Université de Montréal

Modélisation de la structure 3-D des ARN par satisfaction de contraintes

par

Sébastien Lemieux

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

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

Décembre, 1998

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Modélisation de la structure 3-D des ARN
par satisfaction de contraintes

présenté par:

Sébastien Lemieux

a été évalué par un jury composé des personnes suivantes:

Nadia El-Mabrouk,	président-rapporteur
François Major,	directeur de recherche
Guy Lapalme,	membre du jury

Mémoire accepté le: ..10 mars 1999....

Sommaire

L'inférence d'un modèle tridimensionnel (3-D) d'une molécule d'ARN à partir d'informations biochimiques de faible résolution est une tâche complexe. Le développement de méthodes efficaces et objectives de modélisation est essentiel à la compréhension des mécanismes moléculaires impliquant des ARN. Le présent travail propose trois ajouts importants de ce domaine. Dans un premier temps, une définition de l'espace des conformations d'un ARN est établie et la technique de retour-arrière est décrite de façon à permettre une exploration complète de cet espace. Ensuite, un formalisme basé sur la logique floue est présenté. Cette approche permet d'utiliser l'information provenant de séquences multiples et est appliquée pour la modélisation du ribozyme activé par le plomb. Finalement, la flexibilité d'un système de modélisation 3-D utilisant une approche de satisfaction de contrainte est mise en évidence par l'ajout d'un nouveau type de contrainte dans l'engin de recherche, permettant la modélisation efficace de multimères cycliques.

Ces ajouts permettent d'élargir le spectre des informations utiles à la modélisation 3-D des ARN et facilite l'intégration de nouveaux types d'informations à l'intérieur d'une méthode systématique. Les résultats obtenus sur des projets de modélisation spécifiques (ribozyme activé par le plomb et pRNA de ϕ29) permettent de confirmer la pertinence de cette approche.

Mots clés : Modélisation moléculaire, ARN, Satisfaction de contraintes, Logique floue, Espace des conformations.

TABLE DES MATIÈRES

LISTE DES FIGURES

À David Morse,

pour sa démonstration involontaire de l'effet *butterfly* qui résulte aujourd'hui en ce mémoire...

CHAPITRE 1

Introduction

L'initiation en 1990 du projet du génôme humain, et le lancement de plusieurs projets similaires pour d'autres espèces, permit le développement rapide de nouvelles technologies de séquençage de l'ADN. Aujourd'hui plus de trois millions de séquences sont accessibles via diverses banques de données (voir [24]) et une grande proportion des travaux fait en bioinformatique vise l'analyse de cette information (voir [18] pour une présentation générale des algorithmes d'analyse de séquences). Par contre, le véritable intérêt réside dans la compréhension des mécanismes moléculaires et pour ce faire il est nécessaire de pouvoir caractériser la structure tridimensionnelle (3-D) des molécules étudiées.

La structure 3-D d'une molécule est généralement obtenue par l'utilisation de méthodes physiques telles la cristallographie à rayons X ou la résonance magnétique nucléaire (RMN). L'application de ces méthodes est longue et coûteuse mais résulte en des structures précises. Les conditions dans lesquelles ces expériences sont réalisées (formation d'un cristal, absence de certains ions, utilisation d'une séquence incomplète, etc.) peuvent induire un repliement alternatif de la molécule, faussant les résultats obtenus, la structure sera de bonne précision mais sera inexacte. À titre d'exemple, citons le cas de la boucle interne (L3) du ribozyme de l'hépatite δ pour laquelle deux groupes ont rapporté des structures complètement différentes malgré des conditions relativement similaires (voir [20, 28]).

Dans la suite de ce texte, le terme "modèle" sera utilisé pour désigner une structure obtenue par inférence à partir d'un ensemble d'informations de faible résolution (voir la section 1.3). Le terme "structure" sera réservé aux résultats d'une cristallographie à rayons X ou de la RMN (voir [1]).

1.1 Rôles des ARN

Plusieurs molécules sont responsables du bon fonctionnement de toute cellule vivante. La majorité sont des polymères linéaires formés d'une séquence finie de sous-unités. Il existe trois types de ces polymères, soit l'ADN (acides désoxyribonucléique, formée des sous-unités A, T, G et C), l'ARN (acides ribonucléiques, formée des sous-unités A, U, G et C) et les protéines (formée de 20 types d'acides aminés). Le rôle principal de l'ADN est de conserver et de propager l'information génétique, elle adopte de manière générale la structure simple d'une double-hélice (voir [2]). Les protéines sont les principaux acteurs de la vie cellulaire, elles sont responsables de la catalyse de la plupart des réactions chimiques permettant le maintien de l'équilibre de la cellule. Pour ce faire, elles doivent adopter une grande diversité de structures 3-D permettant le positionnement précis des groupements chimiques permettant la catalyse. Les ARN ont un rôle partagé, ils servent entre autres à transférer l'information génétique encodée dans l'ADN du noyau de la cellule vers le cytoplasme (ARN messagers) ou d'autres ARN spécialisés, les ARN de transfert et les ARN ribosomaux, de concert avec plusieurs protéines, permettront la traduction de la séquence d'ARN en une protéine.

La découverte de petits ARN permettant la catalyse rapide de certaines réactions chimiques (voir [6]) transforma la vision de l'ARN comme une forme inerte d'information permettant simplement de faire le transfert de l'information du noyau vers le cytoplasme. Les ARN ayant une fonction catalytique, les ribozymes, doivent, comme les protéines et contrairement à l'ADN, adopter une grande diversité de structures 3-D. Ceci permet d'assurer la spécificité des réactions catalysées et d'éviter que tout ARN (particulièrement les ARN messagers) aient des propriétés catalytiques non-désirables. Les ribozymes sont fréquement utilisés par les virus (VIH, hépatite, mosaïque du tabac, etc.), leurs permettant d'éviter d'avoir recours à la machinerie de traduction de la cellule hôte. L'étude des ARN catalytiques et plus particulièrement la modélisation de leur structure 3-D recèle donc un grand intérêt médical. Le chapitre 2 présente une introduction complète à la modélisation 3-D des structures d'ARN.

1.2 La structure des ARN

L'une des particularités des ARN est la formation, comme dans l'ADN, d'appariements très spécifiques entre les bases azotées de nucléotides éloignés dans la séquence. La figure 1 présente les deux appariements canoniques permettant la stabilisation des double-hélices d'ARN et d'ADN. On appelle structure primaire (1-D) d'un ARN sa séquence, la structure secondaire (2-D) est déterminée lorsqu'on ajoute l'information d'appariement sans obtenir la formation de pseudo-noeuds (voir [25]). La structure 3-D est obtenue lorsqu'il est possible d'assigner une coordonnée spatiale à chaque atome de la molécule. La figure 2 présente ces trois niveaux de structure pour une courte séquence.

L'obtention de la séquence est une tâche relativement facile, les derniers 20 ans ont vu l'apparition d'un grand nombre de méthodes efficaces permettant le séquençage des ARN. Par contre, la détermination de la structure 2-D pose déjà de sérieux problèmes. Plusieurs algorithmes ont été proposés qui permettent d'obtenir de bonnes prédictions de la structure 2-D étant donné une ou plusieurs séquences (voir [8, 17]). La méthode qui est principalement utilisée aujourd'hui est un algorithme de programmation dynamique permettant de déterminer la structure 2-D minimisant l'énergie potentielle des interactions prédites. Lorsque plusieurs séquences sont disponibles on peut aussi incorporer cette information sous la forme d'un alignement des séquences.

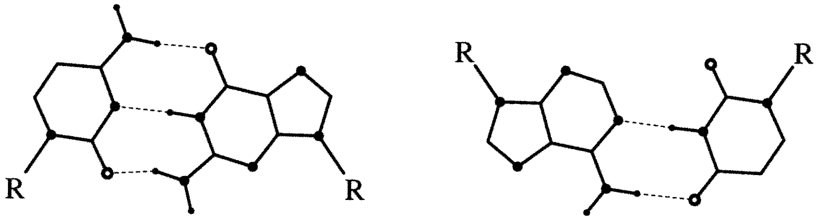


FIGURE 1. Appariements canoniques présents dans les doubles-hélices d'ADN et d'ARN. Ces appariements sont majoritairement responsables de la stabilisation de la structure secondaire. Les lignes en pointillées indiquent la présence de ponts hydrogènes.

À ce jour, aucune méthode ne permet d'inférer la structure tertiaire (3-D) d'un ARN sans avoir recours à de multiples expérimentations. La section 1.3 donne un aperçu des différents types d'information utilisés pour la modélisation des structures présentées aux chapitres 4 et 5.

L'espace des conformations d'un ARN est défini comme étant l'ensemble des modèles 3-D possibles pour cet ARN. Un algorithme de modélisation définit une fonction d'évaluation sur un modèle permettant de retourner une évaluation objective de la valeur du modèle. Le processus de modélisation consiste à explorer l'espace des conformations et à évaluer une suite de modèles, retenant ceux dont l'évaluation indique un niveau de qualité suffisant et se servant de ces évaluations pour orienter la suite de la recherche.

L'information expérimentale peut donc être utilisée à deux niveaux, soit lors de la définition de l'espace des conformations ou intégrée à la fonction d'évaluation. Le chapitre 2 met l'emphase sur les informations qui ont été intégrées à la définition de l'espace des conformations tel qu'implantée dans la version courante du logiciel MC-SYM, tandis que le chapitre 5 montre une application pour laquelle un nouveau type d'information (la contrainte de formation d'un

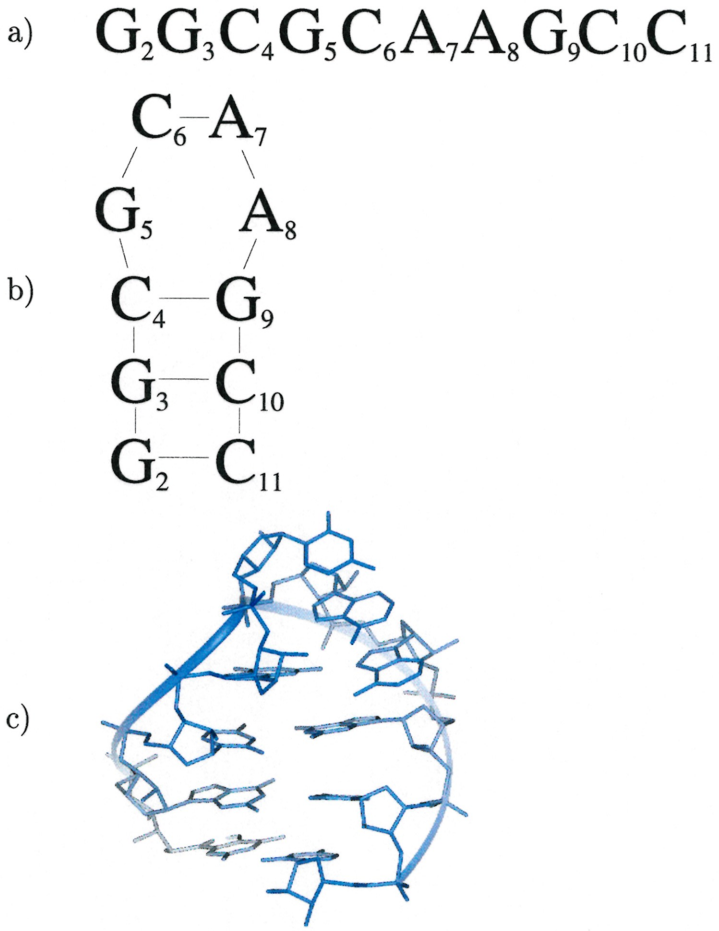


FIGURE 2. Niveaux de structures pour l'ARN. (a) Structure primaire, (b) structure secondaire et (c) structure tertiaire.

héxamère cyclique régulier) a été intégré à la fonction d'évaluation sous forme d'une contrainte.

