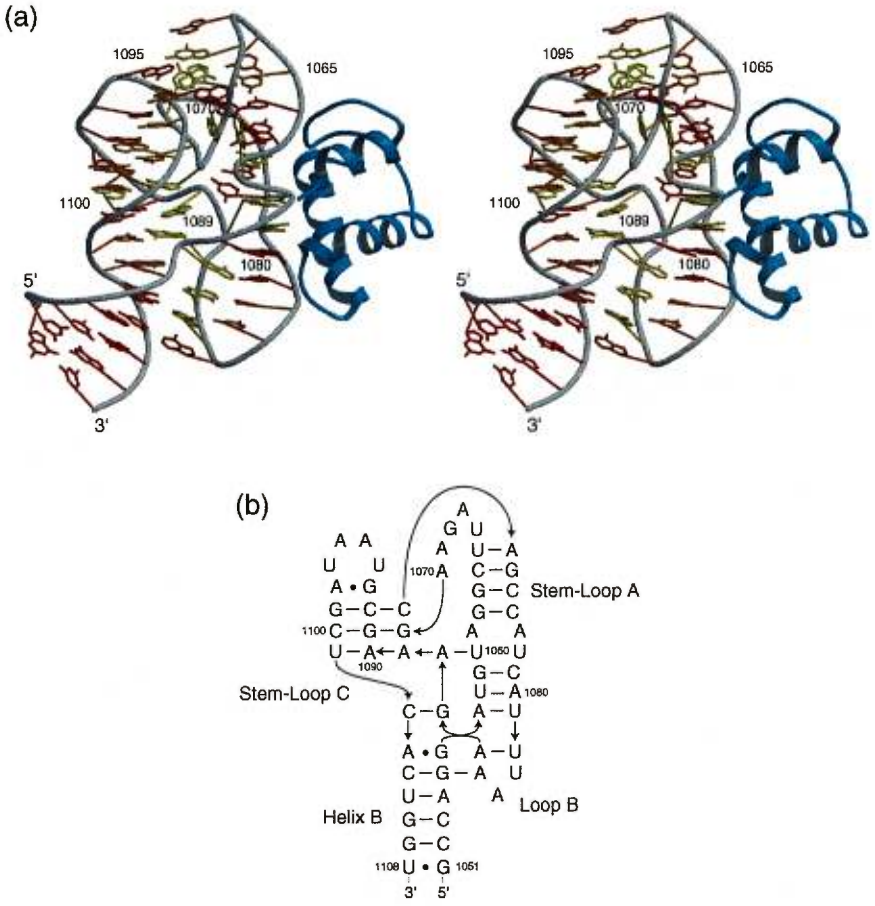


Figure 3.12. *a) Stereo view of the 58-nucleotide domain of rRNA that binds with ribosomal protein L11. The degree of peculiarity of the nucleotide relations is expressed by coloring the involved residues from red (low peculiarity) to yellow (high peculiarity). b) Secondary structure of the binding domain.*

arising in a parallel strand configuration between residues A1057 and A1086 (Fig. 3.13c), the latter being in an uncommon C3'-endo syn conformation, as seen in table I. Also, helix B and loop B are further stabilized by a peculiar one H-bond pairing of type GN2-AN3 between residues G1055 and A1085 (Fig. 3.13b) in which implicated bases are 99% conserved in sequences from the three phylogenetic domains (*Archaea*, *(eu)Bacteria* and *Eucarya*) [33].

Results show that stem-loops A and C interact with each other through various tertiary relations. Among those are two pairings involving nucleotide G1071 which forms a base triple with G1091 and C1100 via rather peculiar base pairings of type VI and CN4-GO6 respectively. This base triple is stacked between two other base triples showing uncommon relations. The first one involves the tertiary pairing of type CN4-GO6 between C1072 and G1099 and a standard Watson-Crick (type XIX) pairing between C1092 and G1099. The second one contains a Watson-Crick pairing (type XX) between U1101 and A1090, the latter being part, along with A1089, of the so called "adenosine platform" motif (Fig. 3.13g), which is an adjacent pairing of type AN6-AN3 that appears to be quite stable given its low peculiarity value. As mentioned by Conn *et al.*, the occurrence of this A-platform was reported by Cate *et al.* [21] at three locations in the crystal structure of *Tetrahymena thermophila* group I intron. Our program also identified a similar motif in the NMR structure of the binding site of small subunit rRNA with ribosomal protein S8 [81]. All these occurrences have been found to lie within a 2 Å distance, suggesting a high stability of the motif. We should note here that all stackings involved in this base triples complex exhibit normal peculiarity values except for G1071 and A1089 where the interaction arises between non-adjacent nucleotides of parallel strands.

This rRNA domain also contains three occurrences of the characteristic U-turn motif found in many RNA structures, a four nucleotide URRN loop that exhibits a hydrogen bond between the 5'U and the phosphate of the 3'N, a stacking between the two R and a very specific unstacked relationship that possesses a low variability for this type of relation. Indeed, 31 occurrences of this relation were identified in different PDB structures with a

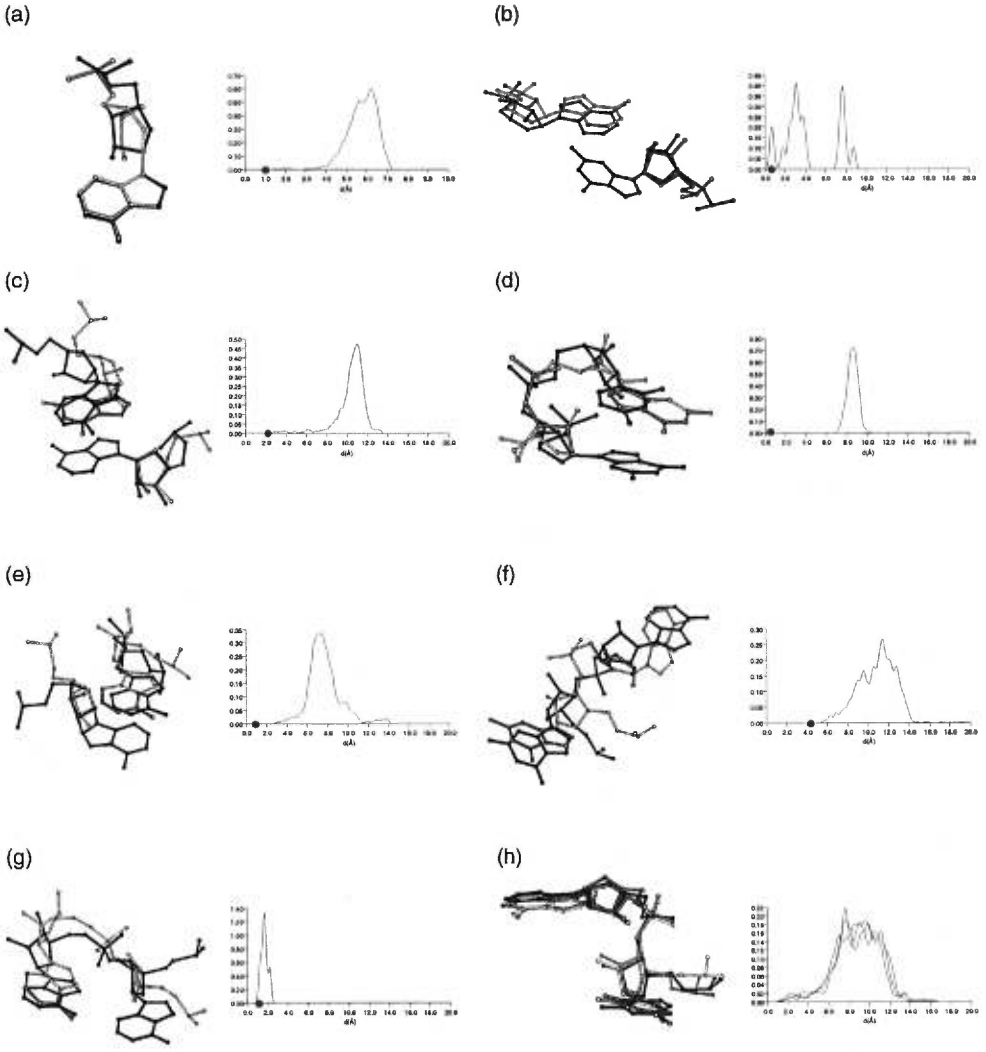


Figure 3.13. *Features found in L11 binding domain. Shown in figures a to g is the superposition of the feature from L11 binding domain with the closest feature found in the MC-Sym database. The distribution of distances between each feature and those of the database is also shown. A red dot indicates the distance to the feature used in the superposition. a) Residue A1088. b) Pairing G1055•A1085. c) Stack A1057-A1086. d) Adjacent stack G1059-U1060. e) Stack A1061-A1070. f) Adjacent A1087-A1088. g) Adjacent pairing A1089-A1090. h) The three occurrences of the U-turn relation found in L11 binding domain and their distribution of distances.*

relative distance of less than 4 Å. Also, the three unstacked relations observed in the domain have a relative distance below 2.7 Å even though they possess a high degree of peculiarity when compared to other adjacent relations (see Fig. 3.13h). They are positioned at the beginning of loop A (1066,1067), B (1083,1084) and C (1094,1095) and serve to reverse the chain direction. One of these relations causes loop A to be inserted deep into the structure, between helices A and C, where it interacts in unusual ways with the surrounding nucleotides. Conn *et al.* already observed that this is an unprecedented configuration for a hairpin loop. Our results indicate that it is effectively the case since five out of the eight adjacent relations exhibit highly peculiar conformations (> 0.9). In addition to the nucleotides G1071 and C1072 involved in two of the base triples described above, the highly conserved G1070 stacks with U1061 (Fig. 3.13e) in an uncommon way to stabilize the structure and expose it's Hoogsteen edge and the Watson-Crick edge of U1061. These two nucleotides, along with exposed A1067 and A1095 from the two adjacent U-turn, are suspected to interact with other components of the ribosome as well as some other molecules [33].

The surface of interaction between the rRNA and the protein L11 is also emphasized by the analysis of the structure. Many peculiar conformations are found in this region that can be attributed either to the electrostatic influence of protein L11 or to the conservation of an uncommon binding site in rRNA used in the recognition of the protein. The two hypotheses are certainly related since it has been shown that mutations at the binding site reduces L11 binding affinity [158] but that the interaction of L11 stabilize the entire rRNA domain [157].

At the core of the binding site are the conserved nucleotides G1059, U1060 and A1088 where the nearly invariant U1060 and A1088 interact in a well formed Hoogsteen base pair. Nucleotide A1088 (Fig. 3.13a) adopt however the unusual C3'-exo syn conformation. To accommodate for the insertion of A1088 into helix A via uncommon relations with G1087 (Fig. 3.13f) and A1089, adjacent relations show high degrees of peculiarity. In particular, residue A1061 bulges out of the helix in an unfamiliar way, where it stacks with G1070

as mentioned above, as does U1078. The resulting is an abrupt change in the direction of the helix axis. Finally, nucleotide U1060 interacts with G1059 in a reverse stacking configuration (Fig. 3.13d) which reverses the polarity of the backbone in the helix at the same time as it exposes it's major groove edge. A similar relation was found by our analysis in the NMR resolved AMP-RNA aptamer complex [77] as well as in the NMR structure of Rev-RRE complex [90] previously mentioned by Conn *et al.* This unusual conformation allows for the formation of a particularly reactive binding site with the exposition of many hydrogen bond acceptors.

In retrospect, one can see that most of the relevant features of the rRNA binding domain of protein L11 are related to uncommon nucleotide conformations and base-base interactions as determined by the annotation program.

3.3.2 Is the database complete?

Given the current content of the database of base-base interactions and nucleotide conformations, one might ask if it is complete, that is if the probing of new structures using X-ray crystallography, nuclear magnetic resonance, or any other experimental method is likely to add more information to the database. This is particularly crucial in the computer modeling program *MC-Sym* since the modeling process is limited by the diversity currently available in the database.

The *MC-Sym* program uses the database of nitrogen base relations extracted from the PDB structures to assemble new structures that satisfy constraints originating both from experimental data and theoretical hypotheses [104]. This approach to building models assumes that the current structural database is sufficiently complete at the level of nitrogen base relations to reproduce any given RNA structure. In practice, it is often difficult to evaluate if the impossibility of building a structure is due to the insufficiency of base relations data or simply to an inconstency in the constraints set. Therefore, it would be interesting to quantify the level of completeness of the base relations database.

The distance metric developed in section 3.2.2.2 can be used for such purpose. We assume that a given relation, M, is present in the database if the database contains a relation, N, such that $d(\mathrm{M}, \mathrm{N}) < c$, where c is a distance cutoff that was arbitrarily fixed to 1.75 Å. To evaluate the completeness of the database, we generated 3000 random single-stranded RNA structures using *MC-Sym*. Each of these starting structures was refined using a 15ps molecular dynamic simulation in which we reduced the temperature from 500K to 0K in the first 10ps. The program sander from the Amber 4.1 suite of programs was used with the Amber 94 force field [120]. All 1–4 electrostatic interactions were scaled by a factor of 1.2 as suggested in [120]. A distance-dependant dielectric model, $\epsilon = 4R_{ij}$, for the Coulombic representation of electrostatic interactions was used, as suggested by [120]. From this protocol, the resulting structures are significantly different from the starting structures since no equilibration was done on the starting structures. This property is of utmost importance since the goal of this protocol is to generate new base relations. All the adjacent relations were extracted from the generated structures and checked against the current database. From this evaluation, we can determine what is the percentage of the base relation space that is covered by the database. It was determined that the stack set has a coverage of 80%, but that the unstacked set has a coverage of only 20%. This observation is consistent with the fact that when building a model with *MC-Sym*, it is much easier to reproduce stacked relation than unstacked ones [92].

In order to increase the diversity of base relations present in the *MC-Sym* database, and thus the precision of the resulting models, we introduced the generated random structures in the database. We needed to evaluate how many random structures would be needed to obtain a complete database. As in the preceding section, all the relations were extracted from the random structures and checked against the database. But after being tested, the relations were inserted in the database. We would expect the coverage values to increase as we test new random structures and increase the size of the relation database. Figure 3.14 shows a running mean of the coverage value, showing that the introduction of base relations extracted from the random structures increases the coverage of both type of relations. In

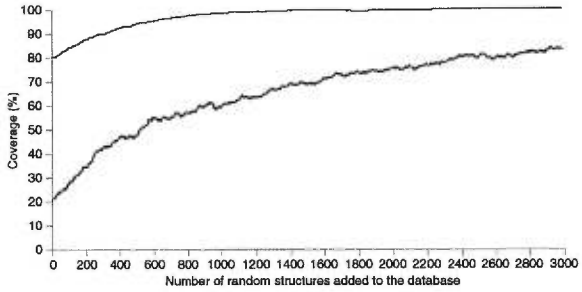


Figure 3.14. *Variation of the coverage of the conformational space with the addition of new random structures to the database. The upper line represents stacking relations whereas the lower line represents non-stacking interactions*

the case of the stack relation a coverage value of 99% was obtain after introducing 3000 random structures. For the unstack relation, a coverage value of 80% was obtained.

