Université de Montréal

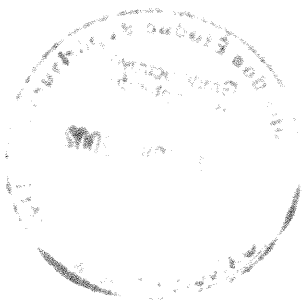
Modélisation automatisée de la structure 3-D des ARNs

par
Sébastien Lemieux

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

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Décembre, 2001



Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Modélisation automatisée de la structure 3-D des ARNs

présentée par:

Sébastien Lemieux

a été évaluée par un jury composé des personnes suivantes:

Yoshua Bengio
président-rapporteur

François Major
directeur de recherche

Max Mignotte
membre du jury

.....
représentant du doyen

Éric Westhof
examinateur externe

TABLES DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'objectif principal de cette thèse est la mise au point d'un système automatisé de modélisation de la structure 3-D des ARNs. Afin d'arriver à cette automatisation, plusieurs étapes intermédiaires s'imposent. D'abord, une analyse systématique des intéractions entre bases azotées est faite et permet d'extraire toutes les relations binaires présentes dans les structures connues. Une attention particulière est apportée aux relations d'appariement puisqu'il n'existait aucune méthode satisfaisante pour les détecter et les classifier, un algorithme précis et efficace est présenté. Le graphe de relation est ensuite introduit comme représentation d'une structure d'ARN et permet l'obtention d'une décomposition de ce graphe en un ensemble de cycles simples (la base minimale de l'espace des cycles). L'application de cette décomposition à la détection de motif est présentée et huit occurences d'un nouveau motif d'intérêt biologique sont identifiées dans la grande sous-unité du ribosome. Finalement, un système de modélisation 3-D est présenté en utilisant la décomposition en une base de l'espace des cycles et la base de donnée de relations binaires extraite des structure connues. La précision du système est évaluée en l'appliquant à deux molécules dont la structure 3-D est connue: une boucle de huit nucléotides extraite de l'ARN ribosomal s'attachant à la protein L11 et la boucle en épingle à cheveux 2555–2580 de la grande sous-unité du ribosome. Dans les deux cas, le système retourne, dans un temps raisonnable, une famille de modèles dont au moins un représente bien toutes les caractéristiques structurelles recherchées.

Mots Clés: Bioinformatique, théorie des graphes, apprentissage non-supervisé, analyse de structure, détection de motifs, optimisation combinatoire.

ABSTRACT

The main goal of this thesis is the development of an automated 3-D modeling system of RNAs. To achieve this automation, several tools were needed to complete the analysis of known structures. A systematic analysis of nitrogen bases interactions is developped and results in the extraction and classification of all binary relations contained in known RNA structures. An accurate and efficient algorithm for base pairing recognition was developped since no satisfying method existed. Then, the graph of relations is introduced as a representation of a RNA structure that allows for its decomposition in a set of simple cycles (the minimal cycle basis). The application of this decomposition to motif detection is presented and eight occurences of a new motif of biological interest are identified in the large ribosomal subunit of *H. marismortui*. Finally, a 3-D modeling system is presented by using the minimal cycle basis decomposition of the graph of relations and the database of binary relations obtained from known structures. Models are built in two steps: first, by optimizing the 3-D structure independently for each cycle; and then by joining these substructures together. The accuracy of the system is evaluated by modeling two molecules with known structures: an eight nucleotides cycle from the ribosomal RNA binding protein L11 and the hairpin 2555–2580 of the large ribosomal subunit of *H. marismortui*. In both cases, the system returns a family of models including at least one that has all the structural features of the known structure.

Keywords: Bioinformatic, graph theory, non-supervised learning, structure analysis, motif detection, combinatorial optimization.

REMERCIEMENTS: *Les travaux présentées dans cette thèse ont été rendues possibles grâce à une bourse de doctorat des IRSC. Les infrastructures du Laboratoire de Biologie Informatique et Théorique (LBIT) et celles du département d'informatique (DIRO) ont aussi été essentielles et bien appréciées. Mes sincères remerciements vont aux membres, présents et passés, du LBIT qui ont su soutenir le dynamisme de ce groupe de recherche et alimenter les discussions enflammées à l'origine de toute passion scientifique. Je tiens aussi à souligner l'excellent travail du groupe de support du DIRO auxquels j'ai sans doute causé bien des soucis dont ils se seraient passé!*

CHAPITRE 1

INTRODUCTION

La détermination en 1953 de la structure de la double hélice d'acide déoxyribonucléique (ADN) [93] permit la naissance de la biologie moléculaire moderne: les caractères génétiques sont encodés dans un polymère biochimique, ce polymère possède une structure précise qu'on peut observer dans certaines conditions, sa structure dicte sa fonction. À partir de ce moment, on comprend que l'étude d'un système vivant peut, et doit, se faire jusqu'au niveau atomique [91]. Par contre, aucune méthode ne permet d'observer une structure biochimique directement au niveau atomique. Au cours des décennies qui suivirent, plusieurs méthodes furent développées pour extraire diverses informations structurelles (distance inter-atomique, densité électronique, angle entre hélices...) et parallèlement pour en dériver un modèle 3-D.

Les travaux présentés dans cette thèse visent l'obtention d'une représentation 3-D d'une molécule étant donnée un certain nombre d'informations structurelles sur celle-ci. Le système mis au point vise la modélisation d'un type très précis de biopolymère: les acides ribonucléiques (ARN). Le système doit être entièrement automatisé et fournir des résultats de qualité pour des ARNs de petites tailles (environ 30 nucléotides) dans des délais raisonnables (quelques heures).

1·1 POURQUOI MODÉLISER LES ARNS?

Le dogme central de la biologie moléculaire indique que l'ADN d'un gène est d'abord transcrit en un ARN messager (ARNm), celui-ci est ensuite traduit en une

protéine qui est la forme effective de ce gène. Plusieurs virus font un usage très varié de l'ARN, par exemple, les retrovirus (dont le VIH) infectent la cellule sous forme d'ARN et d'une enzyme particulière (la retro-transcriptase inverse) permettant de transcrire cet ARN en ADN, la cellule s'occupant de la suite des opérations. L'opération de traduction d'un ARNm en une protéine est effectuée par le ribosome, un complexe regroupant dans plusieurs espèces plus de 4000 nucléotides d'ARN. La molécule permettant d'associer un triplet de base sur un ARNm (un codon) à un acide aminé (sous-unité de la protéine) est l'ARN de transfert (ARNt), une molécule d'environ 70 nucléotides d'ARN dont la structure 3-D a été élucidée à la fin des années '60 [57]. L'ARN est au coeur de la majorité des mécanismes importants de la cellule, la connaissance de sa structure est donc d'une importance primordiale pour l'élucidation de ces mécanismes. D'un point de vue thérapeutique, l'omniprésence des ARNs en fait des cibles de choix pour de nouveaux médicaments.

L'événement déclencheur de l'engouement pour l'étude des ARN est la découverte que ces biopolymères possèdent, à l'instar des protéines, des capacités catalytiques. Les ARNs ne sont donc pas seulement des messagers passifs mais jouent un rôle actif dans la vie cellulaire. En octobre 1989, la Royal Swedish Academy of Sciences décerna le prix nobel de chimie conjointement à Sidney Altmann et Thomas Cech pour la découverte des propriétés catalytiques des ARNs. À l'exception des ARNt et des ARN ribosomaux, les études structurelles visaient essentiellement les protéines. Cech fut le premier à démontrer l'activité catalytique d'un ARN en utilisant l'intron auto-excisable (*self-splicing*) de *Tetrahymena thermophilia*. Depuis, plusieurs dizaines d'autres ribozymes ont été identifiés.

Pour l'instant, toutes les méthodes permettant d'obtenir une structure ou une famille de structures potentielles pour un ARN requièrent une grande expertise et un investissement considérable en temps et matériel. Je propose donc la

mise au point d'une approche informatique entièrement automatisée permettant de transformer une description symbolique minimale d'un ARN en une famille de structures plausibles. À la base, ce projet se voulait une automatisation de la méthode *Mc-Sym* (développée par François Major [63]) mais il s'est rapidement avéré nécessaire d'apporter des modifications majeures à certains concepts sous-jacents à cette méthode. La présente thèse démontre la possibilité d'une telle automatisation en décrivant la mise au point d'un système de modélisation automatisé. Les approches développées en cours de ce projet ont plusieurs retombées importantes dans le domaine de l'analyse de structures 3-D d'ARN. Les chapitres 3, 2 et 4 présentent ces aspects tout en exposant les bases nécessaires à la mise en place du système de modélisation qui sera présenté au chapitre 5.

1·2 Structure d'un ARN

La présente section se veut un survol-éclair des concepts de base permettant de formaliser la structure d'un ARN. Pour une description complète, le lecteur intéressé pourra consulter le livre *Principles of Nucleic Acid Structure* [79]. L'ARN est présent sous forme de polymère, une longue chaîne d'unités semblables répétées. Ces unités sont les nucléotides et sont présents sous quatre formes principales: adénosine (A), cytosine (C), guanosines (G) et uridine (U). Le nucléotide regroupe deux parties: le squelette et la base azotée, les nucléotides successifs étant reliés par le squelette. Le squelette lui-même peut ensuite être décomposé en deux parties: le ribose et le groupement phosphate. D'un type de nucléotide à l'autre, seule la base change. Les types de bases peuvent être regroupés en purines (la guanosine et l'adénosine) ou en pyrimidines (la cytosine et l'uridine). Deux détails différencient chimiquement l'ARN de l'ADN: la présence d'un oxygène supplémentaire sur le ribose (O2') et le remplacement de la thymidine de l'ADN par l'uridine. La figure 1·1 présente cette décomposition

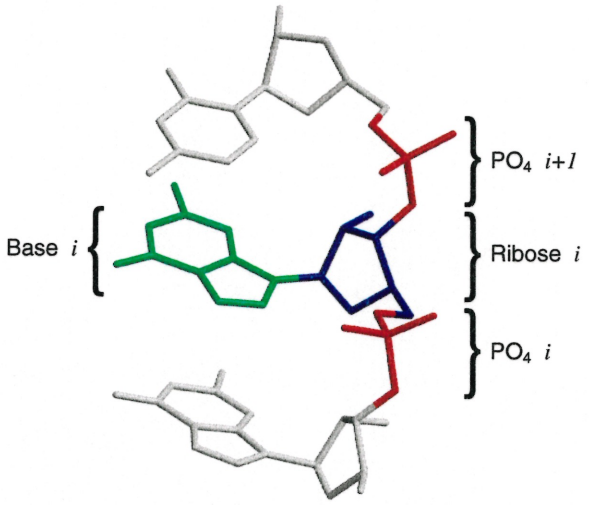


Figure 1·1 *Structure d'une chaine d'ARN. Le nucléotide i est composé de la Base i (vert), du ribose i (bleu) et du groupement phosphate i (rouge). La connectivité de la chaine est assuré par un lien covalent entre le ribose i et le groupement phosphate $i+1$ et entre le groupement phosphate i et le ribose $i-1$. Le ribose est la seule partie flexible de cette chaine.*

de la structure d'une chaine d'ARN.

Le repliement de l'ARN est, en majeure partie, gouverné par la formation d'appariements entre bases azotées stabilisés par des ponts hydrogène (pont-H). Les appariements canoniques sont le C•G et le A•U de type Watson-Crick ainsi que l'appariement de type Wobble G•U (voir figure 1·2). Ces appariements permettent la formation de double-hélices similaires à celle que l'on retrouve dans l'ADN, à la différence que l'appariement Wobble G•T n'existe pas dans l'ADN.

Le repliement de l'ARN diffère substantiellement de celui de l'ADN par la formation de nombreux appariements non-canoniques. La géométrie de ces appariements est très variable et est le sujet de plusieurs études (voir [20, 21, 55, 79]). Le chapitre 2 présentera en plus de détails les caractéristiques de ce type d'intéraction ainsi qu'une méthode permettant de les identifier dans une structure 3-D.

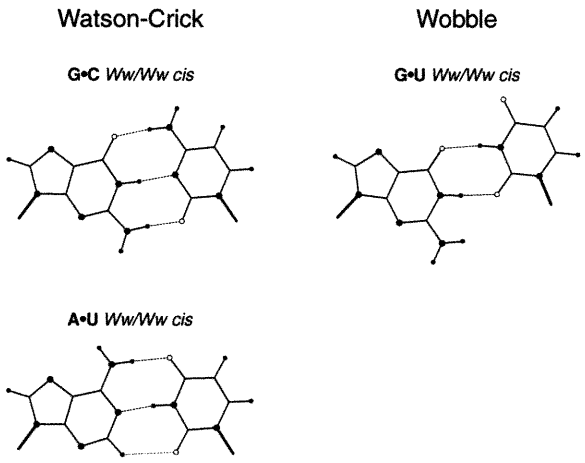


Figure 1·2 *Appariements canoniques dans l'ARN. Le lien glycosyl (entre la base et le ribose) est représenté par une ligne grasse, les atomes d'azote par de grands cercles pleins, les atomes d'oxygène des cercles vides, les atomes d'hydrogène par de petits cercles pleins et les ponts hydrogène par des pointillés.*

1·3 MÉTHODES PHYSIQUES DE DÉTERMINATION

La principale méthode utilisée pour déterminer la structure d'un ARN est la crystallographie par rayons X. L'ARN est mis dans des conditions (température, pression, sels) permettant de faire croître un crystal, organisant les molécules de manière régulière. Le crystal est ensuite bombardé de rayons X et les amplitudes des patrons de diffraction générés permettent de déterminer la structure 3-D de la molécule crystallisée. Dans certaines situations, cette approche permet l'obtention de structures très précises, mais les coûts d'application de cette méthode sont énormes. De plus, les conditions parfois extrèmes permettant la crystallisation risquent d'affecter le repliement de la molécule à étudier, et les artéfacts dûent aux bombardements sont encore peu caractérisés [78]. Cette méthode est applicable à tout ARN peu importe sa taille, sa propension à crystalliser étant le facteur limitant l'utilisation de cette méthode. Les structures de plusieurs ARNs de tailles supérieures à 100 nucléotides ont été déterminées par crystallographie, la structure la plus imposante ayant été déterminé au niveau atomique est la

grande sous-unité du ribosome (code PDB: 1FFK [5]), représentant plus de 2 500 nucléotides.

Une autre méthode couramment utilisée est la résonance magnétique nucléaire (RMN). Dans ce type d'expérience, la molécule demeure en solution et est soumise à de puissants champs magnétiques. Dans de telles circonstances, les protons entrent en vibration à des fréquences dépendant de certains angles et distances dans la molécule. Le spectre de résonance obtenu est ensuite utilisé pour dériver un grand nombre de contraintes de distances, d'angles et de torsions. En général, un algorithme de recuit simulé (dynamique moléculaire) est ensuite appliqué sur une représentation tout atome pour identifier la conformation satisfaisant à toutes ces contraintes. Ce type d'approche est aussi très coûteux en terme de matériel et de temps. L'assignation de résonance à des paires de protons précis est un problème combinatoire complexe et limite l'application de la RMN à de petits ARN, dont la taille est typiquement inférieure à environ 70 nucléotides. Un autre problème inhérent à la RMN est que les contraintes obtenues font toujours référence à des atomes situés à proximité les uns des autres dans l'espace, cette propriété fait que l'accumulation des imprécisions sur les contraintes rend la structure globale particulièrement imprécise. La RMN est en constant progrès, par exemple, la technique de couplage dipolaire permet de fournir des contraintes entre régions distantes de la molécule (voir [92] pour une application à la boucle sarcin-ricin).

Ces méthodes, en plus d'être très coûteuses, soulèvent aussi le problème que les données recueuillies le sont dans des conditions très particulières et rien ne garantie que ces conditions n'affectent pas la conformation adoptée par la molécule. Par exemple, la structure 3-D du ribozyme activé par le plomb (*leadzyme*) a été déterminée par crystallographie [94], par RMN [37] et, au cours de ma maîtrise, par modélisation à partir de données de modifications chimiques [49]. Les trois structures présentent des différences marquées. Malgré

qu'aucune des trois structures ne fourni un modèle expliquant parfaitement toutes les observations biochimiques faites à ce jour, la structure obtenue par modélisation semble la meilleure candidate pour une structure active du ribozyme.

1·4 DESCRIPTION DU PROBLÈME DE MODÉLISATION

D'une manière générale, on appelle *modèle* une représentation abstraite permettant de synthétiser un ensemble d'informations sur un système étudié. Dans le contexte de la biologie structurale un *modèle* est une représentation[1] de la structure satisfaisant aux informations connues sur la molécule étudiée. Le processus de modélisation d'une molécule consiste à construire un ou plusieurs *modèle(s)* de la molécule étudiée qui satisfait(ont) à ces informations. On désigne par *système de modélisation* un programme ou ensemble de programmes qui permet de passer d'une liste d'informations structurales à un ou plusieurs modèle(s) 3-D correspondant à ces informations. Ces systèmes peuvent être décomposés en deux parties: premièrement, la représentation du modèle et la méthode d'exploration dans cette représentation; deuxièmement, le formalisme utilisé pour représenter les informations structurelles et les mécanismes permettant d'évaluer un modèle en fonction de ces informations.

Le système élaboré dans le cadre de cette thèse est quelque peut différent puisqu'il commence d'abord par scinder le modèle à construire en une série de sous-problèmes de plus petite taille et de structure simple (voir chapitre 4) qui seront modélisés en parallèle (voir chapitre 5). Les modèles obtenus pour ces sous-problèmes seront ensuite réassemblés pour donner plusieurs modèles entiers (voir chapitre 5). Pour l'instant, il n'existe aucun mécanisme indépendant dans ce système pour l'évaluation d'informations structurelles. Ces informations sont introduites dans la définition de l'espace à explorer et seront nécessairement

[1]La forme de cette représentation peut aller d'un modèle en bois et fil de fer à la liste explicite des coordonnées 3-D des atomes dans un fichier informatique.

réalisées dans le modèle. Cette approche, par contre, présente le désavantage de contraindre la forme de ces informations.

1·5 Revue des systèmes utilisés présentement

Afin de donner une vision d'ensemble des méthodes utilisées pour modéliser des ARNs, la présente section décrira les principaux systèmes existants. Cette liste ne se veut en aucun cas exhaustive mais vise plutôt à fournir un tour d'horizon de la diversité de ces systèmes en décrivant les plus influants dans le domaine.

A. Simulation physique tout atomes

Les systèmes Amber [16] et CHARMM [10, 23, 61] sont les plus couramment utilisés pour les ARNs. La repésentation utilisée pour le modèle est l'énumération explicite des coordonnées cartésiennes de chaque atome. Les informations structurelles sont encodées sous forme de forces appliquées de telle sorte à rapprocher ou éloigner deux atomes, forcer un angle entre trois atomes ou forcer un angle de torsion entre quatre atomes. La recherche est faite en minimisant la fonction d'énérgie de la molécule ou en simulant la dynamique moléculaire à basse température (équations dynamiques de Newton, voir [66]). Ce type de système de modélisation est couramment utilisé pour construire des modèles à partir d'information de résonance magnétique nucléaire ou de crystallographie par rayons X. Ces systèmes sont aussi souvent utilisés comme étape de raffinement dans d'autre approches.

B. Construction par nucléotides rigides

NAB [60] (Nucleic Acid Builder) permet la construction d'une structure d'ARN en spécifiant directement sa géométrie. Il permet aussi d'interfacer avec un

module de dynamique moléculaire (Amber) et de faire appel à des fonctions de géométrie de distance.

Mc-Sym [52, 63] utilise une base de donnée de relations entre bases azotées et de conformations de nucléotides. Chaque nucléotide est placé en utilisant une relation avec un nucléotide précédant. L'espace de recherche est défini en spécifiant dans le script la taille de l'échantillon à utiliser pour chaque relation et chaque conformation. Un ensemble de contraintes est aussi spécifié dans le script pour représenter les informations stéréo-chimiques (collisions et connectivité) ou expérimentales (proximité, angle, formation de structures cycliques). Un algorithme de retour-arrière est ensuite utilisé pour identifier les structures satisfaisant aux contraintes.

C. Manipulation intéractive

Manip [64], développé dans le laboratoire d'Eric Westhof, permet de manipuler des structures d'ARN en important des fragments de structures connues et en appliquant un protocole de minimisation en temps réel. Afin d'obtenir l'intéractivité recherchée, le champ de force (fonction objectif à minimiser) utilisé est très restreint et ne contient que des critères de distances avec forces harmoniques (la fonction objectif est quadratique). Ce champ de force est beaucoup moins réaliste que celui utilisé dans les méthodes présentées sous la rubrique "Simulation physique tout atomes". Ce logiciel fut utilisé pour modéliser un grand nombre d'ARN: ARNt, ribozyme de l'hépatite δ, intron du groupe I, etc. Un autre système similaire est ERNA-3D [68]. Ce logiciel fut utilisé pour produire l'un des premiers modèle d'une sous-unité complète du ribosome.

Les méthodes de manipulation intéractives ont l'avantage de permettre une totale flexibilité au modélisateur. Lorsque de nouvelles expériences permettent d'obtenir de nouveaux types d'information sur une structure, le modélisateur

pourra tenir compte de ces informations lors de ses manipulations. L'expérience du modélisateur, à tous les niveaux, est accessible au processus de modélisation et aucune formalisation des informations structurelles n'est requise, laissant libre cours à la subjectivité et à l'intuition du modélisateur. Le principal désavantage est que la qualité de la structure résultante dépend de la performance et de l'expérience du modélisateur, il n'existe aucune métrique pour quantifier ces paramètres! L'expérience de modélisation devient aussi non-reproductible et fondamentalement biaisée d'une manière non quantifiable.

D. Representation restreinte

Yammp [85] permet la modélisation d'ARN en remplaçant chaque nucléotide par un nombre restreint de pseudo-atomes. Différents modes permettent de remplacer chaque hélice par un seul pseudo-atome, chaque nucléotide par un pseudo-atome ou encore de placer 3 pseudo-atomes par nucléotides. Le champ de force est dérivé de statistiques faites sur les structures connues ou contruit à partir de critères géométriques ou stéréo-chimiques. L'approche retenue pour explorer l'espace de cette représentation est la dynamique moléculaire (essentiellement un recuit simulé). Puisque cette approche possède l'avantage de produire relativement rapidement des modèles raisonnables pour de très grosses molécules, il fut utilisé pour construire plusieurs modèles préliminaires du ribosome. Un problème fondamental lié à cette approche est le fait que le modèle obtenu n'est pas un modèle tout atomes et ne peut donc être utilisé pour faire des analyses plus poussées de la structure (dynamique moléculaire, étude de liaison, calculs énergétiques). L'utilisation de ces modèles restreints comme échafaudage pour contraindre ou guider une méthode tout atomes est un domaine actif de recherche.

1·6 PRÉSENTATION DES CHAPITRES

Le chapitre 2 (article sous presse à la revue *Nucleic Acids Research*) présente une méthode permettant l'identification et la classification automatique d'appariements dans une structure 3-D d'ARN. La classification adoptée est une extension de celle proposée dans [56]. Les appariements sont classifiés en fonction des groupements chimiques formant les ponts-H, permettant une classification discrète de leur géométrie. Cette outil permet la création d'une base de donnée de tous les types d'appariements présents dans les structures connues d'ARN. Cette information est essentielle à l'obtention d'un échantillonage représentatif pour ce type de relation. Ces relations extraites des structures connues seront ensuite réutilisées lors de la modélisation.

Le chapitre 3 (article publié dans la revue *Journal of Molecular Biology*) présente, entre autre, la mise au point d'une métrique de distance permettant d'évaluer la différence entre deux relations entre bases azotées. Le chapitre présente l'utilisation de cette métrique pour l'analyse de structures 3-D d'ARN et pour l'évaluation de la complétude de la base de donnée de relations. Le rôle de cette métrique dans l'automatisation de la modélisation est de fournir un critère objectif pour la construction d'un échantillonage de relations entre bases azotées. Dans ce travail, j'ai participé au développement de l'ensemble des méthodes présentées et plus particulièrement à la section concernant l'évaluation de la complétude de la base de donnée de relations.

Le chapitre 4 (article en préparation pour la revue *Science*) met en place les bases théoriques permettant la mise au point d'un système automatique de modélisation. La méthode de décomposition en cycles correspond à l'obtention d'une base de l'espace des cycles et permet l'identification de motifs 3-D récurrents dans les structures. Cette méthode est appliquée à l'analyse de la structure de la grande sous-unité du ribosome (code PDB: 1FFK [5]). L'article souligne l'identification d'un nouveau motif 3-D mimant la structure d'une boucle

GNRA à l'aide de deux chaînes d'ARN.

Le système de modélisation résultant des outils présentés aux chapitres précédant est décrit au chapitre 5. On y présente d'abord l'utilisation de la base de l'espace des cycles comme un décomposition d'un problème de modélisation en un certain nombre de sous-problèmes plus simples (modélisation des cycles). Les méthodes présentées à ce chapitre se divisent en deux parties importantes, soit la modélisation de cycles de relations puis la reconstruction d'un modèle complet en combinant les résultats de la modélisation des cycles. Le chapitre est structuré en vu de sa publication dans une revue scientifique.

CHAPITRE 2

CANONICAL AND NON-CANONICAL BASE PAIRING TYPE RECOGNITION IN RNA THREE-DIMENSIONAL STRUCTURES

S. Lemieux et F. Major, *Nucleic Acids Research*, sous presse.

ABSTRACT

In this work, the problem of systematic and objective identification of canonical and non-canonical base pairs in RNA three-dimensional structures was studied. A probabilistic approach was applied, and an algorithm and its implementation in a computer program that detects and analyzes all the base pairs contained in RNA three-dimensional structures were developed. The algorithm objectively distinguishes among canonical and non-canonical base pairing types formed by three, two and one H-bonds, as well as those containing bifurcated and C-H...X H-bonds. The nodes of a bipartite graph are used to encode the donor and acceptor atoms of a three-dimensional structure. The capacities of the edges correspond to probabilities computed from the geometry of the donor and acceptor groups to form H-bonds. The maximum flow from donors to acceptors directly identifies base pairs and their types. A complete repertoire of base pairing types was built from the detected H-bonds of all X-ray crystal structures of a resolution of 3.0 Å or better, including the large and small ribosomal subunits. The base pairing types are labeled using an extension of the nomenclature recently introduced by Leontis and Westhof. The probabilistic method was implemented in MC-Annotate, an RNA structure analysis computer program used to determine the base pairing parameters of the three-dimensional modeling system *MC-Sym*.

Keywords: Hydrogen bond, base pairing types, RNA structure, probabilistic approach, computer algorithm.

2·1 INTRODUCTION

During the past year, two important RNA structures were determined at high resolution by x-ray crystallography: the large and small ribosomal subunits (PDB codes: 1FFK, 1FJG [5,97]). The addition of these two structures does not only confirm important progress that has been accomplished in the field of RNA crystallography, but also marks an important leap in the complexity of the available RNA 3-D structures, and in the difficulty of RNA structure analysis. Up until recently, there was no available tools to extract the useful information of RNA structures automatically, which hinders the effort to fully exploit them. An important paradigm switch in RNA structural analysis is needed, as the observation and discovery processes need to be automated so to provide the speed and objectivity that are necessary to fulfill our hopes towards these structures. A method that automatically identifies hydrogen-bonding patterns among nitrogen bases using the nomenclature proposed in [56] (fully described in [53]).

Hydrogen-bonding (H-bonding) patterns that form between nitrogen bases are particularly important interactions in RNAs. Efforts have been made towards the repository of base pairs from published litterature to show the diversity of nitrogen base pairing types with a particular emphasis on non-canonical ones [69], and a systematic nomenclature has been proposed [56]. From a modeler's perspective, the spatial relations defined by such H-bonding interactions can be used to define the conformational search space of RNA. For instance, in the RNA 3-D modeling software *MC-Sym* (www-lbit.iro.umontreal.ca/mcsym), these spatial relations are learnt from known examples and applied to the construction of new RNA structures [52]. In earlier versions of *MC-Sym* [63], the database was built from base pairs that were identified and annotated using interactive visualization. However, the number of newly determined RNA 3-D structures is such that it has become difficult to maintain the *MC-Sym* database updated simply by continuing to apply such a slow and subjective method. During

the development of an automated RNA 3-D structure annotation program, we realized that no objective method existed for identifying base pairing types in RNA 3-D structures. All currently available ones are limited to the detection of single H-bonds, and therefore, base pairing types must be identified in a further step by visual examination or by using heuristics (as in [59]). All existing methods detect H-bonds from the distance between either the hydrogen or donor atom and acceptor atom, such as in *Manip* [64], and the angle between the hydrogen, donor, and acceptor atoms, such as in the molecular graphics software *insightII* (Biosym/MSI) and HBexplore [59]. The use of such strict parameters is subject to false positives and negatives when applied to RNA 3-D structures that contain distorted base pairs, either due to experimental conditions, density map resolutions, or variations in the application of computer optimization protocols.

We present here a new method that resulted from the search of an automated and objective method for finding and identifying base pairing types in RNA 3-D structures. The probabilistic method provides a degree of certainty for the presence of each H-bond in the structure by considering the formation of H-bonds from competing donors and acceptors. This dependency between H-bonds that share a donor or an acceptor is implemented as a maximum flow problem in a bipartite graph. The decisions are thus taken to maximize the total number of expected H-bonds in a structure without involving a donor or acceptor more than once. The maximum flow problem formulation was adapted to search for an equilibrium solution that suits better the chemical nature of the problem. Base pairs are identified if the total flow, representing the mathematical expectation of the number of forming H-bonds, is higher than a predefined cutoff (typically 0.5). This cutoff can be varied depending on the application and on the desired sensibility of the detection process.

The only *a priori* knowledge used in selecting the parameters of the probabilistic approach is the near aligned geometry of H-bonds. The approach

consists in collecting all local geometries of donor/acceptor pairs, and building a model of this distribution. Using the fact of near aligned geometries, the model is decomposed between H-bond and non-H-bond geometries so that the probability of forming a H-bond is obtained by applying Bayes' theorem. A mixture of Gaussians (with full covariance matrices) was selected as the form of density function for the model, and the parameters of this mixture were optimized using the EM algorithm [18] from a data set extracted from physically determined RNA 3-D structures. The method is robust, reliable, and immune against local distortions due to experimental conditions and computer optimization protocols. The method was implemented in a newly developed RNA 3-D structure analysis computer program that is available on the internet (http://www-lbit.iro.umontreal.ca/). This method was also used to define the base pairing and base stacking parameters of *MC-Sym*, as well as for matching larger RNA 3-D patterns and motifs.

In order to identify a base pairing types, the naming scheme proposed by Leontis and Westhof [53] was used and extended. An algorithm that automatically name a base pairing using the information from the maximum flow optimization is presented. This algorithm was applied to 165 high resolution ($\leq$ 3 Å) X-ray structures in the PDB [8], HR-RNA-SET (see table 2·1 for the list). The collected base pairs were classified, resulting in a complete repertoire of the base pairing types in RNA structures (available at http://www-lbit.iro.umontreal.ca/).

Leontis-Westhof nomenclature:

This nomenclature (fully described in [53]) classifies base pairing types according to three properties: Glycosidic bond oriention (either Cis or Trans), Interacting edges and Local strand orientation (either Parallel or Antiparallel). In our nomenclature, we decided to ignore the Local strand orientation property since it is not defined by the spatial relation between two nitrogen bases.

The glycosidic bond orientation is "defined to be *cis* or *trans* with respect to an axis running parallel to and between the hydrogen bonds of the basepair". In the base pairing types that are used to examplify this definition, the drawn axis appears not to follow this definition making it unclear what is the actual definition used by the authors.

Three edges are defined by the authors on each nitrogen base. The interacting edges of a base pair are the edges of the nitrogen bases on which the H-bonds are forming. The edges are defined by the chemical groups forming H-bonds in each nitrogen base. The names used by the authors to describe the three edges are: Watson-Crick edge, the Shallow-groove edge, and the Hoogsteen (for purines) or 'C-H' edge (for pyrimidine). In [56], they changed the name of the Shallow-groove edge to Sugar edge, making the nomenclature more clearly independent with respect to the helical conformation of the nucleotides involved.