1.3 Types d'informations structurelles

La détermination d'une structure 3-D pour un ARN exige l'obtention d'une grande quantité d'informations. Des méthodes comme la résonance magnétique nucléaire (RMN) ou la cristallographie à rayons X permettent d'obtenir une grande quantité d'informations très précises permettant, dans la plupart des cas, de converger rapidement vers une seule structure 3-D. Par contre, ces méthodes sont très coûteuses et parfois inapplicables dans certains contextes puisque les conditions expérimentales empêchent souvent le repliement correct de la structure. C'est pourquoi il est souvent nécessaire d'utiliser des méthodes fournissant moins d'informations ou des informations dont l'interprétation est floue et non-triviale. Il existe un grand nombre d'expériences biochimiques permettant d'obtenir de telles informations. Afin d'utiliser chacune de ces informations il est nécessaire de définir un formalisme permettant une interprétation non-ambiguë des résultats expérimentaux et l'incorporation de cette information dans un engin de recherche conformationnel.

1.3.1 Sélection *in vitro* (SELEX)

Basée sur le principe de la sélection naturelle, la sélection *in vitro* permet d'obtenir rapidement un ensemble de séquences différentes partageant la même fonction (voir [23]). On génère d'abord un ensemble de séquences aléatoires (par synthèse chimique) parmi lesquelles on sélectionne les molécules préservant la fonction étudiée, on amplifie par utilisation du PCR (*Polymerase chain reaction*) les molécules sélectionnées, puis on resélectionne. Après un certain nombre d'itérations, les molécules résultantes préservent toutes, à divers degrés, la fonction recherchée.

En acceptant l'hypothèse que deux séquences d'ARN relativement similaires et préservant la même fonction doivent préserver la même structure 3-D, on peut utiliser cet ensemble de séquences comme contrainte supplémentaire dans la modélisation. Le chapitre 3 présente un formalisme basé sur la logique floue (voir [4]) permettant d'incorporer ce type d'information dans le processus de modélisation. La méthodologie élaborée est ensuite mise en application au chapitre 4 pour permettre le premier tour de modélisation du ribozyme activé par le plomb.

1.3.2 Mutagénèse dirigée et incorporation de bases modifiées

En modifiant la séquence d'un ribozyme lors du séquençage (ou par l'utilisation de la technique de mutagénèse dirigée), il est possible de vérifier l'importance de chacun des nucléotides pour le maintien de la fonction et de la structure de la molécule. Il est aussi possible de vérifier certaines hypothèses sur la structure 2-D en utilisant des mutations compensatoires, i.e on change un appariement A-U pour un C-G, ou l'inverse.

Une variante de cette approche consiste à utiliser des bases azotées modifiées (par exemple, des déoxyriboses [11]) que l'on incorpore dans la séquence pendant la synthèse. De cette façon, il est possible de vérifier l'importance d'un groupement chimique spécifique (le O2' dans le cas de substitution par des déoxyriboses) dans le maintien de la structure 3-D ou pour l'activité catalytique de la structure. Ce type d'expérience fut utilisé de manière extensive dans le contexte du ribozyme activé par la plomb, les résultats sont rapportés dans [19] et sont à la base de la modélisation présentée au chapitre 4 et publiée dans [26].

1.3.3 Doubles mutants et mise en évidence de multimérisation

Une extension de la méthode de mutagénèse dirigée est présentée au chapitre 5. Elle permet la mise en évidence de la formation de multimères cycliques qui peuvent survenir pour certains pseudo-noeuds (*pseudoknots* en anglais) et dont la formation est essentielle à la fonction biologique de la molécule. L'utilisation de ce type d'information lors de la modélisation est particulièrement problématique parce qu'on ne peut, en général, se permettre de modéliser l'ensemble des unités du multimère. Le chapitre 5 présente l'introduction d'un nouveau type de contrainte dans l'engin de recherche MC-SYM permettant de vérifier efficacement qu'un modèle donné pour une unité permet la formation d'un cycle régulier.

1.4 Approches pour la modélisation 3-D

Il existe une grande variété d'approches pour la modélisation 3-D des ARN. Cette section se veut un rapide survol de ces différentes approches. [1] et [25] présentent une revue extensive de ces méthodes.

1.4.1 Modélisation intéractive

L'approche la plus utilisée pour la modélisation 3-D des ARN est la manipulation interactive des modèles à l'écran (voir [10] pour une application de cette technique à la modélisation 3-D d'un intron du groupe I). Dans cette méthode, l'utilisateur possède un contrôle complet sur les modèles explorés et est entièrement responsable de la qualité du modèle obtenu. On met à profit l'intuition du modélisateur, mais il est impossible de garantir que la recherche est complète et que toutes les contraintes sont respectées dans la structure finale (sauf par inspection visuelle).

Plusieurs logiciels sont présentement disponibles pour permettre ce type d'approche à la modélisation, par exemple: InsightII, VMD et ERNA-3D (voir [22]). Les contraintes et informations de nature experimentale n'ont pas à être représentées dans le système, il est donc possible d'utiliser toutes formes d'informations disponibles à l'usager. Ces programmes utilisent, à plusieurs reprises, des algorithmes de raffinement du modèle permettant d'en optimiser les propriétés stéréochimiques.

1.4.2 Méthodes énergétiques

L'une des premières approches systématiques consiste en la mise au point d'une fonction d'énergie, $f(\mathbf{x})$, où $\mathbf{x}$ représente les coordonnées 3-D d'un modèle. Cette fonction retourne une évaluation du niveau d'entropie d'un modèle donné. L'une des hypothèses de la thermodynamique est que la structure repliée (la structure native) d'un ARN doit être la structure d'énergie minimale parmi l'ensemble des conformations accessibles à la molécule. En sélectionnant une fonction d'énergie différentiable en tout point, il est possible d'utiliser diverses méthodes d'optimisation pour tenter d'identifier la conformation d'énergie minimale correspondant au minimum global de la fonction d'énergie. Les méthodes couramment utilisées sont la minimisation par descente de gradient, par gradients conjugués ou encore des approches comme le recuit simulé ou la dynamique moléculaire (voir [5] pour une présentation de ces méthodes et [9] pour leur application à la dynamique moléculaire).

L'utilisation de ces méthodes sur une représentation contenant tous les atomes d'une molécule donnée résulte en un nombre très élevé de minimums locaux. Les techniques de minimisation d'énergie vont invariablement converger vers le plus proche de ces minimums locaux et le modèle obtenu sera fortement dépendant de la structure initiale tout en étant très différent de la conformation d'énergie minimale recherchée. Afin d'éviter ce type de problème il est possible d'utiliser une représentation réduite des nucléotides en remplaçant plusieurs atomes par un atome fictif, un pseudo-atome (voir [16]). Par exemple, on remplace la base par un pseudo-atome et le groupement phosphate par un autre. Il est ensuite possible de reconstruire une approximation de la fonction d'énergie sur cette nouvelle représentation. Cette simplification réduit le nombre des minimums locaux (ceci correspond au lissage de la fonction d'énergie). Par contre, il est difficile d'obtenir une fonction d'énergie sur une représentation simplifiée qui demeure réaliste. Les modèles obtenus par de telles approches doivent en général être réoptimisés en utilisant leur représentation tout-atome.

1.4.3 Méthodes de recherche discrétisées

Une lacune importante des méthodes présentées jusqu'ici est le fait que la recherche de l'espace des conformations n'est pas complète. On obtient en général un modèle 3-D correspondant à l'ensemble des informations disponibles, mais il est difficile de garantir qu'il n'en existe pas un autre. De plus, les méthodes énergétiques ne garantissent pas que toutes les contraintes ont été respectées. Encore une fois, l'inspection visuelle du modèle est nécessaire pour obtenir une telle garantie. Par contre, si on discrétise l'espace des conformations accessibles au modèle, on peut permettre une exploration exhaustive de cet espace. Ceci permet de garantir que tous les modèles satisfaisant les contraintes soient identifiés.

L'utilisation de treillis pour discrétiser l'espace 3-D exige une représentation simplifiée du modèle (typiquement un pseudo-atome par nucléotide) et la définition d'une fonction d'énergie sur cette représentation. L'approche est particulièrement intéressante pour les études sur les propriétés du repliement des ARN (voir [12] pour une application à l'étude de la dynamique du repliement des protéines). Considérant l'importance des simplifications de la représentation manipulée, cette approche n'est pas utilisée pour la modélisation et rarement appliquée pour les ARN.

Plutôt que de discrétiser l'espace 3-D, il est possible de discrétiser l'espace des relations binaires entre les nucléotides et les conformations de chacun de ceux-ci. L'approche de modélisation utilisée dans ce travail se base sur ce type de discrétisation (voir [13–15]). Le chapitre 2 décrit en détail les concepts de base de cette approche ainsi que les développements réalisés au cours des deux dernières années.

1.5 Présentation des articles

Le chapitre 2 (publié dans le livre *Encyclopedia of Compuational Chemistry*, [27]) présente en détail l'approche de modélisation 3-D développée et utilisée au cours du travail couvert par le présent mémoire. L'emphase est mis sur la définition d'un espace des conformations discretisé permettant d'obtenir un problème de satisfaction de contraintes que l'on résoud de manière efficace par l'utilisation de l'algorithme de retour-arrière. On y introduit le concept de graphe de relations, permettant de représenter de manière flexible la majorité des informations disponibles sur une molécule. L'article présente ensuite deux courts exemples d'application de la méthode permettant d'introduire le lecteur au processus de modélisation 3-D tel que défini pour cette version du logiciel MC-SYM. J'ai directement participé à l'ensemble des travaux présentés dans ce chapitre, à l'exception des exemples de la section 5.