3.4 Discussion

The proposed annotation procedure should simplify the analysis of experimentally determined structures, as well as theoretical models. Using this approach, one can easily probe a RNA or DNA structure in order to qualify each nucleotide conformations or base-base interactions. A comparative approach, in which one evaluates the similarity or difference of each feature from many different occurrences of a given molecule, can also be considered.

We have shown that our structure analysis method yields considerable information on the marginal regions of RNA or DNA structures. As a matter of fact, it can direct attention to those regions that interact with other molecules or that are responsible for the stability of the particular fold of a structure. The analysis of L11 rRNA binding domain support these facts since many key interactions that were identified as structurally or functionnally important were revealed by our approach.

Also, the proposed approach should be useful in targeting conformational features that

are specific to a given structure or to a family of structure. Since RNA functions can be attributed to structural specificity, finding conformational features that deviates from the normally observed features should prove useful in identifying function related conformations and binding sites. In a similar way, finding highly conserved conformations, such as the U-turn motif present in L11 binding domain, could also result in function identification.

Many aspects of this work have implications for the *MC-Sym* molecular modeling program. First, the annotation procedure makes it possible to accurately and automatically build the database. The two distance metrics also allow to efficiently sample the conformational space of nucleic acids when building a new model. With the introduction of the refined random structures in the database, the modeling process should be improved as more of the true conformational space is represented in the database.

It should be noted that, for now, there exists a bias in the database towards standard A-form helices since most of the PDB and NDB structures consist of small and regular stems. The cutoff used to build the database reduces the bias by limiting the "density" of the conformational space. Nevertheless, with the ever increasing size of these nucleic acids repository and of the structure they contain, we should be able to eliminate this bias by adding more conformational diversity to the database of nucleotide conformations and base-base interactions. This should be attained shortly with the venue of rRNA crystal structures, which according to some authors [117] will multiply the size of the structure database by at least one order of magnitude.

Sequence

1051 GCCAGGAUGU AGGCUUAGAA GCAGCCAUCA UUUAAAGAAA GCGUAAUAGC
1101 UCACUGGU

Residue conformations

	Pucker Mode		Glycosyl		Peculiarity
	endo	exo	anti	syn	
G: 1051	C3'		x		0.919
C: 1052	C3'		x		0.885
C: 1053	C3'		x		0.886
A: 1054	C3'		x		0.874
G: 1055		C2'	x		0.892
G: 1056	C3'		x		0.904
A: 1057		C4'	x		0.872
U: 1058	C3'		x		0.850
G: 1059	C3'		x		0.887
U: 1060		C3'	x		0.902
A: 1061	C2'		x		0.936
G: 1062	C3'		x		0.924
G: 1063	C3'		x		0.895
C: 1064	C3'		x		0.815
U: 1065	C3'		x		0.852
U: 1066	C3'		x		0.857
A: 1067		C4'	x		0.875
G: 1068	C3'		x		0.928
A: 1069	C2'			x	0.993
A: 1070		C1'	x		0.950
G: 1071	C3'		x		0.940
C: 1072	C3'		x		0.827
A: 1073	C3'		x		0.906
G: 1074	C3'		x		0.878
C: 1075	C3'		x		0.862
C: 1076	C3'		x		0.873
A: 1077	C3'		x		0.870
U: 1078	C3'		x		0.857
C: 1079	C3'		x		0.849
A: 1080	C3'		x		0.858
U: 1081	C3'		x		0.843
U: 1082		C4'	x		0.842
U: 1083		C2'	x		0.874
A: 1084	C3'		x		0.867
A: 1085	C3'		x		0.861
A: 1086	C3'			x	0.997
G: 1087	C3'		x		0.980
A: 1088		C3'		x	0.998
A: 1089	C2'		x		0.920
A: 1090	C3'		x		0.880
G: 1091		C2'	x		0.887
C: 1092	C3'		x		0.822
G: 1093	C3'		x		0.900
U: 1094	C3'		x		0.882
A: 1095	C3'		x		0.863
A: 1096	C3'		x		0.912
U: 1097	C3'		x		0.931
A: 1098	C3'		x		0.883
G: 1099	C3'		x		0.893
C: 1100	C3'		x		0.818
U: 1101	C3'		x		0.856
C: 1102	C3'		x		0.874
A: 1103	C3'		x		0.895
C: 1104		C2'	x		0.847
U: 1105	C3'		x		0.897
G: 1106	C3'		x		0.880
G: 1107	C3'		x		0.878
U: 1108	C3'		x		0.891

Adjacent relations

	Stacking	Pairing	Peculiarity
G:1051 C:1052	x		0.615
C:1052 C:1053	x		0.723
C:1053 A:1054	x		0.687
A:1054 G:1055	x		0.736
G:1055 G:1056	x		0.859
G:1056 A:1057			0.933
A:1057 U:1058	x		0.704
U:1058 G:1059	x		0.757
G:1059 U:1060	x		0.999
U:1060 A:1061			0.982
A:1061 G:1062			0.957
G:1062 G:1063	x		0.683
G:1063 C:1064	x		0.719
C:1064 U:1065	x		0.793
U:1065 U:1066	x		0.724
U:1066 A:1067			0.940
A:1067 G:1068	x		0.806
G:1068 A:1069	x		0.969
A:1069 A:1070			0.931
A:1070 G:1071			0.964
G:1071 C:1072	x		0.632
C:1072 A:1073			0.966
A:1073 G:1074	x		0.766
G:1074 C:1075	x		0.662
C:1075 C:1076	x		0.706
C:1076 A:1077	x		0.853
A:1077 U:1078	x		0.788
U:1078 C:1079			0.898
C:1079 A:1080	x		0.753
A:1080 U:1081	x		0.730
U:1081 U:1082	x		0.611
U:1082 U:1083	x		0.927
U:1083 A:1084			0.933
A:1084 A:1085	x		0.740
A:1085 A:1086	x		0.999
A:1086 G:1087			0.965
G:1087 A:1088			0.986
A:1088 A:1089			0.942
A:1089 A:1090		AN6–AN3	0.646
A:1090 G:1091	x		0.780
G:1091 C:1092	x		0.723
C:1092 G:1093	x		0.879
G:1093 U:1094	x		0.821
U:1094 A:1095			0.929
A:1095 A:1096	x		0.745
A:1096 U:1097	x		0.942
U:1097 A:1098	x		0.981
A:1098 G:1099	x		0.785
G:1099 C:1100	x		0.645
C:1100 U:1101	x		0.716
U:1101 C:1102	x		0.771
C:1102 A:1103	x		0.687
A:1103 C:1104	x		0.843
C:1104 U:1105	x		0.740
U:1105 G:1106	x		0.778
G:1106 G:1107	x		0.717
G:1107 U:1108	x		0.719

Non-Adjacent relations

	Stacking	Pairing	Peculiarity
G:1051 U:1108		XXVIII	0.717
C:1052 G:1107		XIX	0.634
C:1053 G:1106		XIX	0.619
A:1054 U:1105		XX	0.596
A:1054 G:1106	x		0.681
G:1055 A:1085		GN2–AN3	0.832
G:1055 C:1104		XIX	0.703
G:1056 A:1103		XI	0.514
A:1057 U:1081		XX	0.724
A:1057 A:1086	x		0.997
U:1058 A:1080		XX	0.551
G:1059 C:1079		XIX	0.646
G:1059 A:1080	x		0.743
U:1060 A:1088		XXIII	0.576
A:1061 A:1070	x		0.995
G:1062 C:1076		XIX	0.745
G:1062 A:1077	x		0.770
G:1063 C:1075		XIX	0.633
G:1063 C:1076	x		0.800
C:1064 G:1074		XIX	0.641
U:1065 A:1073		AN6–UO2	0.863
A:1069 A:1073	x		0.999
G:1071 A:1089	x		0.999
G:1071 G:1091		VI	0.907
G:1071 C:1100		CN4–GO6	0.895
C:1072 G:1099		CN4–GO6	0.878
C:1079 A:1088	x		0.975
U:1082 A:1086		XXI	0.394
G:1087 A:1089	x		0.998
G:1087 C:1102		XIX	0.723
G:1087 A:1103	x		0.885
A:1090 U:1101		XX	0.691
A:1090 C:1102	x		0.921
G:1091 C:1100		XIX	0.624
G:1091 U:1101	x		0.884
C:1092 G:1099		XIX	0.615
G:1093 A:1098		XI	0.618

Table I. *Screen shot of the annotation results of L11 ribosomal binding domain.*

Chapitre 4

Comparative structural analysis of nucleic acids

P. Gendron[1], G. Poisson[1], D. Gautheret[2] and F. Major[1]

Abstract

Drug design and phylogenetic analysis of RNA structures can greatly benefit from the study of recurrent nucleotide arrangements, or motifs, since these regions, often conserved through evolution, are generally responsible for the interaction of RNAs with other molecules, be they natural proteins or engineered therapeutic agents. We propose a set of tools and a working schema that allow the identification of conserved motifs in RNA or DNA secondary (2-D) and tertiary (3-D) structure. Using graph isomorphism algorithms and the symbolic representation of 2-D and 3-D structures in the form of structural graphs, these tools make it possible to search for specific motifs, to discover novel motifs, and to compare molecules on a structural basis. Results obtained for the 5S rRNA and 16S rRNA show a large number of common regions among structures of the currently accepted phylogenetic branches (*Archaea*, *(eu)Bacteria* and *Eukarya*). Also, some structural similarity is observed between phylogenetic branches, which allows for the evaluation of the genetic proximity of the individuals of these branches.

[1]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7

[2]CNRS, 31 Chemin J. Aiguier, 13 402 Marseille, Cedex 20, France

4.1 Introduction

RNA's catalytic abilities, first reported by Kruger [84] and Guerrier-Takada [64], opened a new avenue of research in the field of RNA structural determination. This finding led others, like Batey [8], to suggest that RNAs could be compared with proteins. It has long been known that particular protein structural motifs are required for specific functions such as catalytic processes. In light of the finding that RNA also possessed catalytic capabilities, research intensified in recent years to identify conserved structural motifs in RNA that may be linked to precise biological functions.

Structural motifs can be described as a combination of a consensus sequence and the bonds that link each nucleotide in a specific order and conformation. They are classified into two different types [113]. The first type consists of secondary structure (2-D) motifs represented, for example, by helices, bulges and hairpin loops. The second type consists of tertiary structure (3-D) motifs described by the interactions between regions of 2-D structures. A good example of a 3-D motif is the tetraloop-helix interaction that is found in the group I intron.

Finding and analyzing motifs is a crucial step in the process of identifying new drug targets in RNA. Such studies are also important for more theoretical research like the analysis of precise phylogenetic domains. With the rapid growth of nucleic acids databases, some tools have been developed to help search for functional motifs. At present time, the most used tool for this purpose is *RNAMOT* [55]. *RNAMOT* can easily perform searches for structural motifs in RNA sequences. This method offers the advantage of searching RNA sequences for known motifs, but it is also limited by the lack of structural information contained in the linear sequences used as inputs. In fact, the use of *RNAMOT* implies previous knowledge of the motif's secondary structure. *RNAMOT* offers a good means through which known RNA motifs can be located in sequences, but generates a large number of false positive results and can not be used to locate new motifs for which no secondary structure is known. Consequently, knowledge of the RNA's secondary structure included

in the search can greatly accelerate motif discovery and eliminate all false positive results generated by tools that only rely on sequence information.

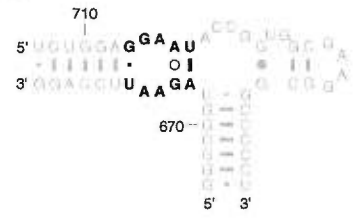
Some tools have been developed to take advantage of the knowledge of a sequence's secondary structure. The program *ESSA* [29] allows a user to visualize the arrangement of stretch of sequences in drawings of secondary structures. Also, the program *Palingol* [14] can search for motifs in the secondary structure of RNA molecules. However, like *RNAMOT*, it can only look for known motifs. Moreover, *Palingol* is unable to locate motifs in *true* three-dimensional space. The ability to locate RNA motifs in three-dimensional space is extremely interesting in light of the need to identify new RNA drug targets. Once a motif is located in three-dimensional space in a specific molecule, its properties (i.e., solvent accessibility, overall shape, environment, etc) can be evaluated.

Here we introduce a package of programs used to search for and analyze conserved motifs in RNA secondary and tertiary structure. This package contains tools for the manual search of known motifs and for the automatic identification of motifs that are conserved and specific to a set of structures. We also present many results on the motif search in secondary as well as tertiary structures, the identification of motifs in the secondary structure of 5S and 16S rRNA of the *Archaea*, *(eu)Bacteria* and *Eukarya* phylogenetic domains, and the comparative analysis of the structural similarity of these domains.

4.2 Methods

Comparative structural analysis is defined here as being the study of RNA secondary or tertiary structure in order to bring to light the different conserved regions that are hopefully responsible for the activity of the given RNAs. Structural motifs, i.e., frequently occuring arrangements of nucleotides, are representatives of these regions. For reviews of motifs as structurally functional components of RNA structures, see [8, 113]. This section describes the different methods used to search for and evaluate conserved motifs as well as domain specific ones.

a)



b)

```
 10 ENERGY = -xxx.xx filename
673 A       0   674   717   673
674 G     673   675   716   674
675 A     674   676     0   675
676 A     675   677     0   676
677 U     676     0   713   677
713 G       0   714   677   713
714 G     713   715     0   714
715 A     714   716     0   715
716 A     715   717   674   716
717 U     716     0   673   717
```

c)

```
sequence( r 674 AGAAU )
sequence( r 714 GGAAU )
adjacent(
  674 677 { connect }
)
non_adjacent(
  673 717 { pairing }
  674 716 { pairing }
  677 713 { pairing }
)
```

Figure 4.1. *a) Part of helices 22-23 of E. coli in which a subgraph is highlighted, b) ct file and c) MC-Sym script.*

4.2.1 Structural graph of relations

A symbolic description using graph-theoretical techniques is used to encode RNA secondary and tertiary structural information and to allow for the detection and extraction of structural motifs. This particular choice of representation makes it possible to, not only apply the elaborated methods on secondary structures, but also to consider three-dimensional ones when combined with a structural annotation program such as the one described in [58]. Hence, the different methods can readily be applied to secondary structures expressed as connect table ("ct") files [160], *MC-Sym* input scripts [85], which allow for 3-D information, or any other format that can be reduced to one of these encodings. Figure 4.1 illustrates the use of these formats to encode parts of a secondary structure.