In their examples, the authors seem to favor the syntax *<Glycosidic bond orientation> <First interacting edge>/<Second interacting edge>*. The original nomenclature is unclear regarding the semantic of the order between the first and second interacting edge.

2·2 RESULTS

Our analysis of RNA 3-D structures led us to three main results. First, we developed a method to automatically identify base pairing types in RNA 3-D structure. Second, we refined an existing nomenclature and implemented its definitions in a computer program. Third, we built a repertoire of the base pairing types found in high-resolution RNA X-Ray structures.

157D	1D96	1E6T	1G2J	1QLN	1ZDJ	333D	429D	4TRA
165D	1D9D	1EC6	1GAX	1QRS	1ZDK	353D	430D	5MSF
1A34	1D9F	1EFO	1GID	1QRT	205D	354D	433D	6MSF
1A9N	1D9H	1EFW	1GSG	1QRU	246D	359D	434D	6TNA
1APG	1DDL	1EHZ	1GTR	1QTQ	247D	361D	435D	7MSF
1AQ3	1DDY	1ET4	1GTS	1QU2	248D	364D	437D	
1AQ4	1DFU	1EUY	1HDW	1QU3	255D	373D	438D	
1ASY	1DI2	1EVP	1HE0	1RMV	259D	377D	462D	
1ASZ	1DK1	1EVV	1HE6	1RNA	280D	397D	464D	
1AV6	1DNO	1EXD	1HMH	1RXA	283D	398D	466D	
1B23	1DNT	1F1T	1HQ1	1RXB	299D	3RAP	468D	
1B7F	1DNX	1F27	1MMS	1SDR	2A8V	3TRA	469D	
1BMV	1DPL	1F7Y	1OFX	1SER	2BBV	402D	470D	
1BR3	1DQF	1F8V	1OSU	1TN2	2FMT	404D	471D	
1BY4	1DQH	1FFK	1QA6	1TRA	2TRA	405D	472D	
1C0A	1DRZ	1FFY	1QBP	1TTT	300D	409D	479D	
1C9S	1DUH	1FG0	1QC0	1URN	301D	413D	480D	
1CSL	1DUL	1FIX	1QF4	1YFG	310D	419D	483D	
1CX0	1DUQ	1FJG	1QF5	1ZDH	315D	420D	485D	
1D4R	1DZS	1G1X	1QF6	1ZDI	332D	421D	4TNA	

Table 2·1 *HR-RNA-SET. The PDB identifiers of the X-ray RNA structures with a resolution of 3.0 Å or better. This list was compiled on February 1st, 2001. Two structures were removed from the list: 1QCU and 406D. These two structures contain multiple models with different chain identifiers and have improper MODEL/ENDMDL tags.*

A. BASE PAIR IDENTIFICATION METHOD

In order to guide the reader through the steps of this method, we exemplified each computation by using a canonical G•C Watson-Crick base pair extracted from positions A79 and B97 of the loop E motif from *E. coli* 5S rRNA (PDB code: 354D [17], Figure 2·1a shows this base pair). The method is divided in three steps:

1. compute the probabilities of H-bonds between each pair of donor and acceptor groups and build a graph representing these interactions;
2. compute the maximum flow in this graph to account for competing donors and acceptors;
3. assign the types of base pairs according to the probabilities of forming H-bonds.

PROBABILITY OF A H-BOND

For each base in the structure, the hydrogens are added according to geometries defined in [16]. Lone pair pseudo-atoms (LP) are added and placed at 1 Å of the oxygen or nitrogen atoms in the direction of the orbital. We use the term donor group to refer to a pair of associated donor and hydrogen atom and the term acceptor group to define a pair of associated acceptor and LP atoms.

Given the list of potential donor and acceptor groups for a 3-D structure, we compute the probability of forming a H-bond from the values of three measurements: the distance between the hydrogen and the LP atoms, the angle between the hydrogen, the donor and the acceptor atoms (refered to as the hydrogen angle), and the angle between the donor and acceptor, and the LP atoms (refered to as the LP angle). Figure 2·2 shows a H-bond with the identification of these three measurements.

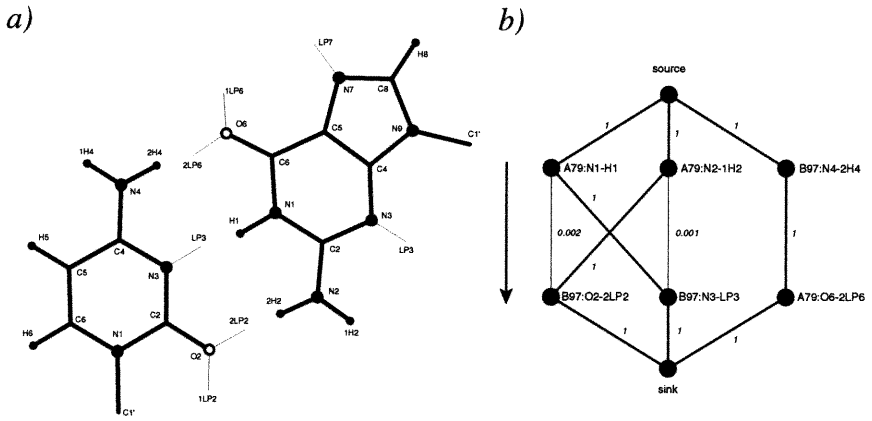


Figure 2·1 *A base pairing and associated graph. a) A canonical G•C Watson-Crick base pair extracted from positions A79 and B97 of the loop E motif from* E. coli *5S rRNA (PDB code: 354D). The thin lines indicate the direction of LP atoms, named using the same convention as for the hydrogen atoms. b) Corresponding graph showing the probabilities associated to this base pair (see table 2·2 for the actual measurements and probabilities). The donor groups are located in the upper row of nodes, and the acceptor groups in the bottom row. The arrow shows the direction of the flow from the source to the sink. The capacities are indicated beside each edge (only edges with capacity above* 10^{-4} *are shown). The thin lines show the edges with no flow after the optimization of the maximum flow. The thick lines between acceptor and donor groups correspond to the selected H-bonds.*

Bases	Acceptor and donor	x_1	x_2	x_3	$P(h\|\mathbf{x})$
A79→B97	C8-H8→O2 (1LP2)	9.971	3.087	2.573	2.127×10^{-20}
→	→O2 (2LP2)	8.239	3.087	0.480	1.346×10^{-22}
→	→N3 (LP3)	7.321	2.869	0.169	2.070×10^{-20}
→	N1-H1→O2 (1LP2)	3.928	0.586	2.719	1.212×10^{-9}
→	→O2 (2LP2)	2.377	0.586	0.628	0.002
→	→N3 (LP3)	1.023	0.076	0.089	0.999
→	N2-1H2→O2 (1LP2)	3.884	2.065	2.051	1.427×10^{-7}
→	→O2 (2LP2)	2.602	2.063	0.119	8.621×10^{-7}
→	→N3 (LP3)	4.090	2.708	0.727	4.252×10^{-9}
→	→O2 (1LP2)	2.580	0.049	2.051	2.688×10^{-8}
→	→O2 (2LP2)	0.968	0.049	0.119	0.999
→	→N3 (LP3)	2.541	0.614	0.727	0.001
B97→A79	N4-1H4→N7 (LP7)	6.359	2.133	1.384	6.505×10^{-14}
→	→O6 (1LP6)	3.831	2.138	1.961	1.985×10^{-7}
→	→O6 (2LP6)	2.651	2.138	0.158	1.005×10^{-6}
→	→N3 (LP3)	8.115	2.720	2.720	3.955×10^{-16}
→	N4-2H4→N7 (LP7)	4.917	0.053	1.384	2.282×10^{-15}
→	→O6 (1LP6)	2.457	0.044	1.961	5.521×10^{-8}
→	→O6 (2LP6)	0.946	0.044	0.158	0.999
→	→N3 (LP3)	6.436	0.635	2.720	2.055×10^{-15}
→	C5-H5→N7 (LP7)	8.599	2.004	1.540	1.216×10^{-18}
→	→O6 (1LP6)	6.101	1.914	2.212	1.220×10^{-13}
→	→O6 (2LP6)	4.599	1.914	0.138	6.995×10^{-12}
→	→N3 (LP3)	9.268	2.455	2.425	6.151×10^{-19}
→	C6-H6→N7 (LP7)	10.184	2.937	1.665	2.350×10^{-20}
→	→O6 (1LP6)	7.835	2.800	2.386	1.077×10^{-15}
→	→O6 (2LP6)	6.145	2.800	0.297	4.896×10^{-16}
→	→N3 (LP3)	9.587	2.927	2.254	7.826×10^{-19}

Table 2·2 *Base pair G:A79•C:B97 of the loop E motif from* E. coli *5S rRNA (354D). The three transformed measurements and the modeled probabilities are shown for each pair of donor and acceptor groups. The values were truncated to the third decimal. The names used to identify LP pseudo-atoms are built using the same rules as the standard PDB hydrogen atoms names.*

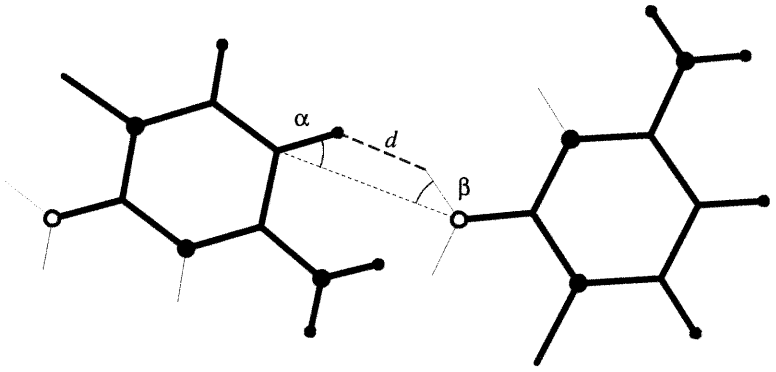


Figure 2·2 *H-bond parameters. The putative H-bond shown is a weak C-H...O. The hydrogen and LP angles are respectively identified by α and β, and the distance between the hydrogen and LP atoms is indicated by d. Nitrogen and hydrogen atoms are shown by filled circle, respectivly large and small. Oxygen atoms are shown with empty circles. Thin lines are used to indicate the direction of the LP pseudo-atoms.*

Our data set is built by extracting these values from all pairs of donor and acceptor groups in HR-RNA-SET (see table 2·1 in Material and Methods for the list of 3-D structures), resulting in a data set $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$, where $\mathbf{x}^i = (x_1^i, x_2^i, x_3^i)$ is a vector containing the distance, the hydrogen angle and the lone pair angle. To reduce the amount of data, we extracted only the values from pairs of residues that contain a pair of atoms at 3 Å of distance or less. The data set contained 1 607 756 data points.

To obtain both flexibility and efficiency, we applied a semi-empirical approach that models the distribution of data points by a sum of Gaussians. Because the geometrical nature of the measurements introduces a bias in the distribution of data points, the raw distributions of the extracted values cannot be directly modeled by a sum of Gaussians. To obtain a proper distribution, a transformation $\mathbf{x}' = F(\mathbf{x})$ was applied to each data point. This process is similar to histogram equalization in computer graphics [95], and allows us to transform any arbitrary distribution into another. Here, we wished to derive a

transformation so that the data points measured from randomly scattered points in space resulted in a uniform distribution, and thus to remove the geometrical bias. Such transformation was obtained by computing the cumulative probability density given the random model for each dimension of the data points. In the case of the distance, the cumulative probability density is proportional to the volume of a sphere of radius x_1. For the angles, the cumulative probability density is proportional to the volume of a spherical cone of angle x_2 (or x_3). The transformation we obtained is given by $F(\mathbf{x}) = [x_1^3, \cos(x_2), \cos(x_3)]$.

However, this transformation was unreasonable to model the distribution as a sum of Gaussians since only a specific range is accessible in each of the three dimensions of the data points ($x_1 \geq 0, 0 \leq x_{2,3} \leq 1$). To solve this problem, a further transformation was applied to the data points so that each dimension was distributed in $[-\infty, \infty]$. The complete transformation is then $F(\mathbf{x}) = (\ln(x_1^3), \operatorname{arctanh}(\cos(x_2)), \operatorname{arctanh}(\cos(x_3)))$. The distributions of the transformed data points are shown in Figure 2·3.

The distribution of transformed data points is modeled as a sum of Gaussians without any constraint on the mean vector and the covariance matrix. This model has the advantage of modeling the dependencies between the dimensions of the distribution. A possible drawback is the increase in the number of parameters, which brings the risk of overfitting the data [9]. However, our data points represent a large sample of the distribution, and in practice we didn't observed overfitting of the data. The parameters of the model (mean vector, covariance matrix and weight for each Gaussian) are optimized using the EM algorithm [9, 18]. To avoid local minima, a variant of the algorithm was used where only 25 000 randomly chosen data points were considered at each iteration. The EM algorithm is known to minimize the negative log-likelihood, and thus to return the parameters that maximize the likelihood of generating the data set. Initial values for the parameters were determined by visual inspection of

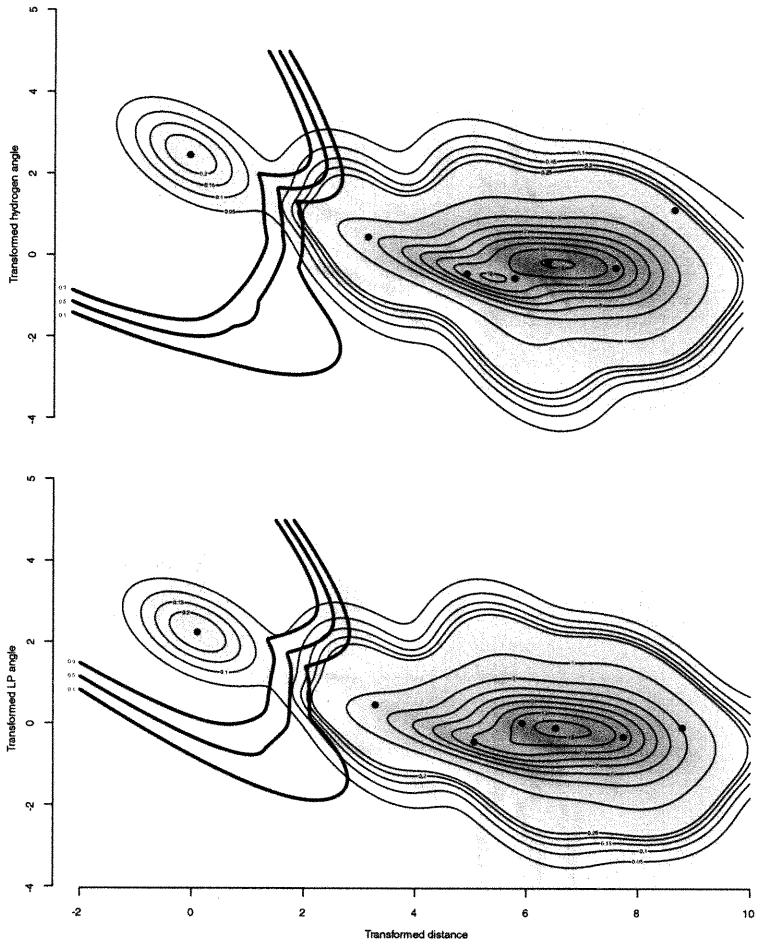


Figure 2·3 *Superimposed 2-D projections of the data set histogram, modeled probability density and surface of decision. The histogram of the data set is shown in shades of grey. The modeled probability density is shown by thin isocontours. Between 0 and 0.25 they were plotted at each 0.05 interval, whereas between 1 and 15 they were plotted at each interval of 1. An integration was done on the axis of projection corresponding to the effect observed by the histogram. The surface of decision is shown with thick lines isocontoured at probabilities 0.1, 0.5 and 0.9. The maximum probability is returned on the axis of projection. The circles represent the optimized mean of the seven Gaussians.*

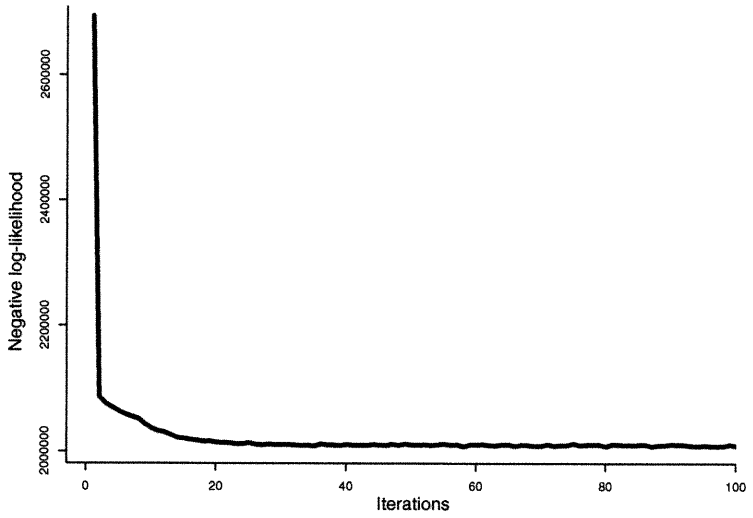


Figure 2·4 *Minimization of the negative log-likelihood for the mixture of seven unconstrained Gaussians on the transformed data set by the EM algorithm. The procedure was stopped after 100 steps, corresponding to 1 hour of CPU time on a PIII-600.*

the data set, and seven Gaussians provided an accurate model of the data set. Figure 2·4 shows the negative log-likelihood of generating the data set with the current parameters as the algorithm progresses. Once the parameters of the model are optimized, a visual inspection of the caracteristics of each Gaussian was sufficient to determine which one(s) is (are) responsible for the data points forming H-bonds. As a result, only one gaussian (the one centered on the smallest distance and angles) represents H-bonds. Table 2·3 shows the initial and final parameters of the seven Gaussians before and after optimization.

The probability that a local geometry, $\mathbf{x}$, forms a H-bond is equivalent to the probability that $\mathbf{x}$ is drawn from the Gaussian describing H-bond geometries, $H = h$, and not from the others. $P(H = h|\mathbf{x})$ can be computed using Bayes theorem:

$$P(H = h \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid H = h)P(H = h)}{p(\mathbf{x})}$$

Gaussian	Initial parameters			Optimized parameters		
	Weight	Mean	Covariance	Weight	Mean	Covariance
1	$\frac{1}{7}$	[0.0, 2.5, 2.5]	I	0.008	[0.101, 2.457, 2.252]	$\begin{bmatrix} 2.801 & 1.049 & 0.890 \\ 1.049 & 2.376 & -0.597 \\ 0.890 & -0.597 & 2.580 \end{bmatrix}$
2	$\frac{1}{7}$	[2.0, 2.5, 1.0]	I	0.010	[8.785, 1.132, -0.074]	$\begin{bmatrix} 0.173 & 0.293 & 0.036 \\ 0.293 & 1.021 & 0.193 \\ 0.036 & 0.193 & 1.751 \end{bmatrix}$
3	$\frac{1}{7}$	[2.0, 1.0, 1.0]	I	0.026	[3.287, 0.449, 0.474]	$\begin{bmatrix} 8.890 & 4.472 & 4.427 \\ 4.472 & 3.168 & 2.614 \\ 4.427 & 2.614 & 3.147 \end{bmatrix}$
4	$\frac{1}{7}$	[2.0, -0.5, 1.0]	I	0.110	[5.923, -0.554, 0.036]	$\begin{bmatrix} 3.190 & 0.842 & 0.863 \\ 0.842 & 0.753 & 0.317 \\ 0.863 & 0.317 & 0.839 \end{bmatrix}$
5	$\frac{1}{7}$	[3.7, 0.0, 0.0]	I	0.121	[5.065, -0.444, -0.425]	$\begin{bmatrix} 11.723 & 13.829 & 11.791 \\ 13.829 & 20.547 & 11.290 \\ 11.791 & 11.290 & 18.297 \end{bmatrix}$
6	$\frac{1}{7}$	[6.5, 0.5, 0.5]	I	0.535	[6.523, -0.165, -0.083]	$\begin{bmatrix} 0.907 & 0.523 & 0.614 \\ 0.523 & 3.271 & 0.548 \\ 0.614 & 0.548 & 3.370 \end{bmatrix}$
7	$\frac{1}{7}$	[8.0, -0.5, 0.5]	I	0.192	[7.736, -0.297, -0.300]	$\begin{bmatrix} 2.190 & 0.417 & 0.438 \\ 0.417 & 1.105 & 0.084 \\ 0.438 & 0.084 & 1.061 \end{bmatrix}$

Table 2·3 *Initial and optimized parameters. The initial parameters of the seven Gaussians are determined manually after examining the distributions of transformed measurements, equal weight and identity covariance are used. The optimized parameters are obtained after 100 steps of the EM algorithm. The values were truncated at the third decimal.*

$$= \frac{p(\mathbf{x} \mid H = h)P(h)}{\sum_{j=1}^{7} p(\mathbf{x} \mid H = j)P(H = j)}, \tag{2.1}$$

where $p(\mathbf{x} \mid H = h)$ is the probability of generating $\mathbf{x}$ from Gaussian h, $P(H = h)$ is the prior probability of forming a H-bond and $p(\mathbf{x})$ is the probability of observing geometry $\mathbf{x}$. Table 2·2 shows the measurements and modeled probability according to equation 2.1 for each pair of donor and acceptor groups for the G•C Watson-Crick base pair extracted from positions A79 and B97 of the loop E motif from *E. coli* 5S rRNA (PDB code: 354D [17]). Figure 2·3 shows the optimized model (thin black lines) superposed with the extracted data (grey shades).

STABLE SET

Consider a specific donor or acceptor group. We define as stable a set of one or more H-bonds that involve this donor or acceptor group if the sum of their associated probabilities (computed independently) is equal to or below 1. Consequently, one can interpret the probabilities as the proportion of time a group is occupied in the formation of each H-bond in a stable set (see Figure 2·5c). The stable set of a given group is chosen in order to maximize the total number of H-bonds in the structure. This is computed efficiently by defining a maximum flow problem on a directed bipartite graph connecting donors to acceptors. The graph, $G = (N, A)$, where N is the node set and A the arc set, is a bipartite graph that contain the set, I, of nodes for the donor groups, and the set, J, of nodes for all acceptor groups. If the probability of forming a H-bond between donor $i \in I$ and acceptor $j \in J$ is greater than 10^{-4}, an arc (i, j) is added to the graph with capacity, u_{ij}, equal to the probability of forming this H-bond. Two special nodes are then added to the graph, s and t, respectively called the source and the sink. Arcs that link the source to all donors, $(s, i) \in A \quad \forall i \in I$, and all acceptors to the sink, $(j, t) \in A \quad \forall j \in J$, are added with a capacity of 1. The maximum number of H-bonds that can form in the molecule is obtain by solving

the maximum flow problem of this graph from node s to t, resulting in values x_{ij} for $i \in I$ and $j \in J$, that indicate the resulting flow.

Algorithms that solve the maximum flow problem return an extremal solution [2]. In the context of H-bond probabilities, an extremal solution means that the algorithm, when faced with a situation where two equivalent H-bonds can form exclusively of one another, will favor the complete formation of one of the H-bonds, and leave the rest of the flow (typically 0) to the other. Since here we are more interested by the equilibrium state of the system, a criterion needs to be added, when allowed (the notation used is the one presented in [2]):

$$x_{ij} \geq x_{ik} \text{ or } x_{ij} = u_{ij} \quad \text{for} \quad i \in I \text{ and } j, k \in J \tag{2.2}$$

$$x_{ij} \geq x_{kj} \text{ or } x_{ij} = u_{ij} \quad \text{for} \quad i, k \in I \text{ and } j \in J. \tag{2.3}$$

This criterion is satisfied by modifying the preflow-push algorithm [30]. As the FIFO variant of the preflow-push algorithm (see [2] for a complete description of the algorithm, and [1] for theoretical and empirical performance comparisons) was selected for its simplicity of implementation, the *push/relabel()* operation was modified in the following way:

procedure *push/relabel(i)*;
begin
  let O be the set of admissible output arcs for node i;
  let n be the size of O;
  sort arcs $(i, j) \in O$ by their r_{ij};
  for $(i, j) \in O$ **do:**
    $\delta \leftarrow \min\{r_{ij}, e(i)/n\}$;
    $x_{ij} \leftarrow x_{ij} + \delta$;
    $e(i) \leftarrow e(i) - \delta$;
    $n \leftarrow n - 1$;
  if $e(i) > 0$ **then**
    let I be the set of admissible input arcs for node i;
    let n be the size of I;
    sort arcs $(i, j) \in I$ by their r_{ij};
    for $(i, j) \in I$ **do:**

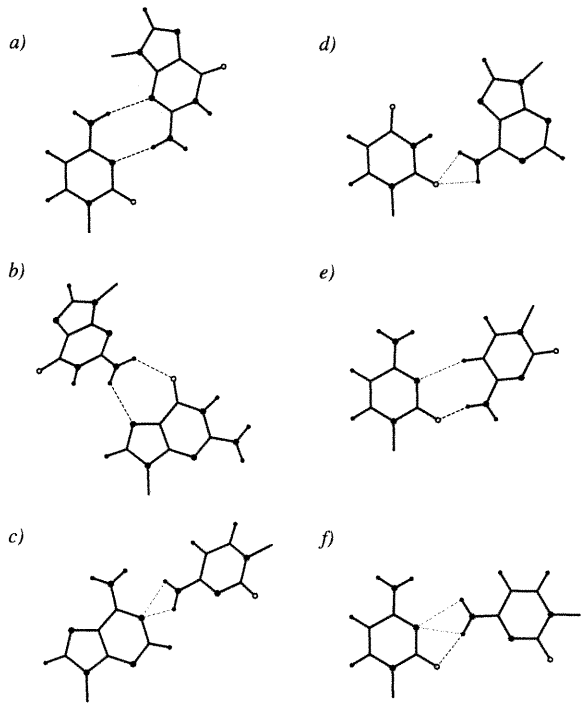


Figure 2·5 *Base pairing type examples. These were found in only one structure of HR-RNA-SET. a) The C•G* Ww/Ss trans *base pair found at positions '9'26:'9'22 and '9'46:'9'43 in 1FFK. b) The G•G* Hh/Bs trans *base pair found at position A260:A265 in 1FJG. c) The A•C* Ww/Bh cis *base pair found at position 38:32 in 1YFG. d) The U•A* Ws/Bh trans *base pair found at positions '0'1116:'0'1246, '0'1244:'0'1118 and '0'2661:'0'2812 in 1FFK. e) The C•C* Ww/Hh trans *base pair found at position '0'1834:'0'1841 in 1FFK. f) The C•C* Ww/Bh cis *base pair found at position '0'937:'0'1033 in 1FFK. The H-bonds are indicated by dotted lines. Empty, small-filled and filled circles are used for oxygen, hydrogen and nitrogen atoms respectively.*

$\delta \leftarrow \min\{r_{ij}, e(i)/n\}$;
$x_{ij} \leftarrow x_{ij} - \delta$;
$e(i) \leftarrow e(i) - \delta$;
$n \leftarrow n - 1$;
if $e(i) > 0$ **then**
$d(i) \leftarrow \min\{d(j) + 1 : (i,j) \in A(i)$ and $r_{ij} > 0\}$;
end;

Figure 2·6 shows the flows resulting from the computation of the stable H-bond set in HR-RNA-SET. In Figure 2·6a, both distributions of capacities and flows are shown. The distribution of Figure 2·6b shows the total flow obtained for every base pairs. The discrete character of this distribution suggests that a cutoff can be applied in the identification of base pairs with at least one H-bond, and thus assuming that a base pair forms only if the total flow between two bases is above or equal to 0.5. This parameter can be adjusted to reflect stringency of the identification process.

B. NOMENCLATURE

Several schemes were proposed to name RNA base pairing types [11,53,79,87]. The proposition from [56], LW, was retained, where a base pair is described by a pair of names that are associated to the faces of the bases involved. This nomenclature has several advantages. First, the names are easy to remember, and there is no need to reference any documentation. Second, the name alone gives a good idea of the base pair geometry. Third, isosteric pairs have the same name.

Despite these advantages, LW cannot differenciate base pairing types that differ by a sliding of the bases along the interacting faces, and especially in the context of single H-bond base pairs. Thus, to increase the precision of LW, we defined LW+, by decomposing the faces in sub-faces. Then, we defined and implemented an objective algorithm to reduce possible identification ambiguities

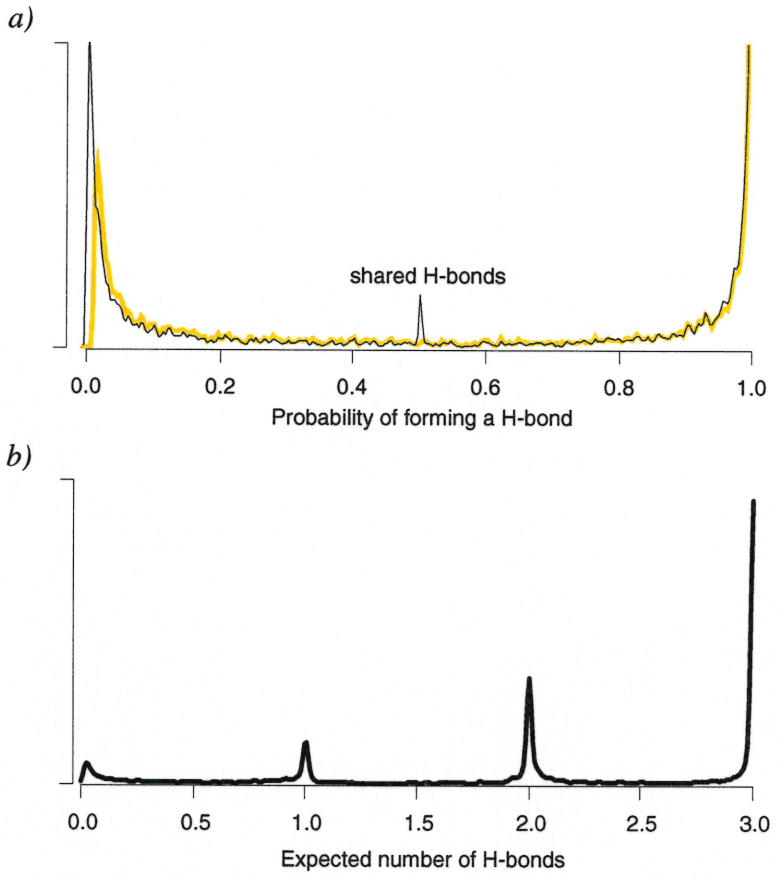


Figure 2·6 *Probability densities for x_{ij}, u_{ij} and the total flow of the base pairs. The probabilities were computed for all base pairs in HR-RNA-SET. Only those with a probability higher than 10^{-4} are plotted. a) The probability density for x_{ij} and u_{ij} are respectively shown with a thin black line and an orange line. The center peak for x_{ij} (the optimized flow) is the result of bifurcated H-bonds. b) The distribution of total flows obtained between every base pairs in HR-RNA-SET. The total flow can be seen as the mathematical expectation of the number of forming H-bonds between two bases. The distribution clearly shows the discrete nature of this value. The area of each peak shows the relative proportion of one, two and three H-bond base pairs.*

to anecdotal occurences. However, the current implementation does not support the detection of water-mediated, ribose- or phosphate-moeity involved base pairs.