Les méthodes de sélection *in vitro* permettent l'obtention d'un grand nombre de séquences de variants actifs pour une structure donnée. Il est possible d'utiliser ce type d'information lors de la modélisation 3-D d'une structure en faisant l'hypothèse que la préservation de la fonction d'une molécule exige la préservation de la structure 3-D. Le chapitre 3 (publié dans le livre *Molecular Modeling of Nucleic Acids*, [21]) présente un formalisme basé sur la logique floue (voir [4]). L'utilisation de la logique floue a été proposée et appliquée au ribozyme activé par le plomb par Abdelmjid Ftouhi. Dans ce travail, j'ai été responsable des sections d'introduction ainsi que la section *Calculating Possibilities.*

Le chapitre 4 (publié dans la revue *RNA*, [26]) présente l'application de ce formalisme à la modélisation du ribozyme activé par le plomb. On y présente aussi l'utilisation de résultats d'expériences d'incorporation de bases azotées modifiées permettant l'obtention d'un modèle 3-D de bonne précision. J'ai directement participé à l'ensemble des travaux présentés dans cet article et en particulier aux calculs de structures 3-D. Pascal Chartrand fut responsable des expériences de modifications chimiques (figure 5 dans l'article).

Le chapitre 5 (publié dans la revue *Molecular Cell*, [29]) est le résultat d'une collaboration avec le groupe de recherche de D. Anderson à l'Université du Minnessota. Les résultats expérimentaux confirment la formation d'un hexamère cyclique lors de l'encapsidation du bactériophage ϕ29 de *B. subtilis.* La modélisation a pu être réalisée en parallèle avec les dernières expérimentations et a exigé l'introduction d'un nouveau type de contrainte permettant de vérifier la possibilité de former un multimère cyclique. Le modèle 3-D obtenu permet de faire plusieurs hypothèses quant aux mécanismes d'action de cette structure au sein du virus. Ces hypothèses sont présentement soumis à des vérifications expérimentales dans le groupe de D. Anderson et théoriques dans le groupe de F. Major, au laboratoire de biologie informatique et théorique du DIRO (Université de Montréal). Mon implication dans ces travaux couvre la modélisation de la structure 3-D et l'implantation de la contrainte de fermeture de multimères cycliques, ainsi qu'une partie des propositions de nouvelles expériences rendues possibles par l'analyse du modèle.

CHAPITRE 2

Representing and Infering RNA Three-Dimensional Structures

S. Lemieux, S. Oldziej et F. Major, *Encyclopedia of Computational Chemistry*, N.L. Allinger et *al.* Eds., John Wiley & Sons, West Sussex, England, 1998.

Nucleic Acids: Qualitative Modeling

Sébastien Lemieux Stanislaw Oldziej François Major
University of Montreal, Canada

Abbreviations CSP = constraint satisfaction problem; MMTV = mouse mammalian tumor virus.

1 INTRODUCTION

Knowledge about RNA three-dimensional structure is essential to RNA function comprehension and manipulation. Due to difficulties associated with physical RNA structure elucidation techniques, such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, it is not surprising that predictive methods are increasingly gaining popularity. Consequent to many genome sequencing projects, many novel RNAs of unknown structure and function are discovered, generating considerable efforts in three-dimensional structure determination by physical methods and in gathering of low and medium resolution structural data, such as those obtained from enzymatic and chemical probing, chemical modifications, sequence analysis and *in vitro* selection. RNA three-dimensional structure prediction methods are, thus, needed.

RNA three-dimensional structure predictions are logical consequences of structural knowledge and data. A prediction summarizes and condenses a large amount of experimental and theoretical data into a more comprehensible format. Predictions must be verified experimentally, and, thus, could be considered as good indicators for the planning of laboratory experiments. Particularly interesting are enzymatic activity data, such as those derived from *in vitro* selection and chemical modification experiments, which allow one to decipher active, as compared to ground state, conformations. Activity data are usually fuzzier and of lower resolution than x-ray crystallography and NMR spectroscopy data, but when used in combination with appropriate RNA three-dimensional structure determination methods could prove extremely informative and predictive.

RNA three-dimensional structure determination methods employ techniques such as distance geometry,[1] molecular mechanics,[2] simulated annealing,[3] interactive computer graphics[4] and other constraint satisfaction methods.[5,6] Several programs based on these techniques are productive and used in specific application field, to determine RNA three-dimensional structures from x-ray crystallography, NMR spectroscopy or low resolution data. It is expected that the next generation of RNA three-dimensional structure determination programs will allow us to enter descriptions of structural data (declarative), and to produce all associated consistent three-dimensional structures (sound and complete). Intuitively, this work could be justified, and motivated, from the fact that interactive computer modeling successfully suggests ways to explore efficiently the conformational space of RNAs, and is producing high precision structures. The resulting programs will automatically select the appropriate method, or combination of methods, according to the nature of the input data. The implementation of such programs could be simplified if a unified model of RNA structural knowledge and data is established. However, since it is computationally hard to address the theoretical flexibility of RNA three-dimensional structures, appropriate conceptualizations and approximations are necessary, implying that completeness and soundness in the biological sense might never be achieved.

The development of RNA three-dimensional structure determination methods requires three essential components. First is a computer representation, or a data structure, of RNA three-dimensional structural knowledge and data. Second is an RNA conformational search space that includes three-dimensional structures consistent with the computer representation. The implementation of a conformational search space includes the following tasks: i) the creation of a set of operators to manipulate the RNA three-dimensional structures; ii) the definition of a metric to evaluate RNA three-dimensional structures; iii) the design of an efficient method for applying the chosen metric; and, iv) the design of an efficient method for generating the next three-dimensional structure to consider. Third is an inference engine which searches the conformational search space for three-dimensional structures that fit input descriptions.

This article presents the most recent development of the MC-SYM research project at the Université de Montréal. MC-SYM is a joint effort between the computer science and biochemistry departments to develop computational methods in RNA three-dimensional structure determination. A first MC-SYM prototype was reported and released in 1991.[5] Since then, the program has been extensively tested and used to determine RNA three-dimensional structures from the use of many different types and sources of structural data. Table 1 shows a list of the main publications[7–20] in which the use of MC-SYM was important. The program is available by anonymous FTP[21] and on the WEB.[22]

The main body of this article is divided into four different sections. In section 2, a computer representation of RNA structural data based on graph theory is introduced. In section 3, an RNA three-dimensional conformational search space is developed from the creation of operators defined from nitrogen base spatial relations (base pairing and base stacking) and rigid nucleotide conformations. In section 4, an RNA three-dimensional structure inference engine based on a backtracking algorithm, which is sound and complete over the conformational search space introduced in section 3. The necessary steps to determine the three-dimensional structure of a GAGA tetraloop from low resolution NMR spectroscopy data[23] are presented to illustrate how one can use the computer technology presented in this article. Finally, in section 5, the three-dimensional structure determination of two small RNAs from low resolution NMR spectroscopy data are presented. The first is the loop 785-797 in 16S ribosomal RNA. The second is the 16S ribosomal RNA symmetrical motif of tandem G•U mismatches.

2 RNA STRUCTURAL DATA

The first step in the development of an RNA three-dimensional structure determination method is to find a computer representation of structural knowledge and data. Important properties of the representation are the flexibility and power to express any set of structural knowledge and data, independently of any particular structure determination method. Think of such an abstract repre-

Table 1 List of projects where MC-SYM was involved in the three-dimensional structure determination. The source of constraints and the main reference to the work are given for each project.

Molecule	Type of constraints	Reference
tRNA	X-ray structure	7
UUCG loop	NMR structure	8
Two transfer RNAs in ribosomal A and P sites	Fluorescence	9
Rev-binding element	NMR, multiple sequence analysis, *in vitro* selection	10
Double symmetrical GA mismatches[a]	X-ray structures, NMR structures	11
16S rRNA	Chemical/photocrosslinks	12
16S rRNA	Crosslinking of 16S rRNA with mRNA	13
P18-P8-P14 junction in ribonuclease P	Covariations	14
Two U1A molecules bound to U1A pre-mnRNA 3'UTR	X-ray structures of U1A RBD-RNA, NMR	15
Group II introns	Chemical modifications	16
Iron responsive element (RNA hairpin loop)	NMR	17
Ribonuclease P RNA	Chemical crosslinks	18
Ribonuclease P RNA	Covariations and chemical protection	19
16S RNA	Chemical crosslinks	20

sentation as the format or profile of a database entry within which any specific instance of an object can be described.

By definition, an RNA three-dimensional structure is a sequence of nucleotides spatially organized in three-dimensional space to achieve the higher-order structure. A graph of relations where the nodes represent the nucleotides and the edges their spatial relations is, thus, a convenient representation. A graph of relations allows one to address the dynamics aspects of RNA three-dimensional structures. Consider the GNRA tetraloop motif (N: any nucleotide; R: a purine) whose three-dimensional structure may fluctuate and change in different contexts (for an example see reference 16). The secondary structure of the GAGA tetraloop in the sarcin-ricin 28 ribosomal subunit is shown in Figure 1a. The three-dimensional structure of this tetraloop was determined from NMR spectroscopy data.[23] The loop three-dimensional structure contains a non-canonical A•G base pair and a sharp turn, with all nitrogen bases being stacked (see Figure 1b). The stabilization of the non-canonical $A_{15}{\bullet}G_{16}$ base pair results from base stacking with the nitrogen base A_{17} and the flanking Watson-Crick base pair. The GNRA tetraloop graph of relations is shown in Figure 2a. The described relations were inferred from secondary structure and NMR spectroscopy data.

3 CONFORMATIONAL SEARCH SPACE

A *conformational search space* is a set of three-dimensional structures that we are interested in searching. The three-dimensional structures of a conformational search space are defined by a series of parameters and its *size* by the product of the numbers of allowed values that can be assigned to each parameter. The three-dimensional structures in a conformational search space are related to each others by *operators* that modify the value of each parameter.