A secondary (or tertiary) structure is represented by a labeled structural graph $G = (V, E)$ where V is a set of vertices representing nucleotides, which can be of any of the types A,C,G,U or T or the IUPAC/IUBMB wild card types Y,R,N,W,S,M,K,B,D,H or V [76]. Optionally, conformational properties may be included in the description of nucleotides when dealing with 3-D structures. A set E of edges describes pairwise nucleotide interactions as a set of relational properties. The *pairing* property refers to the presence of hydrogen bonds between two nucleotides' nitrogen bases regardless of the fact that the actual pairing occurs as a secondary or tertiary interaction. The type of pairing may also

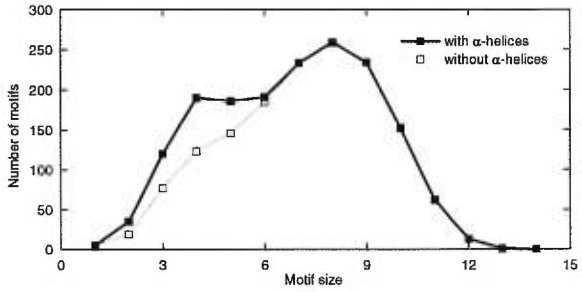


Figure 4.2. *Variation in the number of motifs found in at least 50% of 5S (eu)Bacteria while keeping or not the A-form helices. This shows that there are no conserved A-form helices of size larger than 6 nucleotides.*

be included following the nomenclatures of Saenger [126] and Gautheret [53]. The *adjacent* property is used to qualify relations between nucleotides linked by phosphodiester bonds and can be associated with the *helix* property when the relations lie within helical regions (defined here as consecutive stretches of at least three Watson-Crick or G•U wobble pairings). The *stacking* property can also be used to define more precisely the interactions of adjacent nucleotides in a three-dimensional context as well as describe long-range non-adjacent base stackings.

4.2.2 Formal definition of a motif

In this context, a motif is defined as a recurrent structural subgraph arising in one or many RNA structures. A recurrent subgraph is declared as being a motif if it's significance is higher than a predetermined value. Depending on the desired objective of the study, three definitions of significance can be considered to evaluate motifs [57], that is, first, the absolute number of occurrences, second, the proportion of explored RNA structures in which the subgraph is found, and last, the inverse probability of occurrence of the motif based on a statistical study of the probability of occurrence of each constituents of the motif. Comparative structural analysis of a set of structure mostly benefit from the second

definition whereas the first and third definitions tend to produce motifs that are non domain-specific or too peculiar. In all cases, motifs found exclusively within standard A-form helices should not be considered as such as mentioned in [113]. Figure 4.2 illustrates that A-form helices are not really motifs since they serve no direct biological function and hence were not conserved through evolution. They are instead responsible for the stabilization of the structures' folds which are preserved even if there are nucleotide covariations.

The recognition of motifs is based on a method of classification of the subgraphs found in the RNA structures by means of a graph isomorphism algorithm similar to the one described in [143]. Given two of these subgraphs, $G_A = (V_A, E_A)$ and $G_B = (V_B, E_B)$, the algorithm determines if there exists a one-to-one and onto mapping function, f, from the nucleotides V_A to V_B such that each relation $(i, j) \in E_A$ exists if and only if the corresponding relation $(f(i), f(j)) \in E_B$ is also present. Even though the isomorphism problem can be solved in polynomial time for a few restricted classes of graphs [26, 74], in the general case, no sub-exponential algorithm has been found to solve it. In the case of RNA secondary or tertiary structure, however, the use of labeled vertices and edges significantly decreases the running time of the combinatorial enumeration of the algorithm even though the complexity remains the same [3, 57].

4.2.3 Motif identification

The general philosophy behind the search for motifs uses an incremental exploration of the RNA structures to extract all present subgraphs. The subgraph identification procedure relies on the assumption that a subgraph containing n nucleotides, where $n > 1$, can be recursively defined as a subgraph of size $n - 1$ to which a nucleotide as been added based on the graph of the originating structure. Let a *promising subgraph* of size n be a subgraph satisfying some condition that will allow it to be considered for enlargement to size $n + 1$. This concept can be used in at least three different contexts: the construction of a motif database, the comparison of motif databases and the search of a specific motif.

4.2.4 Database construction

The construction of a database is used to identify motifs that are common to a set $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$ of RNA structures from their respective organisms. Since these structural motifs are conserved through evolution, their identification theoretically points to important functional sites as well as potential RNA drug targets. A database of motifs containing n nucleotides is incrementally built from the databases of motifs of inferior size using a significance criteria that must be satisfied at each step. This particular design permits one to stop (willingly or not) the exploration during step $i < n$ and resume the construction using the database created at step $j = i - 1$ up to size $j = n$. This appears most useful when building large databases for which the program might run for many days and be vulnerable to computer failure.

Another important aspect of the building of motif databases is the identification of the maximal common subgraph; that is we want to eliminate from the database of size i, all motifs that are subgraphs to motifs in the database of size $i + 1$. To accomplish this task, a reference counting technique is used that preserves a motif of size i if no motifs of size $i+1$ refers to it. Using this technique, a motif of size i is found if at least one of it's occurrence cannot be expanded to a higher size while respecting the significance tolerance.

4.2.5 Structural comparison

In this procedure, an existing motif database (containing non-maximal common subgraphs) from a set $\mathcal{S}_1$ of RNA structures is considered. New occurrences of motifs are added to the existing database from a set $\mathcal{S}_2$ of structures. The comparison works in an incremental fashion, adding and extending subgraphs of size i from the new set of structures if they were found promising, that is isomorphic to existing motifs of set $\mathcal{S}_1$. The result is a database indicating the relative significance of motifs depending of the various subsets of structures. Potential RNA drug targets found in a subset of structures can then be verified as being specific or not to that subset.

4.2.6 Specific motif search

In some instances, one might be interested in looking for a particular motif in a set of secondary or tertiary structures. Examples of this would be to identify occurrences of a fragment to be reused in three-dimensional modeling or to verify that a particular conserved motif is specific to the set of structures in which it was originally found. The search of a specific pattern, the query pattern, containing n nucleotides in a secondary structure of m nucleotides, $m > n$, proceeds in the following manner. First, a nucleotide is selected from the query pattern and a list of occurrences of that nucleotide in the secondary structure is built. Then, at each step of an incremental search, a promising subgraph, that is one that occurs both in the query pattern and in the structure, is selected and all it's occurrences are extended. If the promising subgraphs contain n nucleotides, the search succeeds, whereas if no promising subgraph exists, the search fails. Contrary to the database construction and comparison procedures, subgraphs matching the query pattern may contain more relations than the one specified. Also, it is possible to query for patterns that contain the wild card types and for relations that contain only a subset of the properties of those in the actual structure.

4.3 Results and discussion

As a first test case for the proposed methodology, we searched for some specific motifs in secondary and tertiary structures and successfully identified all occurrences of these motifs. First, we looked for all tetraloops in the secondary structures of 269 16S rRNA [68] using the query patterns illustrated in figure 4.3a. By comparison, there were 1538 GNRA and 251 UNCG tetraloops found which represents only 5% of the 256 different possible types of tetraloops, but 60% of the 3022 observed tetraloops (NNNN). This is representative of the already noted stability of these two consensus tetraloops [155].

Then, we performed the search of a more complex, but otherwise well recognized motif, which has been found to be an important part of the structural stability of the three-

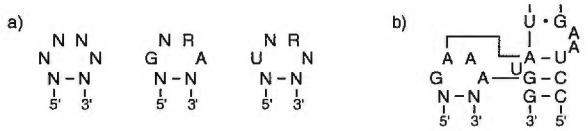


Figure 4.3. *Query pattern for various known motifs. a) Tetraloop. b) GAAA tetraloop-helix interaction.*

dimensional folding of group I introns. This motif, which is illustrated in figure 4.3b, was searched in all Protein Data Bank [13] files so as to demonstrate the ability of our procedure to look for tertiary motifs. Only two occurrences were found, in the P4-P6 domain of the *Tetrahymena thermophila* group I intron (PDB accession number 1GID), which confirms the results of [105].

4.3.1 Structural analysis of 5S ribosomal RNA

Secondary structures of the 5S rRNA from three phylogenetic domains were compared using the methods described previously in order to identify motifs and evaluate the similarity or differences characteristic to each domain. Structures originate from the Gutell database [68] and consist of the 5S rRNA of 11 *Archaea*, 28 *(eu)Bacteria* and 29 *Eukarya*. These structures, which were determined by comparative sequence analysis (CSA) [109], possess 120 nucleotides on average and represent, as a consequence of their small size, one of the most studied RNA molecules. Thus, we can assume that these 2-D structures are well determined.

First, a database of conserved motifs from each domain was built using a significance tolerance of 55% on the proportion of structures in which isomorphic subgraphs must be present. This means that we accept motifs that appear only in a given portion of the structures since perfect conservation of motifs would be too restrictive. Figure 4.4a shows the effect of varying the accepted significance on the quantity of motifs identified. Also, Figure 4.4b illustrates the resulting impact on computation time. The combinatorial nature of

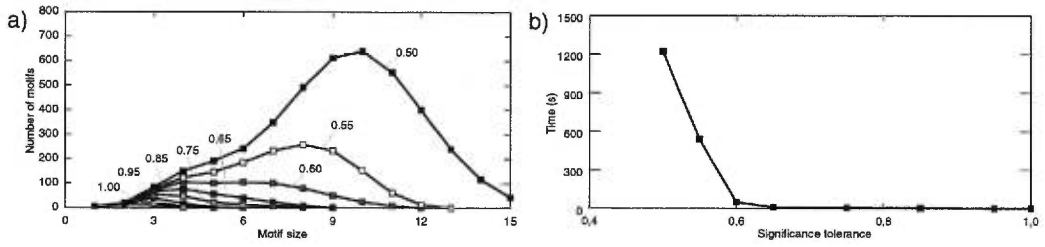


Figure 4.4. *Influence of using different significance tolerance values on a) the number of motifs found and b) the execution time needed to find the motifs in the 28 5S (eu)Bacteria on a 600MHz Pentium III processor.*

the graph isomorphism algorithm, coupled with the large number of individual subgraphs present in secondary structures, is clearly demonstrated by the increase in running time.

By manual inspection of the resulting databases, several interesting motifs were found to be conserved in each domains. These motifs are illustrated in their originating secondary structures in figure 4.5. Since they are preserved in a certain portion of the structures from each domain, they are generally related to important biological functions as they interact in a similar way with proteins or other RNA molecules.

In *Archaea* (see figure 4.5a), motifs a, b and i were found to occur in 73%, 82% and 82% of the structures respectively. Motif a is the well known loop E motif that has already been observed and extensively studied [93] for it is responsible for many interactions with proteins and other parts of the ribosome. Motifs b and i, being highly conserved in this domain, are potentially good candidates in drug design, though they have not yet been thoroughly studied.

In *(eu)Bacteria* (see figure 4.5b), the most recurrent motifs found (d and e) are located in parts of loop C of the secondary structure. This loop, along with the dual-A bulge of helix III (motif i) that occurs in 57% of the structures, was found to be a major anchor region in the interaction of *E. coli* 5S rRNA with *X. leavis* ribosomal protein L5 [130]. Using the specific motif search described in the methods and taking as input the whole loop in which

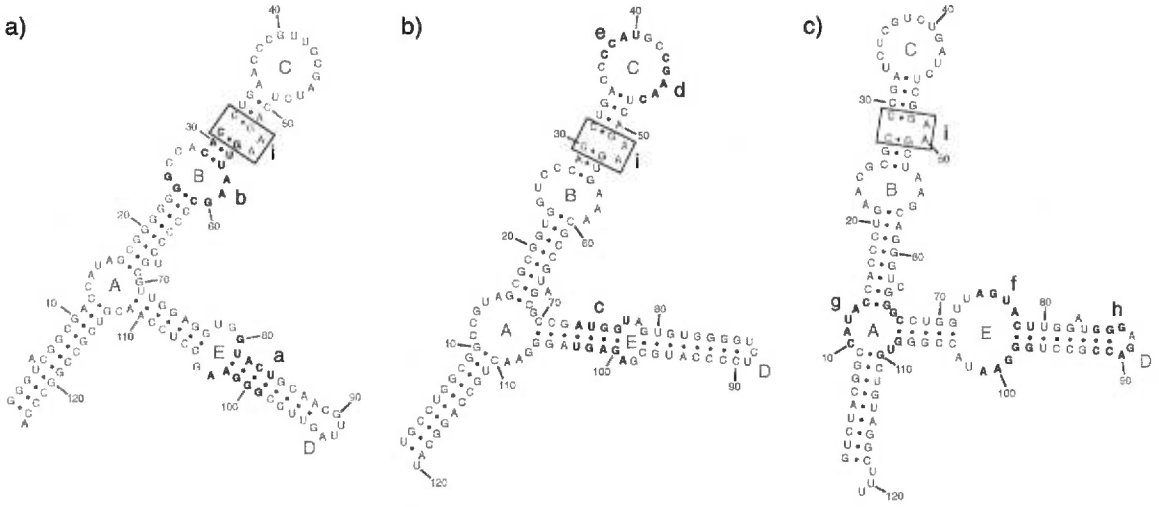


Figure 4.5. *Conserved motifs in a) Archaea Methanothermus fervidus, b) (eu)Bacteria Escherichia coli and c) Eukarya Homo sapiens.*

the wild card nucleotide N was used in lieu of those absent from motifs d and e, we found this particular conformation in 89% of the *(eu)Bacteria.* This illustrates one possible use of the specific search, i.e. to further explore the local environment of the motifs found by building a database of recurrent motifs. The loop E motif (motif c) that binds to protein L25 [100, 138] was also found with a somewhat smaller proportion of occurrence (57%). Thus, even though this motif is important in binding ribosomal proteins, some sequence or variation occurs among structures of the *(eu)Bacteria* domain which would indicate polymorphism in protein recognition.