Figure 2·7 shows the four RNA bases and associated faces. For convenience, the Watson-Crick edge was abbreviated by *W*, the Sugar edge by *S*, and the Hoogsteen/C-H edge by *H*. The sub-face names are indicated by combining face abbreviations, for instance *Ww* corresponds to the central section of the *W* face, and *Hw* to the section of the *H* face that is adjacent to the *W* face. Bifurcated base pairs of LW were renamed by creating small faces at the center of amino and keto groups. These faces are named *Bh* and *Bs*, respectively for the bifurcated base pairs involving the Hoogsteen side amino/keto group and the Sugar side amino/keto group. The C_2-H_2 group of the adenosine was named *Bs* to facilitate the idenfication of isosteric base pairing types (see Figure 2·7). We also introduced a special face, *C8*, for the C_8-H_8 donor group of the purines. The order of the faces is the same as the order of the bases. The *cis* and *trans* semantic for the relative orientiation of the glycosidic bond with respect to the base pair axis are the same as in LW. Note that the local strand orientation and base-sugar conformation are not specified in the base pair notation since they rather belong to nucleotide conformations.

The face involved in a base pairing type is obtained by computing the *contact point*, defined by the weighted mean of the hydrogen and LP atoms of each base. The weights correspond to the calculated probabilities of each H-bond, and as returned by the maximum flow algorithm. The face containing the *contact point* is returned.

To compute the glycosidic bond orientation the *visual contact point* is defined, a variant of the *contact point*, obtained by replacing the LP by the acceptor atoms. The vector between the two *visual contact points*, the *contact vector*, is used as the axis of the base pair, and the glycosidic bonds are attached to its extremities. A *cis* orientation is defined by a torsion around the contact

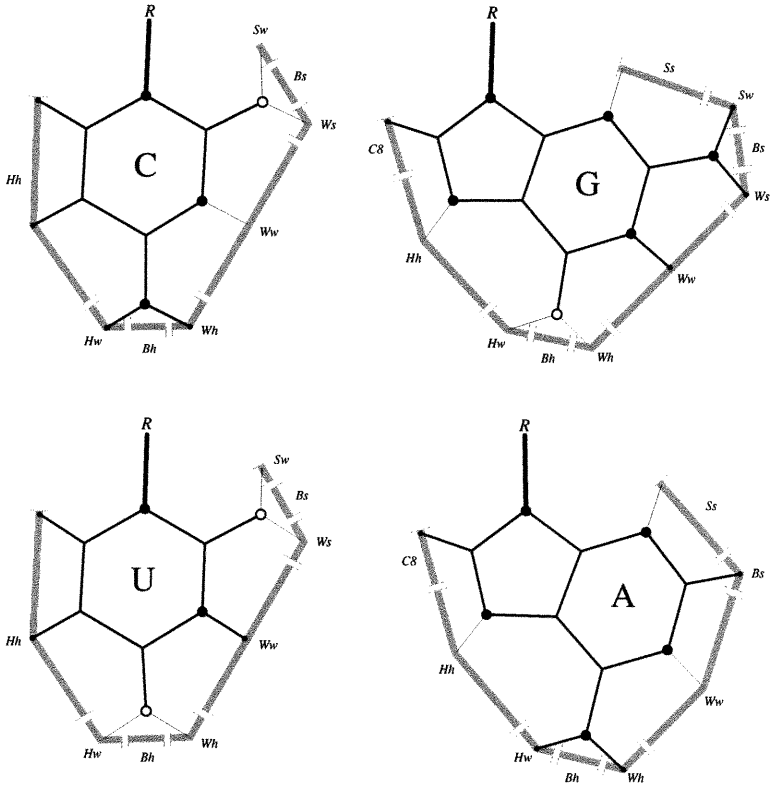


Figure 2·7 *RNA base faces. Nitrogen atoms are shown by a large black circle, hydrogen by a small black circle and oxygen atoms by an empty circle. The LP atoms are shown with thin lines. The ribose moiety is shown by the letter "R".*

vector below 90°, and the *trans* orientation otherwise.

C. Repertoire of base pairing types in RNA

The algorithm presented here allowed us to perform a systematic survey of all of the base pairs in high resolution X-ray RNA structures, and to study their geometrical diversity. For HR-RNA-SET, the complete repertoire was built in less than four minutes on a PIII-600. Figures 2·8 and 2·9 presents 38 base pairing types that occur at least twice in HR-RNA-SET. Because of space constraints, base pairing types that form only one H-bond were not included in this survey. The structure that minimizes the sum of RMSD [43,44] with all other base pairs of the same type is shown. Structure and position information about these specific base pairs is shown in table 2·4. In order to optimize the identification of representative base pairs, the RMSD calculation was limited to the first 200 examples for each base pair type. These results are also available in PDF (Portable Document Format) documents that include the superimposition of all the base pairs of the same type (see various documents about base pair types at our Web site www-lbit.iro.umontreal.ca).

The base pair types that appear in only one structure in HR-RNA-SET were examined. Figure 2·5 shows six, among 86, such examples that we found of particular interest.

Figure 2·5*a* shows a C•G *Ww/Ss trans* that was found in positions '9'26:'9'22 and '9'46:'9'43 of the ribosomal 5S subunit (1FFK). This two H-bonds base pair was not described by [20,21]. The two examples of the 5S subunit of *H. marismortui* are located 23 Å apart, and were found in very different 3-D contexts. The '9'46:'9'43 base pair is a member of a base triplet ('9'46:'9'43:'9'37) that stabilizes a local phosphodiester chain reversal of an unusual 13-nt loop between positions '9'33 and '9'47. The other base pair of this

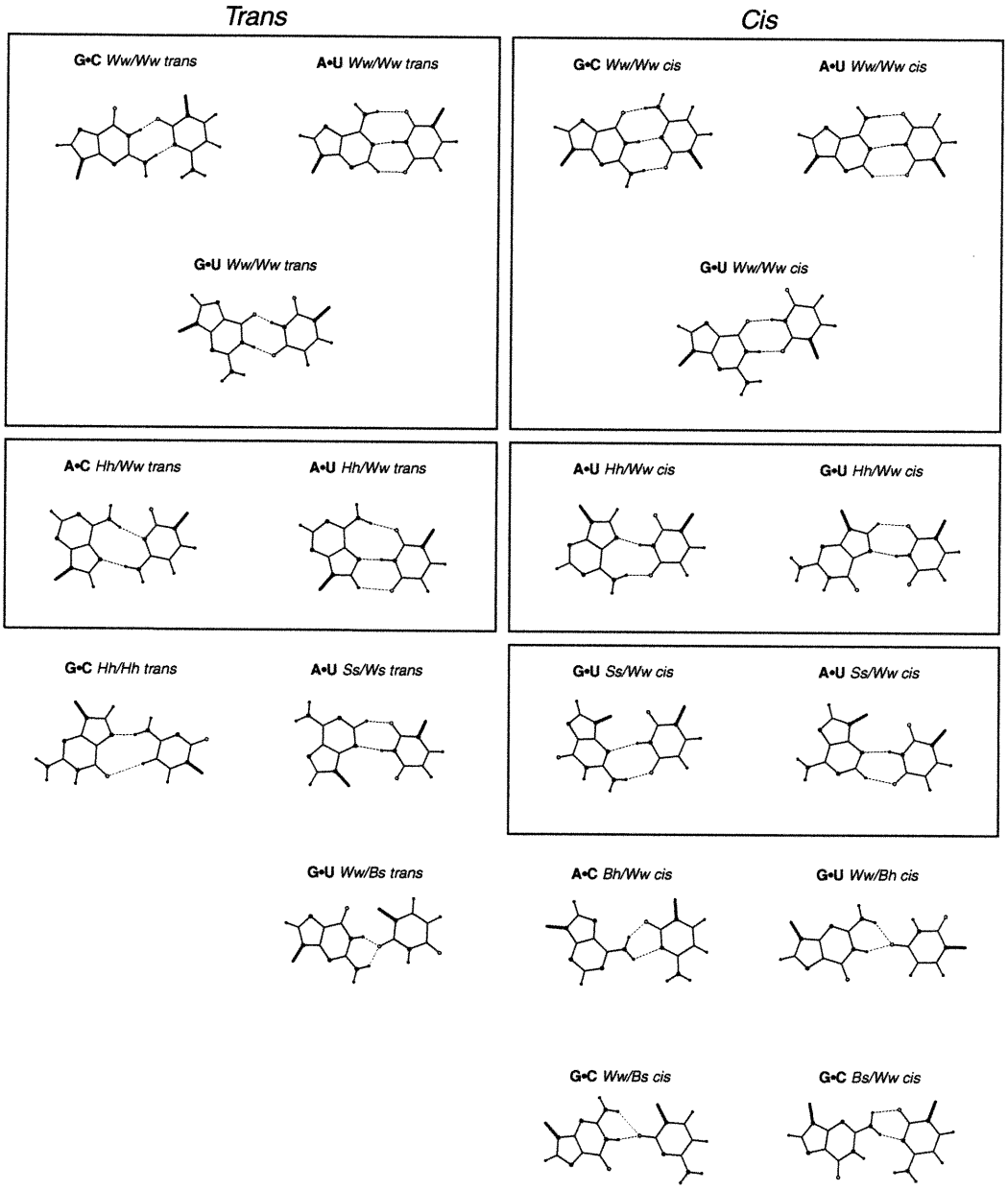


Figure 2·8 *Two H-bonds base pairing types found in HR-RNA-SET. Base pairing types that occur at least twice are shown. The 19 purine•pyrimidines base pairing types are on the left side of the page. The 15 purine•purine base pairing types are on the right side of the page. The 4 pyrimidine•pyrimidine base pairing types are located at the bottom right corner of the page. Base pairing types were classified as either* trans *(left column) or* cis *(right column). Boxes are used to group isosteric base pairing types together.*

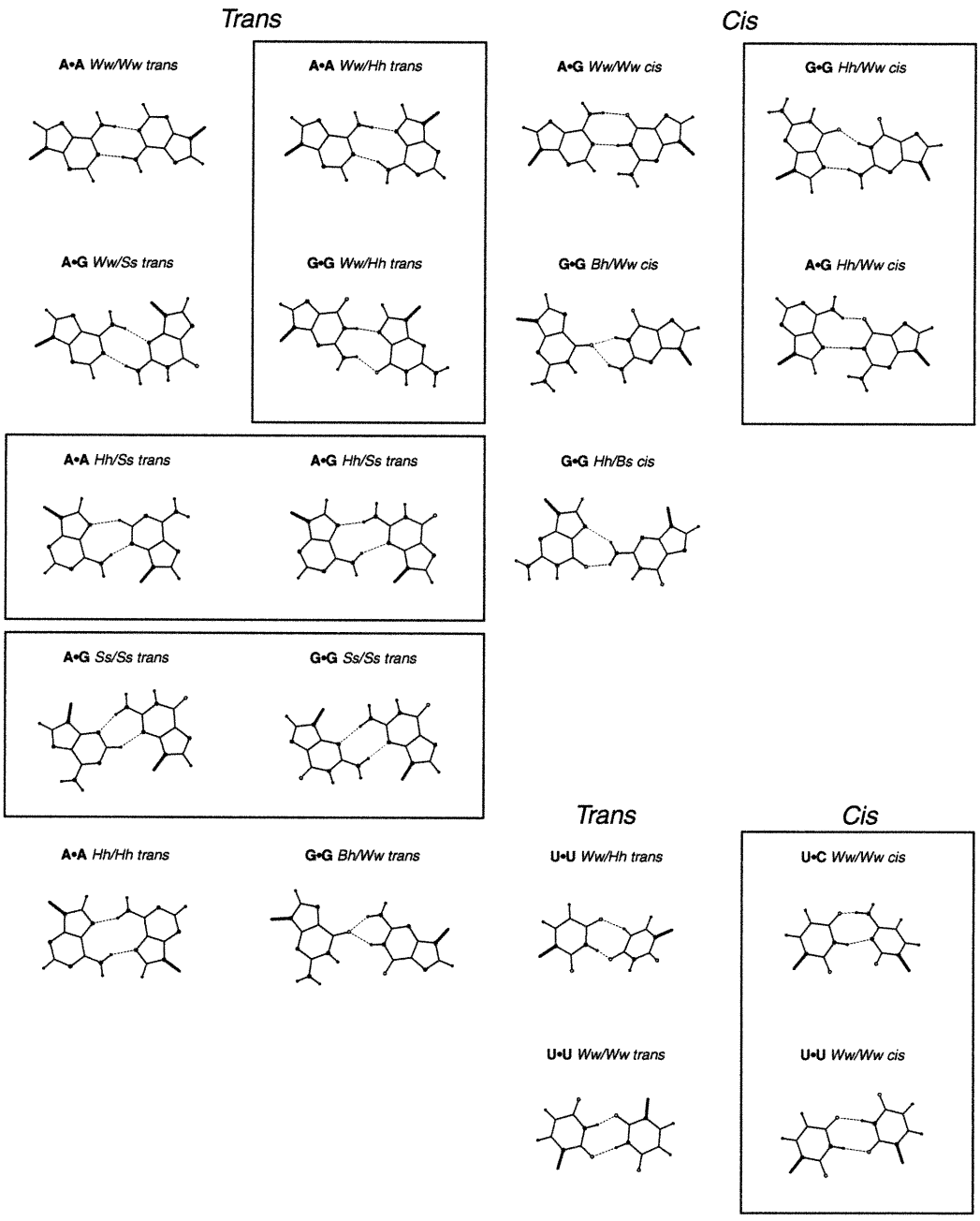


Figure 2·9 *Two H-bonds base pairing types found in HR-RNA-SET (continued).*

Base types	Pairing type		Nb.	Example shown	
		Purine–Purine			
A • A	*Hh/Hh*	*trans*	41	1ASY	S609 – S623
A • A	*Ww/Ww*	*trans*	22	1GID	B151 – B248
A • A	*Hh/Ss*	*trans*	7	1GTS	B22 – B13
A • A	*Ww/Hh*	*trans*	11	1FJG	A411 – A430
A • G	*Ww/Ww*	*cis*	54	1DUL	B161 – B150
A • G	*Hh/Ww*	*cis*	3	1G1X	D665 – E724
A • G	*Hh/Ss*	*trans*	121	1FFK	0 1372–0 2053
A • G	*Ss/Ss*	*trans*	39	1FFK	0 1632–0 1568
A • G	*Ww/Ss*	*trans*	15	1FFK	0 629 –0 2070
G • G	*Hh/Bs*	*cis*	3	3TRA	10 – 45
G • G	*Hh/Ww*	*cis*	25	1ET4	C428 – C410
G • G	*Bh/Ww*	*cis*	5	1D4R	B13 – A16
G • G	*Ww/Bh*	*trans*	2	364D	B76 – C100
G • G	*Hh/Ww*	*trans*	8	1GAX	D921 – D945
G • G	*Ss/Ss*	*trans*	6	1FG0	A2428–A2466
		Purine–Pyrimidine			
A • C	*Bh/Ww*	*cis*	2	364D	C109 – A11
A • C	*Hh/Ww*	*trans*	17	1FJG	A171 – A150
A • U	*Ww/Ww*	*cis*	730	1D4R	B26 – A3
A • U	*Hh/Ww*	*cis*	21	1QA6	D138 – D110
A • U	*Ss/Ww*	*cis*	3	1FFK	0 2083–0 2063
A • U	*Hh/Ww*	*trans*	109	1FJG	A496 – A437
A • U	*Ww/Ww*	*trans*	23	1ASZ	S615 – S648
A • U	*Ss/Ww*	*trans*	3	1FFK	0 761 – 0 645
G • C	*Ww/Ww*	*cis*	2229	1DI2	C2 – D19
G • C	*Ww/Bs*	*cis*	2	1FFK	0 1302–0 1353
G • C	*Bs/Ww*	*cis*	2	1G1X	I588 – I651
G • C	*Ww/Ww*	*trans*	30	1FFY	T15 – T48
G • C	*Hh/Hh*	*trans*	2	1FFK	0 2397–0 2391
G • U	*Ww/Ww*	*cis*	264	1ASZ	R610 – R625
G • U	*Hh/Ww*	*cis*	4	1FG0	A2471–A2278
G • U	*Ww/Bh*	*cis*	2	354D	B102 – A74
G • U	*Ss/Ww*	*cis*	2	1FJG	A362 – A49
G • U	*Ww/Bs*	*trans*	5	1GTR	B18 – B55
G • U	*Ww/Ww*	*trans*	2	1EXD	B915 – B948
		Pyrimidine–Pyrimidine			
U • U	*Ww/Ww*	*cis*	25	280D	C31 – D42
U • U	*Ww/Hh*	*trans*	8	1ET4	E127 – E115
U • U	*Ww/Ww*	*trans*	3	1FJG	A956 – A960
U • C	*Ww/Ww*	*cis*	2	1FFK	0 1702–0 1545

Table 2·4 *The 38 base pairing types in HR-RNA-SET. Each base pairing type was found at least twice in HR-RNA-SET. The example selected for each type for Figures 2·8 and 2·9 is identified in the last column. The four letter code refers to the PDB identifier.*

type, at positions '9'26:'9'22, stabilizes a disordered internal loop. It is worth noting here that a theoretically generated example of this base pair type was included in the *MC-Sym* modeling system [63] since its very first version, as the 119 base pair.

Figure 2·5*b* shows a base pair of type G•G *Hh/Bs trans* found at positions A260: A265 in the structure of *T. thermophilus* 30S ribosomal subunit (1FJG). This two H-bonds base pair was also not described by [20,21], but, again, was theoretically generated for the first version of the *MC-Sym* database. It was referred to as base pair 34. This base pair is flanking a 7-nt loop that interacts with protein S20.

Figure 2·5*c* shows a base pair of type A•C *Ww/Bh cis* found at positions 38:32 of the yeast initiator tRNA (1YFG). Here, we use the term bifurcated to qualify a base pair in which two H-bonds either share the same hydrogen or LP atoms. The equilibrated maximum flow settles the probability of each H-bond to values close to 0.5, expressing the shared nature of the interaction, and hence the pairing *c* of Figure 2·5 is a perfect example of a bifurcated base pair. The base pair of type U•A *Ws/Bh trans* presented in *d* of the same figure is another example of a bifurcated base pair, as found at positions '0'1116:'0'1246, '0'1244,'0'1118 and '0'2661:'0'2812 of 1FFK.

Figure 2·5*e* presents a base pair of type C•C *Ww/Hh trans* found at positions '0'1834:'0'1841 of structure 1FFK. This non-canonical base pair closes a short helix, and stabilizes a bulged out adenosine and a 6-nt loop. The interaction is maintained by a H-bond between the extra cyclic amino of one C to the oxygen of the other base, and by the formation of a weaker C-H...N H-bond. Note that these H-bonds were included in the H-bond data set used to optimize the parameters of the mixture of Gaussians, and although they usually exhibit geometrical parameters slightly different than the other types of H-bonds, they are properly identified by the probabilistic model.

Figure 2·5*f* shows a convoluted network of three partial H-bonds obtained after the resolution of the equilibrated maximum flow problem. The base pair was observed at positions '0'937:'0'1033 in 1FFK, the first non-canonical base pair of a 10-nt internal loop that is adjacent to a G•A sheared tandem. The H-bond network describes a double bifurcated base pair, as the LP atom of N3 is shared between both hydrogens of the extra cyclic amino group, and one of these hydrogens is in turn shared with one of the LPs of the O2 atom.

The probability returned for each H-bond by the maximum flow optimization is such that their sum is maximized, while respecting the stable set property. The base pair is detected and correctly classified by the probabilistic system despite its peculiar geometry.

2·3 DISCUSSION

A. DISTANCE VS PROBABILISTIC MODELS

The most employed distance to recognize H-bonds is the one between the donor and acceptor atoms, d_{D-A}, which is easy to compute and to observe interactively, and it does not require neither the hydrogen or LP atoms. Figure 2·10 presents the distributions of three distances as measured from HR-RNA-SET. The distribution of d_{D-A} on Figure 2·10 (black line) does not contain a clear separation between H-bonds (first peak) and non H-bonds, and thus does not provide a good classification criterion. The distance used in [64], between the hydrogen and the acceptor atoms, d_{H-A}, is a better one, as shown by the green line on Figure 2·10. Massire and Westhof suggested a cutoff at 2.1 Å, but from the distribution in Figure 2·10, a cutoff at 2.4 Å would be a better solution. The 2.1 Å cutoff was retained to reduce the number of false negatives in the context of molecular modeling (personal communication). Finally, as indicated in Figure 2·10, the distance between the hydrogen and LP atoms, d_{H-LP}, among the

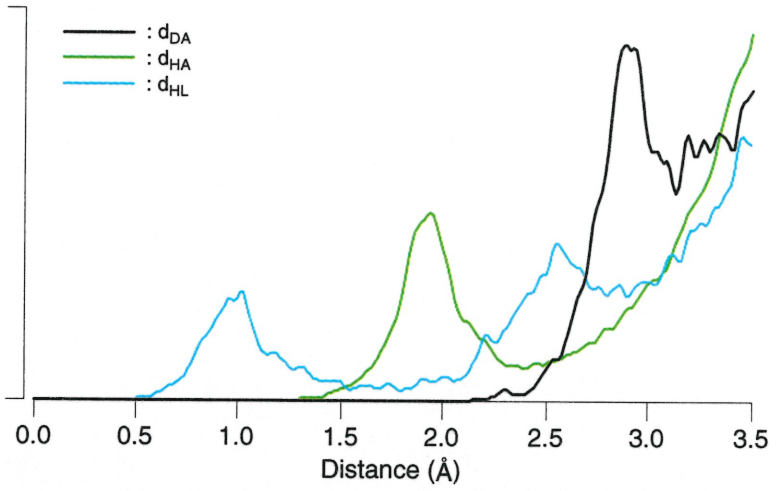


Figure 2·10 *Distance-based parameters. The distributions are computed for all base pairs in HR-RNA-SET. The black line shows the distribution of distances between the donor and acceptor atoms, d_{DA}. The yellow line shows the distribution of distances between the hydrogen and acceptor atoms, d_{HA}. The blue line shows the distribution of distances between the hydrogen and LP atoms, d_{HL}.*

three distances is the best, if only one distance must be used. As indicated from the blue line distribution in Figure 2·10, a cutoff between 1.5 and 1.8 Å would be effective for d_{H-LP}.

In order to quantify the power of using a probabilistic over the strict distance approach, a scattered plot where each dot represents one putative H-bond was created. Figure 2·11 shows that a significant number of H-bonds were assigned a probability 0 by using the probabilistic method, whereas they would have been identified as forming H-bonds using d_{H-A} with a cutoff at 2.1 Å, and as proposed by Massire and Westhof. Moreover, most of the H-bonds that were assigned a probability of 1 using the probabilistic model would have been rejected by the distance method.

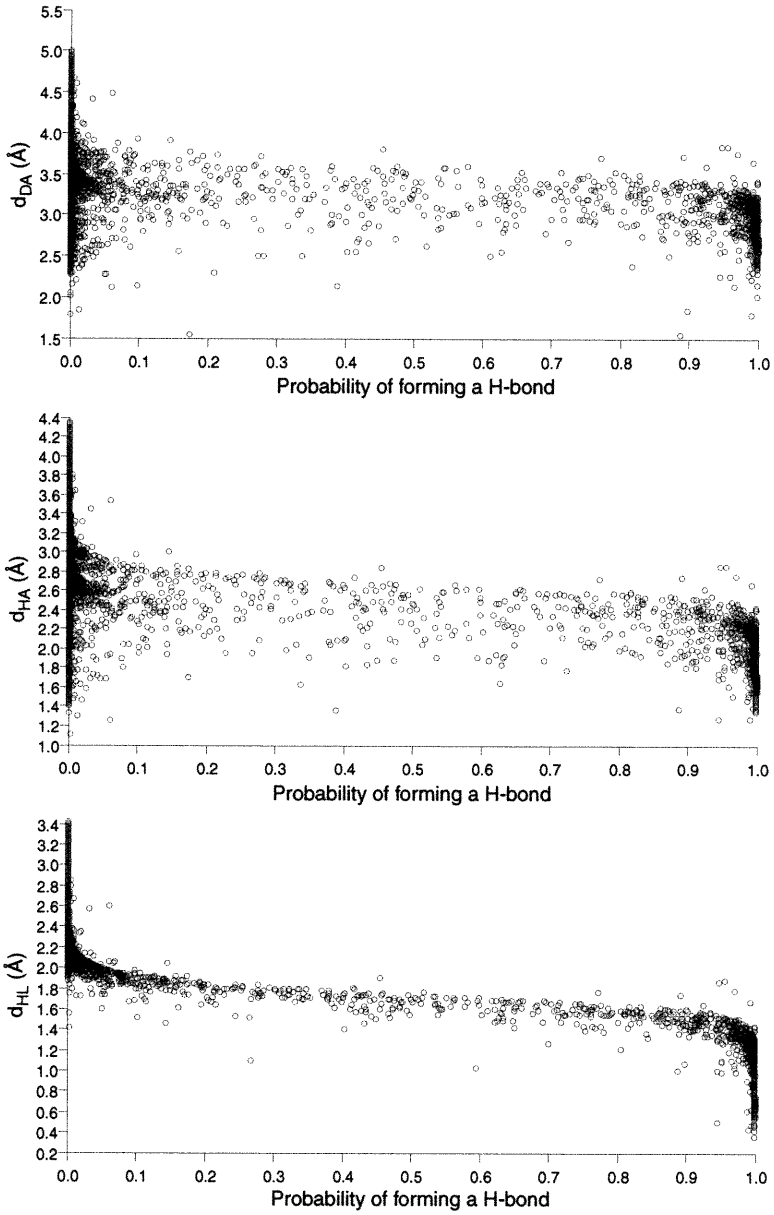


Figure 2·11 *Distance criteria vs. probabilities of forming H-bonds. Each scatter plot shows the correlation between a distance criterion and the probabilities of forming H-bonds. Each dot represents the evaluation of a pair of donor and acceptor groups. The pairs separated by more than 5 Å were not considered.*

B. STRICTNESS PARAMETER

Our probabilistic method returns the mathematical expectation of the number of forming H-bonds between two nitrogen bases. As a default value, two bases are identified as making an "interaction" if the expected number of H-bonds is greater or equal to 0.5. This value can be redefined by the user to reflect the type of interactions that need to be identified. In a context where a structure has been determined imprecisely, the cutoff can be lowered to a value as low as 10^{-4}. However, if only the strong two H-bond base pairs are desired in the output, the value of the cutoff could be raised to as much as 1.8. As an example, during the determination of the 3-D structure of the catalytic core of the hairpin ribozyme, a weak cutoff of (10^{-4}) was used to examine the first generation of thousands of structures that were obtained from secondary structure and low-resolution experimental data. This is a typical first step in RNA 3-D modeling. In several generated structures, the probabilistic method detected a H-bonding pattern that formed a base triple involving two bases in the ribozyme and one base in the substrate. The geometry of the base pairs in the first generation of structures were far from satisfying the strong H-bonding parameters. Nevertheless, this observation was reported to the experimentalists who decided to check for the presence of the triple in the hairpin. The predicted triple was later experimentally determined to form in at least one of the catalytic reaction steps [76]. In the further modeling iterations, a more stringent cutoff, typically 0.5, was used to identify generated 3-D structures that contained "nicer" base pairs.

C. MC-SYM BASE PAIRS

The probabilistic method was applied to the annotation of all available RNA 3-D structures. The identified base pairs were collected and corresponding transformation matrices inserted in the *MC-Sym* RNA 3-D modeling computer

program database. The previous *MC-Sym* databases were built from visual examination of all RNA 3-D structures, a long and subjective process. With the determination of the ribosome structure, a visual annotation would have been hardious. The probabilistic method, on the other hand, is automatic, fast and objective. It completed the base pair recognition process with a throughput of 7042 base pairs per second on a PIII-600. Now, everytime a new RNA 3-D structure is made available to us, the *MC-Sym* database and parameters are entirely updated to address the most recent knowledge brought by the new structure in less than four minutes. The most recent *MC-Sym* database contains ten times more nitrogen base spatial relations than the original version of 1991.

D. Distorsion in RNA structure databases

During the computation of the probabilities of all H-bonds in all available RNA structures, the base pairs that were assigned an expected number of H-bonds near 0.5 were visualized and analyzed. Some of these base pairs pointed us to interesting features of the RNA 3-D structures that are currently in public databases. First, several structures that contain stable Watson-Crick G•C base pairs are distorded, which could be the result of the refinement process where H-bonds are represented by simple harmonic restraints on the distance between the donor and acceptor atoms. The mean distance for H-bonds changes from one structure to another, and can even sometimes reach a value of 3.9 Å, for the H-bond between C:N4 and G:O6 (see for instance 1AOI). We believe this kind of variation can be explained from the use of different force fields and refinement parameters and procedures. Given the observed variations, it becomes obvious that methods based on strict distance and angle values are proned to identification errors, and hence the use of a more flexible approach, such as the one presented here, is strongly recommended for an objective analysis of RNA 3-D structures.

E. Ribosome contribution

When structures of the large and small ribosomal subunits were introduced in the database, it was believed that they would substantially contribute to RNA structural knowledge. During the building of the repertoire of two H-bonds base pairs, we determined that these two structures alone account for 1522 base pairs among a total of 3852 that were indexed, and thus represent 40% of the base pairs in HR-RNA-SET. Furthermore, non-canonical base pairs are often referred as appearing rarely [69], but our analysis revealed that G•C and A•U *Ww/Ww cis*, the canonical Watson-Crick base pairs, account for 77% of the total, where the G•C base pair accounts for 58% alone. This leaves 23% of non-canonical base pairs. If we remove the G•U *Ww/Ww cis* base pair, the wobble, the non-canonical base pairs still represent slightly over 16% of the indexed base pairs in the repertoire. The results of this analysis covers 629 base pairs, excluding those that require a water mediated H-bond or a protonated nitrogen base. The repertoire in Figures 2·8 and 2·9 contains 38 base pairing types that contain at least two H-bonds. Seven base pairing types are formed by one typical H-bond and a weaker C–H ... {O,N}.

F. Nomenclature

Leontis and Westhof have emphasized [56] that their proposed nomenclature has the interesting property of naming all isosteric base pairing types with the same name. This feature is of utmost importance since it allows one to easily describe RNA motifs without having to specify different base pairing types that correspond to sequence variations. This important feature is also a characteristic of LW+, and goes beyond by discriminating base pairing types that differ only by a sliding along the pairing faces.

An important exception to this is the G•U *W/W trans*, which occur in

two different forms that involve two H-bonds of the *W* faces. The first form involves two H-bonds on the *h* side of the *W* face, and the second form involves two H-bonds on the *s* side of the *W* face. Because the contact points represent an average when two H-bonds are present, it is impossible with this approach to modify the face definitions so that these two base pairing types can be differentiated, and without introducing undesired new names for each variation of the classic A•U *Hh/Ww trans* and A•U *Ww/Ww cis*. This is the only ambiguity left in the proposed LW+ nomenclature. The situation could be resolved by introducing an exception, by naming both base pairing types G•U *Wh/Wh trans* and *Ws/Ws trans*. We decided to postpone the implementation of such an exception until proper feedback is obtained from the RNA community.

In LW, the presence of bifurcated H-bonds has to be notified explicitely in the name. This is due to the fact that such base pairs often involve hydrogens or LP atoms from two different faces on one of the bases. The introduction of the *contact points* alleviates this ambiguity, and the addition of the *Bh* and *Bs* faces results in precise names.

The current probabilistic system does not identify water-mediated H-bonds because most of the currently published RNA structures do not contain water molecules, and when they do most of them do not specify the actual positions of the water hydrogen atoms. Identification of water mediated H-bond in an automated manner requires the correct placement of water molecules around the nitrogen bases, which is known to be a difficult problem.

Another limitation of the probabilistic system is that H-bonds involving the $O_{2'}$ group in the ribose moiety are not considered. Again, this is due to the fact that an automated method requires the exact position of the hydrogen atom. The H is free to rotate around the O2' group, and thus the task of computing its optimal position is not trivial, although currently under our investigation.

2·4 CONCLUSION

The probabilistic method introduced here describes the first available algorithm and computer implementation of an automated base pairing type recognition procedure, which also objectively classifies and presents the base pairs of an RNA 3-D structure. The probabilistic method successfully recognized all base pairing types that are present in available RNA 3-D structures, and allowed us to automate their classification. In particular, a complete and well-organized repertoire of observed RNA base pairing types has been made available on the Internet.