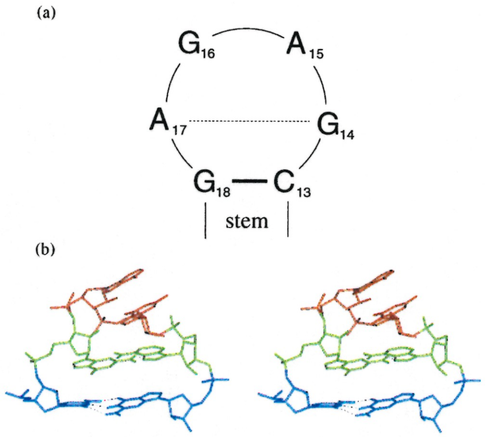


Figure 1 GAGA tetraloop extracted form the sarcin-ricin loop. (a) Secondary structure of the tetraloop. The backbone connectivity is shown with thin lines, the canonical pairing with a thick line and the non-canonical pairing with a thin dotted line. Only the last canonical base pair of the stem is shown. (b) Stereo view of one of the NMR spectroscopy three-dimensional structures. The $C_{13}{\bullet}G_{18}$ Watson-Crick base pair is shown in blue. The non-canonical A•G base pair is shown in green.

3.1 Operators

The RNA three-dimensional structure conformational search space considered here was derived from observed spatial relations among nitrogen bases and rigid nucleotide conformations. The basic operator adds a rigid nucleotide to a partially built three-dimensional structure by applying a spatial transformation to the atomic coordinates of the nucleotide. The spatial transformations indicate how nucleotides are appended to three-dimensional struc-

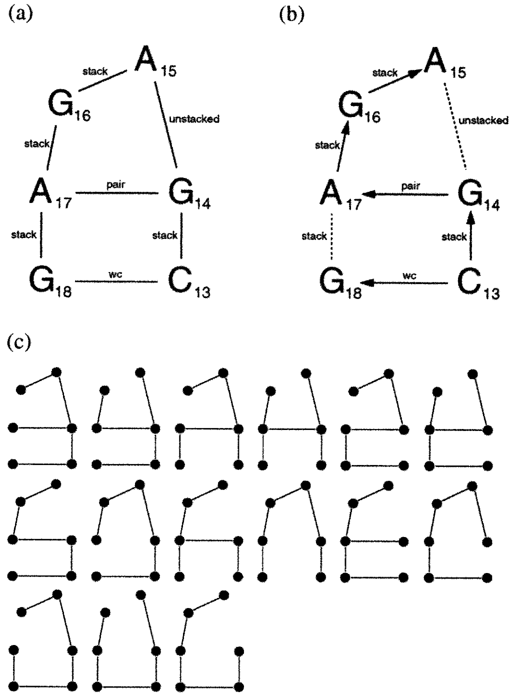


Figure 2 GAGA tetraloop graph of relations and spanning tree topologies. (a) The GAGA graph of relations that was used to rebuild the three-dimensional structure with MC-SYM. The graph includes cycles and the edges are undirected. (b) One spanning tree of the graph in (a). A spanning tree has no cycle and the edges are oriented. The orientation indicates which base is used as a reference and which one is appended to the structure. The edges indicated with dotted lines are not used in the construction. (c) Spanning tree topologies for the GAGA graph of relations in (a). Each topology embeds six different spanning trees.

tures from the nitrogen bases interactions. The rigid nucleotide conformational sets address the flexibility of the ribose and that of the phosphodiester chain. Nitrogen base spatial relations and rigid nucleotide conformations are taken from a database of previously determined and theoretically generated RNA three-dimensional structures and fragments.

The size of a conformational search space defined by spatial relations among nitrogen bases is smaller than a conformational search space that would be defined by backbone torsion angles. Mainly due to the theoretical flexibility of the backbone, a conformational search space defined by backbone torsion angles produces a very large number of three-dimensional structures of high uncertainty in the position and orientation of the nitrogen bases.

Since nitrogen base pairing and stacking are dominant features of RNA three-dimensional structures, it is believed that RNA native conformations are driven from nitrogen base interactions, rather than from the ribose-phosphodiester chain. The imprecision introduced in the ribose-phosphodiester backbone from a conformational sampling based on nitrogen base relations is negligible and several numerical refinement methods, such as energy minimization and molecular dynamics, allows us to fix it. Exhaustive nitrogen base oriented searches for large RNAs are nevertheless impossible in practice. The determination of RNA three-dimensional structures would not be possible in absence of geometrical constraints obtained from experimental data and theoretical inferences, or without reducing the number of known examples of spatial relations and nucleotide conformations.

3.2 Spatial Relations

In classifying spatial relations, one must consider two types of nucleotide pairs: adjacent and non-adjacent. Nitrogen base spatial relations are thus of three types: adjacent-stacked (helical and non-helical), adjacent-unstacked and non-adjacent-paired (see Figure 3). Nitrogen base spatial relations can efficiently be extracted from previously determined three-dimensional structures, stored and organized in a database, and applied to position and orient rigid nucleotide conformations.

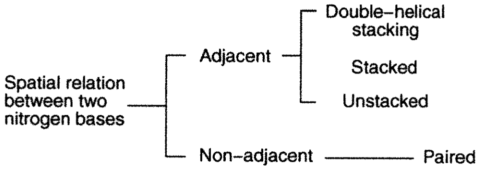


Figure 3 Classification tree of the nitrogen base spatial relations.

A spatial relation is encoded with a homogeneous transformation matrix that corresponds to the transformation, a combination of translation and rotation, of a nitrogen base local referential into another.[24] The local referential of a nucleotide can be thought of as its local axis system and is computed from its atomic coordinates. Consider the arbitrary choice: use the N1 atom as the origin; align the C4 atom with the Y axis; and, align the C6 atom with the X axis (see Figure 4). The local referential of a nitrogen base B1, R_{B1}, and a nitrogen base B2, R_{B2} is now easily expressed in a homogeneous transformation matrix: $T_{B1\rightarrow B2} = R_{B1}^{-1} \times R_{B2}$. A nitrogen base spatial relation between B1 and B2 can be reproduced between any pair of nitrogen bases, say B1' and B2', by applying the homogeneous transformation matrix $R_{B2'}^{-1} \times T_{B1\rightarrow B2} \times R_{B1'}$ to the atomic coordinates of B2' to position and orient B2' with respect to B1'. Symmetrically, $R_{B1'}^{-1} \times T_{B1\rightarrow B2}^{-1} \times R_{B2'}$ applied to atomic coordinates of B1' will position and orient B1' relative to B2', according to $T_{B1\rightarrow B2}$. In this way, any nitrogen base spatial relation found in a three-dimensional structure can be extracted and used afterwards as a building block.

A distance metric between two homogeneous transformation matrices can be defined to evaluate their spatial difference. Such a metric is used to optimize the choice of transformations and to minimize the re-application of similar examples during a conformational search.

3.3 Base Pairing

Hydrogen bonds are weak electrostatic interactions that play an important role in the stabilization of RNA three-dimensional structures. In general, more than 60% of the nitrogen bases in

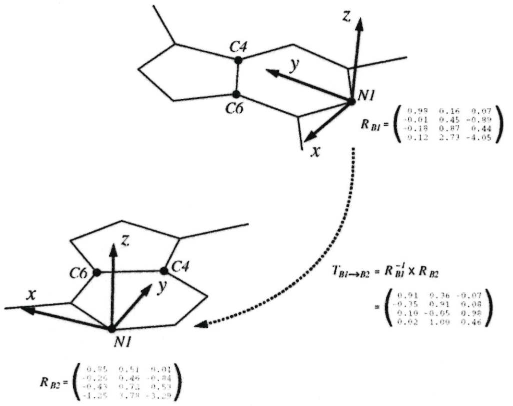


Figure 4 A spatial nitrogen base relation. The local referential, R_{B1} and R_{B2}, are represented by local axis systems. The N1 atom is centered at the origin; the C4 atom is aligned with the Y axis; and, the C6 atom is aligned with the X axis. The spatial transformation of B1 into B2, $T_{B1 \rightarrow B2}$, is given by the matrix product $R_{B1}^{-1} \times R_{B2}$.

an RNA are paired. Two nitrogen bases are *paired* if they share at least one hydrogen bond, even though in many observed examples they are composed of two bonds, and three bonds in the case of the canonical C•G Watson-Crick pair. Base pairs bring together nitrogen bases that are distant in the sequence, and thus, are considered very important for three-dimensional structure determination. Figure 5a shows a Watson-Crick base pair which was extracted from the three-dimensional structure of the frame-shifting pseudoknot of the mouse mammalian tumor virus (MMTV) as determined by NMR spectroscopy.[25] The base pairing network of an RNA defines its *secondary structure*.

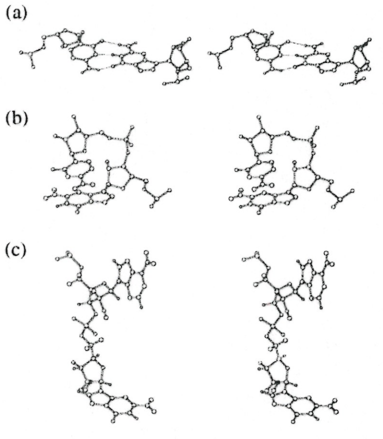


Figure 5 Nitrogen base spatial relations. (a) The Watson-Crick base pairing relation. (b) The base stacking relation. (c) The unstacked relation.

In Figure 6a, the five most distant examples of A•G base pairs of the type XI are shown. The adenines of the pairs were superimposed to evaluate the distribution of the paired guanines. The conformational freedom of a base pair, except for those defined by a single hydrogen bond, is small, as compared to other nitrogen base spatial relations (compare with Figure 6b and 6c). Base pairing information reduces efficiently the conformational search space of an RNA three-dimensional structures and, thus, should be used as much as possible in their determination.

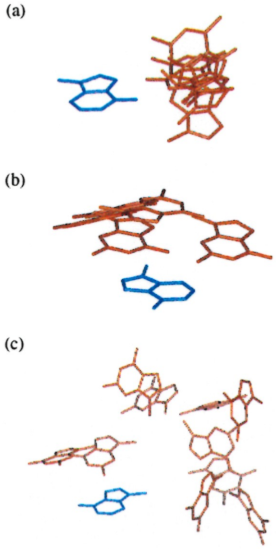


Figure 6 Conformational flexibility of AG spatial relations. All adenines are superimposed and shown in blue, and the guanines are shown in red. (a) The first five examples of the A•G type XI base pairing sorted list. (b) The first five examples of the A/G base stacking sorted list. (c) The first ten examples of the A-G unstacked sorted list.

3.4 Base Stacking

Base stacking is another important interaction in RNA three-dimensional structures. More than 60% of the nitrogen bases stack over another base. Base stacking involve dipole-dipole and dipole-induced dipole interactions (London dispersion), and hydrophobic forces. Although from a thermodynamic point of view these forces are weaker than in hydrogen bonds, the role of base stacking is also important in RNA folding and three-dimensional structure stabilization.