The *Eukarya* domain conceals some interesting motifs as can be seen in figure 4.5c. The loop E motif (motif f) is particularly well conserved in all the domain with a proportion of occurrence of 97%. This region is already known to be part of the binding site of transcription factor TFIIIA and ribosomal protein L5 [1]. The three way junction loop (motif g) is also highly conserved, arising in 76% of the secondary structures. The specific search was again used to determine the identity of the nucleotide at position 9. An input

Motif	*Archaea*	*(eu)Bacteria*	*Eukarya*
a	0.73	0.00	0.00
b	0.82	0.43	0.03
c	0.00	0.57	0.00
d	0.73	1.00	0.17
e	0.55	0.93	0.86
f	0.00	0.04	0.97
g	0.00	0.00	0.76
h	0.00	0.04	0.76
i	0.82	0.57	0.79

Table II. *Proportion of occurrence in each of the three 5S domains of the motifs shown in figure 4.5.*

script containing 9N with the rest of the junction was used to show an 82% occurrence of cytosine and a 18% occurrence of uracyl at this position, a pyrimidine consensus. The C9•G110 Watson-Crick pairing possibly mutated in a U9•G110 in some organism while preserving the associated function. The incomplete loop D (motif h) was found in 76% of the *Eukarya* structures. The missing position 88 was probed by a specific search of the motif and found to contain mostly purines, that is 31% of adenine, 28% of guanine, 7% of cytosine, and 10% of uracyl.

Second, as a complement to this manual inspection of the database, the structural comparison procedure was used to evaluate, within the conserved motifs, those that are domain-specific. Most of the identified motifs appear to be somewhat specific to their domain, as shown in table II, with the exception of the dual-A bulge (motif i), which was found in 66% of all structures, and of the parts of Loop C (motifs d and e). This observation suggests that the other domain-specific motifs could be candidates as targets in the elaboration of new drugs. However, experiments would be necessary to confirm this claim since our comparison method is sensitive to the validity of the secondary structure. In the case of the 5S and 16S databases of secondary structures [68], which were determined by CSA, their is an obvious bias introduced by the determination method toward the conservation of motifs among phylogenetic domains. This does not necessarily reflect reality as some nucleotide interactions might be wrongly assigned or falsely rejected. This is the case for

the loop E region found in *Eukarya* which was experimentally found to include more pairings [140, 152] than those proposed originally and used here. In fact, the addition of these pairings causes the loop E from *Eukarya* to be isomorphic to the one from *Archaea* (but not to the loop E from *(eu)Bacteria*).

The process of establishing conserved functional regions of RNA structures or potential drug targets should not be restricted to the analysis of the secondary structure. In fact, the definition of motifs should include three-dimensional structural criteria that are not encoded in the secondary structure. As mentioned in the methods section, these should include the conformation of the sugar-phosphate backbone, the H-bonding patterns between nucleotides, and the geometry of the stacking, as well as the charge distribution of the solvent-accessible surface.

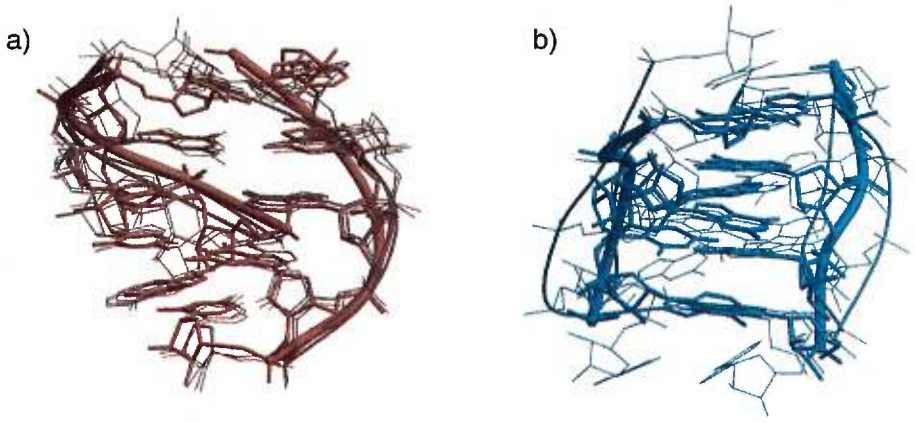


Figure 4.6. *RMS deviation alignment of structural motifs found in the Protein Data Bank. a) Loop E from (eu)Bacteria found in files 1A4D, 1A51, 1D6K, 1DFU, 354D and 364D. b) Loop E from Eukarya found in files 1CO4, 1RRN, 1SCL and 430D.*

Since the loop E motifs from both *(eu)Bacteria* and *Eukarya* are rather conserved and well studied, both loops were specifically searched in the Protein Data Bank using the annotation program described in [58] to generate the symbolic representation of the structures.

Six occurrences of the *(eu)Bacteria* loop were found whereas four *Eukarya* loops were identified. Figure 4.6 shows a superposition of the resulting structures using a minimization of the RMS deviation. All structures were found to lie within 1.61 Å of RMS deviation for *(eu)Bacteria* attesting to the relative stability of the motif. For *Eukarya* however, the structures lie within 4.44 Å of RMS deviation.

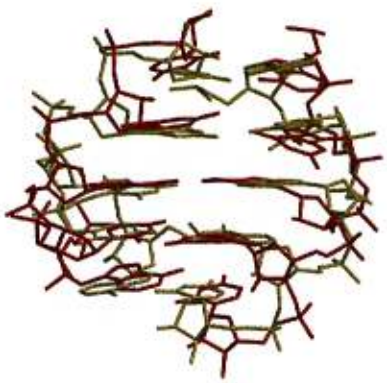


Figure 4.7. *Proposed model for the loop E of (eu)Bacteria (in red) showing a RMS deviation of 1.85 Å with the corresponding motif found in PDB file 354D (in yellow).*

If no occurrences of a given motif were found by our search in PDB files, the modeling program *MC-Sym* [104] could make it possible to generate 3-D models that would allow the thorough study of this potential target and functional site. For example, models were generated for the loop E of *(eu)Bacteria* and refined using the program *CHARMm* [17], and the *CHARMm* 27 forcefield parameters [48]. Ten classes of structures were found for the 347 structures generated. Interestingly, one of these classes contains structures with a RMS deviation of 1.85 Å (nitrogen base atoms only) with the corresponding structures found in PDB. Figure 4.7 shows the superimposition of the structure that is most similar to the ones in PDB. One can see from this example that the 3-D modeling can give more information on the structural properties of motifs that were found in secondary structures.

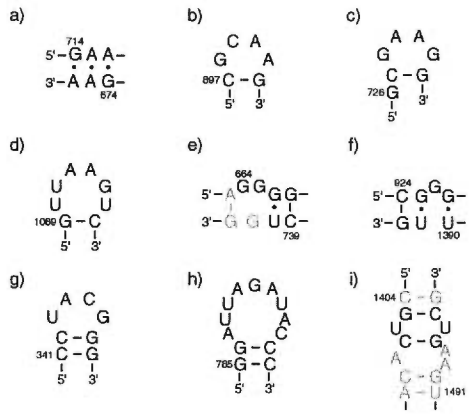


Figure 4.8. *Conserved motifs in the 16S structures of the (eu)Bacteria phylogenetic domain. Nucleotide numbering follows the one from Escherichia coli.*

4.3.2 Structural analysis of 16S rRNA

In a way similar to the analysis of the 5S rRNA, databases were built for each phylogenetic domains of 16S rRNA comprising 20 *Archaea*, 49 *(eu)Bacteria*, and 56 *Eukarya*. However, these structures containing between 1244 and 2741 nucleotides, a significantly larger number of motifs was found in a proportionally longer running time. Consequently, only motifs occuring in at least 75% of the structures of a given domain were conserved at each step of the incremental construction of the databases of motifs of up to 6 nucleotides. Then, the significance tolerance was augmented to 90%, and motifs containing up to 13 nucleotides were found. The total running time was of the order of 12 hours to complete the construction on a 600MHz Pentium III processor. Given the amount of data generated, and for the sake of argument, we present some of the motifs resulting from the comparison of the *(eu)Bacteria* domain relative to the other two domains. Figure 4.8 illustrates some of the most interesting motifs that were identified. It is interesting to note that they consist mostly of hairpin loops, bulges and helices formed of non-canonical pairings. Furthermore, they are also very specific to the *(eu)Bacteria*, being found in smaller proportions in *Archaea*

Motif	*(eu)Bacteria*	*Eukarya*	*Archaea*
a	0.96	0.02	0.00
b	0.98	0.71	0.90
c	0.80	0.09	0.55
d	0.94	0.00	0.70
e	0.96	0.20	0.85
f	0.98	0.00	0.85
g	0.98	0.00	1.00
h	1.00	0.04	0.95
i	0.76	0.82	0.75

Table III. *Conserved motifs in the 16S structures of the* (eu)Bacteria *phylogenetic domain compared to the* Eukarya *and* Archaea *domains.*

and almost never in *Eukarya* (see table III).

In particular, the motif shown in figure 4.8i is part of the A-site of 16S rRNA that contacts the codon-anticodon complex to modulate the fidelity of translation [116]. Since this is an important binding site of the ribosomal RNA, we searched for the larger 13 nucleotide part from *E. coli* shown in the figure to evaluate the occurrence in each domain. The result is a drastic decrease in occurrence of the motif in both the *Archaea* and *Eukarya* domains where it goes down to 15% and 0% respectively. By contrast, it is still observed in 60% of the *Bacteria* confirming the already known fact that it is a good target for antibiotics like neomycin and paromomycin. By indicating that the motif specificity occurs in regions C1407 to G1409 and G1491 to G1494, our study gives more evidence to support the mutagenesis analysis [111, 124] that reveal the importance of these specific nucleotides in the interactions with the antibiotics.

4.3.3 Inter-domain relationship evaluation

The database construction was used, with a significance tolerance of 75% for the 5S and 100% for the 16S in order to evaluate the variation in the number of conserved motifs in each domain. It is interesting to note that the number found in *Archaea* shows a faster increase compared to both rRNA of *(eu)Bacteria* and *Eukarya* (figures 4.9a and 4.10a).

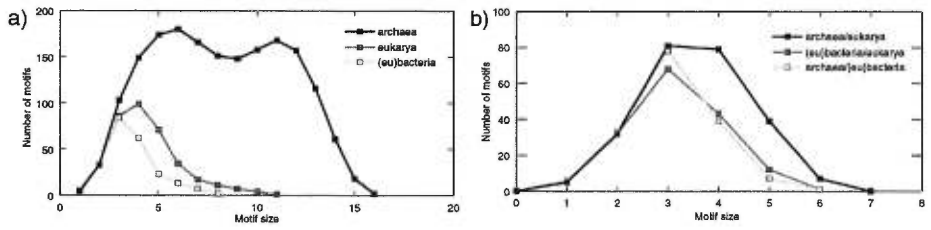


Figure 4.9. *a) Variation of the number of motifs found for the different phylogenetic domain of 5S rRNA using a significance tolerance of 0.75. b) Number of motifs found for different subset of phylogenetic domain showing the structural proximity of the domains.*

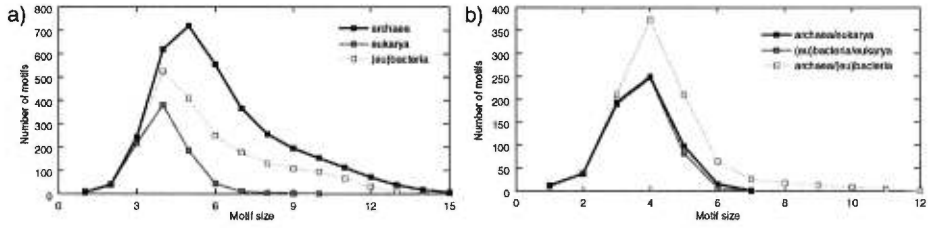


Figure 4.10. *a) Variation of the number of motifs found for the different phylogenetic domain of 16S rRNA using a significance tolerance of 1.00. b) Number of motifs found for different subset of phylogenetic domain showing the structural proximity of the domains.*

A larger amount of conserved motifs indicate higher structural similarity. *Archaea* are primitive life forms and appear to have been subjected to lower genetical drift throughout evolution. The environments in which they live are uninhabitable to most other organisms thereby greatly reducing the evolutionary pressure to accept various sequence and structural mutations to gain an advantage. Consequently, it is not surprising to observe a great deal of structural similarity in the organisms of the *Archaea* domain.

A comparison of the quantity of conserved motifs between each pair of phylogenetic domains was undertaken in the 5S and the 16S rRNA (figures 4.9b and 4.10b). For the 5S,

results demonstrate that the three domains are quite similar. However, it should be noted that *Archaea* and *Eukarya* could be more closely related as shown by the higher curve of figure 4.9b. For the 16S, results clearly show a dominance of the structural similarity of *Archaea* and *(eu)Bacteria.*

In light of these results it is tempting to examine the location of *Archaea* in the Universal Tree. The introduction of rRNA sequencing has changed the view of a world composed solely of two different types of organisms, *Procarya* and *Eukarya.* It was only recently recognized that *Archaea* is a distinct domain from *(eu)Bacteria* [153]). *Archaea* have some unique characteristics such as isopranyl ether lipids and modified tRNA molecules [19]. But what is interesting regarding our results is their resemblance with both *(eu)Bacteria* and *Eukarya* depending on the part of the ribosome considered. The *(eu)Bacteria* and the *Archaea* have a comparable chromosome structure [19] and, more interestingly, sequence similarity in the 5S rRNA [75]. However, the *Archaea* and the *Eukarya* also share much resemblance such as gene similarity [19].

It is important to note that the three-domain classification of living cells has been challenged by some scientists who claim that *Archea* should be treated as a new subdivision of the *Prokarya* [23, 107, 108]. Some researches point to a root of the Universal Tree between the *(eu)Bacteria* and the *Eukarya/Archaea* [5, 153]. Thus, the *Eukarya* and the *Archaea* are sister group. The proximity of these domains is supported by our domain comparison results in the 5S rRNA (figure 4.9b). A closer look at loop E, a functional motif present in all domains, shows that the loop E from *(eu)Bacteria* is significantly different from the loop E motif of *Archaea* and *Eukarya* (figure 4.5, motifs a,c and f). It is important to recall that *Eukarya* loop E motif is isomorphic to the one of *Archeae* if the new pairings from experimental data are included.