The systematic annotation of all RNA 3-D structures, as determined by high-resolution crystallography, provided us with a convincing confirmation that a slightly revised version of the nomenclature proposed by Leontis & Westhof [56] is perfectly suitable to a high-throughput RNA structure analysis context.

2·5 MATERIALS AND METHODS

The software was developped using the *MC-Sym* development library under the Linux operating system, which is publicly available at mccore.sourceforge.net. The code is written in C++, and, therefore, is easily portable to other Unix platforms, such as IRIX and SunOS. The probabilistic method has been integrated to the *MC-Annotate* system [29], and is accessible on the Web. RNA 3-D structures can be submitted for the identification of base pairing types and complete analysis at

www-lbit.iro.umontreal.ca/mcannotate.

The subset of PDB structures used in this work, HR-RNA-SET, is composed of those that contain at least one RNA nucleotide, and that were determined by X-ray crystallography with a resolution of 3 Å or less, as of February 1st, 2001. Table 2·1 shows the list of 3-D structures that are included in HR-RNA-SET. Two files of the initial list were rejected: 1QCU and 406D. Both structures

contain multiple models with different chain identifiers, and do not have proper MODEL/ENDMDL tags. This non-conformity to the PDB syntax precludes us from applying our automated procedure on these two structures. To ensure a complete uniformity of hydrogen atom names, they were removed, if present, and then added using bond lengths and angles from the Cornell *et al.* force field [16]. When appropriate, LP atoms were placed at 1 Å of their atom in the direction of the lone electron pair, as determined by the *sp2* geometry of the base atoms. Names for the LP atoms were assigned by following the standard nomenclature of hydrogen atoms in the PDB, replacing the H by LP.

The EM algorithm was initialized with seven Gaussians, the initial parameters are shown in Table 2·3. To avoid local minima in the optimization, a variant of the EM algorithm was used in which only 25 000 randomly selected data points were considered at each iteration. The algorithm was given 100 iterations, and convergence was confirmed by monitoring the negative log-likelihood as the algorithm progressed (see Figure 2·4). One hour of CPU time was necessary on a PIII/600Mhz to complete the learning process.

For the detection of a stable set of H-bonds, a modified version of the preflow-push algorithm [30] was implemented. The graph of donors and acceptors was first built from the entire 3-D structure, the equilibrated maximum flow was then computed, resulting in a stable set of H-bonds.

ACKNOWLEDGMENTS

We thank Patrick Gendron, Sergei Chteinberg, and Fabrice Leclerc, for providing RNA structure expertise, and Yoshua Bengio for suggesting the use of a mixture of Gaussians. We thank the referees for providing useful comments and suggestions. This work was supported by a grant from the Canadian Institutes of Health Research (CIHR) (MT-14604) to FM. SL holds a Ph.D. scholarship from CIHR.

CHAPITRE 3

QUANTITATIVE ANALYSIS OF NUCLEIC ACID THREE-DIMENSIONAL STRUCTURES

P. Gendron, S. Lemieux et F. Major, *Journal of Molecular Biology* (2001) **308**, 919–936.

ABSTRACT

A new computer program to annotate DNA and RNA three-dimensional structures, MC-Annotate, is introduced. The goals of annotation are to efficiently extract and manipulate structural information, to simplify further structural analyses and searches, and to objectively represent structural knowledge. The input of MC-Annotate is a PDB formatted DNA or RNA three-dimensional structure. The output of MC-Annotate is composed of a structural graph that contains the annotations, and a series of HTML documents, one for each nucleotide conformation and base-base interaction present in the input structure. The atomic coordinates of all nucleotides and the homogeneous transformation matrices of all base-base interactions are stored in the structural graph. Symbolic classifications of nucleotide conformations, using sugar puckering modes and nitrogen base orientations around the glycosyl bond, and base-base interactions, using stacking and hydrogen bonding information, are introduced. Peculiarity factors of nucleotide conformations and base-base interactions are defined to indicate their marginalities with all other examples. The peculiarity factors allow us to identify irregular regions and possible stereochemical errors in 3-D structures without interactive visualization. The annotations attached to each nucleotide conformation include its class, its torsion angles, a distribution of the

root-mean-square deviations with examples of the same class, the list of examples of the same class, and its peculiarity value. The annotations attached to each base-base interaction include its class, a distribution of distances with examples of the same class, the list of examples of the same class, and its peculiarity value. The distance between two homogeneous transformation matrices is evaluated using a new metric that distinguishes between the rotation and the translation of a transformation matrix in the context of nitrogen bases. MC-Annotate was used to build databases of nucleotide conformations and base-base interactions. It was applied to the ribosomal RNA fragment that binds to protein L11, which annotations revealed peculiar nucleotide conformations and base-base interactions in the regions where the RNA contacts the protein. The question of whether the current database of RNA three-dimensional structures is complete is addressed.

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

3·1 INTRODUCTION

The function of ribonucleic acid (RNA) molecules 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 [82]. 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 *Escherichia coli* $tRNA^{GLN}$, were the only available x-ray crystal structures of biologically active RNAs. As a consequence, the reliability of most RNA structure prediction procedure 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) [8] and the Nucleic Acids DataBase (NDB) [7], range from a few nucleotides, such as the lead-activated [37, 94] (PDB codes: 1LDZ and 429D) and the hammerhead [77] (PDB code: 1HMH) ribozymes, to several hundreds of nucleotides, such as the P4-P6 domain of the *Tetrahymena thermophila* group I intron [31] (PDB code: 1GID) and the 23S rRNA [5] (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 [5, 13, 81].

It is clear that understanding and establishing the variety of RNA structures

and activities, as well as manipulating 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 and quantitative analyses of RNA 3-D structures are employed in the validation of new 3-D structures, in the comparison and identification of structural patterns and motifs, in the development of empirical modeling systems, and, more generally, in the studying and learning of structure-function relationships. Only few 3-D structure analysis methods apply specifically to RNA, and all are either based on interactive visualization, which limits analysis to small RNA domains, or on the computation of atomic distances, and bond and torsion angles [4, 46, 47]. Because different torsion angle patterns can result in similar conformations [74], the results of comparative analysis based on torsion angles are not always informative [28].

The unavailability of a quantitative and objective annotation tool prompted us to develop a new computer program, *MC-Annotate*. We evaluated, defined and implemented new computer representations and distance metrics for analyzing and comparing nucleotide conformations and their spatial interactions, hereafter referred to as base-base interactions or, simply, interactions. Base-base interactions stabilize local conformations and determine the folding of the whole structure. For example, the tertiary interaction between U8 and A14 is crucial to the folding of the $tRNA^{PHE}$ (see for instance PDB code: 6TNA) [84] into its characteristic L-shape [62].

The annotation of RNA 3-D structures consists of a preprocessing of the information embedded in their 3-D coordinates. The goals of annotation are to efficiently extract and manipulate structural information, to simplify further structural analyses and searches, and to objectively represent structural knowledge.

These goals were considered during the development of *MC-Annotate*. At first, the structural graph of an input structure is generated. The structural graph encodes geometric information about nucleotide conformations and base-base interactions, atomic coordinates and torsion angles. Then, using the geometric information, symbols are computed and attached to each nucleotide conformation and base-base interaction in the structural graph. Using symbols and numbers for representing an RNA structure, rather than atomic coordinates and torsion angles, simplifies its comparison to other RNAs, as well as the recognition of its motifs. Symbolic annotations are useful to crystallographers and molecular modelers seeking efficient analyses of 3-D structures, and identification of relevant structural features and patterns that could be involved in the activity of their molecules.

MC-Annotate also made possible the creation of databases of nucleotide conformations and base-base interactions, extracted from all available DNA and RNA 3-D structures, which were indexed using the symbolic information. For instance, these databases were employed to update the parameters and conformational sampling of *MC-Sym* [63]. In this article, *MC-Annotate* is examplified by the analysis of the ribosomal RNA (rRNA) fragment that binds to protein L11 [15] (PDB code: 1QA6), which revealed peculiar nucleotide conformations and base-base interactions in the regions where the rRNA contacts with the protein. The symbols generated by *MC-Annotate* were also combined into sets defining the higher-order patterns, or motifs, of the rRNA domain. Using the geometric and symbolic information, a list of matching patterns from other RNAs were identified. Finally, the question whether the current database of DNA and RNA 3-D structures is complete was addressed. A complete database would contain all possible and thermodynamically sound nucleotide conformations and base-base interactions. This was made by generating and measuring the distances of randomly generated examples to those currently in the databases.

3·2 RESULTS

A. LOCAL REFERENTIALS AND HOMOGENEOUS TRANSFORMATION MATRICES

The local referential of a nucleotide, and thus of a nitrogen base, is defined by a Cartesian coordinate system whose position, relative to the base, can be computed from its atomic coordinates (see Figure 3·1). The local referential of a nucleotide can be defined arbitrarily, but must be identical for each type of nucleotide. Let $\mathbf{u}$ be the unit vector between coordinates of atom N1 and C2 in pyrimidines, and N9 and C4 in purines. Let $\mathbf{v}$ be the unit vector between coordinates of atom N1 and C6 in pyrimidines, and N9 and C8 in purines. Then, the unit vector $\mathbf{y}$ of the Cartesian coordinate system lies in the direction given by the sum $\mathbf{u} + \mathbf{v}$, the unit vector $\mathbf{z}$ is oriented along the cross product $\mathbf{u} \times \mathbf{v}$, and the unit vector $\mathbf{x}$, following the right hand rule for a Cartesian coordinate system, is given by $\mathbf{y} \times \mathbf{z}$. The relative positions of local referentials can be expressed using homogeneous transformation matrices (HTM), which were first developed in the field of geometry [65], and later extensively used in computer graphics and robotics. HTMs encode, in the form of a 4x4 matrix, the geometric operations needed to transform objects in 3-D space from one local referential to another. In the base-base interaction context, a HTM describes the spatial relation by a composition of a translation and a rotation between the two local referentials of the involved nitrogen bases.

Let $\mathbf{R}_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_2}$ be the local referentials of nucleotides $\mathbf{b}_1$ and $\mathbf{b}_2$ as expressed relative to the global referential centered at the origin, $(0, 0, 0)$. The spatial relation between $\mathbf{R}_{\mathbf{b}_1}$ and $\mathbf{R}_{\mathbf{b}_2}$ is then given by the HTM $\mathbf{M}_{\mathbf{b}_1 \to \mathbf{b}_2} = \mathbf{R}_{\mathbf{b}_1}^{-1} \mathbf{R}_{\mathbf{b}_2}$ (see Figure 3·1). In a molecular modeling context such as implemented in *MC-Sym* [63], this relation can be reproduced and the atomic coordinates of nucleotide $\mathbf{b}_2'$ relative to nucleotide $\mathbf{b}_1'$ computed by applying

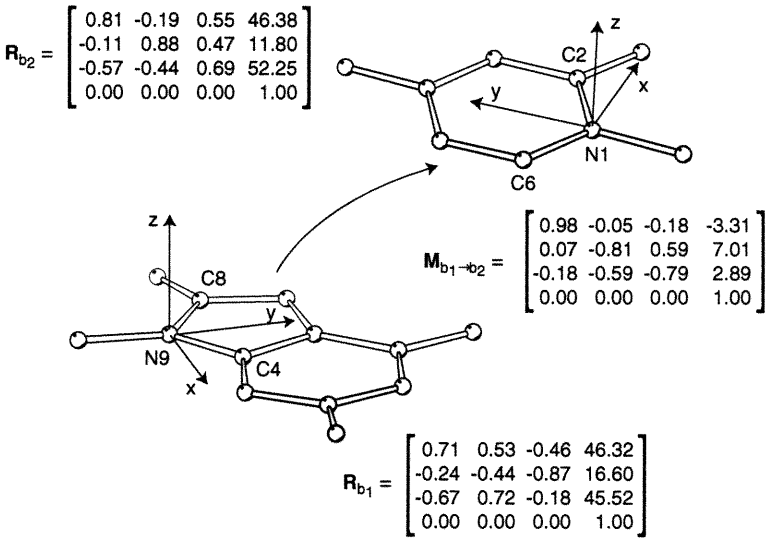


Figure 3·1 *Local referentials and base-base interactions.* $\mathbf{R_{b_1}}$ *and* $\mathbf{R_{b_2}}$ *are the homogeneous transformation matrices representing the local referentials of two nucleotides,* $\mathbf{b_1}$ *and* $\mathbf{b_2}$. $\mathbf{M_{b_1 \to b_2}}$ *encodes the relation between* $\mathbf{R_{b_1}}$ *and* $\mathbf{R_{b_2}}$, *that is the position of* $\mathbf{R_{b_2}}$ *relative to* $\mathbf{R_{b_1}}$.

the transformation obtained by the matrix product $\mathbf{R}_{\mathbf{b}'_1}\mathbf{M}_{\mathbf{b}_1\rightarrow\mathbf{b}_2}\mathbf{R}^{-1}_{\mathbf{b}'_2}$ to the absolute atomic coordinates of $\mathbf{b}'_2$. In a similar way, the atomic coordinates of $\mathbf{b}'_1$ relative to $\mathbf{b}'_2$ can be computed by applying the inverse transformation $\mathbf{R}_{\mathbf{b}'_2}\mathbf{M}^{-1}_{\mathbf{b}_1\rightarrow\mathbf{b}_2}\mathbf{R}^{-1}_{\mathbf{b}'_1}$ to the absolute coordinates of $\mathbf{b}'_1$. It is worth noting here that $\mathbf{M}^{-1}_{\mathbf{b}_1\rightarrow\mathbf{b}_2} = \mathbf{M}_{\mathbf{b}_2\rightarrow\mathbf{b}_1}$, 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 have been extracted between $\mathbf{R}_{\mathbf{b}_2}$ and $\mathbf{R}_{\mathbf{b}_1}$.

B. NUCLEOTIDE CONFORMATIONS

Based on traditional definitions of nucleotide conformations, their symbolic characterization 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 values of the pseudorotation phase angle for furanose rings described by Altona et al. [3] are divided into the ten classes shown in Table 3·1. 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. As accepted by the IUPAC-IUB commision [39], values of χ in the range $[-90°, 90°[$ indicate a *syn* orientation whereas other values indicate a *anti* orientation. Since the other parts of a nucleotide are mostly rigid, the two above properties represent a fair qualitative description of nucleotide conformations. The class of a nucleotide conformation can thus be defined by its sugar puckering mode and nitrogen base orientation around the glycosyl bond. The corresponding symbols assigned by *MC-Annotate* are summarized in Table 3·1.

The distance, $d(\mathbf{b}_1, \mathbf{b}_2)$, between two nucleotide conformations, $\mathbf{b}_1$ and $\mathbf{b}_2$, can be defined by the root mean square deviation (RMSD) between the heavy atoms in the backbone of the two nucleotides, *a posteriori* of optimal

Nucleotide conformation	Set of symbols
1. Type	{A, C, G, U, T}
2. Sugar pucker	{C1'-endo, C1'-exo, C2'-endo, C2'-exo, C3'-endo, C3'-exo, C4'-endo, C4'-exo, O4'-endo, O4'-exo}
3. Orientation around glycosidic bond	{anti, syn}
Base-base interaction	**Set of symbols**
1. Types	$\{A, C, G, U, T\}^2$
2. Adjacency	{adjacent, non-adjacent}
3. Stacking	{stacked, unstacked, helically stacked}
4. Pairing	{paired, unpaired}
a. Relative glycosidic bond orientation	{cis, trans}
b. Interacting edges	$\{W.\text{-}C., Hoogst., Sh.g.\}^2$
c. MC-Sym number	{I, II, ..., XXVIII, 29, 30, ..., 137}

Table 3·1 *Symbols used in classification. Two symbols from the base type and interacting edges are used, one for each nucleotide involved in the base-pairs of two or more H-bonds are in roman, whereas arabic numbers are used for one H-bond base-pairing patterns.*

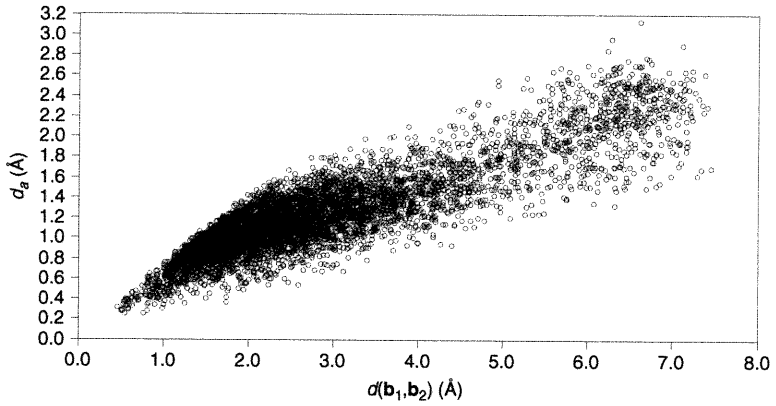


Figure 3·2 *Correlation of the d_a and d distance metrics. The dots represent the RMSD computed using d_a (vertical axis) and d (horizontal axis) between randomly selected pairs of residues from the PDB and NDB.*

superimposition of their local referentials in 3-D space [28]. Our metric is in good correlation with the more standard all-atom superposition and RMSD metric, $d_a(\mathbf{b_1}, \mathbf{b_2})$, performed using the analytical method described by Kabsch [43, 44] (see Figure 3·2). Our metric places the emphasis on the backbone atom positions and orientations relative to the nitrogen base, and is shown in Figure 3·3. Figure 3·4 illustrates two situations in which our metric, d, offers a better evaluation of the structural distance between nucleotide conformations than the d_a metric.

C. Base-base interactions

For the classification of base-base interactions, we considered nitrogen base pairs that involve at least one of the known chemical stabilizing forces, those of two covalently connected nucleotides, base pairing and base stacking. Base-base interactions are thus of five distinct types: adjacent, adjacent-stacked, adjacent-paired, non-adjacent-stacked and non-adjacent-paired (see Table 3·1). Since there is no measurable forces between non-adjacent, non-paired, and

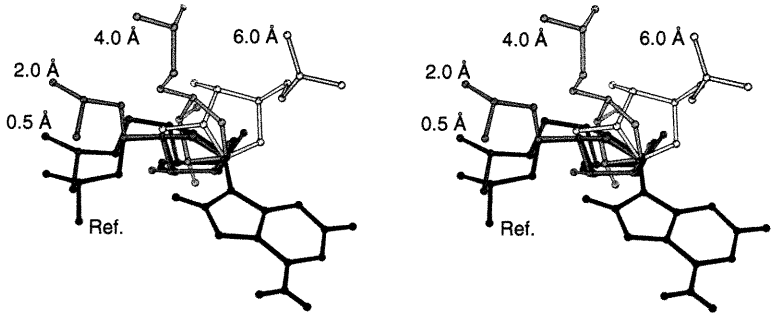


Figure 3·3 *Stereo view of superimposed nucleotide conformations. The RMSD from a reference in black were computed using d. Dark to pale gray variations were used to indicate small to large RMSD.*

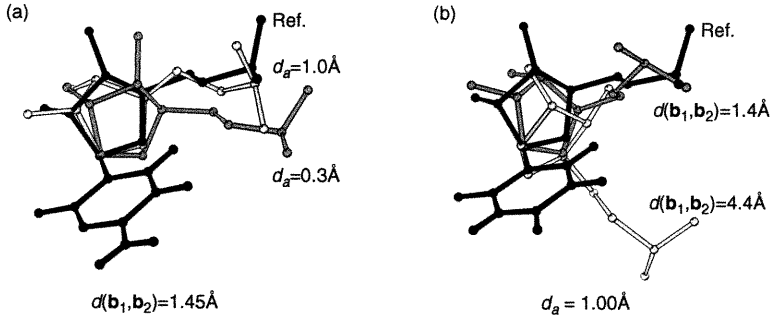


Figure 3·4 *Comparison of the d_a and d distance metrics for nucleotide conformations. a) Two nearly identical nucleotide conformations in light and medium gray, $d = 1.45\text{Å}$, to a reference nucleotide in black. However, their distances differ when computed using d_a, respectively 1.0 and 0.3Å. b) Two different nucleotide conformations in light and medium gray, at respectively 4.4 and 1.4Å of the reference nucleotide in black. Their RMSD computed using d_a is 1.0Å.*

non-stacked nucleotides, they were not considered even though they are more frequent. The adjacency of nucleotides was determined either by using the Protein DataBank (PDB) nucleotide numbering system [8], or a maximum length of 2Å of the O3'-P chemical bond.

Traditional encodings of adjacent base-base interactions use the six backbone torsion angles α, β, γ, δ, ϵ and ζ [79], or the two pseudotorsion angles η and θ [22]. These parameters accurately describe the relative placement of nucleotides linked by a phosphodiester bond. However, it has already been observed that distinct torsion angle combinations can result in similar backbone directions and base orientations. This phenomenon is known as the "crankshaft effect" [36, 71]. Also, non-adjacent base-base interactions, like base pairings that are stabilized by H-bonds and non-adjacent base-base stacking, cannot be accurately parameterized using these angles. Rather, a plethora of rotation and translation parameters have been used to describe these interactions [4, 46, 47]. A simplified and unified encoding scheme for any type of base-base interactions that emerged from the introduction of HTMs is introduced. In order to allow us to effectively compare base-base interactions, a distance metric between two HTMs, $\mathbf{M}_{\mathbf{b}_1\rightarrow\mathbf{b}_2}$ and $\mathbf{N}_{\mathbf{b}'_1\rightarrow\mathbf{b}'_2}$, should possess the following properties:

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

$$d(\mathbf{M}_{\mathbf{b}_1\rightarrow\mathbf{b}_2}, \mathbf{M}_{\mathbf{b}_1\rightarrow\mathbf{b}_2}) = 0 \tag{3.2}$$

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

Figure 3·5 shows a two-dimensional vector analogy of equations 3.1 to 3.3. Equation 3.1 states that the distance metric should obviously be commutative. Equation 3.2 states that a spatial relation should have a null distance with itself, but not with its inverse unless they are equal. Equation 3.3 states that the distance metric should not depend on the direction of application, implicit in the HTM representation.

The simple Euclidean distance in the 16 dimensional space of HTMs does

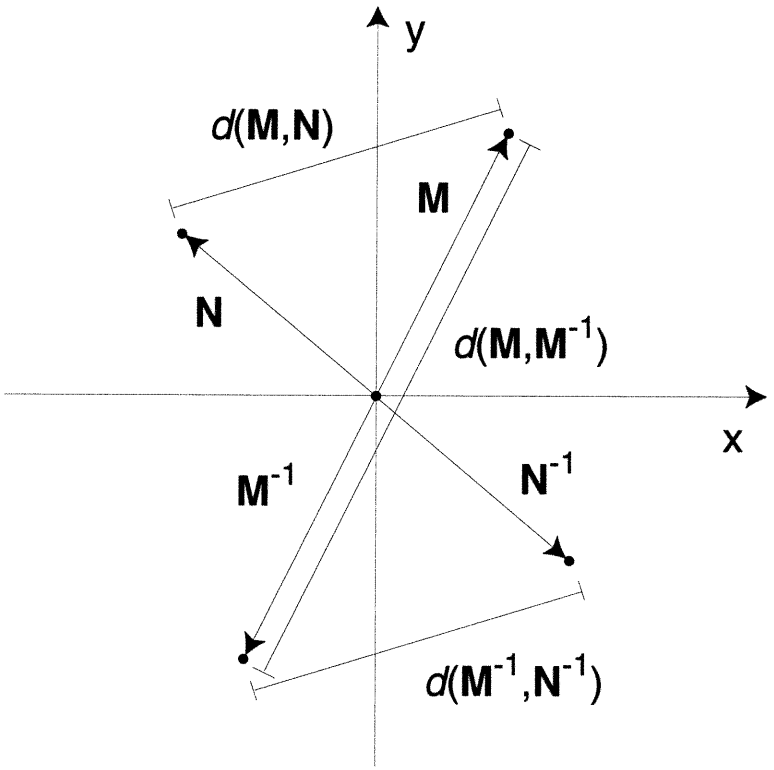


Figure 3·5 *Two-dimensional vectorial representation of the distance metric properties. If* $\mathbf{M}$ *and* $\mathbf{N}$ *are two vectors representing spatial relations, one can see that the distance between the extremities of these vectors is independent of the order in which the distance is computed (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 its inverse will be zero only if they are equal (Equation 3.2).*

not satisfy the above properties since HTMs embed a combination of translation and rotation terms that need to be considered separately. A HTM can be decomposed in the product of two HTMs, $\mathbf{M} = \mathbf{TR}$, where $\mathbf{T}$ contains the translation and $\mathbf{R}$ contains the rotation embedded in the original HTM. Paul [72] showed how to extract the length of the translation, l, as well as the angle θ and the axis of rotation k from matrices $\mathbf{T}$ and $\mathbf{R}$. The strength of a transformation, $S(\mathbf{M})$, regardless of the axis of rotation, is defined by:

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

where α represents a conversion factor applied between the translation and rotation contributions to combine the different units. Figure 3·6 shows that a conversion factor of 30°/Å yields a nice correlation with the RMSD metric, and means that a rotation of 30° around any axis is equivalent to a displacement of 1Å between two nucleotides' local referentials. Using this expression, the distance between two base-base interactions, $d(\mathbf{M}, \mathbf{N})$, can be defined by:

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

which satisfies the requirements of equations 3.1 to 3.3. In equation 3.5, the composition of transformation $\mathbf{MN^{-1}}$ 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, $\mathbf{M^{-1}N}$ can be interpreted 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 3·6 shows that $d(\mathbf{M}, \mathbf{N})$ is roughly equivalent to the more standard RMSD, d_a, calculated after the optimal global superimposition of the atomic coordinates of the two pairs using the analytical method described by Kabsch [43, 44]. However, our distance metric better discriminates between two spatial relations that differ by a rotation of the nitrogen bases. Figure 3·7 shows two situations where the more standard RMSD metric incorrectly interprets the

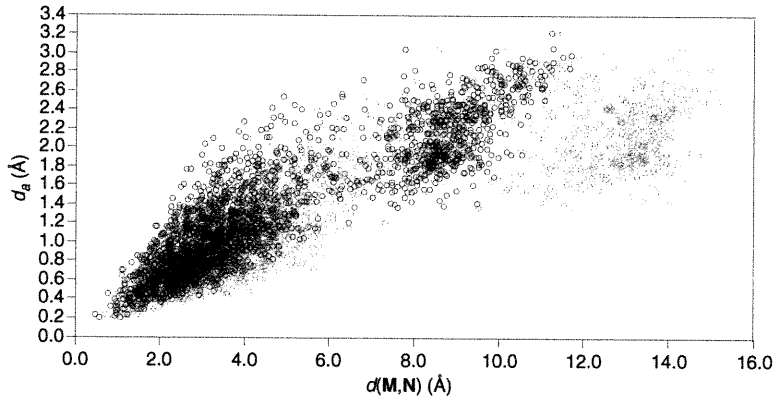


Figure 3·6 *Correlation between standard RMSD, d_a, and the local referential metric for base-base interactions, $d(\mathrm{M},\mathrm{N})$, with different factors of α. All dots represent distances of base-stacking interactions. The gray dots represent distances obtained using a factor $\alpha = 15°/Å$, whereas the black dots represent distances obtained with a factor $\alpha = 30°/Å$.*

distances because of nitrogen base rotations, whereas our metric returns a better evaluation of the distances.

Although HTMs are perfectly suited to uniformly encode base-base interactions, the information they contain is too compact to identify the type of spatial relations they encode without reproducing them in 3-D space, and evaluating other parameters. For this reason, the symbolic annotations of base-base interactions are determined from atomic coordinates.

D. Base pairing

Hydrogen bonds (H-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 interactions in the folding and stabilization of DNA and RNA molecules. H-bonds are directional due to the orbital shape of the electron density distributions, and thus favor planar

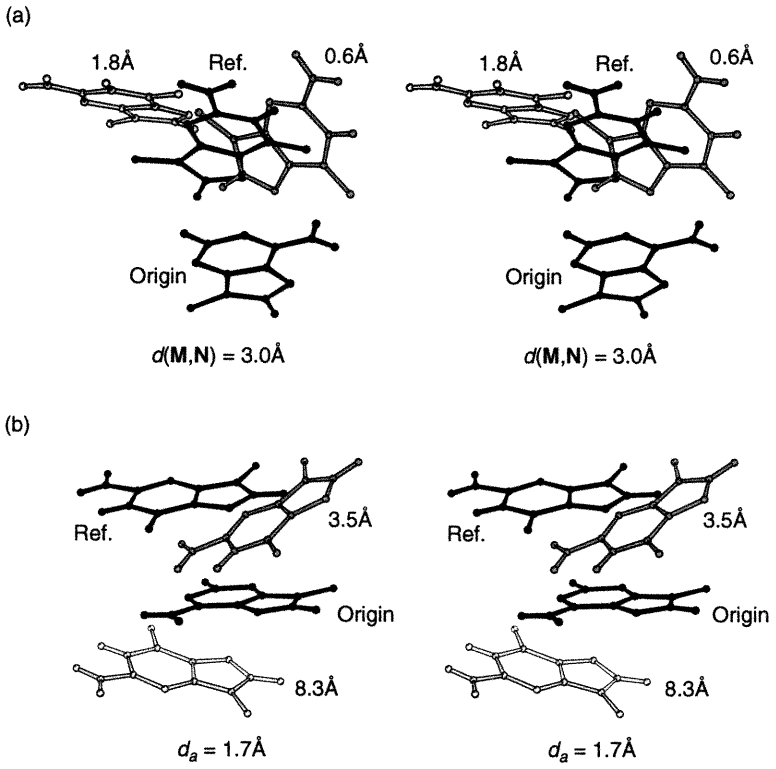


Figure 3·7 *Stereo views of superimposed base stacking interactions. a) Two base stacking interactions, at equal distance of 3.0Å to the reference stacking interaction in black, but at* $d_a = 1.8$, *and 0.6Å respectively. b) The all-atom superimposition does not allow us to distinguish between two different base-stacking interactions when the differences come from nitrogen base rotations. Two base stacking interactions at respectively 3.5 and 8.3Å, due to a rotation of 180° of the nitrogen base, is not detected by* $d_a = 1.7$*Å.*

base pair geometries formed by at least two H-bonds.

Most base pairing types are planar and subject to base stacking forces within the helical regions where they are found. Base pairing between two nucleotides can be determined using the empirical method developed in our group, which yields a symbolic classification of the possible H-bonding types. For a given pair of nucleotides, the list that contains all possible types, involving two or more H-bonds as defined by Donohue [20, 21], and involving one H-bond as defined by Gautheret and Gutell [27], $\mathcal{H} = \{H_1, H_2, \ldots, H_n\}$, is considered. A base pair of geometry $\mathbf{X}$ satisfies the pairing H_k if $P(H_k \mid \mathbf{X}) > P(H_i \mid \mathbf{X})$, $H_i \in \mathcal{H}$, $i \neq k$, that is, the probability of forming pairing H_k given geometry $\mathbf{X}$ is higher than any other pairing, and $P(H_k \mid \mathbf{X}) \geq c$, that is, the probability of pairing X_k given geometry $\mathbf{X}$ is higher than cutoff, c, that was empirically fixed to 0.3. An empirical approximation of the probability of observing base pairing type H, given the geometry $\mathbf{X}$, $P(H \mid \mathbf{X})$, can be obtained by considering H to be the set of donor/acceptor pairs that should form H-bonds in the given geometry, and $\overline{H}$ to be the set of H-bonds that should not form. This approximation can be obtained by the following product:

$$P(H \mid \mathbf{X}) = \Pi_{h_i \in H} P(h_i \mid \mathbf{x}_i) \cdot \Pi_{h_i \in \overline{H}} (1 - P(h_i \mid \mathbf{x}_i)), \tag{3.6}$$

where $P(h \mid \mathbf{x})$ represents the probability of forming H-bond h given local geometry $\mathbf{x}$. The local geometry of a pair of donor/acceptor, $\mathbf{x}$, is 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. We obtain an approximation of the probability of h by multiplying the probabilities associated to each above parameter, which is computed by the following function:

$$P(h \mid x) = \begin{cases} x < \mu & 1 \\ \text{else} & e^{-\left(\frac{x-\mu}{2\sigma^2}\right)^2} \end{cases} \tag{3.7}$$

The constants μ and σ were obtained empirically for each parameter by visual observation of the histogram.