In general, base stacking occurs between two nitrogen bases that are adjacent in the sequence. Base stacking is a dominant feature of double-helical regions where the planes of the two nitrogen bases are almost perfectly parallel, and their overlapping is maximized. A vertical distance in the range of 3.3 to 3.6Å separates the two bases. Stacked bases in single-stranded regions do not necessarily keep the nitrogen bases parallel, and their overlapping is not optimal. The vertical distance between non-adjacent stacked bases can move up to approximately 4.5Å. Figure 5b shows an example of base stacking, as extracted from the three-dimensional structure of MMTV frame-shifting pseudoknot determined by NMR spectroscopy.[25]

In Figure 6b, the five most distant examples of A/G base stacking are shown. The adenines of the pairs were superimposed to evaluate the distribution of the stacked guanines. Mainly because of weaker stabilizing forces, the conformational freedom of two stacked bases is larger than that of paired bases. The most important variation comes from the base overlapping. Nevertheless, base stacking information reduces as well the conformational search space of RNA three-dimensional structures and, thus, should be used as much as possible in their determination.

3.5 Unstacked Bases

Two adjacent nitrogen bases are not necessarily stacked. In general, an unstabilized nitrogen base bulges out of the helical stem. An example of an unstacked relation is shown in Figure 5c, as extracted from the three-dimensional structure of MMTV frameshifting pseudoknot determined by NMR spectroscopy.[25] Nitrogen bases can also stack with only one partner. Consider the U-turn motif, such as observed in yeast tRNA^{Phe}. The U-turn occurs between a cytosine and a guanine where the latter is stabilized from base stacking with its 3' adenine. No base stacking can be measured between the guanine and its 5' uracil. The nitrogen base of this uracil, on the other hand, stabilizes its 5' cytosine with stacking.

In Figure 6c, the ten most distant examples of A-G unstacked bases are shown (see section 2.2.7 for details about the distance metric and classification of spatial relations). The adenines were superimposed to evaluate the distribution of the unstacked guanines. The guanines are located all over around the adenines, covering almost all stereochemical possibilities. The absence of stabilizing forces in bulged-out nitrogen bases implies a large conformational freedom that is constrained by the backbone flexibility only. The conformational space defined by the unstacked interaction is large and should be avoided in RNA three-dimensional structure determination, unless a specific unstacked relation, such as in the U-turn, is addressed. Unstacking can simply be verified by using a distance constraint between the O3' and P atoms of the two adjacent nucleotides.

3.6 Nucleotide Conformations

Although the conformational space of RNA three-dimensional structures is mainly constrained by nitrogen base spatial relations, the ribose and phosphate groups still allow one to prune several inconsistent conformations. In MC-SYM, the appending of nucleotides in a three-dimensional structure is performed by applying directly the nitrogen base transformations to the atomic coordinates of rigid nucleotide conformations.

The geometrical classification is an important step to generate a good sampling of nucleotide conformations, as well as to build an appropriate database for MC-SYM. The flexibility of nucleotide conformations is conferred by seven free torsion angles: six in the backbone (α, β, γ, δ, ϵ and ζ) and one around the glycosyl bond linking the sugar pucker and nitrogen base. The χ torsion adopts two main values: the *anti* which places the bulk of the nitrogen base oriented in the direction opposite to the sugar group; and, the *syn* which places the bulk of the nitrogen base oriented towards the sugar group.

Another important parameter is the sugar puckering mode which is strongly dominated by the C3'- and C2'-endo conformations, as found respectively in the A- and B-form double-helix. The cyclic furanoside sugar can be either puckered with a single atom pointing out of the plane formed by the four others, by up to 0.5Å, or twisted with two adjacent atoms out of a plane formed by the other three. Atoms out of the three or four atom plane on the side of the C5' are said to be *endo*, and those on the opposite side are said to be *exo*. Different combinations of the eight torsion angles correspond to similar nucleotide conformations and implies that a classification based on the eight angles is not possible; this is known as the "crankshaft" effect.[26] The next section describes how nucleotide conformations are classified in the structural database to optimize conformational sampling.

3.7 Structural Database

The exploitation of nitrogen base spatial relations and nucleotide conformations necessitates efficient storage and retrieval from a database. The entities in this structural database must map the structural knowledge and data representation described in section 2.1, that is, the graphs of relations. The database queries to determine the three-dimensional structure of an RNA are direct consequences of the selected spanning tree from the graph of relations. Some edges in the spanning tree can be directly mapped to the spatial relations described in section 2.2.3, 2.2.4 and 2.2.5.

In the MC-SYM structural databases, the spatial relations and nucleotide conformations are stored in lists. The lists are sorted so that any sublist of the first n elements represents the most efficient sampling of the addressed space. This property is achieved by selecting as the first element of a list, the one that minimizes the sum of the square distances with all other elements. This element is considered the most "common" example of the given relational or conformational list. The next element to insert in the list is the one that maximizes the difference with all previously included elements. This sorting method supposes the existence of distance metrics to evaluate the difference between two spatial relations or nucleotide conformations. The following metrics gave good results in practice.

The simplest method to measure the distance between two homogeneous transformation matrices is to sum the squares of the differences between the elements, resulting in the Euclidean distance metric. A variant of the Euclidean metric consists in weighting the translation and rotation components. By reducing the weight of the translation, it was possible to improve the conformational sampling of the unstacked nitrogen base relations. Another metric consists in computing the RMS deviation between two vectors obtained from the transformation of the canonical orthonormal vector, ($< 1,0,0 >, < 0,1,0 >$ and $< 0,0,1 >$), by the two matrices to be compared.

A simple metric to measure the distance between two nucleotide conformations is to compute the RMS deviation between the heavy atoms in the backbone of the two nucleotides when their nitrogen bases are superimposed.[8] A variant of this metric consists in calculating the distance between the phosphate atoms only. This allows one to avoid the potential counterbalancing effect of the sugar pucker modes and backbone torsion angles.

The MC-SYM structural database was built from the extraction of spatial relations and nucleotide conformations in previously determined RNA three-dimensional structures. A computer program that automatically detects the three types of spatial relations described above was implemented. The program analyzes hydrogen bonding patterns, quantifies nitrogen base stacking and computes the parameters of nucleotide conformations.

4 INFERENCE ENGINE

In searching a conformational search space, we have a particular *goal* in mind which describes what is searched for. In molecular structure determination, the goal corresponds to a subset of three-dimensional structures that satisfy what is described in the RNA graph of relations, including stereochemical and thermodynamics rules, and experimentally determined and theoretically inferred geometrical and chemical constraints. Most search methods proceed by systematically applying the set of operators (see section 2.2.1) and verifying whether the resulting three-dimensional structures are elements of the goal. Search methods guided by metrics are called *heuristics* and are based on techniques that are supposed to perform well in practice, although they provide no guarantees they will find the goal.

In computer three-dimensional structure determination, a metric on a conformational search space allows us to compute some measure of the *value* of a given conformation. The best example of such a metric is the model potential energy function. This metric can direct the search based on the assumption that applying operators to a conformation estimated to be closer to the goal will lead to it more rapidly than applying operators to more distant conformations. The model potential energy function serves as an *evaluation function* which estimates the conformations' values according to thermodynamics criteria. In practice, however, partly due to the local minima problem, the model potential energy function is employed to refine three-dimensional structures rather than to select the ones that will be used in further iterations of the determination task.

The view of the RNA three-dimensional structure determination problem described in this article matches the definition of a more general problem in computer science: the constraint satisfaction problem (CSP) (see reference 27 for a review). This correspondence was first noted by Major et al. in 1991.[5] Their MC-SYM search algorithm was based on a classical technique employed to solve discrete CSPs: the backtracking. This algorithm was most popularized when it started to be employed as the inference engine of logic programming languages such as Prolog. The computational complexity of the backtracking algorithm was analyzed by Haralick in 1980,[28] and only allows us to solve small problems. The backtracking search space grows exponentially with the number of variables, the number of allowed values for each variable; however, it decreases exponentially with the constraints. The backtracking search procedure is sound and complete.

In MC-SYM, the evaluation function is boolean. It accepts or rejects a three-dimensional structure whether or not all geometrical constraints are satisfied. Since nucleotides are assigned sound conformations, the verification of steric conflicts consists in checking inter-nucleotide collisions and O3'-P inter-nucleotide connections. The evaluation function has two components: the first for steric conflicts and O3'-P connections, and the other for problem specific user constraints.

The backtracking algorithm organizes the search space as a tree where each node corresponds to the application of an operator. At each application, the consistency of the partial structure is evaluated. If consistent, the next operator is applied and the process continues. If inconsistent, this node and attached branches are pruned from the search tree and the algorithm "backtracks" to the previous node.

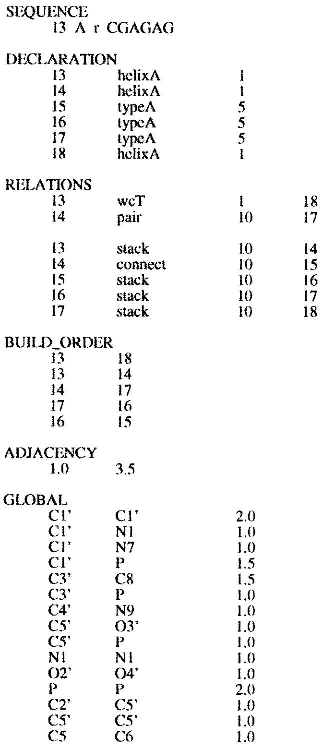
A spatial relation from the RNA graph of relations can either be used as a set of geometrical constraints or employed directly in an operator to position and orient a nucleotide. The latter use of spatial relations prunes the conformational search space more efficiently. However, it is not possible to use all spatial relations in this way without introducing a costly optimization problem. In a graph of N nucleotides, only $N-1$ spatial relations can be applied directly; all others become geometrical constraints that must be verified by the evaluation function. The spatial relations that are chosen to be applied directly determine a spanning tree of the relational graph. Figure 2b shows one possible spanning tree for the GNRA tetraloop. The GNRA tetraloop relational graph contains 15 different spanning trees (see Figure 2c). Spatial relations are symmetric and, therefore, the spanning trees cannot be differentiated by the orientation of the edges.

In MC-SYM, the selection of a spanning tree is left to the user. In making that choice, it is suggested to introduce first the nucleotides that impose the most constraints. For instance, there is less structural flexibility and more known examples of paired than stacked nucleotides and of stacked than unstacked nucleotides (see section 2.2.1). For instance in the GAGA tetraloop, it is more costly to use the edge $A_{15} \xrightarrow{unstacked} G_{14}$ than the edges $A_{15} \xrightarrow{stack} G_{16}$ or $A_{17} \xrightarrow{stack} G_{16}$. In this case, it is easy to replace the unstacked relations by an O3'-P distance constraint to insure the connectivity of the phosphodiester backbone.