Eukarya and *Archaea* share a great deal of similarity in most aspects of DNA replication and transcription, RNA translation and protein synthesis. The most important dissemblance lies in the 16S and 23S rRNA sequence. However, it is mostly in the sequences of 16S and 23S rRNA that *Archeae* and *(eu)Bacteria* are similar [5]. This could explain our

apparent contradictory results in the comparison of the 16S and of the 5S of the phylogenetic domains. It is clear from the result shown in figure 4.10b that a big change occurs relative to the 5S data of figure 4.9b where high similarity between *Eukarya/Archaea* shifts to high similarity between *Archaea/(eu)Bacteria*.

4.4 Conclusion

We have shown, in the preceding sections, the use of many graph-theoretical techniques in the context of probing RNA secondary and tertiary structures for regions of high structural and functional importance. Many structural motifs were identified among which, some directly refer to regions of high interest for the different activity of the ribosomal RNA and for drug targets. Also, the methods have permitted us to bring further evidence for the generally accepted division of species in various phylogenetic domains.

Also, with the use of molecular modeling programs like *MC-Sym* [104] or *MA-NIP* [105], it will be possible to, not only propose theoretical 3-D structures for identified motifs, but also to use predetermined motifs as structural components for larger models. During the analysis of RNA targets, the generation of theoretical 3-D models for motifs can be used to evaluate the conformational variability of the motif and it's interactions with other molecules (docking).

With the advent of more RNA three-dimensional structure determined either via NMR or X-ray crystallography, the methods described here, together with the annotation procedure described in [58], should prove most useful to identify new potential target sites for drug design. Indeed, crystal structures of 16S ribosomal RNAs are scheduled to appear in the PDB database in the next year [117].

One limitation of the current methodology comes from the imprecision of the RNA secondary structures used. Indeed, some interactions were found to be missing in structures predicted by theoretical techniques like CSA. Thus, the effectiveness of the methods should improve proportionally with the refinement of secondary structure prediction.

Chapitre 5

Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis

Isabelle Barrette[1,2], Guylaine Poisson[1,2], Patrick Gendron[2] and François Major[2]

Abstract

The human prion gene contains five copies of a 24-nucleotide repeat that is highly conserved among species. The folding free energies of the human prion mRNA, and in particular of the repeat region, were analyzed and indicate a biased codon selection and the presence of RNA patterns. Pseudoknots, similar to the one predicted by Wills in the human prion mRNA, were found in the repeat region of all available prion mRNAs found in GenBank, but not those of avians and the red slider turtle. An alignment of prion mRNAs, which share low sequence homology, shows several covariations to maintain the pseudoknot pattern. The presence of pseudoknots in yeast Sup35p and Rnq1 suggests the acquisition of a pseudoknot back to the prokaryotic era. A three-dimensional model of the human prion pseudoknot highlight protein and RNA interaction domains, which suggest a possible effect in prion protein translation. The possible role of pseudoknots in prion diseases is discussed as individuals with extra copies of the repeat develop the familial form of Creutzfeldt-Jakob disease.

---

[1]These authors contributed equally to the work.

[2]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7

5.1 Introduction

Approximately 15% of all documented human Creutzfeldt-Jakob disease (CJD) cases are due to inherited mutations in the prion gene [32]. A subset of these cases are due to insertional mutations, where four to nine extra copies of a 24-nucleotide (nt) repeat in the prion protein gene are implicated with the development of the disease [32, 83, 86, 62]. Normally, the human prion gene contains five copies of this repeat, which are highly conserved among species in prion genes [156], and have even been found in the yeast prion Sup35p [97]. The repeats might have appeared through unequal crossover events [62], and vary in number among species and individuals. Recently, it was demonstrated that the cloning in mice of a mouse homologue of the human prion gene containing four extra repeats caused the occurrence of a prion disease, whose symptoms resembled those of human familial CJD [31]. Moreover, it was later shown that the expression of a prion protein containing 14 copies of the repeat in mice resulted in apoptosis of cerebellar granule cells [30]. However, the exact function of these repeats has remained elusive. The oligopeptides resulting from the repeats have been shown to bind copper, and it was suggested that prion protein diseases could be due to defective copper metabolism [18]. Recently, however, the prion's role in copper metabolism was challenged by finding that overexpression of the prion protein did not alter copper levels in mice [146].

Thermodynamic analyses of the 24-nt repeat region suggest the presence of several hairpin-loop structures [88, 151], but more interestingly, Wills proposed the presence of a RNA pseudoknot (pk) (Fig. 5.1a) [150]. Pks are known to interfere in translation speed, and to cause frameshifting [42]. However, it is now known that frameshifting does not occur in prion diseases, but rather that the eight strains of the scrapie form of the prion protein, PrPSc, are composed of the same polypeptide sequence [127].

In an attempt to identify functional RNAs in prion genes, we performed an extensive analysis of the folding free energies (FFE) of the prion mRNA. These analyses are based on the work by Seffens and Digby who reported that messenger RNAs (mRNA) contain frag-

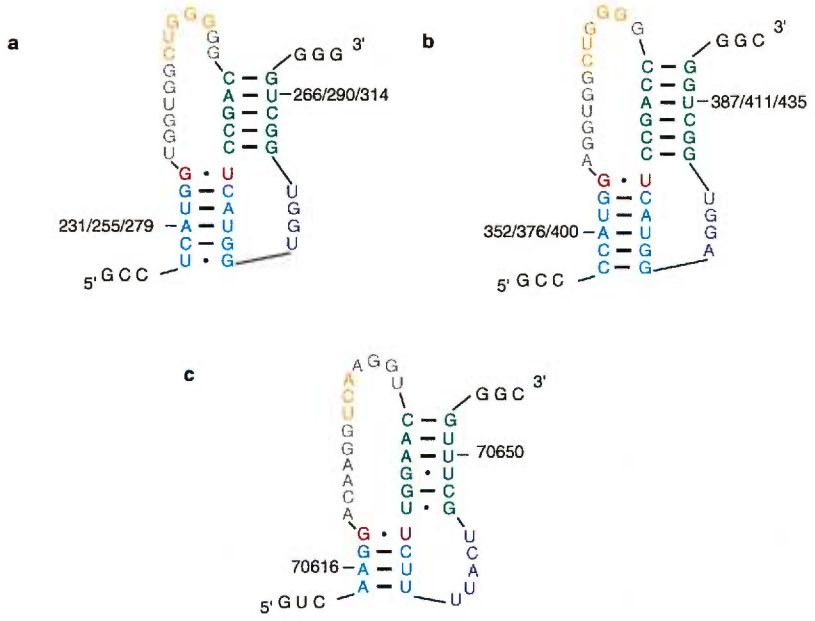


Figure 5.1. *Prion pseudoknot secondary structures. The nucleotides are coloured in the following manner: the G•U wobble base-pair is in red, the major loop is in grey, the minor loop is in violet, and the stems are in cyan and green. a) The pseudoknot described by Wills [151] in humans. The CUGGG motif is in yellow. b) The pseudoknot found in bovine. The CUGGG motif is in yellow. c) The pseudoknot found in yeast Rnq1. The UCA motif is in yellow.*

ments of greater negative folding energies than their shuffled or codon choice randomized counterparts [131]. Forsdyke performed the same type of studies on DNA, and observed similar results [49]. In some cases, the codon choice in mRNAs and DNA is biased to conserve functional elements, which may participate in regulatory processes [42, 39, 98].

We developed a computer program, a RNA Structural Pattern Finder, which was used to confirm the presence of codon choice biased regions in the human prion mRNA, and, in particular, in the 24-nt repeat region that was suggested by Wills to contain a pk. More importantly, using the computer program *RNAMOT* [55], a prion pk descriptor was developed, which allowed us to find similar pks in the repeat region of all prion mRNA sequences cur-

rently contained in GenBank, but not those of avians and the red slider turtle. The mRNA sequences are not homologous in the repeat region as was demonstrated using BLAST [2]. An alignment of the mRNA sequences in the repeat region shows several covariations to maintain the pseudoknot pattern. Finally, a three-dimensional model of the human prion pk was built using the molecular modeling package *MC-Sym* [104]. The model suggests the exposure to the solvent of reactive chemical groups, and in particular of the common UNR (U-turn) motif in the major loop of the pk, suggesting interactions with protiens and/or other RNAs.

5.2 Materials and Methods

5.2.1 Analysis of the folding free energies (FFE) of human PrP mRNA

The total folding free energies (FFE) of a sequence results from base order and composition [49], and represents the minimum of its free energy profiles [131]. The FFE of a sequence is determined from base pairing and stacking energies of its most stable predicted secondary structure. For instance, we compute the FFE of a sequence by using the dynamic programming algorithm and thermodynamic parameters developed by Rivas et al. [125]. To compute the base order dependent FFE of a sequence, we compare its FFE with the FFEs of sequences obtained from randomly permuting its base order. The average FFE ($\overline{\text{FFE}}$) of the permuted sequences reflects the base composition dependent FFE. The base order dependent FFE is given by the difference between the $\overline{\text{FFE}}$ of the permuted sequences and the FFE of the native sequence, $\Delta\text{FFE} = \overline{\text{FFE}} - \text{FFE}$. Positive values of ΔFFE indicate that the choice of codons in a particular region is biased in order to accommodate functional elements [49].

We developed a Structural Pattern Finder (SPF) computer program for detecting the regions of a RNA sequence that significantly deviate from the $\overline{\text{FFE}}$s of its shuffled sequences. The RNA sequence is divided into n windows of m nts, which overlap the preceding window by l nts. Each window is folded and its FFE measured. Then, all of the windows are

shuffled x times and folded again. The 2415-nt human prion mRNA (GenBank accession number NM_000311) was selected for the study. Forty-five windows of 200 nts overlapping the preceding ones by 150 nts were defined and shuffled 500 times. The statistical significance was tested for the biases observed in the calculated FFEs between the native human prion mRNA and the 500 randomized sequences.

5.2.2 Pseudoknot motif search

RNAMOT [55] was used to search for pks in eighty-seven mRNA sequences taken from GenBank. The *RNAMOT* descriptor was developed and generalized from the one predicted by Wills [150] in human prion mRNA (see Fig. 5.1a) and is shown in Fig. 5.2. The descriptor was constructed keeping only the features important for function and structural stability. For instance, the presence of the UNR motif was required in Loop I. This motif is thought to play an important role in protein-RNA and RNA-RNA interactions [51]. To regulate the allowed length of Loop I, zero to seven nucleotides were permitted 5' to the UNR motif, while one to five nucleotides were required 3'. No sequence restraints were applied for Loop II. Only the length was regulated by requiring the presence of a minimum of four nucleotides, but allowing a maximum of two extra nucleotides. In Stem I, a G•B base pair was required, as this type of base pair has been shown to be important in RNA-protein interactions [72]. A minimum of 4 base pairs with the possibility of adding 2 extra base pairs was chosen in order to maximize the stability of the structure for both Stem I and Stem II.

The accession numbers of the searched sequences are: AF003087, AF009181, AF015603, AF090852, AF113937, AF113938, AF113939, AF113941, AF113942, AF113943, AF113944, AF113945, AF117309, AF117310, AF117311, AF117312, AF117313, AF117314, AF117315, AF117316, AF117317, AF117318, AF117319, AF117320, AF117321, AF117322, AF117323, AF117324, AF117325, AF117326, AF117327, AF117328, AF117329, AF157954, AF157955, AF157956, AF157957, AF157958, AF157959, AF157960, AJ223072, AJ245488, D50093,

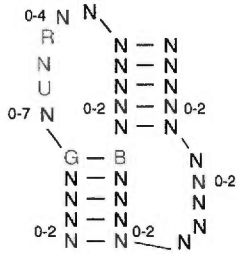


Figure 5.2. *RNAMOT descriptor based on the pseudoknot found in the human prion mRNA. The nucleotides are colored in the following manner. Orange represents the motifs important for function. Green represents the flexible nucleotides: N is for any base, R for purine and B for either C, G or U.*

K02234, L07623, M13685, M21129, M33958, M61145, M95404, NC_001135, NM_000311, U08291, U08292, U08293, U08294, U08295, U08296, U08297, U08298, U08299, U08300, U08301, U08302, U08303, U08304, U08305, U08306, U08307, U08308, U08309, U08310, U08311, U08312, U08952, U21210, U28334, U75382, U75383, U75384, U75385, U75386, U75387, U75388, U75389, X74759, Y09760, Y09761. The rate of false positives was computed by searching for prion pks in 1000 randomly generated sequences of 120 nucleotides. The prion pk was found in 90 random sequences, giving a false positive rate of approximately 0.09.

5.2.3 Modeling of PrP mRNA pseudoknot

The pk secondary structure (see Fig. 5.1a), predicted by Wills [150], was used to generate three-dimensional models using the *MC-Sym* molecular modeling computer program [104] on a dual Intel 600 MHz PentiumIII with 1 Gb of RAM. The models were built by sections and the resulting structural graph is shown in Fig. 5.3[3]. Standard A-RNA type helices were assumed for the two stems, and the wobble hydrogen-bonding pattern was assigned

[3]The *MC-Sym* molecular modeling computer program is available on the Web at www-lbit.iro.umontreal.ca/mcsym.

to the G•U base-pair at the end of the first stem. The nucleotides in the minor loop region were assigned C3'-endo sugar pucker conformations combined with anti orientations of their glycosyl bond torsion angle. The nucleotides in the major loop were allowed to adopt any type of conformations. However, to model the UNR motif, distance constraints were applied between N3 of uridine and the phosphate of the nucleotide immediately following the motif. Moreover, the nucleotides 5' to the UNR motif were required to be stacked upon each other. Different combinations of nucleotides in the stacked conformation were tried and evaluated using energy minimization. The potential energy of the models was minimized using the program *CHARMm* [17], and the *CHARMm* forcefield parameters 1997 [101]. The minimization was performed using a distance-dependent constant dielectric ($e = 4r$). The obtained structures were then used to calculate the electrostatic solvation free energy using the program *UHBD* (28) with a grid of 1003 points separated by 0.8 Å. The Poisson-Boltzmann linear equation was applied with a bulk salt concentration of 400 mM. The sums of the potential energies obtained in *CHARMm* and the electrostatic solvation free energies calculated in *UHBD* were used to compare the models.

5.3 Results

5.3.1 Folding free energies (FFEs)

The randomized human prion mRNA FFEs are normally distributed, as shown in Fig. 5.4a. The fourth window of the native sequence, containing the repeat region, has a FFE of -92.7 kcal/mol, which is one of the lowest energies obtained. The $\overline{\text{FFE}}$ of the shuffled sequences for the fourth window, -83.8 $\pm$ 0.4 kcal/mol, is approximately two standard deviations away from the native sequence, indicating an evolutionary biased codon selection. The error margin comes from the 500 shuffled sequences at a confidence level of 95%. Several other regions were under an evolutionary biased codon selection (Fig. 5.4b), but the fourth window has one of the largest ΔFFEs obtained.