The class of a base pair is usually defined by its relative glycosidic bond orientation and the interacting faces of the two bases [54]. But since these two attributes can define more than one base pairing type, the H-bonds involved in the pattern must also be present. To simplify the classification of base pairing types, identification numbers were introduced in *MC-Sym*, and are used in *MC-Annotate* to define their classes [45]. Roman numerals indicate the two (or three) H-bonds pairings identified by Donohue [20, 21], whereas arabic numerals indicate the bifurcated and single H-bond pairing patterns generated by Gautheret [27]. Table 3·1 summarizes the different parameters of base pair classification.

E. BASE STACKING

Vertical nitrogen base stacking is a significant stabilizing interaction of DNA and RNA 3-D structures, which plays a major role in their folding and complexation. Stacking occurs more frequently between adjacent, but also non-adjacent, nucleotides, mostly in double-stranded helical regions. The stabilization of base stacking involves London dispersion forces [34], and interactions between partial charges within the adjacent rings [80]. Evidences for hydrophobic forces between bases in solution [86], as well as a contradictory nonclassical hydrophobic effect [70], have been observed. However, these interactions were not characterized and parameterized such that they could define precise energy parameters that could be used for the detection of base stacking [83].

In order to include examples with large deviations from ideal parameters, we employed a geometrical approach that uses relaxed ranges of the values defined in the Gabb et al. method [25]. It has been shown that many inconsistencies exist in the atomic coordinates of RNA structures. The deviations measured in

NMR spectroscopy and x-ray diffraction structures can be due to variations in the refinement protocols and force fields, as well as to artifacts resulting from the determination processes [19,96].

Stacking between two nitrogen bases is considered if the distance between their rings is less than 5.5Å, the angle between the two normals to the base planes is inferior to 30°, and 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°. The class of a stacking interaction is defined by the nucleotides involved in it (see Table 3·1).

F. Structural databases

Nucleotide conformations and base-base interactions identified and annotated by *MC-Annotate* in all available DNA and RNA 3-D structures (see Materials and Methods) were stored in databases. Nucleotide conformations and base-base interactions originating from newly determined structures can thus easily be compared to all others, for instance to detect peculiar and similar regions.

G. Peculiarity

From the two distance metrics defined above, peculiarity factors were defined to identify specific nucleotide conformations and base-base interactions, as well as to detect possible stereochemical errors in a given RNA 3-D structure without having to visualize it. Each nucleotide conformation and base-base interaction is thus evaluated relatively to all other examples of its class, and its peculiarity, or adversely conformity, can be assessed using the peculiarity factor.

The *degree of peculiarity* of a feature $\mathbf{v_i}$ is a measure, within its conformational space c, of how scarce the space surrounding $\mathbf{v_i}$ is. A kernel-based

density estimation method is therefore used with a Gaussian kernel function centered on each feature to evaluate its contribution to the peculiarity factor at a given point in conformational space. The degree of peculiarity is then given by:

$$Q_c(\mathbf{v_i}) = \mathbf{1} - \frac{\mathbf{1}}{\mathbf{n}\sqrt{\mathbf{2}\pi\sigma_\mathbf{c}^\mathbf{2}}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{e}^{-\frac{\mathbf{1}}{\mathbf{2}}\left(\frac{\mathbf{d(v_i,v_j)}}{\sigma_\mathbf{c}}\right)^\mathbf{2}}, \tag{3.8}$$

where the sum is taken over all features $\mathbf{v_j}$ (nucleotides or base-base interactions) of the same class as $\mathbf{v_i}$ in the databases. The standard deviation σ_c determines the "size" of the Gaussian kernel, that is, the extent to which a distant feature $\mathbf{v_j}$ in the space of conformations contributes to the peculiarity at point $\mathbf{v_i}$. Since the size and density of conformational space varies within each class of features, we compute an unbiased estimate of the standard deviation in each class as follows:

$$\sigma_c = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{v_i} - \mu_\mathbf{c})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{v_i} - \mathbf{v_j})\right)^2}, \tag{3.9}$$

where the difference $(\mathbf{v_i} - \mathbf{v_j})$ is simply one of our distance metrics $d(\mathbf{v_i}, \mathbf{v_j})$ defined above. In order to more easily compare peculiarity values, we express them as relative peculiarities using the expression:

$$Q'_c(\mathbf{v_i}) = \frac{\mathbf{Q_c(v_i)} - \mathbf{Q_c^{min}}}{\mathbf{Q_c^{max}} - \mathbf{Q_c^{min}}} \tag{3.10}$$

Peculiarity values range from 0 to 1 with high values indicating high degrees of peculiarity.

H. STRUCTURAL GRAPHS

A structural graph is a computer representation of nucleic acid structures in which nodes correspond to nucleotides, and edges to spatial interactions between pairs of nucleotides. The first level of annotation stores the atomic coordinates, torsion angles, and nitrogen base spatial interactions encoded by HTMs. The second level of annotation computes and attaches the symbols that characterize the nucleotide conformations and base-base interactions, as described in Table 3·1. *MC-Annotate*

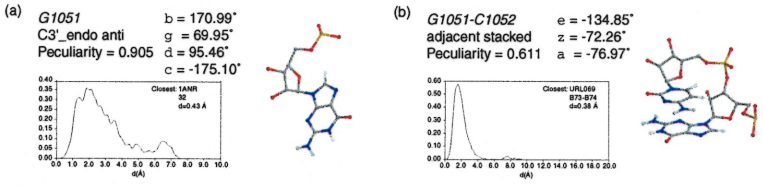


Figure 3·8 MC-Annotate *output for each node and edge of a structural graph. a) The node for G1051 in the rRNA domain binding protein L11.* MC-Annotate *computed the distribution of distances between the G1051 conformation and all others examples in the same class, attached examples of the same class (not shown) and displayed the closest example among them (here, nucleotide 32 in 1ANR), and returned its peculiarity value. b) The edge for the base stacking interaction between G1051 and C1052 in the rRNA domain binding protein L11.* MC-Annotate *assigned the appropriate base stacking symbol, computed the distribution of distances with all other examples of the same class, attached examples of the same type (not shown) and displayed the closest among them (here the base stacking between nucleotides B73 and B74 in URL069), and returned its peculiarity value.*

generates the structural graph of any given DNA or RNA 3-D structure in the PDB format. This encoding of nucleic acids structures and the distance metrics, which relate nodes and edges, permit the use of many graph theoretical algorithms to explore higher-order nucleotide arrangements, such as motifs, both on the geometric and symbolic standpoints. Figure 3·8 shows the annotations attached to the nucleotide conformations and base-base interactions.

I. The rRNA domain binding protein L11

To illustrate the various possibilities of *MC-Annotate*, we have thouroughly analysed the structure of the rRNA domain binding protein L11. The x-ray crystallography structure of this rRNA domain contains several non-standard conformations and structural distorsions, which are due to its interaction with

the protein. The domain contains 58 nucleotides of the large subunit of the ribosome of *E. coli*, and forms a complex with the ribosomal protein L11 (PDB code: 1QA6) [15]. This highly conserved domain of the rRNA is an important functional site and a potential target for antibiotics. The role of protein L11 can be structural to maintain the unusual fold of the rRNA domain. A careful analysis of the geometrical features of this domain is therefore an essential step to understand its function, and to characterize its interactions with protein L11 and antibiotics. Here, using *MC-Annotate*, we provide further evidence and a quantification of the unusual structural features reported by Conn *et al.* [15].

The results of the annotation procedure are shown in Table 3·2, which shows the symbolic description of each nucleotide and base-base interaction, as well as their associated degrees of peculiarity. All the relationships originally predicted by Gutell [33] using comparative sequence analysis (CSA) on the large subunit rRNA are present in the crystal structure, and were identified by our computer program. Using a graph isomorphism algorithm, we were able to identify and localize conserved motifs that occur both in the L11 binding domain and in many other PDB structures (see Table 3·3).

J. Completeness of the base-base interaction database

Given the current database of base-base interactions, one might ask if it is complete, that is if the probing of new structures using x-ray crystallography, NMR spectroscopy, or any other experimental method is likely to provide more structural information, and thus improve our knowledge about RNA structure. The answer to this question is particularly crucial to the development of empirical molecular modeling computer programs, and in particular to *MC-Sym*, whose accurary and precision of the construction process highly depend on the diversity of examples currently available in the database. *MC-Sym* uses the database of base-base interactions to assemble new RNA structures that satisfy experimental

Description	Pattern	Molecule	PDB id.	Position
1. A-Platform	A_{1089}•A_{1090}	P4-P6 group I intron domain	1GID	A171•A172, A218•A219, A225•A226
		LSU rRNA group I intron	1GRZ	A171•A172
		SSU rRNA S8 binding site	1BGZ	17•18
		LSU rRNA	1FFK	0.59•0.60, 0.441•0.442, 9.51•9.52
2. Base triple	C_{1072}•G_{1099}•C_{1092}	LSU rRNA group I intron	1GRZ	A260•A212•A109
		SUNY rRNA group I intron T4	1SUN	92•62•37
		LSU rRNA	1FFK	9.113•9.66•9.15
3. U-turn	URRN	HIV-1 RNA hairpin loop	1BVJ	A11-A14
		Tobramycin-RNA aptamer complex	2TOB	12-15
		Hammerhead rybozyme	301D	A40-A50-A60-A70
		LSU rRNA	1FFK	0.392-0.395, 0.2598-0.2601
		U2 snRNA stem loop IIa	2U2A	9-12
4. Non-adjacent stack	A_{1061}/A_{1070}	Yeast $tRNA^{ASP}$	2TRA,3TRA, 1ASY,1ASZ, 486D	9/46 R609/R646, S609/S646 A9/A46, E9/E46
		SSU rRNA S15,S6,S18 binding site	1EKC	I665/J733
		LSU rRNA	1FFK	0.191/0.204, 0.1684/0.1691

Table 3·2 *Motifs in the rRNA domain binding protein L11. The position indicated in the patterns refer to position in the rRNA binding element. The molecules in which patterns of the same class were found are listed. The pattern matching was made according to the classification symbols defined in Table 3·1, as well as distance comparison of the HTMs of the edges in the target pattern, defined by the rRNA domain, and matches in the other molecules. A distance cutoff of 2 Å for each relation was used during pattern matching.*

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		0.881
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		0.729
C:1053 A:1054	x		0.695
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		0.764
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		0.794
U:1065 U:1066	x		0.728
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		0.712
C:1076 A:1077	x		0.855
A:1077 U:1078	x		0.788
U:1078 C:1079			0.898
C:1079 A:1080	x		0.759
A:1080 U:1081	x		0.730
U:1081 U:1082	x		0.617
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		41	0.646
A:1090 G:1091	x		0.780
G:1091 C:1092	x		0.723
C:1092 G:1093	x		0.880
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		0.720
U:1101 C:1102	x		0.774
C:1102 A:1103	x		0.696
A:1103 C:1104	x		0.843
C:1104 U:1105	x		0.743
U:1105 G:1106	x		0.784
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		57	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		46	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		123	0.895
C:1072 G:1099		122	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 3·3 *Annotation results for the rRNA domain that binds to protein L11.*

and theoretical constraints [63]. An important assumption is the completeness of the database, which guarantees the construction of any possible RNA structures. In practice, it is often difficult to evaluate if the problems encountered in building a particular structure is caused by the use of a restricted conformational sampling, due to insufficient base-base interaction examples, or to inconstancies in the constraints. Therefore, it is of utmost importance to quantify the completeness of the base-base interaction database.

The distance metric developed in section C can be used to address the completeness of the database. We assume that a given spatial relation, M, is present in the database if the database contains a spatial relation, N, such that $d(\mathrm{M}, \mathrm{N}) < \mathrm{c}$, where c is a distance cutoff that was arbitrarily fixed to 1.75Å. To evaluate the completeness of the database, 3000 single-stranded RNA structures were generated according to the protocol described in Materials and Methods. All adjacent base-base interactions were extracted from the random structures, and the base-base interactions that were not present in the database were counted, S_a. The probability of finding a new interaction can be estimated by S_a divided by the total number of randomly generated base-base interactions, S. Inversely, $1 - \frac{S_a}{S}$, represents the base-base interaction coverage of the database. The stacking interaction database covers 87% of all possible stacking interactions, and the non-stacking interaction database covers 26% of all possible non-stacked adjacent interactions. In order to increase the diversity of base-base interactions in the database, and thus improve the precision of the *MC-Sym* construction procedure, the randomly generated structures were introduced in the database.

We then evaluated the number of such new random structures required to obtain a complete coverage, that is to reach a point where $S_a = 0$, for the stacked and non-stacked interactions. As above, all base-base interactions were compared to the existing examples of the database. However, after being compared, the randomly generated examples were inserted in the database. Figure 3·9 shows

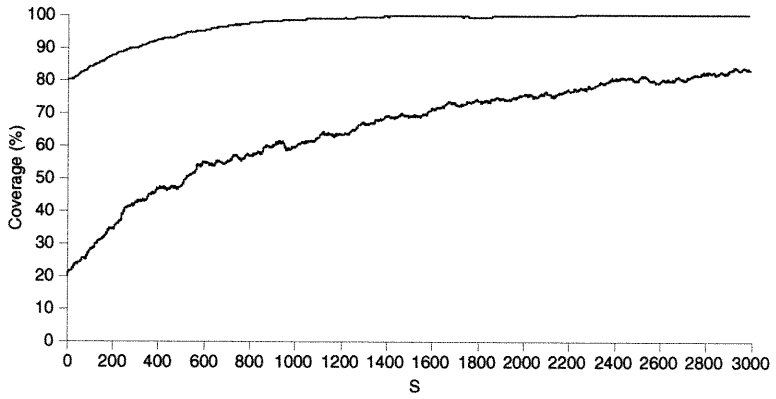


Figure 3·9 *Database coverages. The curves show coverage increases of the conformational space as new randomly generated structures are added to the database. The upper curve is for the base stacking interaction database, and the lower curve is for the non-stacking base-base interactions.*

that as new examples are inserted, the coverages of stacked and non-stacked interactions increase, as expected. In the base stacking interaction database, coverage goes up to 99%, after the insertion of only half of the randomly generated interactions, whereas in the case of non-stacked base-base interactions, coverage reaches an asymptotic point near 89%.

3·3 Discussion

A. The rRNA domain binding protein L11

Figure 3·10a shows the 3-D structure of the rRNA domain that binds protein L11. As indicated by the yellow color, a large portion of the structure is peculiar. The RNA fold is centered around a four-branch loop connecting, in clockwise order, stem-loop A, loop B, stem-loop C, and helix B (see Figure 3·10b). In the 3-D structure, stem-loops A and C are parallel and adjacent. Helix B and loop B are parallel to each other, and antiparallel to stem-loops A and C (see Figure 3·10a). The large deviations of the Watson-Crick base pairs A1057•U1081 ($Q'_c = 0.75$)

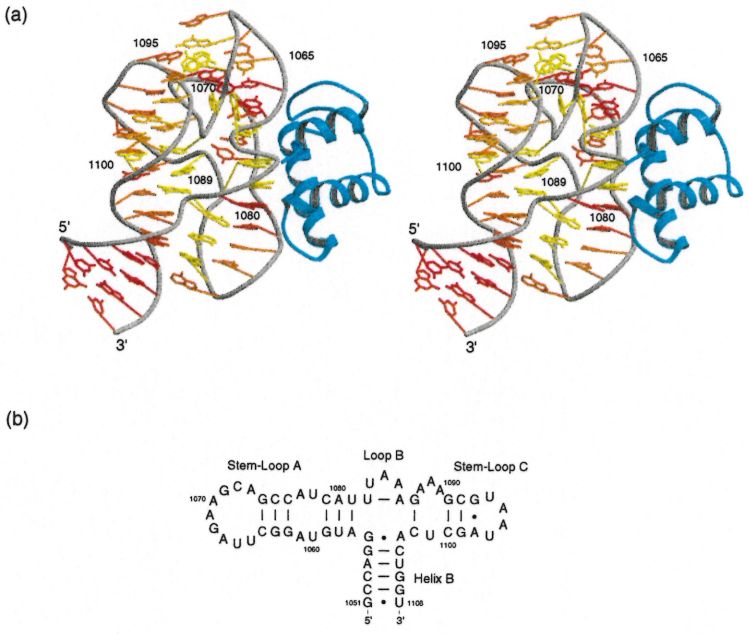


Figure 3·10 *The rRNA domain binding protein L11. a) Stereo view of the crystal structure of the 58-nucleotide domain. The degrees of peculiarities of the base-base interactions are shown in red for low peculiarity values, and yellow for high peculiarity values. b) Secondary structure. Watson-Crick base pairs are indicated by lines. Non Watson-Crick base pairs are indicated by dots.*

and G1087•C1102 ($Q'_c = 0.64$), and the reversed Watson-Crick A1082•U1086 ($Q'_c = 0.80$, Figure 3·11b) at the branching of the loop suggest that energetic strains are introduced by this particular arrangement of the domain.

The structure of the domain is stabilized by many tertiary interactions between the four regions, and shows a large interior core where most bases are stacked and paired. The surface contains several phylogenetically conserved bases that are exposed to the solvent, and available to make interactions. An unusual stacking interaction can be observed between residues A1057 and A1086 ($Q'_c = 0.97$, Figure 3·11c) located in two parallel strands, where A1086 adopts the uncommon C3'-endo syn conformation ($Q'_c = 0.61$). Helix B and loop B are also stabilized by a non-peculiar single H-bond base pairing between G1055 and A1085 ($Q'_c = 0.11$). These positions are 99% conserved among the sequences of the three phylogenetic domains (*Archaea*, *(eu)Bacteria* and *Eucarya*) [15].

Results show that stem-loops A and C interact with each other through several tertiary base pairs. A base triple formed by the base pairs G1071•G1091 ($Q'_c = 0.86$) and G1071•C1100 ($Q'_c = 1.00$) via peculiar base pairing interactions of types *cis* Watson-Crick/Hoogsteen (VI) and *trans* Hoogsteen/C-H (123) respectively, stacks between two other base triples [32]. The first base triple involves the C1072•G1099 ($Q'_c = 0.71$, type *trans* Watson-Crick/Hoogsteen (122)) and the Watson-Crick C1092•G1099 ($Q'_c = 0.30$) base pairs. The second base triple contains the Watson-Crick A1090•U1101 ($Q'_c = 0.60$) and A1089•A1090 ($Q'_c = 1.00$, type *trans* Hoogsteen/C-H (41)) base pairs. A1090 and A1089 are part of the so called "adenosine platform" motif (Figure 3·11g), a very stable base pairing formed by two consecutive adenosines. Three occurences of A-platforms were found in the crystal structure of the P4-P6 domain of *T. thermophila* group I intron [12], and five more were detected by *MC-Annotate* (see Table 3·3). All occurrences of A-platform motifs are within a distance of 2Å of each other, indicating an energetically stable motif. The base stacking observed in the base

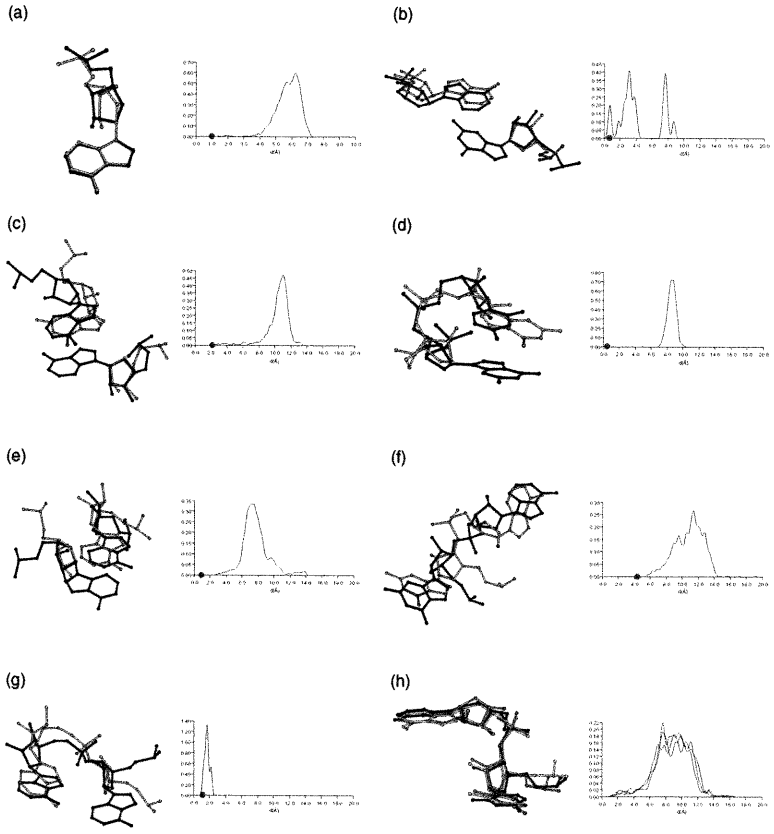


Figure 3·11 *Peculiar nucleotide conformations and base-base interactions of the rRNA domain binding protein L11. Superimposition of the conformations and interactions extracted from the rRNA domain (in black) with their closest example in the* MC-Annotate *databases (in gray), and their distribution of distances with all other examples of the same class. The red dots in the distribution curves indicate the distance of the conformation or interaction with the one used its closest. a) Residue A1088. b) Base pairing G1055•A1085. c) Base stacking between A1057 and A1086. d) Adjacent stacking between G1059 and U1060. e) Base stacking between A1061 and A1070. f) Adjacent non-stacked bases from A1087 and A1088. g) Adjacent base pairing between A1089 and A1090. h) The three occurrences of the relation UR in the URRN motif, also called U-turn, that were found in the rRNA domain binding protein L11: U1066-A1067, U1083-A1084, and U1094-A1095. Also shown are their distance distributions to the non-stacked interactions in the database.*

triple complex is normal, except for the peculiar stacking between G1071 and A1089 ($Q'_c = 0.99$).

The rRNA domain also contains three occurrences of the characteristic U-turn motif found in many RNA structures (see Table 3·3). The U-turn motif is a structured URRN loop that is stabilized by a H-bond between the 5'U:N3 and the phosphate of the 3'N, as well as stacking between the Rs. The motif has a very specific UR non-stacking interaction that possesses a low variability for this type. Among all of the non-stacked base-base interactions in the PDB, only 1% are closer than 4Å to the three UR interactions of the rRNA domain, which among them share distances inferior to 2.7Å. They nevertheless exhibit relatively high degrees of peculiarity with other adjacent non-stacking interactions (Figure 3·11h), indicating the importance of specializing further each class of base-base interactions. They are positioned at the beginning of loop A (U1066-A1067), B (U1083-A1084) and C (U1094-A1095), and turn the direction of the phosphodiester chain. One of the U-turn allows loop A to be inserted deep inside the structure between helices A and C, where the loop interacts in unusual ways with the surrounding nucleotides. Conn *et al.* reported that this is an unprecedented configuration for a hairpin loop. Our results quantify this observation since five out of the eight adjacent interactions adopt peculiar conformations ($Q'_c > 0.5$). In addition to the nucleotides G1071 and C1072, involved in two of the base triples described above, the phylogenetically conserved G1070 stacks with U1061 ($Q'_c = 0.86$; Figure 3·11e) in a peculiar way to stabilize the structure and to expose to the solvent the Hoogsteen edge of G1070 and the Watson-Crick edge of U1061. These two nucleotides, as well as the solvent exposed A1067 and A1095 found in the two nearest U-turns, are suspected to interact with other components of the ribosome and possibly other molecules [15]. Other occurrences of this particular type of base-stacking were found in the yeast $tRNA^{ASP}$ crystal structures, and in the small and large rRNA subunits.

Many peculiar conformations were found in the region near the protein binding site, indicating a structural role of the protein in the stabilization of these unusual base-base interactions. These distorsions can be attributed to the electrostatic influence of protein L11, or to the formation of an uncommon, but specific, protein binding site in the domain. On one hand, mutations at the binding site of the rRNA reduce L11 binding affinity [99]. On the other hand, the protein interactions stabilize the entire rRNA domain [98].

The phylogenetically conserved nucleotides G1059, U1060 and A1088 are at the heart of the protein binding site where the nearly invariant U1060 and A1088 interact in a well formed Hoogsteen base pair ($Q'_c = 0.36$), although A1088 adopts an unusual C3'-exo syn conformation ($Q'_c = 1.00$, Figure 3·11a). A peculiar adjacent relation between nucleotides G1087 and A1088 ($Q'_c = 0.87$), in stem C, is necessary to accommodate the insertion of A1088 into helix A (Figure 3·11f). Also, nucleotide A1061 bulges out of helix A in an unfamiliar way to accomodate stacking with G1070, as mentioned above. U1078 also bulges out of the helix. These bulges introduce an abrupt change in the helix axis. Nucleotide U1060 interacts with G1059 in a reverse stacking configuration ($Q'_c = 1.00$, Figure 3·11d), which reverses the polarity of the backbone in the helix and exposes its major groove edge to the solvent. We found a similar relation in the NMR structure of the AMP-RNA aptamer complex [40], as well as in the NMR structure of the Rev-RRE complex [48], as observed by Conn *et al.*. This unusual conformation creates a particularly reactive binding site by exposing many H-bond acceptors to the solvent.

In summary, most of the relevant features of the rRNA domain that binds protein L11 involve uncommon nucleotide conformations and base-base interactions. Here, *MC-Annotate* was useful for targeting and quantifying these conformational features. Since function can be attributed to structural specificity, finding conformational features deviating from the norm can be useful for

identifying structure-function relations. In a similar way, finding highly conserved conformations, such as those defining the U-turn motif, can also results in structure-function information.

B. Base-base interaction database

The protocol used to evaluate the base-base interaction database coverage has allowed us to quantify the stacked and non-stacked interactions. The stacking interaction database covers 87% of all possible stacking interactions, and the non-stacking interaction database covers 26% of all possible non-stacked adjacent interactions. This observation was consistent with the fact that *MC-Sym* builds stacked regions more easily than non-stacked ones, such as unstructured loops [52].

It was assumed that we could generate all stereochemically sound stacked and non-stacked base-base interactions by applying the molecular dynamics protocol to single-stranded oligonucleotides. In fact, since some adjacent base-base interactions are stabilized by secondary and tertiary structures, it is very difficult to reproduce some unusual examples using molecular dynamics. It is highly probable that these structures were not generated using our protocol, and indicate that our results might correspond to overestimates of the database completeness. It is also important to note that the completeness was only quantified for adjacent interactions, and thus non-adjacent (paired and stacked) interactions appearing in larger structures are difficult to explore using this approach.

These results are particularly important in the interpretation of the absolute peculiarity values obtained in the annotations since they are artificially raised in the context of incomplete sets. In particular we should expect systematically higher peculiarity for non-stacked interactions than for stacked ones. Peculiarity factors should only be used to compare interactions from equally complete sets.

With the introduction of newly determined 3-D structures in the *MC-Annotate* databases, we expect a steady decreasing in peculiarity values, reflecting the fact that each structure increases the amount of structural knowledge contained in the database. However, the peculiarity values of unsound nucleotide conformations and base-base interactions will increase.

C. Conclusion

The proposed annotation procedure simplifies the analysis of experimentally and theoretically determined structures. In particular, *MC-Annotate* allows one to classify the nucleotide conformations and base-base interactions, and to detect marginal regions that could indicate interactions with other molecules, or new sites that are responsible for structure and/or function. The results of the analysis of the rRNA domain binding L11 support these facts. The key interactions that were identified by the analysis correspond to actual structures involved in protein binding. Therefore, peculiarity values can be used to identify original and new structural features, as well as potentially faulty ones, and offer a fast approach to verify the presence of similar features in other RNA structures.

The interpretation of peculiarity values depend on the parameters used to classify the conformations and base-base interactions, and of the standard deviation of the Gaussian distributions. The distribution in each class varies, and therefore the interpretation is different for each class. In practice, we consider that peculiarity values higher than 0.75 for nucleotide conformations or base-base interactions indicate regions of interest. The peculiarity values should never be interpreted independently of the distributions they are calculated from.

Many aspects of this work have implications for the *MC-Sym* molecular modeling computer program. First, the annotation procedure makes it possible to accurately and automatically build nucleotide conformations and base-base

interactions databases. The two distance metrics allowed us to define a more efficient sampling of the conformational space of nucleic acids. With the introduction of random structures in the database, the modeling process was improved and now reflects more accurately the actual conformational space of nucleic acids.

Obviously there exists a bias towards standard A-form helices in the databases since a large portion of the PDB structures contain only small DNA and helical RNA fragments. Nevertheless, with the ever increasing size of the nucleic acids repository, any bias should disappear as the conformational diversity increases. In fact, a significant addition was made to structure database and consequently to our knowledge on RNA structure with the recent introduction of rRNA crystals. For instance, between August 2000 and January 2001, the base-base interaction database coverage augmented by factors of 6% and 7%, respectively for the stacking and non-stacking interactions.

3·4 Materials and Methods

MC-Annotate was developped under a Linux environment using the *MC-Sym* software development library. The code was developed in C++, and is therefore portable to several other computer architectures such as IRIX and SunOS among others. To accomodate as many users as possible, the computer program was made available through the Web, which eliminates any need for a local installation. The Web application provides a user-interface allowing one to easily browse the results of a particular annotation. DNA and RNA 3-D structures in the PDB format can be submitted at www-lbit.iro.umontreal.ca/mcannotate/.

The Results were obtained using a database of nucleotide conformations and base-base interactions from 1630 RNA and DNA 3-D structures, taken from the PDB [8] database, as of January 2001, as well as from a limited number

of personal contributions. Some files from the PDB were not used since they did not conform to the PDB format specifications. Examples of faulty PDB files include those containing multiple models without the ENDMDL tags, and files with misidentified or incomplete nucleotides. Hydrogen atoms were added, to the PDB and NDB structures lacking them, prior to the analysis of bond lengths and angles as described in the Cornell et al. force field [16]. The resulting databases contain 108 240 nucleotide conformations (8.9% from rRNA structures), and 140 645 base-base interactions (10.8% from rRNA structures). The considered rRNA structures were 1FFK, 1FFZ, 1FG0, 1FJF, 1FJG, 1FKA, 1G1X, and 1HR0.

To evaluate the completeness of the database, the following protocol was designed to obtain a large sample of random conformations. 3000 single-stranded RNA structures were generated using *MC-Sym*, and were refined using a 15ps molecular dynamic simulation in which we reduced the temperature from 500K to 0K in the first 10ps. The oligomer "AACGCAUAGGUCCUUGA" was used with a sampling of five non-stacked base-base interaction examples from the *MC-Sym* database and the ideal A-RNA type nucleotide conformation for each position, defining a conformational search space of $5^{16} = 1.5 \times 10^{11}$ possible conformations. No distance constraints were used, except for a 1Å cutoff for steric clashes. Each generated structure had at least an RMSD of 5Å with any other, when computed for the nitrogen base atoms only, to guarantee a sufficient structural diversity. The program sander from the Amber 4.1 suite of programs was used with the Amber 94 force field [73] to optimize the stereo-chemical parameters of the generated structures. All 1–4 electrostatic interactions were scaled by a factor of 1.2 as suggested in [73]. A distance-dependant dielectric model, $\epsilon = 4R_{ij}$, for the Coulombic representation of electrostatic interactions was used, as suggested by [73]. As expected, the generated structures were significantly different from the starting *MC-Sym* structures since no equilibration was applied. This property was important since the goal of this protocol was to generate new base-base interactions. Here, 11 121 stacked and 36 879 non-stacked interactions

were generated using this protocol.