Once a spanning tree has been chosen (see Figure 2b), the MC-SYM search engine requires the order in which the selected edges (operators) are to be applied. The order does not influence the results but will affect the execution time. For the GAGA tetraloop, the edges can be ordered in the following way: 1) $C_{13} \xrightarrow{wc} G_{18}$; 2) $C_{13} \xrightarrow{stack} G_{14}$; 3) $G_{14} \xrightarrow{pair} A_{17}$; 4) $A_{17} \xrightarrow{stack} G_{16}$; and, 5) $G_{16} \xrightarrow{stack} A_{15}$.

The production of three-dimensional structures from the spanning tree presented in Figure 2b is made by writing a script describing the structural information (see Figure 7), introducing a construction order that corresponds to the selected spanning tree, and finally running the MC-SYM inference engine. The SEQUENCE section introduces the sequence information. Each SEQUENCE record contains: an identifier referring to the first nucleotide in the sequence; a chain identifier; the type of the molecule (here "r" for RNA); and, the sequence of nucleotides. Each DECLARATION record contains: the sequence identifier of the nucleotide; the name of the rigid nucleotide conformational list; and, the number of conformations to extract from the conformational list. "helixA" corresponds to a list of one typical double-helical nucleotide conformation. "typeA" corresponds to the list of C3'-endo, *anti* conformations. Each RELATION record contains: the sequence identifier of the first nucleotide; the name of a nitrogen base spatial relational list; the number of transformations to extract from the transformational list; and, the sequence identifier of the second nucleotide. "wcT" corresponds to a typical Watson-Crick double-helix base pairing. "pair" corresponds to the list of all base pairing patterns between the two nucleotides. "stack" corresponds to the list of all base stacking relations between the two nucleotides. "connect" corresponds to the interleaved list of base stacking and unstacked relational lists. The BUILD_ORDER section indicates the order in which the three-dimensional structures are built. The first nucleotide identifier in each line indicates the reference nitrogen base. The following identifiers refer to the nucleotides to append to the one previously positioned in the list. In the script of Figure 7, nucleotide 13 is used as the global reference. 18 is placed from 13; 14 is placed from 13; 17 is placed from 14; 16 is placed from 17; and, 15 is placed from 16. The ADJACENCY and GLOBAL sections define respectively the following constraints: the O3'-P bond distances which ensure correct closure in the backbone, and the steric collisions given as a set of atomic pairs and distances which ensure that none of the atomic pairs will be positioned closer than the indicated distances. This represents a heuristic function to detect steric conflicts in substitution of full van Der Waals calculations. The lower and higher bounds for the O3'-P bond distance is declared in the ADJACENCY section. The GLOBAL section serves for the introduction of collision constraints. The first two fields indicate two atoms involved in the distance constraint and the third field indicates their lowest acceptable distance. The set of GLOBALS defined in the script of Figure 7 was developed over the years and could be considered as a good heuristic, although steric conflicts can sometimes be observed in MC-SYM generated three-dimensional structures (see the MC-SYM Homepage for more details about the script syntax[22]).

The conformational search space defined by the script of Figure 7 contains 1.25×10^6 different three-dimensional structures. The exploration requires approximately 1 second of CPU on a 180 MHz R5000 Silicon Graphics O2 workstation. MC-SYM explored 2400 complete structures where only 52 loops satisfied the constraints (see Figure 8a). The generated three-dimensional structures were compared to the NMR structure of Figure 1b. The RMS deviation of the superimposed nitrogen bases was calculated for each structure and was found to range between 1.3 and 2.4Å (see Figure 8b). This example shows how a graph of relations can be translated into accurate three-dimensional structures using MC-SYM. In the next section two additional MC-SYM applications are presented.



```
SEQUENCE
        13  A  r  CGAGAG

DECLARATION
        13          helixA              1
        14          helixA              1
        15          typeA               5
        16          typeA               5
        17          typeA               5
        18          helixA              1

RELATIONS
        13          wcT                 1           18
        14          pair                10          17

        13          stack               10          14
        14          connect             10          15
        15          stack               10          16
        16          stack               10          17
        17          stack               10          18

BUILD_ORDER
        13          18
        13          14
        14          17
        17          16
        16          15

ADJACENCY
        1.0         3.5

GLOBAL
        C1'         C1'                 2.0
        C1'         N1                  1.0
        C1'         N7                  1.0
        C1'         P                   1.5
        C3'         C8                  1.5
        C3'         P                   1.0
        C4'         N9                  1.0
        C5'         O3'                 1.0
        C5'         P                   1.0
        N1          N1                  1.0
        O2'         O4'                 1.0
        P           P                   2.0
        C2'         C5'                 1.0
        C5'         C5'                 1.0
        C5          C6                  1.0
```

Figure 7 A MC-SYM script for the GAGA tetraloop. The script represents the spanning tree in Figure 2b. All distances are given in Å.

5 APPLICATIONS

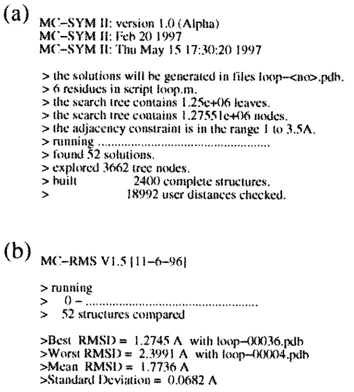
5.1 Loop 785-797

The loop 785-797 from the small subunit of 16S ribosomal RNA is one of the key elements involved in the production of proteins. The sequence of this loop is highly conserved across different organisms. Figure 9a shows the secondary structure of the loop. A coarse description of a set of NMR data[29] proposes canonical Watson-Crick base pairing for C_{797}•G_{785} and C_{796}•G_{786}, and unknown base pairing patterns for A_{787}•C_{795}, A_{788}•U_{794} and A_{790}•U_{793}. The NMR data indicate that all bases are stacked, except for U_{789} which flips out of main helical stem.

An MC-SYM script was written to reflect the RNA graph of relations corresponding to the qualitative evaluation of the NMR data (see Figure 9b). The base pairing patterns that were used in the exploration of the loop conformational search space for the three unknown base pairing patterns are shown in Table 2, for the A•U pairs, and Table 3, for the A•C pairs. All nucleotide conformations were assigned C2'- and C3'-endo sugar pucker modes and *anti* glycosyl bond torsion.

The conformational search space size of the selected spanning tree is 10^{23} corresponding to 3375 different combinations of base pairing patterns. MC-SYM generated 33998 consistent three-dimensional structures, composed of 45 different combinations of base pairing patterns (see Table 4). The RMS deviation among the 45 classes vary from 2.9 to 7.0Å, for all atoms except hydrogens. The potential energies vary from -72.5 to -28.6 Kcal/mol. Note that potential energy values are not necessarily good indicators of the quality or soundness of three-dimensional structures and cannot be used to select any candidate structure from this set. Nevertheless, five of these three-dimensional structures have lower potential energies, varying from -72.5 to -70.0 Kcal/mol: 75-48-42, 75-XX-XX, 75-XX-48, 75-48-XX, and 77-48-XX. Figure 10 shows one of these low energy conformations (75-XX-48). The A•C(75) base pairing pattern involves a protonated cytosine.

This example shows the efficiency of MC-SYM in identifying a small number of possible combinations from qualitative data and stereochemical constraints. 45 classes could be considered a large number of possibilities but, in fact, very few additional constraints would allow to reduce this number considerably. An NMR ex-

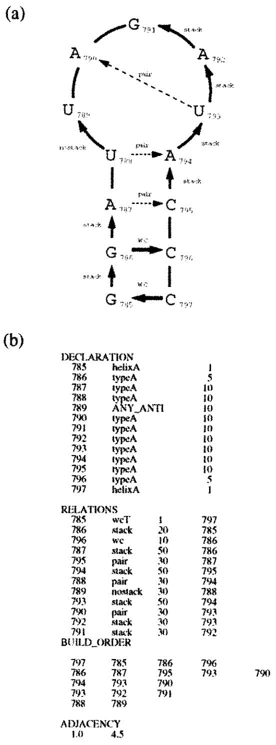


```
(a) MC-SYM II: version 1.0 (Alpha)
    MC-SYM II: Feb 20 1997
    MC-SYM II: Thu May 15 17:30:20 1997

    > the solutions will be generated in files loop-<no>.pdb.
    > 6 residues in script loop.m.
    > the search tree contains 1.25e+06 leaves.
    > the search tree contains 1.27551e+06 nodes.
    > the adjacency constraint is in the range 1 to 3.5A.
    > running ................................................
    > found 52 solutions.
    > explored 3662 tree nodes.
    > built              2400 complete structures.
    >                   18992 user distances checked.


(b) MC-RMS V1.5 [11-6-96]

    > running
    >    0 - ................................................
    >   52 structures compared

    >Best  RMSD =  1.2745 A  with loop-00036.pdb
    >Worst RMSD =  2.3991 A  with loop-00004.pdb
    >Mean  RMSD =  1.7736 A
    >Standard Deviation =  0.0682 A
```

Figure 8 MC-SYM results for the GAGA tetraloop. (a) MC-SYM summary of the execution. (b) MC-SYM summary of the RMS deviation analysis. The comparison was made with one of the NMR spectroscopy structures.

periment can be designed to measure the chemical exchange of a proton with solvent. For instance in the 785-797 loop, such an experiment could be designed for the amido protons of A_{788}. A low exchange rate would mean that the protons are involved in hydrogen bonding which would allow one to eliminate all structural classes that do not contain a hydrogen bond involving this amido proton. This would reduce the number of models from 45 to only 14, since only the base pairing of types XX and XXI satisfy this constraint.