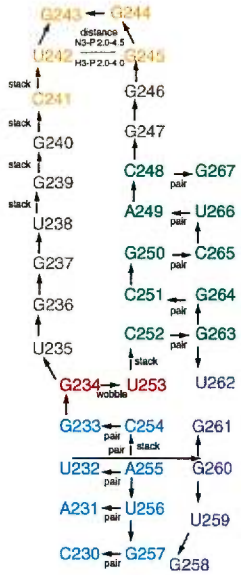


Figure 5.3. *Structural graph used to build the three-dimensional model of the human prion mRNA pseudoknot. Nodes represent the nucleotides, numbered according to the human prion mRNA sequence (GenBank accession number NM_000311). Edges represent relation beetween nucleotides. The construction pathway follow the arrows.*

5.3.2 Motif search

Using our *RNAMOT* descriptor, the pk was located in all 76 mammalian prion genes contained in GenBank, as well as in the yeast prion genes Sup35p and the recently discovered Rnq1 [136]. The only species in which no pks were found are the nine available avians and the red slider turtle [133], which possess genes significantly different from the others [156, 133]. The secondary structures of human, bovine, and yeast pks are shown in Fig. 5.1. An alignment of the 78 mRNA sequences in the pk region is shown in Fig. 5.5, where the pseudoknots are classified into 34 groups based on sequence only. The pseudoknots can also be grouped into nine different classes, depending on species and the sequence of the pseudoknots. From these nine classes, the likelihood of finding by chance

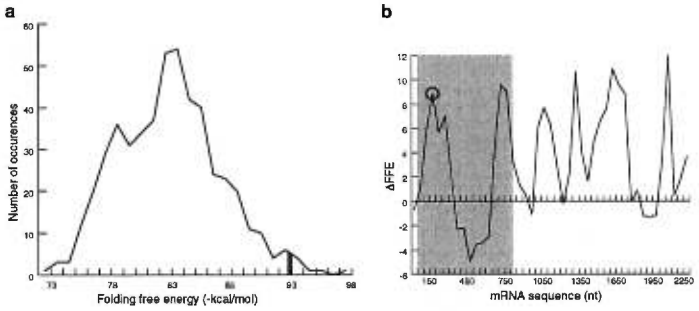


Figure 5.4. *FFE distribution of the human prion gene. a) Distribution of the FFEs of the shuffled sequences from the fourth 200-nt window encompassing nucleotides 150 to 350. The red bold line represents the FFE of the fourth window in the native sequence. b) ΔFFEs of the 45 windows. The ΔFFEs of windows in the coding sequence, nucleotides 50 to 811, are shown in the grey area. The ΔFFE peak of the fourth window is circled in blue.*

the prion pk in these 78 genes, computed from the *RNAMOT* rate of false positives (see Materials and Methods), was evaluated to be less than 1/109. The use of the nine classes to calculate the rate of false positives gives a better estimate as not all the sequences are independent events. When the pseudoknot sequences were used in BLAST [2], the human pseudoknot sequence could not match that of the two yeast pseudoknot sequence. In fact, it was impossible to locate all of the pseudoknots found using *RNAMOT* [55] when BLAST was used and only the sequence of the pseudoknots was examined.

5.3.3 Three-dimensional modeling of the pseudoknot (pk)

Using different scripts, the molecular modeling computer program *MC-Sym* [104] generated multiple models of the human pk for each script. After visual inspection and energy minimization with *CHARMm* [17, 101] and *UHBD* [38], the model shown in Fig. 5.6 was selected (available at www-lbit.iro.umontreal.ca/en/archives) and was generated using the

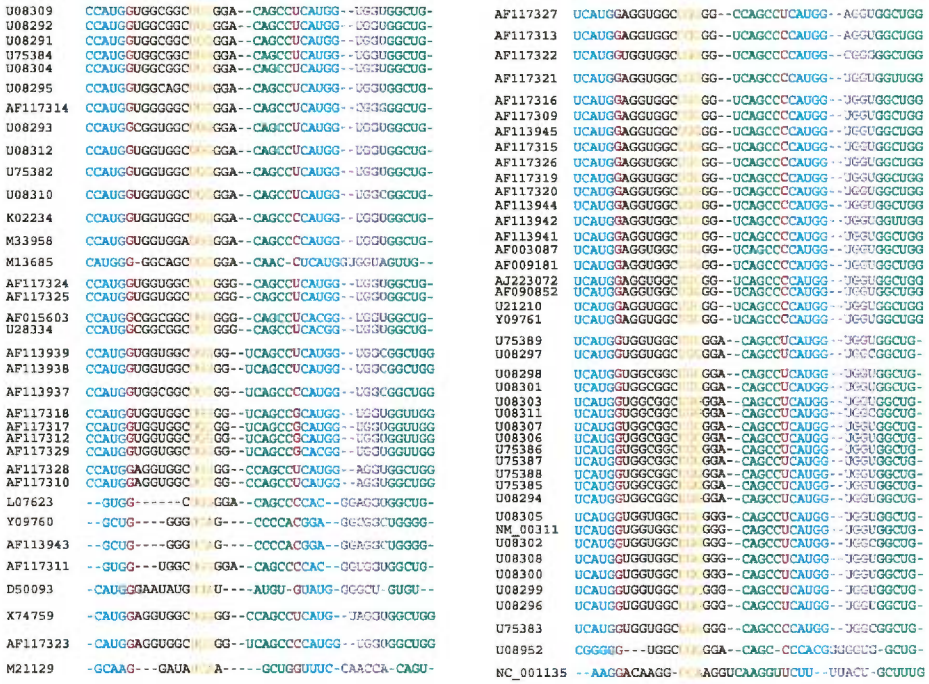


Figure 5.5. *Alignment of the pseudoknot found in all 78 sequences. The nucleotides are coloured following the rules of Fig. 5.1.*

script shown in Fig. 5.3. In this model, there is a hydrogen bond between the uridine of the UNR motif and the phosphate of the nucleotide following the R. The nucleotides preceding the UNR motif are also stacked. It is clear from Fig. 5.6 that the UNR motif, the stacked nucleotides, as well as the G•U base pair are all solvent-accessible.

5.4 Discussion

The FFE analysis suggests that the human prion mRNA sequence was biased through evolution to conserve structural elements (Fig. 5.4b). From the results of our analyses, the region containing the 24-nt repeats possesses one of the highest ΔFFEs obtained, thereby supporting the presence of a functional RNA pseudoknot, first proposed by Wills [150].

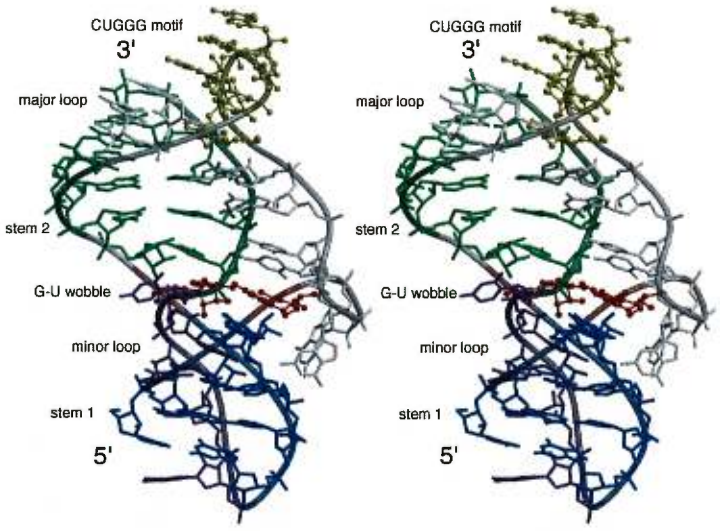


Figure 5.6. *Stereo view of the pseudoknot three-dimensional model of lowest energy. The ribbon joins the phosphates. The ribbon and nucleotides are coloured following the rules of Fig. 5.1.*

Using *RNAMOT*, this pseudoknot could be located in 78 different prion sequences, including those of yeast. The presence of a pk in yeast Rnq1 and Sup35p is extremely interesting. First, note the striking similarity between the pks in yeast Rnq1 and in human PrPc (compare Fig. 5.1a and 5.1c). Rnq1 has a prion domain that does not contain the 24-nt repeat. This suggests that the pk carried out through the repeat region, was probably acquired from yeast. The finding of a pk in Rnq1 also supports our low likelihood of finding the pks by chance in all mRNAs. The prion domain in Rnq1 was determined in the primary mRNA sequence from mutagenesis experiments [136].

An infrequent polymorphism occurs in the 24-nt repeats (see the alignment of the 78 mRNA sequences in Fig. 5.5), where the G•U wobble base-pair is substituted by a G•C Watson-Crick base-pair, and in rare occasions by a G•G base-pair. In particular, the G•U and G•G base-pairs at a terminal position of a stem have been observed in RNA-RNA and RNA-protein interactions [72]. The UNR motif found in the major loop of the prion pks

can adopt the U-turn motif known to be involved in RNA loop-loop interactions [51]. The UNR U-turn motif is stabilized by a hydrogen bond between the U and the phosphate of the nucleotide following the R ([67], Fig. 5.6). This particular arrangement exposes the acceptor and donor groups of the NR and following base to the solvent (Fig. 5.6), such as in the anticodon loop of tRNAs ([67], Fig. 5.6). The CUGGG motif in the human prion pk was also found in the loop of HIV TAR RNA [151], and Tat, p68 and galectin-3 have been shown to interact with the human prion mRNA [129].

The YUNR and UNR motifs are also good targets for antisense sequences [151, 51]. We were interested in searching and finding in the 3'UTR regions of the prion protein mRNAs (results not shown) antisense hairpins complementary to their associated prion pk UNR motifs in human, bovine, and yeast Rnq1. Although highly speculative, the co-acquired antisense could inhibit the formation of pk-protein complexes that would interfere during translation. Thanaraj and Argos noticed that alpha helices were preferentially coded by mRNA regions, where the rate of translation is fast, whereas beta strands and coils were coded by regions where the rate is slow [142]. In fact, a correlation exists between the use of certain codons in the mRNA and the topological features of the resulting protein [142]. This observation is interesting in light of the conformational change occurring in the prion protein, where the mainly alpha-helical protein, PrPc, attains a beta-sheet rich conformation in its infectious form, PrPSc. In this case, the pk may serve as a potential therapeutic target since it has been shown that the symptoms of prion diseases may be related to the concentration of prion protein expressed [31].

However, although our observations suggest the presence of pks in the mRNAs of prion protein genes, their actual folding and possible interference in PrPc translation are yet to be demonstrated experimentally. Nevertheless, it is tempting to suggest that these pks could be involved in the conformational changes of PrPc into PrPSc, which arises at the endoplasmic reticulum where translation occurs [71]. In particular, pk-protein complexes could form and interfere during translation and lead to the folding of the prion protein into its pathogenic form.

5.5 Conclusion

The finding of a RNA pk in 78 prion mRNA sequences, including those of yeast Sup35p and Rnq1, suggests its involvement in familial forms of CJD. The presence of a RNA pk in primitive species such as yeast, and in particular in yeast Rnq1, suggests its conservation through evolution since the procaryotic era, an interesting phenomenon regarding CJD, and more generally prion protein diseases. Using bioinformatics, a re-evaluation of the role of RNAs in familial CJD is warranted, and incisive experiments are suggested. We are planning in using our computer program SPF, which combined with *RNAMOT*, will help us identify other structured RNA in the genes implicated in other amyloid diseases, such as Alzheimer's.

Acknowledgments

We thank Laurent David for his help with *CHARMm* and *UHBD* and we thank Anthony Kusalik and Stephen Michnick for reviewing the manuscript. This work was supported by a grant from the Medical Research Council of Canada (MT-14604) to FM. GP holds a Ph.D. scholarship from the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

Chapitre 6

Conclusion

L'ensemble des travaux présentés dans ce mémoire apportent une contribution non négligeable à l'étude objective de la structure des acides ribonucléiques, de la séquence à la structure tertiaire. Ces travaux rendent possible l'identification et l'analyse de motifs structuraux à tous les niveaux de structure de l'ARN.

Les chapitres 2 et 3 proposent l'utilisation d'un formalisme basé sur la théorie des graphes pour la description symbolique de l'organisation structurale des ARN. Cette représentation laisse entrevoir de nombreuses possibilités algorithmiques empruntées à l'analyse et à l'optimisation de réseaux. Dans le cas présent, la recherche de sous-structures communes maximales à l'aide d'un algorithme d'isomorphisme de graphe permet de déterminer, de façon automatique, des sites fonctionnels conservés. L'analyse comparative effectuée au chapitre 4 montre la divergence évolutive des ARN ribosomaux des trois domaines phylogénétiques, en mettant en évidence des motifs structuraux communs, ainsi que d'autres spécifiques à chacun. Parmi ceux-ci, on retrouve certains motifs déjà reconnus pour interagir de façon particulière avec des antibiotiques, ce qui confirme la validité de la méthode proposée.

Outre l'analyse de motifs structuraux, le chapitre 3 propose une méthode de description automatique des différentes composantes d'une structure tridimensionnelle d'ARN. Cette méthode s'avère être un outil important d'analyse de nouvelles structures générées à la fois de façon expérimentale et théorique. De plus, la comparaison des composantes des ARN permet de cibler les relations déviant de la normale, lesquelles sont en général responsables du repliement tridimensionnel particulier des ARN et de leurs interactions avec d'autres molécules biologiques.

En jumelant la détection de motifs à l'annotation des structures tridimensionnelles, il est possible d'intégrer des motifs au processus de modélisation des ARN. De cette façon, le modélisateur peut tirer parti de l'information structurale contenue dans ces motifs, et ainsi réduire considérablement le temps d'exploration des conformations des structures. D'ailleurs, les engins de modélisation *MC-Sym* [104] et *MANIP* [105] permettent déjà d'intégrer ce genre d'information structurale.

Enfin, le travail présenté au chapitre 5, en plus de permettre l'identification de régions à haut potentiel structural dans les séquences d'ARN, montre l'importance de l'étude du repliement et de la structuration des ARN dans la recherche sur certaines maladies. Ce travail laisse entrevoir une nouvelle voie de recherche en ce qui concerne des maladies associées à des protéines particulières. En effet, la structure des ARN responsables de l'expression de ces protéines semble jouer un rôle important sur la forme finale des protéines. Dans le cas de la protéine impliquée dans la maladie de Creutzfeldt-Jakob, il a été montré que tous les cas familiaux de cette maladie faisaient intervenir une modification de l'ARN messager correspondant. La création d'agents thérapeutiques agissant au niveau de l'ARN plutôt que de la protéine est donc une voie possible.