Acknowledgments This work was supported by a grant from the Medical Research Council of Canada (MT-14604) to FM. SL holds a Ph.D. scholarship from the Medical Research Council of Canada. PG holds a M.Sc. scholarship from the Fonds pour la formation de Chercheurs et l'Aide à la Recherche du Québec.

CHAPITRE 4

A NEW MOTIF IN THE LARGE RIBOSOMAL SUBUNIT IS REVEALED BY GRAPH THEORY

S. Lemieux et F. Major, *Science*, à soumettre.

ABSTRACT

Representation of RNA structures as graphs has been used both for modeling 3-D structures and predicting secondary structure. By extracting the minimal cycle basis of this graph, fundamental blocks of the RNA structure are isolated and can be compared to discover redundant motifs. The application of this technique to the 3-D structure of the large ribosomal subunit has led to the identification of a novel RNA 3-D motif similar to the GNRA tetraloop but that is formed by two independent strands. The structural environment of this motif suggests that it plays a role in the stabilization of tertiary contact by binding the minor groove of an adjacent helix.

4·1 INTRODUCTION

RNA structure information can be stored in a graph of relations (GOR), a general computer representation, where nucleotide information is attached to the nodes of the graph and nucleotide interactions are defined in the edges. The GOR is a general data structure in the sense that it allows one to store application specific information, and in particular to manipulate any of the common RNA structure abstractions: the sequence (primary structure), base pair set (secondary structure), and three-dimensional (3-D) atomic coordinates (tertiary structure). The GOR is an appropriate representation for RNA modeling system development [49, 52, 62, 63, 75, 100], and for systematic analysis of RNA 3-D structures [29].

The GOR is an undirected and unweighted graph. It allows pseudoknots, tertiary interactions, base triples and inter-strand stacking, and consequently it is neither outerplanar or subcubic [58]. RNA GORs are sparse graphs since steric repulsion limits the number of neighbors.

A new RNA structure abstraction suggested by the GOR representation is a subset of cycles of nucleotide interactions. A cycle in a GOR is a circular path of at least three edges. We contend that RNA function and structure are built from a minimum cycle basis (the smallest set of cycles allowing to rebuild the complete graph through cycle composition) [35] of the GOR: the elements of this basis are the fundamental building blocks of RNA structure. We studied a minimum cycle basis of the large ribosomal subunit (LRS), and discovered a new four-nucleotide motif that mimicks the structure of the GNRA tetraloop in internal loops, and systematically binds to the RNA minor groove of another stem.

The GOR of the LRS was computed from its x-ray crystal structure by using the algorithm described by Gendron et al. [29] and a new base pair detection algorithm developed in our laboratory. The GOR is then decomposed in a minimal

cycle basis using Horton's algorithm [38]. The cycles of this minimal cycle basis are made as compact as possible by always including the edges of the shortest path between any pair of nodes. The cycles were compared using an objective distance metric and classified using a hierarchical clustering algorithm. The motifs were defined as the recurrent cycles with high similarity.

4·2 MINIMAL CYCLE BASIS OF A GOR

When using the GOR to analyze a RNA 3-D structure, the spatial relation between nitrogen bases are extracted from the structure and stored in the edges of the graph. Homogenous transformation matrices (HTM) are used to represent these spatial relations. They are encoded as 4×4 matrices, are used to represent a combination of translation, rotation, sheer and scale [72]. Figure 4·1 shows a typical graph of relations for the hairpin at positions 2555–2580 in the LRS of *H. marismortui*, PDB code: 1FFK [5].

We decomposed LSR in substructures corresponding to the simple cycles of the GOR. We also seeked that these cycles be as compact as possible and that they are a sufficient representation of the entire molecule. This can be formalized by defining that a cycle contains a short-circuit if the shortest path between two nodes of the cycle in the graph is not part of this cycle, which would result in the possibility to form two shorter cycles. Thus, cycles of the smallest basis of the cycle space do not contain short-circuits (this can be deduced from Horton's algorithm [38]), and with respect to the composition operator on cycles, $\oplus$, they are sufficient to generate the complete cycle space of the molecule. This representation contains the complete structural information. This is done by introducing the concept of coherency for both the GOR and a simple cycle and then showing that the coherency of the minimal cycle basis implies the coherency of the GOR.

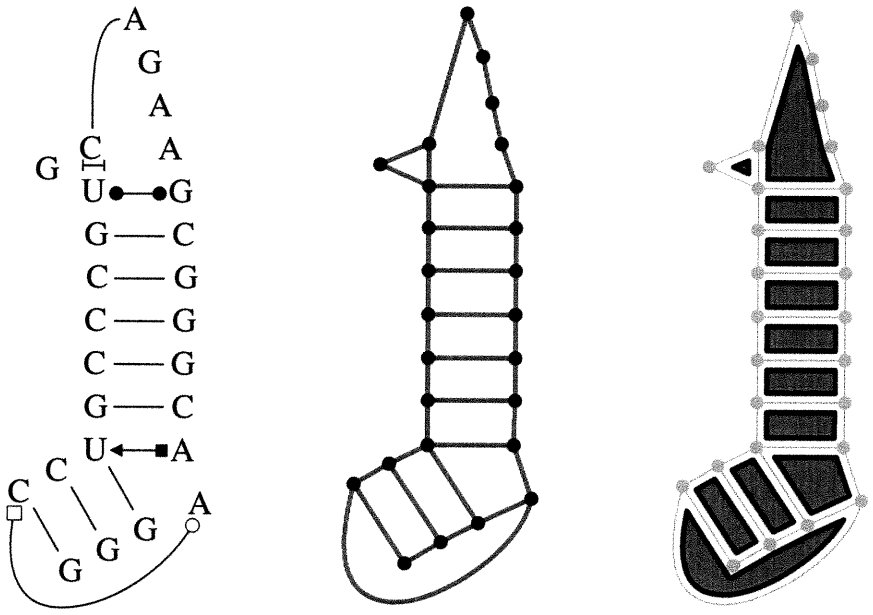


Figure 4·1 *Decomposition of hairpin 2555–2580 of the large ribosomal subunit in a minimal basis of the cycle space. (left) Graph of relation. (center) Corresponding undirected graph. (right) Minimal basis of the cycle space.*

We define that a cycle of relation is coherent if the product of HTMs assigned to its edges is equal to the identity matrix, I. A GOR is coherent if all its cycles are coherents. As a consequence of this definition, the coherency is a property that ensures that a given assignment of HTMs can be used to rebuild the 3-D structure. A subset of coherent cycles that can be used to infer the coherency of the complete molecule is sufficient to represent this 3-D structure. To manipulate cycles in graphs, we will use the cycle composition operator, $\oplus$, which is defined as the symetric difference between the two sets of edges representing the cycles. If both cycles are simple cycles and share a single contiguous segment of edges, the result of the composition will also be a simple cycle. The cycle space is known to be closed with respect to this operator [38]. The coherency of a cycle, c, can be infered from the coherency of a subset of cycles, $\{c_1, c_2, \ldots, c_k\}$, if $c = c_1 \oplus c_2 \oplus \ldots \oplus c_k$. Without loss of generality, let's assume that we have two coherent simple cycles $c_1 = (a, b)$ and $c_2 = (b, c)$

sharing a single contiguous segment of edges, b, and we want to show that their composition $c_3 = c_1 \oplus c_2$ is also coherent. We assume that the HTMs H_x represent the composition of transformations assigned to edges of segment x. Since c_1 and c_2 are coherent, we know that $H_b H_a = I$ and $H_c H_b = I$. From this, we obtain that $H_a = H_c$, resulting in $H_c^{-1} H_a = I$ and showing the coherency of cycle $c_3 = (c^{-1}, a)$. A cycle basis of graph G, $\mathcal{B}(G)$, is defined as a set of cycles that can act as a basis for the cycle space of G with respect to operator $\oplus$ [35]. This property ensures that any cycle in G can be obtain by a composition of cycles in $\mathcal{B}(G)$. The coherency of a cycle basis of graph G implies the coherency of G and thus is sufficient to describe the molecule represented by G. The minimal cycle basis provides the smallest set of cycles, with respect to the sum of edges (in the current implementation, no weights are assigned to edges so this criterion correspond to the total number of edges), sufficient to represent the 3-D structure of the RNA. It could also be argued that the ideal decomposition would be the one that minimizes the length of the longest cycle, but [14] have shown that both criteria would result in the same set of cycles.

4·3 MOTIF DETECTION

We propose a new hierarchical organization of RNA structure based on the graph representation where the first level is the nucleotide (nodes of the graph), the second level is the binary relations (two nodes linked by an edge), the third level is the cycles of a minimal cycle basis of the GOR, and the fourth level is the complete GOR. The first level of structure is used when analyzing sequence motifs, torsion angle motifs or pseudo-torsion motifs [22]. The second level is considered when non-covalent interactions, including tertiary interactions, are analyzed, as in MC-Annotate [29]. Cycles of the minimal basis are thus the next step in the understanding of the network of molecular interactions observed in RNA structures. We propose that such cycles should be regarded as fundamental

building blocks to identify structural motifs in RNA 3-D structures and we present a method to identify redundant building blocks. A distance metric between two HTMs, $d(M_1, M_2)$, was defined in [29]. Building on this definition, we propose the following metric between two cycles, c^1 and c^2:

$$
\begin{aligned}
d(c^1, c^2) &= \min\left(\min_{p=0}^{n-1} f(c^1, c^2, p), \min_{p=0}^{n-1} r(c^1, c^2, p)\right) \qquad (4.1)\\
f(c^1, c^2, p) &= \sqrt{\sum_{i=1}^{n} d(c_i^1, c_{(i+p) \bmod n}^2)}\\
r(c^1, c^2, p) &= \sqrt{\sum_{i=1}^{n} d\left(\left(c_{n-i+1}^1\right)^{-1}, c_{(i+p) \bmod n}^2\right)}
\end{aligned}
$$

where c^1 and c^2 are two cycles of equal length, n, c_j^i represents the j^{th} HTM of cycle c^i, and p corresponds to the phase used to superpose the two cycles. This distance metric identifies the optimal superposition of both cycles with respect to the cartesian distance based on the HTM distance metric. Computing the distance between two cycles of length n is done with a running time in $O(n)$. The motif analysis of a RNA is done by applying hierarchical clustering on the distances obtained from this metric.

4·4 RESULTS

The LRS was completely annotated as described in [29] but using an improved method for base pairs detection described in [51]. The results were encoded as a GOR consisting of 2 828 nodes and 4 642 relations. The decomposition of the GOR in a minimal cycle basis using Horton's algorithm is a CPU intensive process that require a running time in $O(n^7)$, where n is the number of nucleotides in the molecule. Despite this daunting worst case running time, the algorithm returns the minimal cycle basis after 28 minutes 20 seconds of CPU time (PIII-600) and the process occupies 148 Mb of memory to store the cycles generated in the intermediate step of the algorithm [38]. The basis contains 1 816

Length	#	%	$\sum$ #	$\sum$ %
3	572	31.4%	—	—
4	905	49.8%	1477	81.3%
5	123	6.8%	1600	88.1%
6	58	3.2%	1658	91.3%
7	29	1.6%	1687	92.9%
8	20	1.1%	1707	94.0%
9	15	0.8%	1722	94.8%
10	11	0.6%	1733	95.4%
11	5	0.3%	1738	95.7%
12	10	0.6%	1748	96.3%
13	2	0.1%	1750	96.8%
14	0	0.0%	1750	96.8%
15	9	0.5%	1759	96.9%
>15		3.1%	1816	100.0%

Table 4·1 *Distribution of cycle lengths in a minimal cycle basis of the GOR of the large subunit of the ribosome. # and % report the number and proportion of cycles of a given length, while $\sum$# and $\sum$% indicate their cumulative values.*

cycles, the sum of cycle length is 8 710 and the longest cycle contains 66 edges. Table 4·1 shows the distribution of the cycle length in this minimal basis. On a general graph, the algorithm has a worst case running time in $O(n^7)$ but since the worst case of this algorithm is encountered for complete graphs, we expected more reasonable running time when using it with sparse RNA graphs. Tests made on random sparse graphs (results not shown) suggest that the topological properties of the graph greatly influence both the running time and the amount of memory needed. But currently, no formalism exists to fully describe the properties of RNA GORs, thus it is very difficult to obtain an upper bound for the worst case running time of this algorithm on RNA GORs. Since this molecule is known as the largest structural RNA, the current algorithm for the decomposition has achieved a practical efficiency in the context of RNA GORs.

Analysis of the resulting minimal basis of the cycle space of the LRS was performed by computing for each cycle length the distance between each pair of cycles. From the resulting matrix, hierarchical clustering (using the nearest

neighbor algorithm with maximum distance criterion) was applied to obtain a classification of the cycles. Clusters were visually identified and the substructures they represent were extracted and superimposed for 3-D visualization of the putative motifs. This tedious task was applied for cycles of length 4, Figure 4·2 shows the classification obtained where two clusters are identified for further analysis.

Both motifs identified on Figure 4·2 were further investigated in the following way: for each motif, we extracted all occurences from the LRS and superposed them using three atoms and pseudo-atoms per base. The atoms used were the N9 for purine or N1 for pyrimidines, a pseudo-atom at 1 Å of the N{1,9} atom in the direction of the C1'-N{1,9} vector, and another pseudo-atom at 1 Å of the N{1,9} in the direction corresponding to the normal of the nitrogen base plane. These three atoms were selected to compare substructures with different sequences and without relying on the backbone conformation (motifs were selected for the similarity of their nitrogen bases relations and we wished to emphasize this property in the superposition as well).

The GNRA tetraloop motif presented in Figure 4·3 is one of the most studied RNA motif. This loop is thermodynamically very stable [42] and its main function is to form tertiary interactions by binding to specific tetraloop receptor motifs. Thus, there is no surprise in easily identifying 10 nearly identical occurences of this motif in the LRS. Two sequences are of particular interest since they do not conform to the usual GNRA sequence motif associated with this type of tetraloop. Substructures #11 and #13 on figure 4·3 respectively use sequences UCAC and CAAC, forming non-canonical base pairs U•C and C•C, which are both isosteric to the sheared G•A present in the standard GNRA structure. Substructure #11 appears on the surface of the ribosome and may only contribute to the formation of a thermodynamically stable hairpin. On the other hand, substructure #13 appears to stabilize a tertiary interaction with a distant

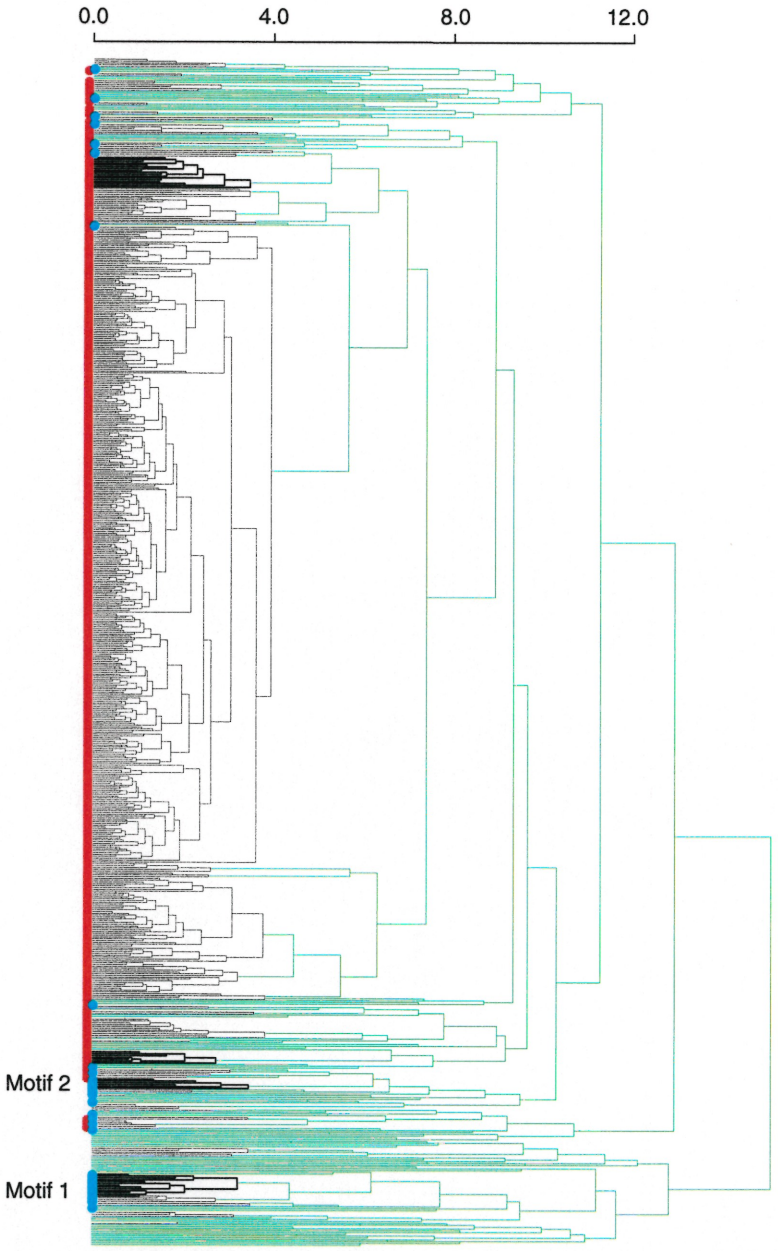


Figure 4·2 *Hierarchical clustering of four edges cycles in the large ribosomal subunit. Clusters shown with black lines are grouped within a maximal distance of 4.0 and should be considered as putative motifs. Two motifs are identified with bold lines and will be presented in more details in Figures 4·3 and 4·4. Red dots correspond to stacked base pairs, while blue dots represent GNRA-like tetraloop motifs, defined as one base pair, two base stacking and a covalent bond.*

stem, effectively showing that this type of function is not limited to tetraloop with a GNRA sequence. Both substructures are shown in yellow and using thicker sticks on Figure 4·3 to emphasize the structural similarity they share with other GNRA tetraloops. This result suggests that the GNRA tetraloop motif has more sequence flexibility than its name implies, and perhaps should be renamed to avoid possible confusion.

The motif presented in Figure 4·4 is of particular interest. The main difference between the GNRA tetraloops presented in Figure 4·3 and this motif is that the topmost nucleotide (second position in the GNRA tetraloop) is flipped on the axis of the nitrogen base, displacing the backbone on the other side of the motif. What is striking is the constancy of the environment in which it is found: the eight occurences identified in the LRS always contact a minor groove of a Type-A helix on the side of the motif corresponding to the Watson-Crick face of the third base. Despite the fact that there appears to be no direct chemical interaction between the motif and the minor groove of the associated helix, the systematic conservation of the structural environment of this motif is a strong indication that it is selected to promote this kind of tertiary contact. Another observation on this motif is the strong sequence conservation among its occurences, the sequence GA.AA is favored in seven of the eight occurences with one exception where the sheared G•A is replaced by an isosteric G•U one H-bond base pair.

4·5 Discussion

Three main issues remain with the proposed method. First, there is no formal definition of what consitutes a motif in RNA structures, redundancy is the most commonly admitted feature of a motif and the one used here. But, ideally, the concept of motif implies that the observed feature is *unusually* recurrent, raising the possibility that it has been selected and that it plays an important role (either

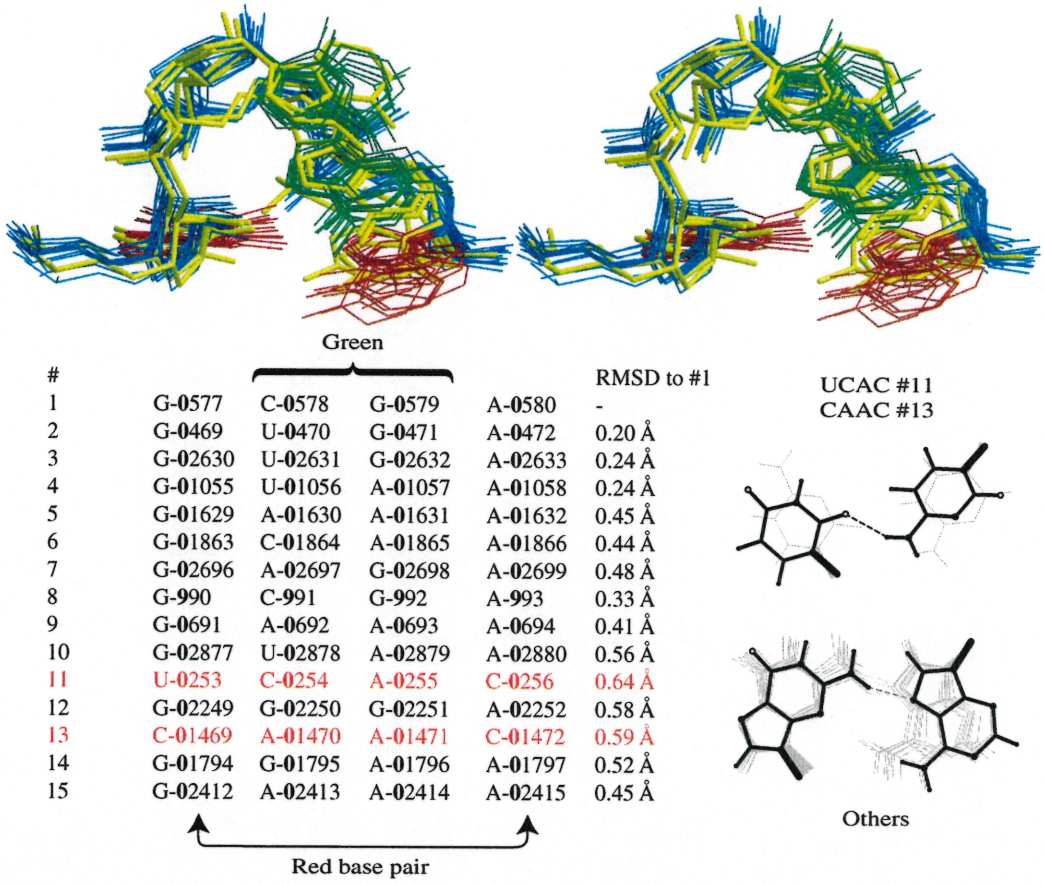
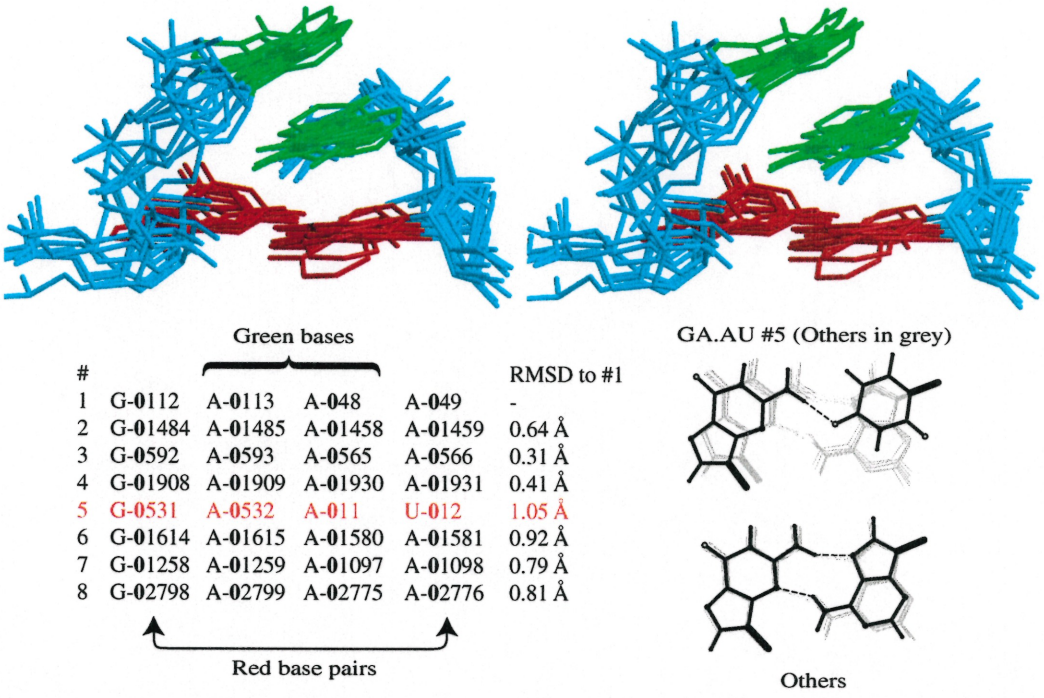


Figure 4·3 *Motif 1: GNRA-like tetraloops from the large ribosomal subunit. The stereo shows the superposition of 10 tetraloops adopting a similar GNRA-like conformation. The backbone is shown in blue, the base pair in red and the two other nitrogen bases in green. Two loops with non-GNRA sequences (CAAC and UCAC) are adopting a GNRA-like conformation and are shown with thick yellow sticks. The table shows the positions of each occurence of this motif in the large ribosomal subunit and the RMSD to substructure #1. The base pairs closing the loops are shown on the lower right, the pyrimidine•pyrimidine base pairs of the CAAC and UCAC loops are shown separately.*



#					RMSD to #1
1	G-0112	A-0113	A-048	A-049	-
2	G-01484	A-01485	A-01458	A-01459	0.64 Å
3	G-0592	A-0593	A-0565	A-0566	0.31 Å
4	G-01908	A-01909	A-01930	A-01931	0.41 Å
5	G-0531	A-0532	A-011	U-012	1.05 Å
6	G-01614	A-01615	A-01580	A-01581	0.92 Å
7	G-01258	A-01259	A-01097	A-01098	0.79 Å
8	G-02798	A-02799	A-02775	A-02776	0.81 Å

Figure 4·4 *Motif 2: Revisiting the GNRA, but without forming a loop. The stereo shows the superposition of 8 occurences of this motif in the large ribosomal subunit. The view angle and color code are similar to Figure 4·3. The superposition of the base pairs are presented on the lower right panel. All substructures use the sequence GA.AA except substructure #5 that uses GA.AU, forming an isosteric G•U base pair.*

structural or catalytic) in its host organism. In sequence motif detection, the ratio of observed/expected occurences is commonly used to reflect the stringency of the selection process. With RNA 3-D structures, there is still no way to estimate the expected frequency of a given motif. A solution to this problem lies in the derivation of a random model for these cyclic substructures that would be used to obtain an expected frequency of occurence. The second issue is that motifs identified using the cycles from a minimal basis of the cycle space are often partial motifs. This is due to the fact that, frequently, the motifs are built from several of these cycles. The method presented here should be used as a tool to rapidly identify motif *seeds*. The third issue is that it is possible that the minimal basis of the cycle space is not unique and in the context of motif detection it would be preferable to use the union of all minimal cycle bases of the GOR. This union is called the set of relevant cycles of a graph and an efficient algorithm has already been proposed [90]. We refrained from implementing this algorithm since the algorithm is no more polynomial and in fact, the set of relevent cycles is not guaranteed to be of polynomial cardinality (this situation arises naturally when a RNA GOR forms a pseudoknot or some type of tertiary contact).

By showing that the decomposition of a RNA GOR in its smallest basis of the cycle space is still sufficient to infer the coherency of a 3-D structure we have provided a genuinely novel representation of RNA molecules. We have shown that this representation can be useful for structure analysis, but it also offer new perspectives for the modeling of RNA 3-D structures. Indeed, the cycles are the smallest non-trivial modeling problems and, in this respect, provide a natural way to tackle the problem of modeling the 3-D structure of an entire RNA molecule. On an other scale, properties of cycles of HTMs were used as a post-processing constraint in the *MC-Sym* program to obtain a 3-D model of the regular hexamer of the prohead RNA of *B. subtillis* Φ29 bacteriophage [100].

Here we proposed a new level of analysis for RNA 3-D structures that can

be applied to motifs larger than a single relation. We applied this technique to the 3-D structure of the LRS and identified two non-GNRA tetraloop adopting conformations strickingly similar to the standard GNRA tetraloop and a novel motif that mimicks the GNRA tetraloop but without forming a loop. Most importantly, this decomposition isolates the fundamental building blocks of RNA structures and is thus a critical step in identifying structural motifs in 3-D structures.

ACKNOWLEDGMENTS

This work was supported by a grant from the Canadian Institutes of Health Research (CIHR) (MT-14604) to FM. SL holds a Ph.D. scholarship from CIHR.

CHAPITRE 5

AUTOMATIC 3-D MODELING OF RNA USING THE MINIMAL CYCLE BASIS DECOMPOSITION

S. Lemieux et F. Major, manuscrit en préparation.

The minimal cycle basis decomposition of a graph of relations (GOR) was introduced in [50]. This decomposition allows for a new definition of the RNA 3-D modeling problem, consisting in two steps: first, we obtain a decomposition and model each cycle independently; second, solutions obtained for each cycle are combined to rebuild a complete 3-D structure. The modeling of a cycle consists in finding assignment of HTM to the edges so that the coherency of the cycle is maintained. This step can be computed in parallel for each cycle and is conveniently distributed on a cluster of workstations. Results of this step could be archived and retrieved when identical cycles are requested. The second step is to decide which assignment to retain for each cycle and rebuild the 3-D structure from these asssignments. Two cycle assignments sharing a segment of edges will be considered compatible if the distance between each pair of corresponding HTMs is below a cutoff. The identification of a set of cycle assignments is implemented as a heuristic search method using the distance between corresponding HTMs [29] as a minimization criterion.

On top of retaining only coherent assignments, two types of constraints have to be verified in both steps of the algorithm. By assuming that spatial relations between nitrogen bases are sufficient descriptors of the 3-D structure we can avoid to explore the conformational space of the sugar moiety (conformational

sampling in MC-SYM [52]). On the other hand, the phosphate groups (PO_4) provide an important steric constraint and should be included in the model as soon as possible. Since the phosphate group can be considered as a rigid body, we included its local referential in the database of relations, this addition being done only for adjacent relations. The two types of constraints are the possibility to rebuild a ribose moiety linking adjacent phosphate groups and a nitrogen base (called the ribose constraint) and the absence of atomic collision (called the collision constraint) sometime refered as the steric clash. Both contraints will be presented with further details and their implementation will be presented in the following sections.

Before describing the details of the modeling engine, we will first discuss the various implications of applying the minimal cycle basis decomposition for the modeling of RNA 3-D structures. Then, we will introduce the constraints that are considered and the optimization methods used to solve the modeling problem. Finally, results will be presented for two test molecules: the first one is an eight nucleotides cycle from the ribosomal RNA binding protein L11 (PDB code: 1QA6, [15]) and the second one is the hairpin 2555–2580 of the large ribosomal subunit of *H. marismortui* (PDB code: 1FFK, [5]). For brevity, these two molecules will be respectively refered as the L11 loop and the hairpin 2555–2580.

5·1 THE MINIMAL CYCLE BASIS FOR MODELING

A RNA 3-D structure can be represented as a graph of relations (GOR) where each node represents a nucleic acid base and edges represents HTM between the two connected nodes. This representation have been used for RNA modeling [52], and for the analysis of RNA 3-D structures [29]. With this type of representation, the structure of a RNA can be manipulated by modifying the relative orientation and translation of nitrogen bases, and by considering each rigid nucleic acid

base as a local referential, these manipulations are easily carried out by using homogenous transformation matrices (HTM) to represent the relation [72]. HTMs will be limited to rigid body transformations, and thus will only represent a combination of translation and rotation. The inverse of an HTM always exists and can be obtained in constant time. Sequences of translations and rotations can be composed by multiplying the correspondant HTMs. The modeling program MC-SYM [52, 63] successfully uses this representation to infer 3-D structures for several RNA molecules [49, 76, 100].