It can be questioned whether these 45 classes represent the actual conformational space of this motif, that is, the set of three-dimensional structures which do not violate any stereochemical rule. We believe that MC-SYM probably produces a subset of the actual conformational space due to missing nitrogen base relations in the database. This problem could have been avoided by generating theoretical examples, as introduced in section 2.2.7 concerning the structural database, but it would have been costly given the size of this loop. Figure 10 shows one of the MC-SYM generated and AMBER 4.1[2] optimized structure. The energy minimization protocol consists of: i) 2000 steps with fixed coordinates in all nitrogen bases, and, thus, only the backbone atoms were modified; ii) 3000 steps with distance and torsion angle constraints to preserve the original geometry of the hydrogen bonds; and, iii) 20000 unconstrained steps. All optimization steps were performed *in vacuo*.



```
DECLARATION
   785        helixA              1
   786        typeA               5
   787        typeA              10
   788        typeA              10
   789        ANY_ANTI           10
   790        typeA              10
   791        typeA              10
   792        typeA              10
   793        typeA              10
   794        typeA              10
   795        typeA              10
   796        typeA               5
   797        helixA              1

RELATIONS
   785        wcT        1       797
   786        stack     20       785
   796        wc        10       786
   787        stack     50       786
   795        pair      30       787
   794        stack     50       795
   788        pair      30       794
   789        nostack   30       788
   793        stack     50       794
   790        pair      30       793
   792        stack     30       793
   791        stack     30       792
BUILD_ORDER

   797        785       786       796
   786        787       795       793        790
   794        793       790
   793        792       791
   788        789

ADJACENCY
   1.0        4.5
```

Figure 9 The 785-797 loop of 16S ribosomal RNA. (a) Secondary structure deduced from NMR data and a spanning tree. The arrows indicate the selected edges in the spanning tree. (b) MC-SYM script corresponding to the spanning tree in (a). The "ANY_ANTI" list contains combinations of C3'-, C2'-endo and *anti* nucleotide conformations.

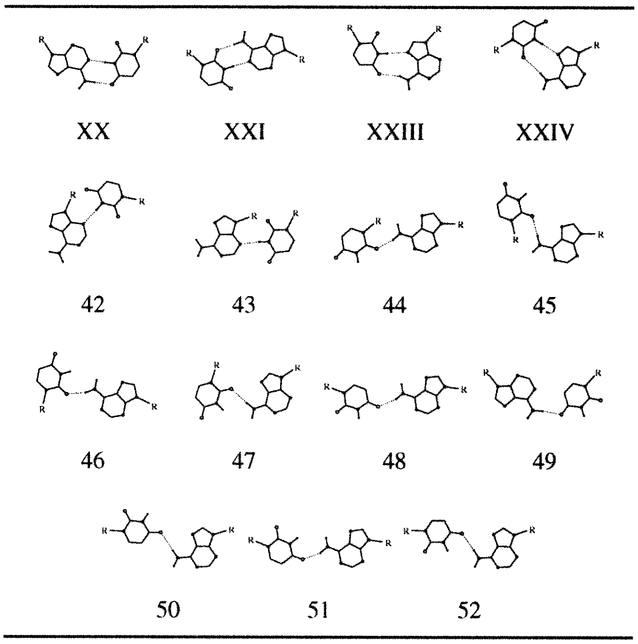
5.2 Double G-U Mismatches

This application exemplifies how theoretical methods were used to extend the MC-SYM database. The symmetrical tandem of G-U mismatches is a motif that was identified in ribosomal RNAs. The secondary structure and a selected spanning tree of the motif are shown in Figure 11. The list of possible G•U base pairing patterns that involve two hydrogen bonds are shown in Table 5. Two of these patterns, XXVII (reverse Wobble) and XXVIII (Wobble), were described in Saenger[30] and both were found in RNA and DNA x-ray crystallography and NMR spectroscopy three-dimensional structures.[31-33] The two others, XXX and XXXI, were created from hydrogen bonding theory. All nucleotide conformations in the motif were assigned C2'- and C3'-endo sugar pucker modes and the glycosylic torsion the *syn* and *anti* values.

When using the experimentally observed base pairing examples, MC-SYM determines that only the Wobble base pairing pattern is consistent to a symmetrical tandem of G•U mismatches. This fits the experimental data where the reverse Wobble pattern was only observed in a single mismatch and the Wobble pattern was observed in both single and double mismatches. The RMS deviations of the MC-SYM built three-dimensional structures with the x-ray crystallography structure are in the range of 2.3 to 2.6Å, for all atoms except hydrogens. The RMS deviation between one of the MC-SYM structures refined using AMBER 4.1[2] and the x-ray crystallography structure is less than 1.5Å.

The currently available RNA three-dimensional structures do not contain examples of the XXX and XXXI base pairing patterns. One way to overcome this problem is to use theoretical methods such as quantum and molecular mechanics to generate examples of the missing patterns. The conformational space of a nucleotide was defined by its seven torsion angles (see section 2.2.6). Steps of 60^o from 0^o were assigned to the β, γ and ϵ torsion angles. The values 0^o (*syn*) and 180^o (*anti*) were assigned to the χ torsion angle. Finally, the combination values {60-60, 60-180, 150-

Table 2 A•U base pairing patterns. The nitrogen atoms are indicated with small bullets and the oxygen atoms with circles. The riboses are indicated by the letter "R".

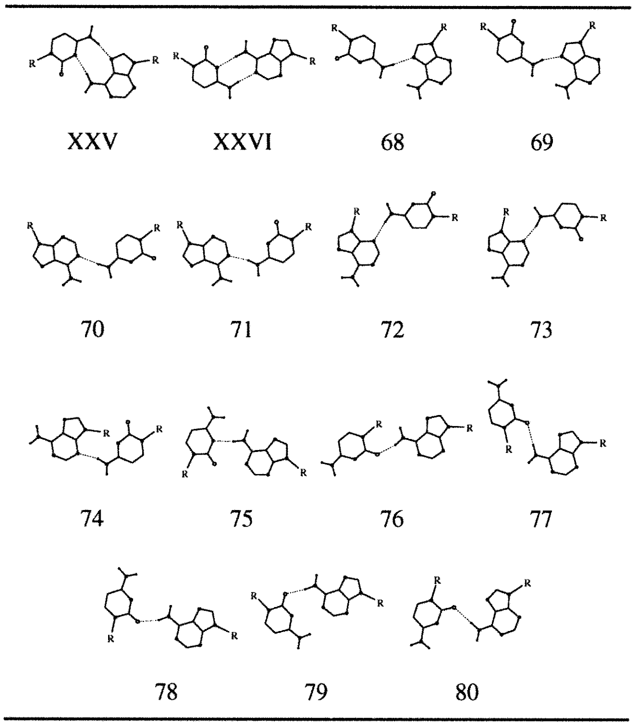


240, 180-120, 270-90, 270-180 and 270-300} were respectively assigned to the α and ζ torsion angles.[30] This defines a nucleotide conformational search space of 3024 pairs of adjacent nucleotides. Experimental examples of base pairings were used when available and idealized planar examples were generated otherwise. The geometry of all obtained motif conformations were optimized using the AMBER 4.1 force field.[2] For the first 5000 steps, the selected torsion angles were allowed to change by ± 30° according to their initial values. 60000 geometrically optimized conformations were obtained, creating a theoretical database from which base pairing and stacking relations were incorporated in the MC-SYM database.

Using the new database, MC-SYM produced consistent three-dimensional structures that contained the reverse Wobble, the Wobble and the XXXI G•U base pairing patterns. The three MC-SYM three-dimensional structures of symmetrical tandem of G•U mismatches, refined using AMBER 4.1,[2] are shown in Figure 12. In the case of class structures that include the reverse Wobble base pairing pattern, the guanosines were assigned *syn* conformations. The base pairs are not coplanar, a twist angle of 26° was measured. The base stacking pattern is also distorted from ideal parallel relative orientations (see Figure 12a). The class structures that contained the Wobble base pairs were similar to the results presented in the above paragraph (see Figure 12b).

The class containing the XXXI base pairing pattern (see Figure 12c) displays cross-strand stacking, similar to what was observed in the three-dimensional structure of symmetrical tandem of A•G mismatches, as determined by NMR spectroscopy[34] . Base stacking is almost perfectly parallel, suggesting a good stabilization of the three-dimensional structure. The uracils were assigned *syn* conformations and the base pairing patterns are not perfectly coplanar. The base pairing pattern XXXI involves two hydrogen bonds: a proton of the amino group in the guanine pairs with the O2 in the uracil (not preserved in the optimized structure shown in Figure 12c); and, the amid proton in the uracil pairs with N7 in the guanine. The stabilization of the structure is rather based on the formation of two additional hydrogen bonds: a proton of the amino group in the guanine pairs with one of the phosphate's oxygens in the uracil of the opposite strand; and, the O4 of the uracil pairs with a proton of the 2' hydroxyl group in the guanine of the opposite strand.

Table 3 A•C base pairing patterns. The nitrogen atoms are indicated with small bullets and the oxygen atoms with circles. The riboses are indicated by the letter "R".



6 RELATED ARTICLES

- CAC041
- CAM001
- CAM023
- CAN010
- CAN011
- CAN013

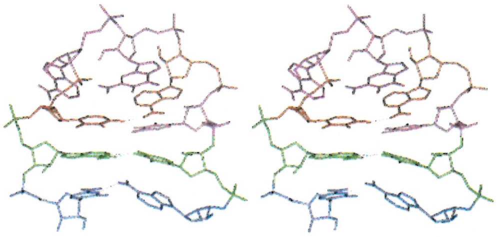
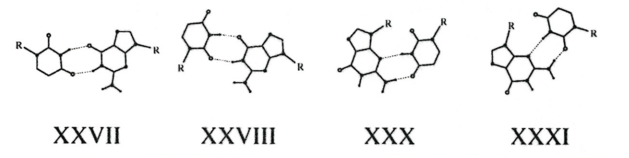


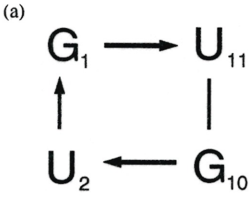
Figure 10 Stereoview of a three-dimensional structure of class 75-XX-48 for the 785-797 loop of 16S ribosomal RNA. The hydrogen atoms covalently bonded to carbon atoms were removed. The A_{787}•C_{795} base pair is shown in blue; the U_{788}•C_{794} base pair is shown in green; the A_{790}•C_{793} base pair is shown in red; and, the three unpaired nitrogen bases U_{789}, G_{791} and A_{792} are shown in magenta.

Table 4 The 45 combinations of base pairing patterns found in the set of loop 785-797 three-dimensional structures generated by the MC-SYM program.

75-44-42	75-48-42	75-XX-42
75-44-44	75-48-44	75-XX-44
75-44-48	75-48-48	75-XX-48
75-44-XX	75-48-XX	75-XX-XX
76-44-42	76-48-42	76-XX-44
76-44-44	76-48-44	76-XX-48
76-44-48	76-48-48	76-XX-XX
76-44-XX	76-48-XX	
		77-XX-44
77-44-42	77-48-42	77-XX-48
77-44-44	77-48-48	77-XX-XX
77-44-48	77-48-XX	
77-44-XX		77-XXI-44
	XXV-44-44	
78-48-42	XXV-44-XX	XXV-48-44
78-48-48		XXV-48-48
78-48-XX	XXV-XX-XX	XXV-48-XX

Table 5 G•U base pairing patterns involving two hydrogen bonds. The nitrogen atoms are indicated with small bullets and the oxygen atoms with circles. The riboses are indicated by the letter "R".