En définitive, les outils développés dans ces travaux devraient permettre de localiser de nouveaux sites, dans la structure des ARN, responsables de fonctions biologiques précises. Ces outils devraient aussi aider à l'élucidation et à la compréhension du lien entre la structure et la fonction des ARN, en plus d'accroître nos capacités d'actions sur ces molécules particulières.

Références

[1] L.A. Allison, P.J. Romaniuk et A.H. Bakken. RNA-protein interactions of stored 5S RNA with TFIIIA and ribosomal protein L5 during *xenopus* oogenesis. *Developmental Biology*, **144**:129–144, 1991.

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

[3] P.J. Artymiuk, H.M. Grindley, A.R. Poirrette, D.W. Rice, E.C. Ujah et P. Willett. Identification of β-sheet motifs, of ψ-loops, and of patterns of amino acid residues in three-dimensionnal protein structures using a subgraph-isomorphism algorithm. *Journal of Chemical Information and Computer Sciences*, **34**:54–62, 1994.

[4] M.S. Babcock, E.P.D. Pedneault et W.K. Olson. Nucleic acid structure analysis. *Journal of Molecular Biology*, **237**:125–156, 1994.

[5] S.L. Baldauf, J.D. Palmer et W.F. Doolittle. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences (USA)*, **93**:7749–7754, 1996.

[6] N. Ban, P. Nissen, J. Hansen, P.B. Moore et T.A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**:905–920, 2000.

[7] I. Barrette, G. Poisson, P. Gendron et F. Major. Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis. sous presse, *Nucleic Acids Research*.

[8] R.T. Batey, R.P. Rambo et J. Doudna. Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition*, **38**:2326–2343, 1999.

[9] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp et D.L. Wheeler. GenBank. *Nucleic Acids Research*, **28**:15–18, 2000.

[10] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**:573–580, 1999.

[11] G. Benson et M.S. Waterman. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Research*, **22**:4828–4836, 1994.

[12] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A.R. Srinivasan et B. Schneider. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal*, **63**:751–759, 1992.

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

[14] B. Billoud, M. Kontic et A. Viari. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Research*, **24**:1395–1403, 1996.

[15] C.-I. Branden et J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., 1999.

[16] A.T. Brint et P. Willett. Algorithms for the identification of three-dimensional maximal common substructures. *Journal of Chemical Information and Computer Sciences*, **27**:152–157, 1987.

[17] B.R. Brooks, R.E Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan et M. Karplus. CHARMm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, **4**:187, 1983.

[18] D.R. Brown, K. Qin, J.W. Herms, A. Madlung, J. Manson, R. Strome, P.E. Fraser, T. Kruck, A. von Bohlen, W. Schulz-Schaeffer, A. Giese, D. Westaway et H. Kretzschmar. The cellular prion protein binds copper in vivo. *Nature*, **390**:684–687, 1997.

[19] J.R. Brown et W.F. Doolittle. *Archaea* and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Review*, **61**:456–502, 1997.

[20] M.P.S. Brown. A stochastic context free grammar used to predict secondary interactions. http://www.cse.ucsc.edu/research/compbio/ssurrna.html.

[21] J.H. Cate, A.R. Gooding, E. Podell, K. Zhou, B.L. Golden, A.A. Szewczak, C.E. Kundrot, T.R. Cech et J.A. Doudna. RNA tertiary structure mediation by adenosine plateforms. *Science*, **273**:1696–1698, 1996.

[22] J.H. Cate, M.M. Yusupov, G.Z. Yusupova, T.N. Earnest et H.F. Noller. X-ray crystal structures of 70S ribosome functional complexes. *Science*, **285**:2095–2104, 1999.

[23] T. Cavalier-Smith. Bacteria and eukaryotes. *Nature*, **356**:570, 1992.

[24] T. Cech. RNA splicing: Three themes with variations. *Cell*, **34**:713–716, 1983.

[25] R. Cedergren et F. Major. Modeling the tertiary structure of RNA. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 37–75, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[26] J. Chen. A linear-time algorithm for isomorphism of graphs of bounded average genus. *SIAM Journal on Discrete Mathematics*, **7**:614–631, 1994.

[27] J.-H. Chen, S.-Y. Le et J. Maizel. Prediction of common secondary structures of RNAs: A genetic algorithm approach. *Nucleic Acids Research*, **28**:991–999, 2000.

[28] A. Cheng, R.S. Stanton, J.J. Vincent, A. van der Vaart, K.V. Damodaran, S.L. Dixon, D.S. Hartsough, M. Mori, S.A. Best, G. Monard, M. Garcia, L.C. Van Zant et K. M. Merz. ROAR 2.0. Rapport technique, The Pennsylvania State University, 1999.

[29] F. Chetouani, P. Moestié, P.Thébault, C. Gaspin et B. Michot. ESSA: an integrated and interactive computer tool for analysing RNA secondary structure. *Nucleic Acids Research*, **25**:3514–3522, 1997.

[30] R. Chiesa, B. Drisaldi, E. Quaglio, A. Migheli, P. Piccardo, B. Ghetti et D.A. Harris. Accumulation of protease-resistant prion protein (PrP) and apoptosis of cerebellar granule cells in transgenic mice expressing a PrP insertional mutation. *Proceedings of the National Academy of Sciences (USA)*, **97**:5574–5579, 2000.

[31] R. Chiesa, P. Piccardo, B. Ghetti et D.A. Harris. Neurological illness in transgenic

mice expressing a prion protein with an insertional mutation. *Neuron*, **21**:1339–1351, 1998.

[32] J. Collinge, J. Beck, T. Campbell, K. Estibeiro et R.G. Will. Prion protein gene analysis in new variant cases of creutzfeldt-jakob disease. *Lancet*, **348**:56, 1996.

[33] G.L. Conn, D.E. Draper, E.E. Lattman et A.G. Gittis. Crystal structure of a conserved ribosomal protein-RNA complex. *Science*, **284**:1171–1174, 1999.

[34] D.G. Corneil et C.C. Gotlieb. An efficient algorithm for graph isomorphism. *Journal of the ACM*, **17**:51–64, 1970.

[35] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell et P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acide, and organic molecules. *Journal of the American Chemical Society*, **117**:5179–5197, 1995.

[36] C.C. Correll, A. Munishkin, Y-L. Chan, Z. Ren, I.G. Wool et T.A. Steitz. Crystal structure of the ribosomal RNA domain essential for binding elongation factors. *Proceedings of the National Academy of Sciences (USA)*, **95**:13436–13441, 1998.

[37] M. Costa et F. Michel. Frequent use of the same tertiary motif by self-folding RNAs. *EMBO Journal*, **14**:1276–1285, 1995.

[38] M.E. Davis, J.D. Madura, B.A. Luty et J.A. McCammon. Electrostatics and diffusion of molecules in solution: simulations with the university of houston brownian dynamics program. *Computer Physics Communications*, **62**:187–197, 1991.

[39] M.H. de Smit et J. van Duin. Control of prokaryotic translational initiation by mRNA secondary structure. *Progress in Nucleic Acids Research*, **38**:1–35, 1990.

[40] R.E. Dickerson, K. Grzeskowiak, M. Grzeskowiak, M.L. Kopka, T. Larsen, A. Lipanov, G.G. Privé, J. Quintana, P. Schultze, K. Yanagi, H. Yuan et H.-C. Yoon. Polymorphism, packing, resolution, and reliability in single-crystal DNA oligomer analyses. *Nucleosides and Nucleotides*, **10**:3–24, 1991.

[41] D.E. Draper. Themes in RNA-protein recognition. *Journal of Molecular Biology*, **293**:255–270, 1999.

[42] D.E. Draper, T.C. Gluick et P.J. Schlax. Pseudoknots, RNA folding, and translational regulation. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 415–436, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[43] C.M. Duarte et A.M. Pyle. Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology*, **284**:1465–1478, 1998.

[44] R. Durbin, S. Eddy, A. Krogh et G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[45] S.R. Eddy. Multiple alignment using hidden markov models. Dans C. Rawlings et al., editeur, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120, Menlo Park, 1995. AAAI Press.

[46] N. El-Mabrouk et M. Crochemore. Boyer-moore strategy to efficient approximate string matching. Dans D. Hirschberg et G. Myers, editeurs, *Lecture Notes in Computer Science*, CPM, 7th annual symposium, pages 24–38, Laguna Beach, California, June 1996.

[47] N. El-Mabrouk et F. Lisacek. Very fast identification of RNA motifs in genomic DNA. application to tRNA search in the yeast genome. *Journal of Molecular Biology*, **264**:46–55, 1996.

[48] N. Foloppe et A.D. MacKerell. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, **21**:86–104, 2000.

[49] D.R. Forsdyke. A stem-loop kissing model for the initiation of recombination and the origin of introns. *Molecular Biology and Evolution*, **12**:949–958, 1995.

[50] D. Fourmy, S. Yoshizawa et J.D. Puglisi. Paromomycin binding induces local

conformational change in ths A-site of 16S rRNA. *Journal of Molecular Biology*, **277**:333–345, 1998.

[51] T. Franch, M. Petersen, E.G.H. Wagner, J.P. Jacobsen et K. Gerder. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. *Journal of Molecular Biology*, **294**:1115–1125, 1999.

[52] H.A. Gabb, S.R. Sanghani, C.H. Robert et C. Prévost. Finding and visualizing nucleic acid base stacking. *Journal of Molecular Graphics*, **14**:6–11, 1996.

[53] D. Gautheret et R.R. Gutell. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Research*, **25**:1559–1564, 1997.

[54] D. Gautheret, D. Konings et R.R. Gutell. G.U base pairing motifs in ribosomal RNA. *RNA*, **1**:807–814, 1995.

[55] D. Gautheret, F. Major et R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *CABIOS*, **6**:325–331, 1990.

[56] D. Gautheret, F. Major et R. Cedergren. Modeling the three-dimensionnal structure of RNA using discrete nucleotide conformationnal sets. *Journal of Molecular Biology*, **229**:1049–1064, 1993.

[57] P. Gendron, D. Gautheret et F. Major. Structural ribonucleic acid motifs identification and classification. Dans *High Performance Computing Systems and Applications*. Kluwer Academic Press, 1998.

[58] P. Gendron, S. Lemieux et F. Major. Quantitative analysis of nucleic acid three-dimensional structures. sous presse, *Journal of Molecular Biology*.

[59] P. Gendron, G. Poisson, D. Gautheret et F. Major. Comparative structural analysis of nucleic acids. soumis, *Nucleic Acids Research*.

[60] J-F. Gibrat, T. Madej et S.H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**:377–385, 1996.

[61] B.L. Golden, A.R. Gooding, E.R. Podell et T.R. Cech. A preorganized active site in the crystal structure of the tetrahymena ribozyme. *Science*, **282**:259–264, 1998.

[62] L.G. Goldfarb, P. Brown, W.R. McCombie, D. Goldgaber, G.D. Swergold, P.R. Wills, L. Cervenakova, H. Baron, C.J. Gibbs et D.C. Gajdusek. Transmissible familial Creutzfeldt-Jakob disease associated with five, seven, and eight extra octapeptide coding repeats in the PRNP gene. *Proceedings of the National Academy of Sciences (USA)*, **88**:10926–10930, 1991.

[63] J. Gorodkin, L.J. Heyer et G.D. Stormo. Finding the most significant common sequence and structure motif in a set of RNA sequences. *Nucleic Acids Research*, **25**:3724–3732, 1997.

[64] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace et S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **3**:849–857, 1983.

[65] A.P. Gultyaev, F.H. van Batenburg et C.W. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, **250**:37–51, 1995.

[66] R.R. Gutell. Comparative sequence analysis and the structure of 16S and 23S rRNA. Dans Dahlberg A. et Zimmerman B., editeurs, *Ribosomal RNA. Structure, evolution, processing, and function in protein biosynthesis*, pages 111–128, Boca Raton, Florida, 1996. CRC Press.

[67] R.R. Gutell, J. Cannone, D. Konings et D. Gautheret. Predicting U-turns in ribosomal RNA with comparative sequence analysis. *Journal of Molecular Biology*, **300**:791–803, 2000.

[68] R.R. Gutell, S. Subashchandran, M. Schnare, Y. Du, N. Lin, L. Madabusi, K. Muller, N. Pande, N. Yu, Z. Shang, S. Date, D. Konings, V. Schweiker, B. Weiser et J.J. Cannone. Comparative sequence analysis and the prediction of RNA structure, and the web. en préparation.

[69] S. Hanlon. The importance of london dispersion forces in the maintenance of deoxy-

ribonucleic acid double helix. *Biochemical and Biophysical Research Communications*, **23**:861–867, 1966.

[70] M.E. Harris, D.N. Frank et N.R. Pace. Structure and catalytic function of the bacterial ribonuclease P ribozyme. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 309–337, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[71] R.S. Hedge, J.A. Mastrianni, M.R. Scott, K.A. DeFea, P. Tremblay, M. Torchia, S.J. DeArmond, S.B. Prusiner et V.R. Lingappa. A transmembrane form of the prion protein in neurodegenerative disease. *Science*, **279**:827–834, 1998.

[72] T. Hermann et E. Westhof. Non-watson-crick base pairs in RNA-protein recognition. *Chemistry and Biology*, **6**:R335–R343, 1999.

[73] C.G. Hoogstraten, P. Legault et A. Pardi. NMR solution structure of the lead-dependent ribozyme: evidence for dynamics in RNA catalysis. *Journal of Molecular Biology*, **284**:337–350, 1998.

[74] J.E. Hopcroft et J.K. Wong. Linear time algorithm for isomorphism of planar graphs. Dans *Proceedings of the 6th Annual ACM Symposium on the Theory of Computing*, pages 172–18, Seatle, WA, 1974.

[75] H. Hori et S. Osawa. Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. *Molecular Biology and Evolution*, **4**:445–472, 1987.

[76] IUPAC et IUBMB. *Biochemical Nomenclature and Related Documents*. Portland Press, 1992.

[77] F. Jiang, R.A. Kumar, R.A. Jones et D.J. Patel. Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature*, **382**:183–186, 1996.

[78] V. Juan et C. Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *Journal of Molecular Biology*, **289**:935–947, 1999.

[79] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, **A32**:922–923, 1976.

[80] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, **A34**:827–828, 1978.