The 3-D modeling of RNA molecules corresponds to the exploration of possible assignments of HTMs to edges of the GOR and the verification that these assignments follow some predefined stereo-chemical criteria. Since most of the stereo-chemical criteria (Van der Walls and electrostatic energies, bond lengths, angles and torsion, etc.) can only be applied on the all-atoms representation of the molecule, it is essential to check that the assignments of HTMs allows to build the 3-D structure, a property that will be defined as the coherency of the assignment. A straightforward way to solve this problem is to compute the product of HTMs for every possible cycle of the graph for each assignment, which would require exponential amount of time in the worst case just to test one specific assignment. The exploration of all possible assignments being a combinatorial process in itself, it is essential to avoid another embedded combinatorial process for the simple validation of each assignment. We propose to use the decomposition of the GOR in its minimal cycle basis for this purpose and we will show that if the coherency property is respected for the cycle basis, it will be respected for any cycle. The extraction of a subset of cycles will be run prior to the modeling process and will require a worst case running time in $O(n^7)$, where n is the number of nucleotide in the molecule, allowing for the verification that an assignment is consistent with a time in $O(n^2)$ (corresponding to the maximum length of the minimal basis). Despite a daunting worst case running time in $O(n^7)$, the algorithm performs in reasonable amount of time on

the largest RNAs [50]. We will also see that this decomposition suggests an efficient parallelization of the modeling process.

In the molecular modeling program MC-SYM, coherency of the HTMs assignment is guaranteed since only a spanning tree of the graph is used to rebuild the 3-D structure. Since no cycle is present in the set of edges used, any assignment can consistently be converted to a 3-D representation. This approach results in a new problem, determining the spanning tree that is the most likely to yield the desired results... This problem was conveniently left to the modeler. Currently, there exists no systematic way of determining this spanning tree, and this appears as the most challenging problem in the design of an automated RNA modeling system. In 1995, Turcotte [89] proposed and implemented a modeling approach using quadratic numerical optimization applied on the HTMs. The results obtained in this work insist on the necessity to split the GOR in subgraphs (called groups) in order to achieve feasibility, but no rationale was proposed to automate this grouping step. Here, we show that the decomposition of a GOR in its minimal cycle basis [14, 38] can play a similar role, strongly limits the topology of these subgraphs and is readily implemented.

A cycle of relations is said to be coherent if the product of HTMs assigned to its edges is equal the identity matrix, I. This definition can be weakened by requiring that the distance between the HTM, T, and the identity matrix, I, is below a predefined cutoff: $d(T, I) < \epsilon$. If ϵ is choosen to be small, the resulting approximation can be neglected. A distance metric between two HTMs has already been presented in [29]. The coherency of a GOR or of a cycle confirms that it is possible to build a consistent 3-D structure from this assignment of HTM to edges.

By using the smallest basis of the cycle space, we assume that the sum of edges that is minimized in the Horton algorithm [38] also results in a basis that will result in the smallest modeling problem. For modeling, the space to explore

depends on the length of cycles in this decomposition. Since each edge will be assigned a set of possible HTMs, the cardinality of the space, $|S|$, associated with a basis, $B(G)$, will be:

$$|S| = \prod_{c \in B(G)} \left(\prod_{e \in c} s(e) \right),$$

where $s(e)$ is the number of HTMs used to sample the spatial relation corresponding to edge e. By assuming a constant sampling for each edge ($s(e) = s$ for any $e \in E$), we obtain:

$$\begin{aligned} |S| &= \prod_{c \in B(G)} s^{|c|} \\ \log_s |S| &= \sum_{c \in B(G)} |c| \end{aligned}$$

Since seeking the basis that will result in the smallest search space, $|S|$, is the same as searching for the decomposition with smallest $\log_s |S|$, the smallest cycle basis as defined in [38] results in the basis that will minimize the amount of work for subsequent modeling steps.

A rather technical complication that arises from using HTM to represent spatial relations on an undirected graph is that HTMs are specific to one direction, they should be applied in the same direction that they were extracted, suggesting a contradiction with the undirected property of the GOR. Since the inverse of a HTM is easily obtained, it is always possible to reverse the HTM to fit the direction of application. In the rest of this paper, and only when appropriate, arrows will be used on edges to indicate the direction to use for the encoded HTM.

5·2 CONSTRAINTS

Two types of constraints are essentials to obtain a stereo-chemically sound model of a RNA. First, one has to make sure the covalent structure of the RNA molecule is respected. Second, the model should be free of atomic collisions.

Other types of constraints will eventually be implemented in the system to encode various types of informations that can be obtained on RNA molecules.

A. RIBOSE CONSTRAINT

The ribose constraint can be tested during the optimization of independant cycles for riboses that are completely determined by the optimized cycle. Other riboses can only be placed during the assembly of adjacent cycles. In both situations, the fixed position of the phosphate groups ensures that this optimization can be done independently for each ribose without affecting the rest of the backbone. This constraint is tested by optimizing the five free parameters of the ribose moiety (four torsion angles and the pseudorotation angle [79], the ribose pucker amplitude is kept fixed) in order to minimize the RMSD to anchoring atoms: O3', O5' and both P (see figure 5·1). For this numerical optimization we prefered, for simplicity, an optimization method that does not require the derivative of the RMSD with respect to the torsion angles (and specially with respect to the pseudo-rotation angle). We also wished to avoid the problem related to the inverse kinematics of cyclic robot arms (see [72] for a discussion of this problem), and thus did not investigate the possibility of an analytic solution. Methods presented in section 8.4 of [6] were implemented and tested. They were compared by measuring the average CPU time spent to optimize a ribose conformation. To our surprise, the fastest method turned out to be the cyclic coordinates descent method without the use of the linear search, essentially the simplest approach. This optimization is relatively expensive (140.1 ribose/sec on a PIII-600) and is only applied once a coherent cycle assignment is obtained.

A ribose is considered correctly placed if the sum of squared distances to anchor atoms is below a cutoff of 0.5 Å. The rate of rejection due to this constraint is difficult to evaluate since it depends greatly on the nature of the cycle. When tested on the eight nucleotides loop of the rRNA binding protein

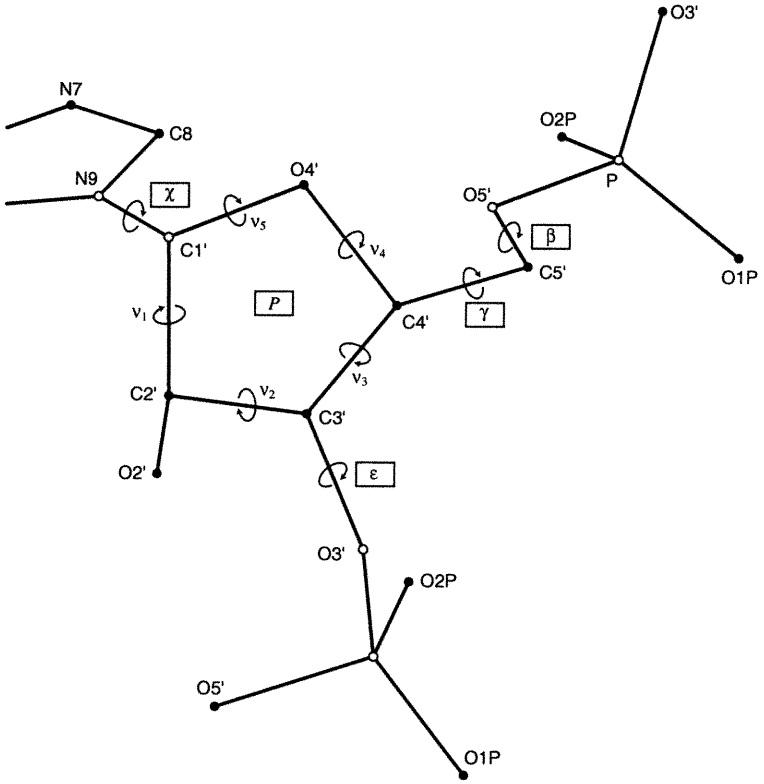


Figure 5·1 *Ribose structure used to rebuild the RNA backbone. Atoms are identified by circles, empty circles represent anchor atoms already positioned by the placement of nitrogen bases and phosphate groups. Free torsion angles are shown with circular arrows, degrees of freedom are identified by boxes.* P *represents the pseudorotation angle and is used to determine torsion angles* ν_{1-5}.

L11, an average rejection rate of 50% was observed for all riboses. This result suggests that for long cycles requiring the placement of numerous riboses this constraint could become the limiting factor to produce valid cycle assignments. The probability that a structure is rejected by this constraint can be approximated at 0.5^n where n is the number of riboses to place. To avoid this problem, it is possible to increase the rejection cutoff, lowering the rejection rate, but resulting in more distorted structures. It would also be possible to precompute the ribose parameters for each pair of adjacent relation. The time and memory needed for this step depends on n, the number of riboses to place, and s, the number of HTMs that are considered at each relation, and are both bound by $O(ns^2)$. The test structure used required the placement of 7 riboses and we estimated the acceptation ratio to be around one structure out of 128. In practice, this aspect of the system do not seem problematic since most cycles in a RNA strcutures are of length below 5 and contains the placement of between 0 and 3 riboses [50].

B. Collision constraint

The collision constraint is first tested during the cycle optimization by adding a penalty term for each collision between rigid objects (either nitrogen bases or phosphate group). A collision is detected if the distance between two heavy atoms is below 2 Å, corresponding to a hard-sphere potential. This verification is performed only for rigid objects not belonging to the same relation (extracted relations are assumed to be sound). Once a coherent cycle is obtained and smoothed (see next section), the collision constraint is retested, this time rejecting solutions where at least one collision occurs.

The collision constraint is tested by verifying that the distance for each pair of atoms between the two rigid objects is over 2 Å. A major drawback of this aproach is that the 3-D structure need to be explicitly rebuilt before testing the constraint. It would be possible to accelerate the collision detection using

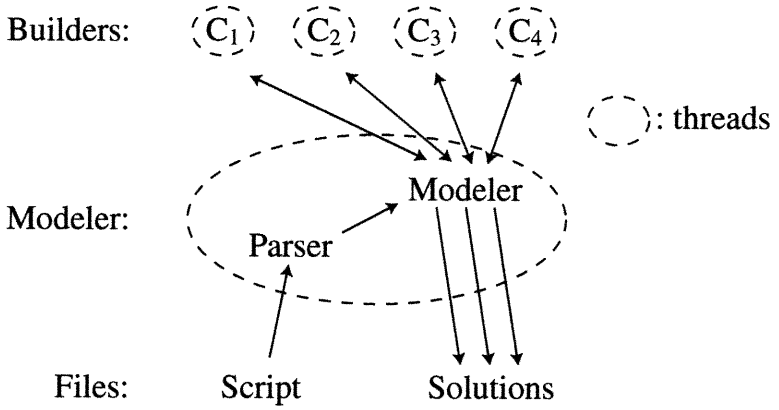


Figure 5·2 *Schematic representation of the modeling process. The process is divided in two types of threads, the builders are computing valid 3-D structures for each cycle and sending these structures to the modeler. The modeler receives cycle structures from the builders and assembles them to produce complete structures. All file I/O occur in the modeler thread.*

sophisticated collision detection algorithms, but since each rigid object contains a small number of atoms (around 10 for bases and 5 for phosphate groups), the $O(n^2)$ running time is not problematic.

5·3 AUTOMATED MODELING

Figure 5·2 shows a schematic of the overall process. The cycle optimization method is applied by the builder threads, one is created for each cycle in the structure and their optimizations are all carried out in parallel. A modeler thread is responsible to receive, through shared memory, the cycle structures optimized by the builders and assemble them in complete structures. The modeler acts as a front-end to the user, reading the input script file and writing the output structures in all-atoms PDB format.

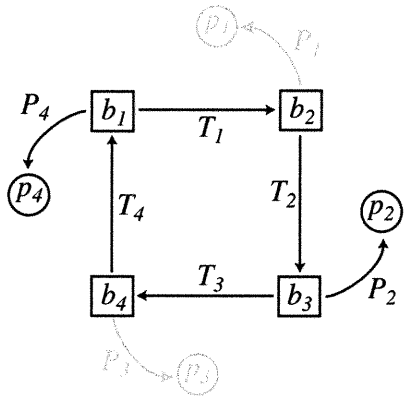


Figure 5·3 *Example of a cycle with four relations.* b_i *and* p_i *represent respectively the nitrogen bases and phosphate groups,* T_i *represent the HTM used to place* b_{i+1} *with respect to* b_i *and* P_i *are HTMs used to place* p_i *with respect to* b_{i+1}. T_i *and* P_i *are stored and retrieved in pair from the database of relations. Elements shown with circle represent the phosphate groups and are optional.*

A. Cycle optimization method

This section describes the algorithm used by the builder threads to build coherent structures for each cycle of the molecule. As an example, we will first suppose the optimization of a hypothetical cycle of four relations, two of which assign a phosphate group (see figure 5·3). This cycle can represent a typical RNA base pair tandem. In the following text, T_i represents the HTM assigned to relation i and T_{kji} is equivalent to $T_k T_j T_i$. For a given assigment of HTMs to edges of the cycle, the product of these HTMs will sometime yield a residual transformation refered as the "error", E, of this assignment.

For a serie of n HTMs, it is possible to compute the rebuilding error of the cycle, $S(E)$, in n different ways. The function $S()$ is the *strength* of a transformation and is defined in [29]. As an example:

$$E_4 T_4 T_3 T_2 T_1 = I \qquad \text{or} \qquad T_4 T_3 E_2 T_2 T_1 = I$$

0:	T_1	$\underline{T_2}$	T_3	T_4
1:	$\boxed{T_{21}}$	$\boxed{T_{32}}$	T_{43}	T_{14}
2:	$\boxed{T_{4321}}$	$\boxed{T_{1432}}$	$\boxed{T_{2143}}$	$\boxed{T_{3214}}$

Table 5·1 *Example of the array used to compute the minimal error of rebuilding (MRE) in $O(n)$. Boxed positions indicates what needs to be recomputed when T_2 is modified.*

$$E_4 = (T_4T_3T_2T_1)^{-1} \qquad E_2 = (T_2T_1T_4T_3)^{-1}$$

$$S(E_4) \neq S(E_2)$$

We will use the notation E_i to represent the residual transformation after composing HTMs starting with T_i. For a given assignment, the rebuilding sequence minimizing the rebuilding error should always be used. The optimization problem thus consists in finding a cycle assignment that minimizes the minimal rebuilding error (MRE). To solve this problem, we implemented a probabilistic Monte Carlo algorithm in which at each iteration a random relation is randomly changed and if the new MRE is lower than the current, the modification is conserved. If the MRE gets below 1.25, the cycle assignment is accepted.

Naively, the modification of a single assignment requires to recompute in n different ways the rebuilding error to find the new MRE, resulting in an $O(n^2)$ complexity to obtain the MRE at each iteration. It is possible to accomplish this operation in $O(n)$ by using an array containing the values of intermediate computations and allowing their reuse. Figure 5·1 shows this array for our four bases example. Each position of this array is defined by:

$$M_{ij} = \begin{cases} \text{if } i = 0 & T_j \\ \text{else} & M_{(i-1)((j+2^{i-1}) \bmod n)} \cdot M_{(i-1)j} \end{cases}$$

Figure 5·1 shows the HTMs that have to be recomputed when T_2 is modified.

From this, we obtain that on row i, 2^i positions need to be recomputed. Since the array has $\log_2 n$ rows and that the update of one position requires only one matrix multiplication, we obtain that the number of required multiplications is:

$$\begin{aligned}\left(\sum_{i=0}^{\log_2 n} 2^i\right) - 1 &= 2^{(\log_2 n)+1} - 2 \\ &= 2 \cdot 2^{\log_2 n} - 2 \\ &= 2 \cdot n - 2 \\ &\in O(n)\end{aligned}$$

To initialize the array, $n \log_2 n$ matrix multiplications have to be computed. This technique assumes that the cycle length is a power of 2. In other cases, the array has to be extended to the next power of 2 by setting $M_{0j} = I$ for $j \geq n$. This extension does not affect the complexity of the algorithm.

A cycle assignment is accepted if its MRE is below a cutoff, ϵ, that was fixed at 1.25. This leaves us with the problem that this assignment, even if it is close to, is not coherent. The simplest solution to obtain a coherent cycle is to put all the error in one of the HTM that is adjacent to the position used for the MRE. This method results in substantial changes on the selected HTM when ϵ is increased. The approach implemented consists in redistributing the MRE equally accross all HTMs of the cycles. The *smooth* procedure is applied in an iterative way following these steps:

1. Select position p corresponding to the MRE.
2. Obtain HTM E_p corresponding to the residual transformation when multiplying the HTMs by starting at position p.
3. $T_p \leftarrow (\lambda \cdot E_p)T_p$, where $0 < \lambda < 1$ is a small constant.
4. Continue until the current MRE is below 0.05.

The multiplication of a HTM by a scalar is computed by transforming the HTM to a translation vector and a rotation quaternion. Scalar multiplication is then applied to this representation and a new HTM is then computed. This technique is the classical approach to interpolate between two HTMs in computer animation. The results presented in this chapters were computed with $\lambda = 0.01$. Once the assignment is smoothed, the 3-D structure has to be rebuilt and a final check is done to verify that there is no atomic collision and that all ribose can be rebuilt correctly.

B. MERGING CYCLE STRUCTURES

Assembling cycle structures to obtain a complete solution for the molecule also results in a problem where the order of the operations has an effect on the quality and quantity of the results. To avoid delegating this task to the user or having to design some heuristic to find the ideal sequence of operations, we implemented a stochastic algorithm based on the one presented in [24].

The modeling process is composed of successive generations, during which each builder contributes one cycle structure to a pool of substructures. Once all these structures are collected for a given generation, the modeler thread randomly selects pairs of substructures and tries to agglomerate them using superposition of shared bases between the two substructures. The number of selected pairs was arbitrarily set to twice the number of cycles in the structure. At the end of each generation, structures corresponding to an agglomeration of all cycles are considered complete and removed from the pool. The detailed steps followed at each generation are:

1. Pool the builders until one 3-D structure is obtained for each of them.
2. Execute $2n$ steps of agglomeration, where n is the number of cycles in the structure. One substructure is selected from the pool and tested for

combination with every other substructure. The pair minimizing the RMSD of the superposed shared bases is combined if the RMSD is below 1.25 Å. The RMSD corresponding to the best superposition is obtained using the algorithm developed by [43,44].

3. A structure where all the cycles are assigned is considered completed and removed from the pool.

C. EVALUATION

The performance of the algorithm to optimize cycle models was assessed by using an eight nucleotides cycle (positions C142-C149) from the ribosomal RNA binding protein L11 (PDB code: 1QA6, [15]) having the following sequence: CGUAAUAG. Nitrogen bases 1 and 8 are forming a Watson-Crick base pair, nucleotides 1 to 3 are stacked and 4 to 8 are also stacked. The more classical anticodon example was avoided because of the presence of modified nitrogen bases and its over-representation in the database.

To evaluate the accuracy of the complete method we selected the hairpin 2555-2580 of the large ribosomal subunit of *H. marismortui* (PDB code: 1FFK [5]). This structure was randomly selected as being small, but structurally not trivial to model. The important features of this structure are an anticodon-like loop resulting in a seven node cycle; two bulged nucleotides including one involved in a base triple to form a tertiary interaction; and two base triples forming non-canonical base pairs. The graph of relations corresponding to this structure and its 3-D conformation are presented in figure 5·4. The decomposition of this graph of relations in a minimal basis of its cycle space is shown in figure 5·5. Finally, figure 5·6 presents the script used to build the hairpin 2555–2580 of the large ribosomal subunit. The scripts encodes in a textual form the information contained in the graph of relations.

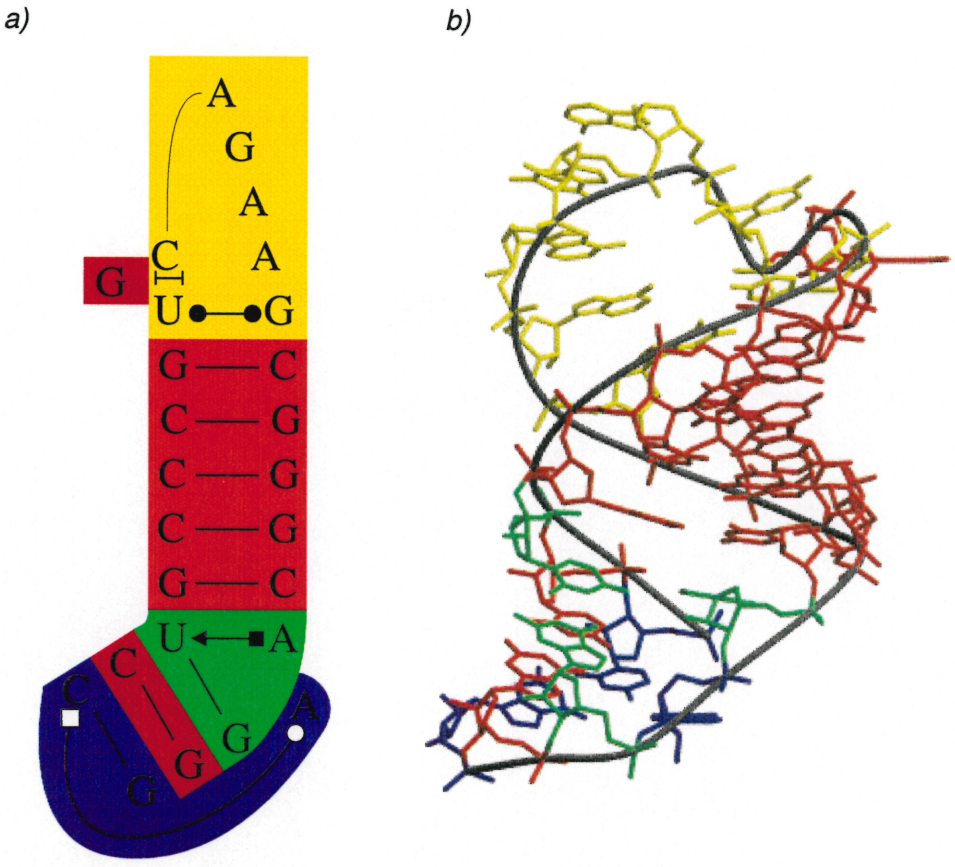


Figure 5·4 *Hairpin 2555-2580 of the large ribosomal subunit of* H. marismortui. *On both panel, yellow is used to indicate the cycle of relations corresponding to the loop, green for the central base triple and blue for the terminal one. a) Graph of relations of the structure. Symbols used to indicate pairings and stacking are defined in [56]. b) 3-D structure extracted from 1FFK [5].*

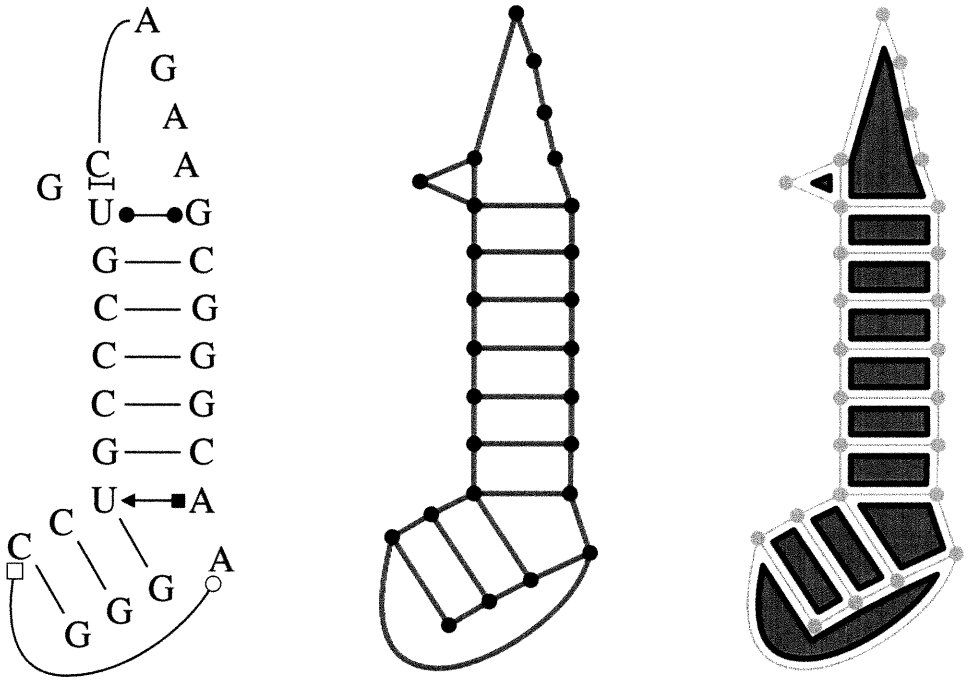


Figure 5·5 *Decomposition of hairpin 2555-2580 in a minimal basis of its cycle space. (left) Graph of relations. (center) Corresponding undirected graph. (right) minimal basis of the cycle space.*

```
ensemble_name ("1ffk_2555")

sequence (
  rna A2555
  CCUGCC CGUGC AGAA GCG GGCAAGGG
)

pair (
  A2561 A2572 { Ww/Ww cis }
  A2562 A2571 { Ww/Ww cis }
  A2563 A2570 { Ww/Ww cis }

  A2560 A2573 { Ww/Ww cis }
  A2559 A2574 { Ww/Ww cis }
  A2558 A2575 { Ww/Ww cis }
  A2557 A2576 { Hw/C2 trans }
  A2555 A2577 { Hw/Ww trans }
  A2557 A2578 { Ww/Ww cis }
  A2556 A2579 { Ww/Ww cis }
  A2555 A2580 { Ww/Ww cis }
)

stack (
  range A2555-A2557
  range A2558-A2563
  range A2566-A2576
  range A2578-A2580

  single A2563 A2565
)
```

Figure 5·6 *Example script used to model hairpin 2555–2580 of the large ribosomal subunit. The nomenclature used to describe base pairs is described in [51]. The syntax of the script is similar to the* MC-SYM *syntax (http://www-lbit.iro.umontreal.ca/mcsym), but without specifying ribose conformation, sampling size and building order.*

5·4 RESULTS

Results are presented in two steps corresponding to the analysis of the cycle optimization and the algorithm for the assembly of cycle structures.

A. CYCLE OPTIMIZATION

After 17 hours of cpu time (PIII-600), 1572 structures were generated for the eight nucleotides cycle from the rRNA binding protein L11, resulting in a throughput of 1.54 structures/min. To estimate the redundancy of this set of structures, each one was compared to all preceeding ones (using a RMSD criterion on the nitrogen bases). By assuming that at a RMSD of 1.0 Å two structures are considered identical we obtain a redundancy rate of below 1%. This suggests that no redundancy filter should be used on the generated structures for the cycles. The structure most similar, according to the RMSD criterion, is at 1.10 Å of the X-ray structure. The mean RMSD to the X-ray structure is 2.50 Å and the maximum RMSD observed is 4.57 Å.

B. CYCLE ASSEMBLING

The complete algorithm was tested on hairpin 2555–2580 of the large ribosomal subunit. 200 generations of the algorithm were completed in 1 hour 45 minutes of real time on a two CPUs linux workstation (PIII-600). The average processor usage is 1.3 showing that the absence of load-balancing between builder threads greatly hurts the gain that are made by parallelism. For the first 100 generations, the distribution of substructure sizes are presented in figure 5·7.

From this, 6 structures were obtained with an average RMSD to the X-ray of 2.94 Å, the best one has a RMSD of 1.78 Å. Resulting structures are shown on figure 5·8, superposed on panel (b) with the best model shown separately on

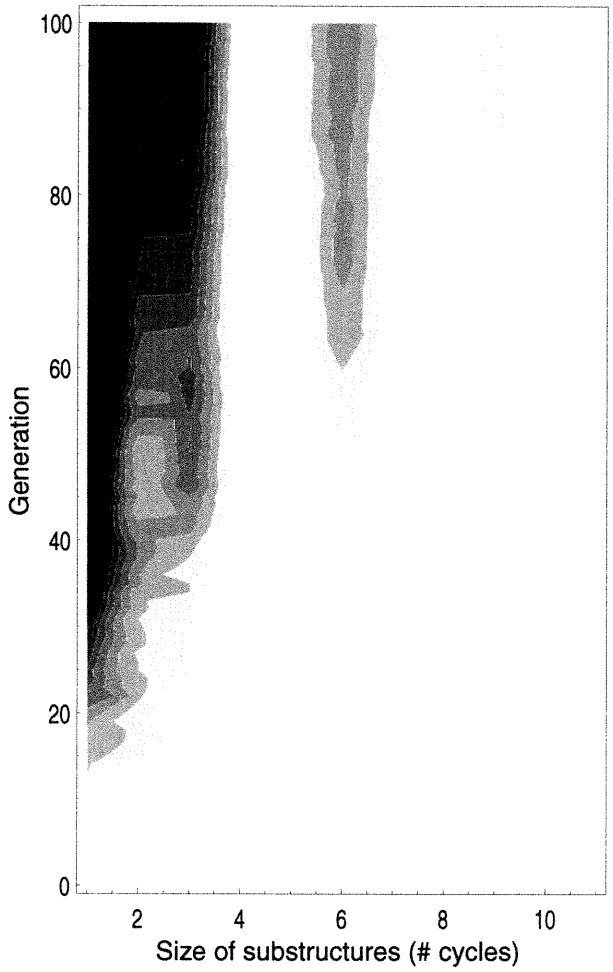


Figure 5·7 *Distribution of the substructure sizes at each generation. The size of a substructure is given by the number of cycles assigned. Important accumulations of structures of size 1, 3, 6 and 9 occur during the time of this run. At generation 100, the pool is mainly composed of substructures of size 1, 2 and 3.*

panel (d) beside the X-ray structure extracted from 1FFK. In the best structure obtained, all structural features of the X-ray (direction of the bases in the loop, base triples, relative orientation of the base triples) are perfectly reproduced.

5·5 DISCUSSION

The use of a Monte Carlo probabilist algorithm has several advantages when compared to the backtrack algorithm used in *MC-SYM*. The first one is that the search can be launched in parallel on several machines since only final results need to be exchanged. By implementing clever load balancing algorithm to equally share the cycle building effort across the builders, we should obtain an acceleration that is linear in the number of CPU used. The approach also guarantees that all relations used to model the structure are extracted from the database of known 3-D structures and not only relations belonging to a spanning tree of the GOR as in MC-SYM. Finally, the most interesting advantage of this method is the complete automation of the process.

As compared to *MC-SYM*, the results obtained are easier to justify structurally since every relation is extracted from the database of known structures and they do not depend on the arbitrary choice of a spanning tree on the GOR. No decision is required from the user, except the amount of resources to allow for a given problem. On the other hand, many disadvantages are expected: first, there is the loss of reproducibility of the modeling experiment due to the use of a heuristic search method; second, is the difficulty to determine the required amount of resources to allocate for a given problem; third, is the daunting task of identifying optimal values for the various cutoffs and parameters that would work well with a variety of structures.

The cutoff to accept a cycle assignment before smoothing was arbitrarily set to 1.25. This does not reflect the fact that we should expect longer cycles to

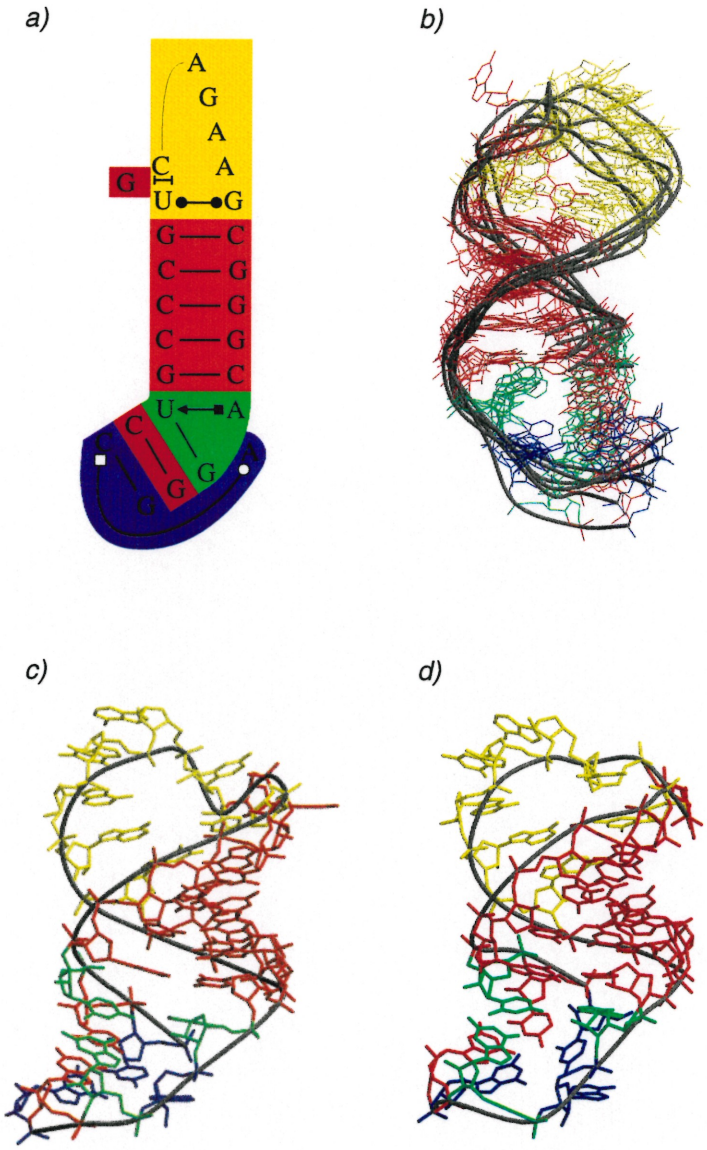


Figure 5·8 *Models built for hairpin 2555–2580 of the large ribosomal subunit of* H. marismortui *(PDB code: 1FFK, [5]). a) The graph of relations. b) Superposition of the six models. Each structure is superposed to minimize the RMSD to the first structure. c) The X-ray structure. d) The best structure (RMSD: 1.78 Å). The color code is maintained accross the four panels to emphasize the structural elements of the molecule.*

accumulate larger errors and that this error would be easier to redistribute across these cycles. This cutoff should depend at least on the cycle length and possibly on the nature of the edges involved in the cycle, but it is still unclear how this dependence should behave.

Another problem is to decide when a substructure should be removed from the pool. The current policy is to remove a substructure only when it is successfully agglomerated. This utterly conservative approach results in the accumulation of "dead-end" substructures, on which it is impossible to build further. These substructures slowly take over the complete population of the pool and they significantly slow down the process of identifying possible pairs of substructures for agglomeration. During the modeling of small RNA structures like the 2555–2580 motif we used, this accumulation did not present a real problem since satisfactory solutions are obtained before the pool gets clogged. Also, the bottleneck of the current design is the production of cycle structures for the most "complex" cycle, corresponding to the builder with the lowest throughput. The modeler is thus spending most of its time waiting for this builder to produce its structure. We tried to remove the necessity by the modeler of waiting to obtain a structure from each builder but it results in a largely inhomogenous pool that gets completely filled with substructures for the simplest cycles. A dynamic load balancing algorithm should take care of assigning the task of generating cycles to builders, dispatching one builder per CPU instead of one builder per cycle.

The current implementation uses shared memory for exchanging information between the modeler and builder threads. Replacing this by a message passing approach and transforming the threads in processes would permit distribution of the processes across a cluster of workstations, providing affordable parallelism.

Larges cycles are inherently difficult to model, they indicate a lack of information on this region of the structure or a very flexible region of the

molecule. But, in some particular cases, a large cycle is the result of a long range tertiary interaction (as in a pseudoknot) and will result in the tight binding of other structural elements. In this latter situation, the assignment of HTMs to the edges of the large cycle is strongly constrained by the resulting assignment to all other neighboring cycles. Exploiting this information would be essential to efficently model this type of structures. A possible approach would be that each builder shares with adjacent cycles informations on which HTMs should be prefered for homologous edges, depending on its own solutions. Modeling of cycles would not be independent anymore and this would preclude reusing the substructures obtained from modeling similar cycles.

An interesting variation on the collision constraint consists in using Van der Waals and electrostatic energies from a molecular mechanics forcefield instead of the arbitrary penalty used. The Van der Waals force is a repulsion term and could be directly substituted to the collision penalty. The advantage of this approach is that a finer evaluation of the affinity between the two rigid objects can be obtained at a minimal cost. This approach was implemented, using the Amber 4.1 forcefield [73], in a similar context with excellent results (unpublished results). The introduction of the electrosatic term biased the search toward more compact and thus stable 3-D structures. The use of an energetic evaluation function during the cycle optimization should produce more plausible 3-D structures that would eliminate most of the artifacts produced by the *MC-SYM* algorithm.

The target application for this algorithm is the high throughput modeling of small structures (typically internal loops or stem-loops). This type of modeling is of great interest for the rational design of RNA targeted drugs. The algorithm proposed can be used to build a library of possible conformations for the target that will be used for drug screening using standard small molecule docking techniques [26,41,67].

CHAPITRE 6

CONCLUSION

L'objectif principal étant la réalisation d'un système automatisé de modélisation des ARNs, les résultats présentés au chapitre 5 démontrent que le système mis au point atteint cet objectif pour des ARNs de taille relativement faible (environ 20-30 nucléotides). L'impossibilité d'appliquer l'approche à des molécules de taille plus importante est essentiellement d'ordre pratique puisque le système délègue la modélisation de chaque cycle à un "thread", la modélisation d'une structure de la taille d'un ARN de transfert engendrerait une trentaine de processus. De plus, la méthode utilisée pour assembler les cycles en une structure complète n'est fiable que si le nombre de cycle à joindre est faible, permettant l'obtention de structures satisfaisantes avant que le bassin de sous-structures soit envahie de structures "cul-de-sac". Les diverses étapes entourant la modélisation comme la collecte d'information dans les structures connues ou l'analyse des structures générées ont aussi gagné en automatisme par l'introduction des outils développés aux chapitres 2 et 3.

De manière spécifique, les travaux présentés dans cette thèse comme des étapes intermédiaires à l'obtention d'un système de modélisation automatisé ont aussi permis le développement de plusieurs outils d'analyse ayant des impacts au-delà du domaine de la modélisation. Entre autre, le chapitre 2 présente l'élaboration d'une méthode d'identification des appariements dans les ARNs en utilisant la nomenclature de Leontis et Westhof [56]. L'algorithme développé a été utilisé pour construire la base de donnée de relations (dont celle utilisée dans la version courante de *MC-SYM*), pour l'annotation de structure 3-D, pour la

visualisation de pont-H. Le système d'annotation présenté au chapitre 3 permet de mettre en relation une structure à analyser avec le reste des structures connues. Lors de l'analyse d'une nouvelle structure, cet outil permet de mettre l'emphase sur les régions de la molécule qui présentent des conformations inhabituelles et donc possiblement responsables de fonctions particulières. La décomposition d'un graphe de relations en une base minimale de l'espace des cycles a été développée dans le but de mettre au point un système automatisé de modélisation. Par contre, au cours de son développement, cette approche s'est révelée pouvoir jouer un rôle important dans la découverte de motifs, les résultats présentés au chapitre 4 en témoignent.

Plusieurs résultats présentés au cours de cette thèse sont des avancées importantes dans le domaine de la biochimie structurales des ARNs. Par example, la mise au point d'un système permettant l'identification automatique des appariements a permis l'obtention d'un repertoire complet des types d'appariements observés dans les structures determinées par crystallographie par rayons X. Un tel repertoire est une ressource convoitée depuis plus de 40 ans et certains groupes ont entrepris à plusieurs reprises sa réalisation (voir [69]) par visualisation interactive de toutes les structures connues, un travail colossal mais voué a un inévitable échec.

Le chapitre 4 présente aussi l'identification d'un nouveau motif, la pseudo-boucle GN:RA. L'identification de ce motif par visualisation interactive aurait demandé un travail long et minutieux sans aucune garantie d'exhaustivité. À la lumière des intéractions entre chacune des occurences de ce motif dans la grande sous-unité du ribosome avec le sillon mineur d'une double-hélice, il semble évident que cette conformation est sélectionnée pour sa capacité à former ce type d'intéraction. À notre connaissance, l'existence de ce motif n'a toujours pas été rapportée dans la littérature scientifique. La grande conservation des séquences formant ce motif permet aussi de l'ajouter au repertoire du

modélisateur, permettant d'inférer une conformation à partir d'une séquence.

6·1 MAUX ET REMÈDES ...

Les approches présentées dans ce travail souffrent toutes d'imperfections que nous rappelerons ici, en énumérant quelques remèdes possibles.

A. PONT-H GÉNÉRALISÉS

Lors de l'identification d'appariements entre bases azotées (cf. chapitre 2), l'impossibilité de détecter des intéractions impliquant le ribose ou le groupement phosphate limite l'utilisation généralisée de cette approche. Cette contrainte est essentiellement dûe au fait que la méthode requiert la présence explicite des atomes d'hydrogène et de pseudo-atomes donnant la direction des paires d'électrons libres. Puisque, sur les ribose, certains de ces atomes sont mobiles (par exemple le HO2' chez les ARNs) et sont en général absent des structures à analyser, il faudrait inclure le positionnement de ces atomes dans l'optimisation du nombre de pont-H formés. Inclure les torsions responsables du positionnement de ces atomes et pseudo-atomes lors de l'optimisation sort définitivement du contexte d'un problème de flot maximum et le problème d'optimisation résultant risque d'être complexe et coûteux à résoudre. Par contre, la résolution de ce problème de détermination des hydrogènes libres offrirait la possibilité d'inclure pratiquement tous les types de molécule dans l'analyse d'appariements et particulièrement l'identification de contacts ARN-protéine impliquant la formation de pont-H entre les deux molécules.

B. Expansion des motifs

Le concept de motif est lié à celui d'évolution par sélection naturelle. On suppose que si une caractéristique structurelle est sélectionnée positivement, on devrait pouvoir l'observer d'une manière inhabituellement fréquente. La méthode permettant l'identification de motifs présentée au chapitre 4 souffre du fait qu'elle ne considère que la fréquence d'observation d'un motif et qu'elle ne tente d'aucune façon de voir si cette redondance est naturelle ou non. À ce titre, la méthode détecte comme motif très significatif les tandems de quatre bases correspondant à l'empilement de deux appariements Watson-Crick.

Le domaine de l'identification de motif est très étudié au niveau des séquences d'ADN et de protéine. L'approche la plus courante consiste à calculer le ratio du nombre attendu d'occurences du motif sur le nombre observé et de faire l'hypothèse que l'intensité de la sélection est proportionnelle à ce ratio. Le calcul d'un nombre attendu d'occurences implique la mise au point d'un modèle probabiliste permettant de prédire le nombre d'instances possibles pour un motif, ce qui n'a toujours pas été développé dans le domaine des structures d'ARNs.

C. Répartition de la tâche de modélisation des cycles

L'architecture utilisée par le système de modélisation présenté au chapitre 5 implique une exécution parallèle des processus effectuant la modélisation des cycles. Chacun de ces processus ayant une tâche d'une difficulté variable (et impossible à déterminer *a priori*) leur taux de production de nouvelles structures est inégal. Par contre, l'approche stochastique pour l'assemblage des cycles ne fonctionne bien que si le bassin de sous-structures à sa disposition est suffisamment homogène. Cette nécessité est à la base de la contrainte d'obtenir à chaque génération une seule structure pour chaque cycle. Puisqu'il existe toujours un cycle plus complexe que les autres, l'utilisation des processeurs disponibles est

affecté par le temps nécessaire pour que ce processus obtienne sa structure.

Ce problème peut-être résolu en modifiant l'architecture du système de telle sorte qu'il y ait un processus par processeur s'occupant de la construction des cycles et que la tâche de construire chacune des structures soit considérée comme une unité de travail qui peut-être transferée d'une processeur à l'autre ou dupliquée sur plusieurs processeurs. L'optimisation d'un cycle peut être vue comme un processus de Poisson, ce qui permet de dire que si on consacre deux processeurs pour un cycle, le temps moyen pour obtenir une structure optimisé sera divisé par deux [88]. Cette dernière remarque suggère une politique séquentielle pour la distribution des tâches dans laquelle tous les processeurs tentent successivement d'optimiser chacun des cycles. Dès qu'un processeur obtient une structures, tous passent au cycle suivant.

6·2 DÉVELOPPEMENT FUTURS

A. PARALLÉLISATION DISTRIBUÉE

D'un point de vue pratique une amélioration simple au système de modélisation serait de transformer les processus légers (*threads*) construisant les cycles par des processus indépendants. Cette modification permetterait de distribuer automatiquement ces derniers à travers une grappe de stations de calcul linux en utilisant le logiciel MOSIX.

B. RÉUTILISATION DES CYCLES

Comme il a été démontré au chapitre 4, les cycles d'un ARN sont, en majorité, de courte taille. Puisque le nombre de types de relations et le nombre de types de bases est limité (quatre types de bases, quatre à cinq types de relations de manière générale), on s'attend à ce que dans une molécule de taille raisonnable plusieurs

cycles aient la même description. Une telle situation permettrait de n'utiliser qu'un "builder" pour ces cycles similaire et simplement dupliquer les structures produites dans le bassin de sous-structures.

Une extension possible de ce concept est la mise au point d'une cache permettant de conserver les résultats de modélisation des cycles. Lors de la modélisation d'une nouvelle molécule, si un cycle a la même description qu'un cycle précédemment modélisé, il serait possible de réutiliser les solutions trouvées. On pourrait facilement imaginer une cache centrale partagée par plusieurs groupes de modélisation pour diminuer le temps de calcul. L'espace nécessaire pour sauvegarder une solution correspond à 2 MTHs (*HTM*) et 5 paramètres de riboses (environ 100 octets) par arête du cycle.

C. UNE STRUCTURE, PLUSIEURS SÉQUENCES

Un extension proposée depuis longtemps pour *MC-SYM* est la possibilité de modéliser plusieurs séquences en même temps. En effet, dans plusieurs projets de modélisation la séquence de la molécule est connue dans différents organismes. Puisque la molécule a la même fonction dans chacun de ces organismes, on fait l'hypothèse que la structure est conservée malgré la variation de séquence. Dans ce contexte, il serait intéressant de contraindre la modélisation à trouver des structures qui peuvent se réaliser dans un maximum des séquences observées (idéalement toutes!). Une telle approche a été utilisée au cours de la modélisation du ribozyme activé par le plomb [49] (réalisé au cours de ma maîtrise), mais le formalisme proposé devait s'appliquer *a posteriori* d'une expérience de modélisation indépendante sur chacune des structures. Avec le système proposé il serait possible de se servir de la décomposition pour détecter les régions constantes dans certaines séquences et ne lancer qu'un "builder". Ceci est obtenu en implantant la réutilisation des cycles. Le processus d'assemblage des cycles aurait la responsabilité lors de l'agglomération de deux sous-structures

de s'assurer qu'une structure équivalente peut être construite pour chacune des séquences. Cette opération pourrait s'appliquer par un simple critère de RMSD sur les bases en utilisant les pseudo-atomes décris au chapitre 4.

6·3 LE MOT DE LA FIN...

D'une manière plus générale, la représentation obtenue en utilisant la base minimale de l'espace des cycles pour décomposer un graphe de relation est propice à un renouveau important de l'analyse et de la modélisation des structures d'ARN. Les travaux présentés dans cette thèse ne font, à mon avis, qu'effleurer la surface des possibilités de cette approche. L'arrivée pratiquement simultanée de ce nouveau point de vue et des structures du ribosome ne laisse, pour l'instant, qu'entrevoir les fascinantes révélations que l'ARN s'apprête à livrer.

CHAPITRE 7

GLOSSARY

5S subunit: Smallest subunit of the ribosome (in the structure 1FFK, it is represented as chain '9'). See **Ribosome**.

Adjacent/Non-adjacent: Term used in the annotation of RNA 3-D structure to define the property of two nucleotide of being covalently bonded or not. The term refers to the adjacency of the two nucleotides in the primary structure, the sequence.

Backbone: Part of the biopolymer that links the subunit to one another. In the RNA, the backbone is composed of an alternance of phosphate group and ribose sugars, the variable part of the nucleotides is linked to the C1' atom of the ribose.

Cis/Trans: Describe the relative orientation of both glycosidic bond with respect to the axis of the base pair.

Covalent bond: A permanent chemical bond between two atoms. The set of covalent bond forms the chemical graph of a molecule.

GNRA Tetraloop: A GNRA tetraloop is one of the best known motif in RNA 3-D structures. It is formed by the stabilization of a short helical stem closed by a four nucleotide loop corresponding to the sequence [G][ACGU][AG][A]. The term now commonly refer to the well characterized 3-D structure that is adopted by such RNA in which position 1 and 4 of the loop (the G and A) form a sheared G.A base pair and the other two nitrogen bases stack on top of the paired A.

H. marismortui: (ref. to 1FFK)

Large Ribosomal Subunit (LRS): The ribosome is divided into two main subunit, refered to as the small and large ribosomal subunit. In bacteria, the large ribosomal subunit is also called the 50S subunit. It is further divided into the 23S and 5S subunit. A crystal structure of the large ribosomal is available in the PDB database with identifier 1FFK. The large ribosomal subunit is the largest RNA structure to be determined by X-ray crystallography.

Major groove: (cf. Minor groove)

Minor groove: The formation of a RNA double helix results in a screw-like structure. The Cis orientation of the glycosidic bonds in the canonical base pairs forming the helix create two distincts grooves each sides of the helix. The minor groove is the shallower of the two, exposing the nitrogen bases Hoogstein edges to the outside of the helix. This is often the surface of the RNA helix that is used as an anchor for other molecules as it provides specificity (access to the bases). The major groove is found on the opposite side of the helix and is composed of the sugar edge of the bases and the riboses on each side.

Non-canonical base pair: Any base pair type that is not the standard Watson-Crick G.C or A.U. Depending on the authors, the wobble G.U base pair type is sometimes included in the set of canonical base pair types.

Paired/Unpaired: These terms are used to describe if two nucleotides are stabilized by the formation of one or more hydrogen bond between their respective nitrogen bases.

Pseudorotation angle: The complete description of the ribose pucker mode require the specification of 5 torsion angles (angles and distances between atoms are fixed). This parametrization is highly redundant and can be reduced to only two free parameters. The most important one is the pseudorotation angle (obtained

by sum of the 5 torsion angles of the ribose), the second one is the pucker amplitude (which is highly constrained in the ribose moiety of RNA.

Ribose: Cyclic carbohydrate composed of 5 carbons. Ribose and deoxyribose form the basis of the RNA and DNA backbone.

Ribose moiety: Invariant part of a RNA nucleotide composed of a ribose sugar. In base pairs diagrams, the ribose moiety is often replaced by a "R".

Ribose pucker mode: The ribose is a five-membered ring sugar. In solution, it adopts two major conformation. The first one is an enveloppe-like conformation in which for the of 5 carbons are in the same plan and the fifth on is outside the plane. The twist conformation is obtained when three adjacent carbons are aligned in the same plane and the two remaining carbons lie each side of this plane. The pucker describe the conformation in which the ribose is.

RMSD: Root Mean Squared Deviation. This is a common measure of dissimilarity between to 3-D structure of molecules. Implicitly, the reported measure is the one obtained after applying a rigid-body transformation to one of the two molecules in order to obtain the minimal RMSD. The process of transforming one of two molecule to minimize the RMSD is refered to as **superposition**.

Sheared base pair type: One of the most frequently observed non-canonical base pair type. It is obtained by hydrogen bonding of the Watson-Crick edge of an adenosine to the Hoogstein edge of a guanosine. It is frequently observed as a tandem G.A/A.G, often refered as the GA tandem mismatch.

Stacked/Unstacked/Helically stacked: Stacking is an important force stabilizing the 3-D structure of RNA. These three terms are used to characterize the relation between two nucleotides with respect to the stacking. Helically stacked refers to the type of stacking that is observed in RNA double-helices of type A,

corresponding to the most stable form of RNA structure.

Sugar moiety: See **Ribose moiety**.

BIBLIOGRAPHIE

[1] R. K. Ahuja, M. Kodialam, A. K. Mishra, and J. B. Orlin. Computational investigations of maximum flow algorithms. *European Journal of Operational Research*, 97:509–542, 1997.

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, NJ, 1993.

[3] C. Altona and M. Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. a new description using the concept of pseudorotation. *Journal of American Chemical Society*, 94:8205–8212, 1972.

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

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

[6] M. S. Bazaraa and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, NY, 1979.

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

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig,

I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[9] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, NY, 1995.

[10] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[11] M. E. Burkard, D. H. Turner, and I. Tinoco Jr. The interactions that shape RNA. In R. F. Gestland, J. F. Atkins, and T. R. Cech, editors, *The RNA World*, pages 233–264. Cold Spring Harbor Press, 1999.

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

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

[14] D. M. Chickering, D. Geiger, and D. Heckerman. On finding a cycle basis with a shortest maximal cycle. *Information Processing Letters*, 54(1):55–58, 1995.

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

[16] W. D. Cornell, P. Cieplak, C. I. Bayley, I. R. Gould, K. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. Kollman. A

second generation force field for the simulation of proteins and nucleic acids. *Journal of the American Chemical Society*, 117:5179–5197, 1995.

[17] C. C. Correll, B. Freeborn, P. B. Moore, and T. A. Steitz. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, 91:705–712, 1997.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[19] 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, and H.-C. Yoon. Polymorphism, packing, resolution, and reliability in single-crystal DNA oligomer analyses. *Nucleosides, Nucleotides & Nucleic Acids*, 10:3–24, 1991.

[20] J. Donohue. Hydrogen-bonded helical configurations of polynucleotides. *Proceedings of the National Academy of Sciences*, 42:60–65, 1956.

[21] J. Donohue and K. N. Trueblood. Base pairing in DNA. *Journal of Molecular Biology*, 2:363–371, 1960.

[22] C. M. Duarte and A. M. Pyle. Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology*, 284(5):1465–1478, 1998.

[23] N. Foloppe and A. D. Mackerell Jr. 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(2):86–104, 2000.

[24] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

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

[26] E. J. Gardiner, P. Willett, and P. J. Artymiuk. Protein docking using a genetic algorithm. *Proteins*, 44:44–56, 2001.

[27] D. Gautheret and R. R. Gutell. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Research*, 25(8):1559–1564, 1997.

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

[29] P. Gendron, S. Lemieux, and F. Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology*, 308:919–936, 2001.

[30] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum flow problem. *Journal of the ACM*, 35:921–940, 1988.

[31] B. L. Golden, A. R. Gooding, E. R. Podell, and T. R. Cech. A preorganized active site in the crystal structure of the tetrahymena ribozyme. *Science*, 282(5387):259–264, 1998.

[32] R. R. Gutell, J. J. Cannone, D. Konings, and D. Gautheret. Predicting U-turns in ribosomal RNA with comparative sequence analysis. *Journal of Molecular Biology*, 300(4):791–803, 2000.

[33] 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, and J. J. Cannone. Comparative sequence analysis and the prediction of RNA structure, and the web. Manuscript in preparation.

[34] S. Hanlon. The importance of london dispersion forces in the maintenance of deoxyribonucleic acid double helix. *Biochemical and Biophysical Research Communications*, 23:861–867, 1966.

[35] F. Harary. *Graph Theory*. Addison-Wesley, 1969.

[36] S. R. Holbrook, J. L. Sussman, R. W. Warrant, and S.-H. Kim. Crystal structure of yeast phenylalanine transfer RNA, II: Structural features and functional implications. *Journal of Molecular Biology*, 123(4):631–660, 1978.

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

[38] J. D. Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing*, 16(2):358–366, 1987.

[39] IUPAC-IUB Joint Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of conformations of polynucleotide chains. *European Journal of Biochemistry*, 131:9–15, 1983.

[40] F. Jiang, R. A. Kumar, R. A. Jones, and D. J. Patel. Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature*, 382(6587):183–186, 1996.

[41] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267:727–748, 1997.

[42] F. M. Jucker, H. A. Heus, P. F. Yip, E. H. M. Moors, and A. Pardi. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *Journal of Molecular Biology*, 264:968–980, 1996.

[43] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923, 1976.

[44] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34:827–828, 1978.

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

[46] R. Lavery and H. Sklenar. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *Journal of Biomolecular Structure and Dynamics*, 6(1):63–91, 1988.

[47] F. Leclerc, J. Srinivasan, and R. Cedergren. Predicting RNA structures: the model of the RNA element binding rev meets the NMR structure. *Folding & Design*, 2(2):141–147, 1997.

[48] P. Legault, J. Li, J. Mogridge, L. E. Kay, and 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(2):289–299, 1998.

[49] S. Lemieux, P. Chartrand, R. Cedergren, and F. Major. Modeling active RNA structures using the intersection of conformational space: Application to the lead-activated ribozyme. *RNA*, 4:739–749, 1998.

[50] S. Lemieux and F. Major. On the minimal cycle basis of nucleic acids graph of relation and its use for 3-D motifs detection. *Bulletin of Mathematical Biology*, 2001. to be submitted.

[51] S. Lemieux and F. Major. Recognition of base pairing types in RNA three-dimensional structure. *Journal of Molecular Biology*, 2001. To be submitted.

[52] S. Lemieux, S. Oldziej, and F. Major. Nucleic acids: Qualitative modeling. In N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, and

P. R. Schreiner, editors, *Encyclopedia of Computational Chemistry*, West Sussex, England, 1998. John Wiley & Sons.

[53] N. B. Leontis and E. Westhof. Conserved geometrical base-pairing patterns in RNA. *Quaterly Reviews of Biophysics*, 31(4):399–455, 1998.

[54] N. B. Leontis and E. Westhof. Conserved geometrical base-pairing patterns in RNA. *Quaterly Reviews of Biophysics*, 31(4):399–455, 1998.

[55] N. B. Leontis and E. Westhof. Recurrent rna motifs: Analysis at the basepair level. In *RNA Biochemistry and Biotechnology*, pages 45–61. Kluwer Academic, 1999.

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

[57] M. Levitt. Detailed molecular model for transfer ribonucleic acid. *Nature*, 224:759–763, 1969.

[58] J. Leydold and P. F. Stadler. Minimal cycle bases of outerplanar graphs. *Electronic Journal of Combinatorics*, 5:R16, 1998.

[59] K. Lindauer, C. Bendic, and J. Suhnel. HBexplore — a new tool for identifying and analysing hydrogen bonding patterns in biological macromolecules. *Computer Applications in the Biosciences*, 12(4):281–9, 1996.

[60] T. Macke and D. A. Case. Modeling unusual nucleic acid structures. In N. B. Leontis and J. SantaLucia, Jr., editors, *Molecular Modeling of Nucleic Acids*, pages 379–393, Washington, DC, 1998. American Chemical Society.

[61] A. D. Mackerell Jr. and N. K. Banavali. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

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

[63] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 253(5025):1255–60, 1991.

[64] C. Massire and E. Westhof. MANIP: An interactive tool for modelling RNA. *Journal of Molecular Graphics and Modelling*, 16(4–6):197–205, 1998.

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

[66] J. A. McCammon and S. C. Harvey. *Dynamics of proteins and nucleic acids*. Cambridge University Press, Cambridge, Great Britain, 1987.

[67] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998.

[68] F. Mueller and R. Brimacombe. A new model for the three-dimensional folding of *escherichia coli* 16S ribosomal RNA. I. fitting the RNA to a 3D electron microscopic map at 20 Å. *Journal of Molecular Biology*, 271(4):524–544, 1997.

[69] U. Nagaswamy, N. Voss, Z. Zhang, and G. E. Fox. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Research*, 28(1):375–376, 2000.

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

[71] W. K. Olson. Computational studies of polynucleotide flexibility. *Nucleic Acids Research*, 10:777–787, 1982.

[72] R. P. Paul. *Robot manipulators: mathematics, programming, and control.* MIT Press, Cambridge, MA, 1981.

[73] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. R. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and 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.

[74] D. A. Pearlman and 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.

[75] R. Pinard, K. J. Hampel, J. E. Heckman, D. Lambert, P. A. Chan, F. Major, and J. M. Burke. Functional involvement of g8 in the hairpin ribozyme cleavage mechanism. *EMBO Journal*, 20(22):6434–6442, 2001.

[76] R. Pinard, D. Lambert, N. G. Walter, J. E. Heckman, F. Major, and J. M. Burke. Structural basis for the guanosine requirement of the hairpin ribozyme. *Biochemistry*, 38(49):16035–16039, 1999.

[77] H. W. Pley, K. M. Flaherty, and D. B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372(6501):68–74, 1994.

[78] R. B. Ravelli and S. M. McSweeney. The 'fingerprint' that x-rays can leave on structures. *Structure*, 8:315–328, 2000.

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

[80] A. Sarai, J. Mazur, R. Nussinov, and R. L. Jernigan. Origin of DNA helical structure and its sequence dependence. *Biochemistry*, 27(22):8498–8502, 1988.

[81] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 å resolution. *Cell*, 102:615–623, 2000.

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

[83] J. Sponer and 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(2):277–292, 1993.

[84] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and S. H. Kim. Crystal structure of yeast phenylalanine tRNA. I. Crystallographic refinement. *Journal of Molecular Biology*, 123:607–630, 1978.

[85] R. K. Tan and S. C. Harvey. YAMMP: Development of a molecular mechanics program using the modular programming method. *Journal of Computational Chemistry*, 14:455–470, 1993.

[86] I. Tazawa, T. Koike, and 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(1):33–38, 1980.

[87] I. Tinoco Jr. Structure of base pairs involving at least two hydrogen bonds. In R. F. Gestland, J. F. Atkins, and T. R. Cech, editors, *The RNA World*, pages 603–607. Cold Spring Harbor Press, 1993.

[88] K. Trivedi. *Probability and statistics with reliability, queuing, and computer science applications*. Prentice-Hall, Englewoods Cliffs, NJ, 1982.

[89] M. Turcotte. *Génération et traitement de contraintes relationnelles pour la modélisation des acides nucléiques*. PhD thesis, Université de Montréal, 1995.

[90] P. Vismara. Union of all the minimum cycle bases of a graph. *Electronic Journal of Combinatorics*, 4:R9, 1997.

[91] A. Wada. Bioinformatics — the necessity of the quest for 'first principles' in life. *Bioinformatics*, 16(8):663–664, 2000. (editorial).

[92] J. J. Warren and P. B. Moore. Application of dipolar coupling data to the refinement of the solution structure of the sarcin-ricin loop rna. *Journal of Biomolecular NMR*, 20(4):311–323, 2001.

[93] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A structure for deoxyribonucleic acid. *Nature*, 171:694–967, 1953.

[94] J. E. Wedekind and D. B. McKay. Crystal structure of a lead-dependant ribozyme revealing metal binding sites revelant to catalysis. *Nature structural biology*, 6(3):261–268, 1999.

[95] A. R. Weeks Jr. *Fundamentals of Electronic Image Processing*. Spie/IEEE, 1998.

[96] H. Weissig and P. E. Bourne. An analysis of the protein data bank in search of temporal and global trends. *Bioinformatics*, 15(10):807–831, 1999.

[97] B. T. Wimberly, D. E. Brodersen, W. M. Clemons Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407:327–339, 2000.

[98] Y. Xing and D. E. Draper. Cooperative interactions of RNA and thiostrepton antibiotic with two domains of ribosomal protein L11. *Biochemistry*, 35(5):1581–1588, 1996.

[99] Y. Xing, D. GuhaThakurta, and D. E. Draper. The RNA binding domain of ribosomal protein L11 is structurally similar to homeodomains. *Nature Structural Biology*, 4(1):24–27, 1997.

[100] F. Zhang, S. Lemieux, X. Wu, D. St-Arnaud, C. T. McMurray, F. Major, and D. Anderson. Function of hexameric RNA in packaging of bacteriophage ϕ29 DNA *in vitro*. *Molecular Cell*, 2(1):141–7, 1998.