[81] K. Kalurachchi et E.P. Nikonowicz. NMR structure determination of the binding site for ribosomal protein S8 from *escherichia coli* 16S rRNA. *Journal of Molecular Biology*, **280**:639–654, 1998.

[82] S.H. Kim, F.L. Suddath, G.J. Quigley, A. McPherson, J.L. Sussman, A.H. Wang, N.C. Seeman et A. Rich. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**:435–440, 1974.

[83] S. Krasemann, I. Zerr, T. Weber, S. Poser, H. Kretzschmar, G. Hunsmann et W. Bodemer. Prion disease associated with a novel nine octapeptide repeat insertion in the PRNP gene. *Brain Research: Molecular Brain Research*, **34**:173–176, 1995.

[84] K. Kruger, P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling et T.R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, **31**:147–157, 1982.

[85] Laboratoire de Biologie Informatique et Théorique, Université de Montréal, http://www-lbit.iro.umontreal.ca/mcsym/. *MC-Sym 3.1 - User Manual*, 2000.

[86] J.L. Laplanche, K.H. Hachimi, I. Durieux, P. Thuillet, L. Defebvre, N. Delasnerie-Laupetre, K. Peoch, J.F. Foncin et A. Destee. Prominent psychiatric features and early onset in an inherited prion disease with a new insertional mutation in the prion protein gene. *Brain*, **122**:2375–2386, 1999.

[87] R. Lavery et H. Sklenar. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *Journal of Biomolecular Structure and Dynamics*, **6**:63–91, 1988.

[88] R. Lück, G. Steger et D. Riesner. Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *Journal of Molecular Biology*, **258**:813–826, 1996.

[89] F. Leclerc, J. Srinivasan et R. Cedergren. Predicting RNA structures: the model

of the RNA element binding rev meets the NMR structure. *Folding and Design*, **2**:141–147, 1997.

[90] P. Legault, J. Li, J. Mogridge, L.E. Kay et J. Greenblatt. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell*, **93**:289–299, 1998.

[91] S. Lemieux et F. Major. Probabilistic recognition of base pairing patterns in RNA three-dimensional structures. *Nucleic Acids Research*, **28**, 2000.

[92] S. Lemieux, S. Oldziej et F. Major. Nucleic acids: Qualitative modeling. Dans N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer et P.R. Schreiner, editeurs, *Encyclopedia of Computational Chemistry*, West Sussex, England, 1998. John Wiley & Sons.

[93] N.B. Leontis et E. Westhof. A common motif organizes the structure of multi-helix loops in 16S and 23S ribosomal RNAs. *Journal of Molecular Biology*, **283**:571–583, 1998.

[94] N.B. Leontis et E. Westhof. Conserved geometrical base-pairing patterns in RNA. *Quaterly Reviews of Biophysics*, **31**:399–455, 1998.

[95] N.B. Leontis et E. Westhof. Recurrent RNA motifs. Dans J. Barciszewski et B.F.C. Clark, editeurs, *RNA biochemistry and biotechnology*, pages 45–61. Kluwer Academic Publishers, 1999.

[96] M.-Y. Leung et T.E. Yamashita. Applications of the scan statistic in DNA sequence analysis. Dans N. Baoalrishnan et J. Glaz, editeurs, *Recent developments in scan statistics and applications*, pages 269–286. Birkauser Publishers, 1999.

[97] J.-J. Liu et S. Lindquist. Oligopeptide-repeat expansions modulate protein-only inheritance in yeast. *Nature*, **400**:573–576, 1999.

[98] H.D. Love, A. Allen-Nash, Q. Zhao et G.A. Bannon. mRNA stability plays a major role in regulating the temperature-specific expression of a tetrahymena thermophila surface protein. *Molecular and Cellular Biology*, **8**:427–432, 1988.

[99] T.M. Lowe et S.R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**:955–964, 1997.

[100] M. Lu et T.A. Steitz. Structure of escherichia coli ribosomal protein l25 complexed with a 5S rRNA fragment at 1.8 Å resolution. *Proceedings of the National Academy of Sciences (USA)*, **97**:2023–2028, 2000.

[101] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, **102**:3586–3616, 1998.

[102] F. Major et D. Gautheret. Computer modeling of RNA three-dimensional structure. Dans Robert A. Meyers, editeur, *Molecular Biology and Biotechnology: A Comprehensive Desk Reference*, pages 847–850, New York, NY, 1995. VCH Publishers Inc.

[103] F. Major, D. Gautheret et R. Cedergren. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proceedings of the National Academy of Sciences (USA)*, **90**:9408–9412, 1993.

[104] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion et R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, **253**:1255–1260, 1991.

[105] C. Massire et E. Westhof. MANIP: An interactive tool for modelling RNA. *Journal of Molecular Graphics*, **16**:197–205, 1999.

[106] E.A. Maxwell. *Methods of Plane Projective Geometry Based on the Use of General Homogeneous Coordinates*. Cambridge University Press, Cambridge, England, 1946.

[107] E. Mayr. A natural system of organisms. *Nature*, **348**:491, 1990.

[108] E. Mayr. Two empires or three? *Proceedings of the National Academy of Sciences (USA)*, **95**:9720–9723, 1998.

[109] F. Michel et M. Costa. Inferring RNA structure by phylogenetic and genetic analyses. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 175–202, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[110] F. Michel et E. Westhof. Modelling the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, **216**:585–610, 1990.

[111] H. Miyaguchi, H. Narita, K. Sakamoto et S. Yokoyama. An antibiotic-binding motif of an RNA fragment derived from the A-site-related region of *escherichia coli* 16S rRNA. *Nucleic Acids Research*, **24**:3700–3706, 1996.

[112] H. Moine, B. Ehresmann, C. Ehresmann et P. Romby. Probing RNA structure and function in solution. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 77–115, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[113] P.B. Moore. Structural motifs in RNA. *Annual Review of Biochemistry*, **68**:287, 1999.

[114] F. Mueller et R. Brimacombe. A new model for the three-dimensional folding of E.coli 16S ribosomal RNA: I. Fitting the RNA to a 3D electron microscopic map at 20 Å. *Journal of Molecular Biology*, **271**:524–544, 1997.

[115] L.F. Newcomb et S.H. Gellman. Aromatic stacking interactions in aqueous solution: Evidence that neither classical hydrophobic effects nor dispersion forces are important. *Journal of the American Chemical Society*, **116**:4993–4994, 1994.

[116] H.F. Noller. Ribosomal RNA and translation. *Annual Review of Biochemistry*, **60**:191–227, 1991.

[117] H.F. Noller, 2000. Personnal communication.

[118] H.F. Noller et C.R. Woese. Secondary structure of 16S ribosomal RNA. *Science*, **212**:403–411, 1981.

[119] R.P. Paul. *Robot Manipulators: Mathematics, Programming and Control.* MIT Press, Cambridge, 1981.

[120] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.R. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel et P. Kollman. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Computer Physics Communications*, **91**:1–41, 1995.

[121] D.A. Pearlman et S.H. Kim. Conformationnal studies of nucleic acids: III. Empirical multiple correlation functions for nucleic acid torsion angles. *Journal of Biomolecular Structure and Dynamics*, **4**:49–67, 1986.

[122] P. Penczek, N. Ban, R.A. Grassucci, R.K. Agrawal et J. Frank. Haloarcula marismortui 50s subunit-complementarity of electron microscopy and X-ray crystallographic information. *Journal of Structural Biology*, **128**:44–50, 1999.

[123] H.W. Pley, K.M. Flaherty et D.B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, **372**:68–74, 1994.

[124] M.I. Recht, S. Douthwaite, K.D. Dahlquist et J.D. Puglisi. Effect of mutations in the A site of 16S rRNA on aminoglycoside antibiotic-ribosome interaction. *Journal of Molecular Biology*, **286**:33–43, 1999.

[125] E. Rivas et S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, **285**:2053–2068, 1999.

[126] W. Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, USA, 1984.

[127] J. Safar, H. Wille, V. Itri, D. Groth, H. Serban, M. Torchia, F.E. Cohen et S.B. Prusiner. Eight prion strains have PrP(Sc) molecules with different conformations. *Nature Medicine*, **4**:1157–1165, 1998.

[128] A. Sarai, J. Mazur, R. Nussinov et R.L. Jernigan. Origin of DNA helical structure and its sequence dependence. *Biochemistry*, **27**:8498–8502, 1988.

[129] U. Scheffer, T. Okamoto, J.M.S. Forrest, P.G. Rytik, W.E.G. Muller et H.C. Schroder. Interaction of 68-kDa TAR RNA-binding protein and other cellular proteins with prion protein-RNA stem-loop. *Journal of Neurovirology*, **1**:391–398, 1995.

[130] J.B. Scripture et P.W. Hubert. Analysis of the binding of xenopus ribosomal protein L5 to oocyte 5S rRNA. *Journal of Biological Chemistry*, **270**:27358–27365, 1995.

[131] W. Seffens et D. Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research*, **27**:1578–1584, 1999.

[132] S.T. Sigurdsson, J.B. Thomson et F. Eckstein. Small ribozymes. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 339–376, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[133] T. Simonic, S. Duga, B. Strumbo, R. Asselta, F. Ceciliani et S. Ronchi. cDNA cloning of turtle prion protein. *FEBS Letters*, **469**:33–38, 2000.

[134] R.W. Simons et M. Grunberg-Manago, editeurs. *RNA Structure and Function*. Cold Spring Harbor Laboratory Press, New York, 1998.

[135] T.F. Smith et M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**:195–197, 1981.

[136] N. Sondheimer et S. Lindquist. Rnq1: An epigenetic modifier of protein function in yeast. *Molecular Cell*, **5**:163–172, 2000.

[137] J. Sponer et J. Kypr. Theoretical analysis of the base stacking in DNA: Choice of the force field and comparison with the oligonucleotide crystal structure. *Journal of Biomolecular Structure and Dynamics*, **11**:277–292, 1993.

[138] M. Stoldt, J. Wöhnert, O. Ohlenschläger, M. Görlach et L.R. Brown. The NMR structure of the 5S rRNA E-domain-protein L25 complex shows preformed and induced recognition. *EMBO Journal*, **18**:6508–6521, 1999.

[139] J.L. Sussman, S.R. Holbrook, R.W. Warrant, G.M. Church et S.H. Kim. Crystal

structure of yeast phenylalanine tRNA. I. Crystallographic refinement. *Journal of Molecular Biology*, **123**:607–630, 1978.

[140] A.A. Szewczak et P.B. Moore. The sarcin/ricin loop, a modular RNA. *Journal of Molecular Biology*, **247**:81–98, 1995.

[141] I. Tazawa, T. Koike et Y. Inoue. Stacking properties of a highly hydrophobic dinucleotide sequence, N6, N6-dimethyladenylyl(3' leads to 5')N6, N6-dimethyladenosine, occurring in 16–18-S ribosomal RNA. *European Journal of Biochemistry*, **109**:33–38, 1980.

[142] T.A. Thanaraj et P. Argos. Protein secondary structural types are differentially coded on messenger RNA. *Protein Science*, **5**, 1996.

[143] J.R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM*, **23**:31–42, 1976.

[144] G. Ferbeyre V. Bourdeau, M. Pageau, B. Paquin et R. Cedergren. The distribution of RNA motifs in natural sequences. *Nucleic Acids Research*, **27**:4457–4467, 1999.

[145] D. Voet et J.G Voet. *Biochemistry*. John Wiley & Sons, Inc., 1995.

[146] D.J. Waggoner, B. Drisaldi, T.B. Bartnikas, R.L. Casareno, J.R. Prohaska, J.D. Gitlin et D.A. Harris. Brain copper content and cuproenzymes activity do not vary with prion protein expression level. *Journal of Biological Chemistry*, **275**:7455–7458, 2000.

[147] J.D. Watson et F.H.C. Crick. A structure for deoxyribose nucleic acid. *Nature*, **171**:737–738, 1953.

[148] J. E. Wedekind et D. B. Mckay. Crystal structure of a lead-dependent ribozyme revealing metal binding sites relevant to catalysis. *Nature Structural Biology*, **6**:261–268, 1999.

[149] H. Weissig et P.E. Bourne. An analysis of the protein data bank in search of temporal and global trends. *Bioinformatics*, **15**:807–831, 1999.

[150] P.R. Wills. Potential pseudoknots in the PrP-encoding mRNA. *Journal of Theoretical Biology*, **159**:523–527, 1992.

[151] P.R. Wills et A.J. Hughes. Stem loops in HIV and prion protein mRNAs. *Journal Acquired Immuno Deficiency Syndrome*, **3**:95–97, 1990.

[152] B. Wimberly, Gabriele Varani et I. Tinoco. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry*, **32**:1078–1087, 1993.

[153] C.R. Woese, O. Kandler et M.L. Wheelis. Towards a natural system of organisms: proposal for the domains *archaea*, *bacteria*, and *eucarya*. *Proceedings of the National Academy of Sciences (USA)*, **87**:4576–4579, 1990.

[154] C.R. Woese et N.R. Pace. Probing RNA structure, function, and history by comparative analysis. Dans Gesteland R.F. et Atkins J.F., editeurs, *The RNA World*, pages 91–118, Plainview, New York, 1993. Cold Spring Harbor Laboratory Press.

[155] C.R. Woese, S. Winker et R.R.Gutell. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences (USA)*, **87**:8467–8471, 1990.

[156] F. Wopfner, G. Weidenhöfer, R. Schneider, A. von Brunn, S. Gilch, T.F. Schwarz, T. Werner et H.M. Schätzl. Analysis of 27 mammalian and 9 avian PrPs reveals high conservation of flexible regions of the prion protein. *Journal of Molecular Biology*, **289**:1163–1178, 1999.

[157] Y. Xing et D.E. Draper. Cooperative interactions of RNA and thiostrepton antibiotic with two domains of ribosomal protein L11. *Biochemistry*, **35**:1581–1588, 1996.

[158] Y. Xing, D. GuhaThakurta et D.E. Draper. The RNA binding domain of ribosomal protein L11 is structurally similar to homeodomains. *Nature Structural Biology*, **4**:24–27, 1997.

[159] B.N. Zeiler et R.W.Simons. Antisense RNA structure and function. Dans R.W. Simons et M. Grunberg-Manago, editeurs, *RNA Structure and Function*, pages 437–464, Plainview, New York, 1998. Cold Spring Harbor Laboratory Press.

[160] M. Zuker, D.H. Mathews et D.H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. Dans J. Barciszewski et B.F.C. Clark, editeurs, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43. Kluwer Academic Publishers, 1999.

[161] M. Zuker et D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, **46**:591, 1984.