(b)

```
DECLARATION
  1   ANY_ANY 30
  2   ANY_ANY 30
 10   ANY_ANY 30
 11   ANY_ANY 30

RELATIONS
  2   XXVII        *      10
  1   connect    300       2
 10   XXVII        *       1

BUILD_ORDER
 10   2       1    11

ADJACENCY
1.0   3.5
```

Figure 11 Symmetrical tandem of G•U mismatches. (a) Secondary structure. The arrows indicate the selected edges in the spanning tree. (b) MC-SYM script corresponding to the spanning tree in (a). The "ANY_ANY" list corresponds to combinations of C3'-, C2'-endo, *anti* and *syn* nucleotide conformations. The XXVII relational list corresponds to the reverse Wobble G•U base pairing pattern (see Table 5). The "*" indicates that all examples of a given list are tested by MC-SYM.

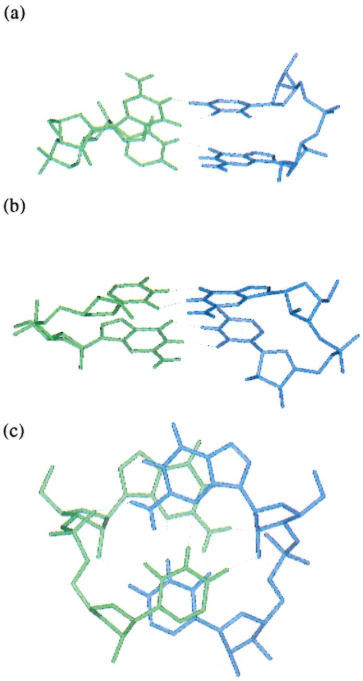


Figure 12 Stereoviews of three three-dimensional structures of symmetrical tandem of G•U mismatches generated by MC-SYM. (a) Class containing the reverse Wobble base pairing pattern. (b) Class containing the Wobble base pairing pattern. (c) Class containing the theoretical XXXI base pairing pattern, and the cross-strand stacking motif.

7 REFERENCES

1. L.M. Blumenthal, 'Theory and applications of distance geometry', Chelsea, NY, 1970.
2. D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, D.M. Ferguson, G.L. Seibel, U.C. Singh, P.K. Weiner, P.A. Kollman, 'AMBER 4.1', University of California at San Franciso, 1995.
3. S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi, *Science*, 1983, **220**, 671–680.
4. F. Michel and E. Westhof, *J. Mol. Biol.*, 1990, **216**, 585–610.
5. F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion and R. Cedergren, *Science*, 1991, **253**, 1255–1260.
6. R.B. Altman, B. Weiser and H.F. Noller, in 'Constraint satisfaction techniques for modeling large complexes: application to the central domain of 16S ribosomal RNA', eds., R. Altman, D. Brutlag, P. Karp, R. Lathrop and D. Searls, Proceedings ISMB, AAAI Press, Menlo Park, 1994, pp. 10–18.
7. F. Major, D. Gautheret and R. Cedergren, *Proc. Natl. Acad. Sci. (USA)*, 1993, **90**, 9408–9412.
8. D. Gautheret, F. Major and R. Cedergren, *J. Mol. Biol.*, 1993, **229**, 1049–1064.
9. T. Easterwood, F. Major, A. Malhotra and S. Harvey, *Nucl. Acids Res.* 1994, **22**, 3779–3786.
10. F. Leclerc, R. Cedergren and A.D. Ellington, *Nature Struct. Biol.* 1994, **1**, 293–300.
11. M. Foucrault and F. Major, in 'Symbolic generation and clustering of RNA 3-D motifs', eds., C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak, Proceedings ISMB; AAAI Press, Menlo Park, 1995, pp. 121–126.
12. P. Minchew, S. Joy, R. Bhangu and P. Wollenzien, *Nucleic Acid Symposium Series*, 1995, **33**, 68–69.
13. P. Wollenzien, D. Juzumiene, T. Shapkina and P. Minchew, *Nucleic Acid Symposium Series*, 1995, **33**, 76–78.
14. J. Brown, J.M. Nolan, E.S. Haas, M.A. Rubio, F. Major and N.R. Pace, *Proc. Natl. Acad. Sci. (USA)*, 1996, **93**, 3001–3006.
15. L. Jovine, C. Oubridge, J.M. Avis and K. Nagai, *Structure*, 1996, **4**, 621–631.
16. D.L. Abramovitz, R.A. Friedman and A.M. Pyle, *Science*, 1996, **271**, 1410–1413.
17. L.G. Laing and K.B. Hall, *Biochemistry*, 1996, **35**, 13586–13596.
18. M.E. Harris, A.V. Kazantsev, J.-J. Chen and N.R. Pace, *RNA*, 1997, **3**, 561–576.
19. T.R. Easterwood and S.C. Harvey, *RNA*, 1997, **3**, 577–585.
20. C. Wilms, J.W. Noah, D. Zhong and P. Wollenzien, *RNA*, 1997, **3**, 602–612.
21. MC-SYM FTP anonymous site: ftp.umontreal.ca (/pub/lbit/).
22. MC-SYM Homepage: http://www-lbit.iro.umontreal.ca/MCSYM/.
23. A.A. Szewczak, Y.L. Chan, P.B. Moore and I.G. Wool *Biochimie* **1991**, 73, 871–877.
24. R.P. Paul, 'Robot Manipulators: Mathematics, Programming, and Control', MIT Press, Cambridge, 1981.
25. X. Chen, H. Kang, L.X. Shen, M. Chamorro, H.E. Varmus and I. Tinoco Jr., *J. Mol. Biol.*, 1996, **260**, 479–483.
26. D.A. Pearlman and S.H. Kim, *J. Biomol. Struct. Dyn.*, 1986, **4**, 49–67.
27. V. Kumar, *AI magazine*, 1992, **13**, 32–44.
28. R.M. Haralick, *Artificial Intelligence*, 1980, **14**, 263–313.
29. L. KangSeok, S. Varma, J. SantaLucia Jr. and P.R. Cunningham, *J. Mol. Biol.*, 1997, **269**, 732–743.
30. W. Saenger, 'Principles of Nucleic Acid Structure', Springer-Verlag, New-York, 1984.
31. T. Brown, G. Kneale, W.N. Hunter and O. Kennard, *Nucl. Acids Res.*, 1986, **14**, 1801–1809.
32. C. Cheong, G. Varani and I. Tinoco Jr., *Nature*, 1990, **346**, 680–682.
33. J.A. McDowell and D.H. Turner, *Biochemistry*, 1996, **35**, 14077–14089.
34. J. SantaLucia Jr. and D.H. Turner, *Biochemistry*, 1993, **32**, 12612–12623.

CHAPITRE 3

Computer RNA Three-Dimensional Modeling from Low-Resolution Data and Multiple Sequence Information

F. Major, S. Lemieux et M. Ftouhi. *Molecular Modeling and Structural Determination of Nucleic Acids*, N.B. Leontis et J. Santa Lucia Eds., American Chemical Society Books, Washington, D.C., 1997.

Computer RNA Three-Dimensional Modeling From Low Resolution Data and Multiple-Sequence Information

François Major, Sébastien Lemieux and Abdelmjid Ftouhi

Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal
Montréal, Québec, Canada H3C 3J7
major@iro.umontreal.ca

The problem of modeling three-dimensional structures of ribonucleic acids is expressed in terms of the constraint satisfaction problem. Three-dimensional structures are represented by constraint graphs, where vertices represent nucleotides and edges represent structural constraints. A formalism to help rationalize a series of modeling experiments in the context of low resolution and multiple-sequence data was developed. From secondary structure and low resolution data, several structural hypotheses corresponding to different constraint graphs can be derived. In presence of several structurally related sequences, the application of three-dimensional modeling to each sequence and hypotheses produces a sequence-structure relation that can be analyzed using fuzzy set theory, given the imprecision and uncertainty involved in the modeling process.

The popularity of computer modeling of RNA three-dimensional structure can be explained by the desire to rapidly understand the function of newly discovered RNAs and by the difficulties of applying high resolution structure determination techniques, such as X-ray crystallography and nuclear magnetic resonance spectroscopy. Computer modeling implies the interpretation of experimental data, the formation of structural hypotheses, and the building of three-dimensional models. Such models offer a simultaneous view of many aspects of the molecule and allow one to design more precise and incisive experiments which, in turn, generate new structural data and hypotheses leading to new modeling experiments. Thus, models are dynamic objects that represent the quantity and quality of structural knowledge on a molecule at a given time. The iterative use of modeling and low resolution experimental methods should converge on a highly defined and accurate model.

FM is a fellow of the Canadian Genome Analysis and Technology program and the MRC of Canada. This work has been supported by the MRC of Canada.

Most three-dimensional modeling projects begin with primary and secondary structure, low resolution data and multiple-sequence data *(1)* from which many different structural hypotheses can be derived. One way to support a structural hypothesis consists in building a consistent three-dimensional model compatible with each available sequence *(2, 3, 4)*. A systematic verification consists in building all possible models for each active sequence. This creates a relation, $R \subseteq S \times H$, where R, the set of relations, is a subset of binary relations between S, the set of sequences, and H, the set of structural hypotheses; $(S_i, H_j) \in R$ if and only if the sequence S_i generates at least one three-dimensional model that satisfies the structural hypothesis H_j. Each structural hypothesis, H_j, is associated with a set, E_{H_j}, that contains all three-dimensional models consistent with H_j. Computer programs such as MC-SYM *(5)* which transform constraint graphs into three-dimensional models can be used, although the formalism presented here is independent of any particular modeling method.

Uncertainty in modeling lies in the fact that a three-dimensional model can either support a structural hypothesis or can be the result of modeling artifacts. Computer modeling is subject to imprecision in the low resolution data, subjectivity in the generation of three-dimensional models, and uncertainty in the formation of structural hypotheses. The theory of possibility, based on fuzzy logic, is used to classify structural hypotheses according to their likelihood to contain multiple-sequence data consistent conformations based upon the sequence-structure relation, R.

In this article we present the constraint graph representation used by MC-SYM to transform structural data into three-dimensional models. Then, we discuss the sequence-structure relation and the theory of possibility to assign plausibility coefficients to each structural hypothesis. Finally, we discuss the application of this technique to the lead-activated ribozyme and indicate how modeling was used iteratively with experimentation to derive its active structure.

RNA Conformational Space

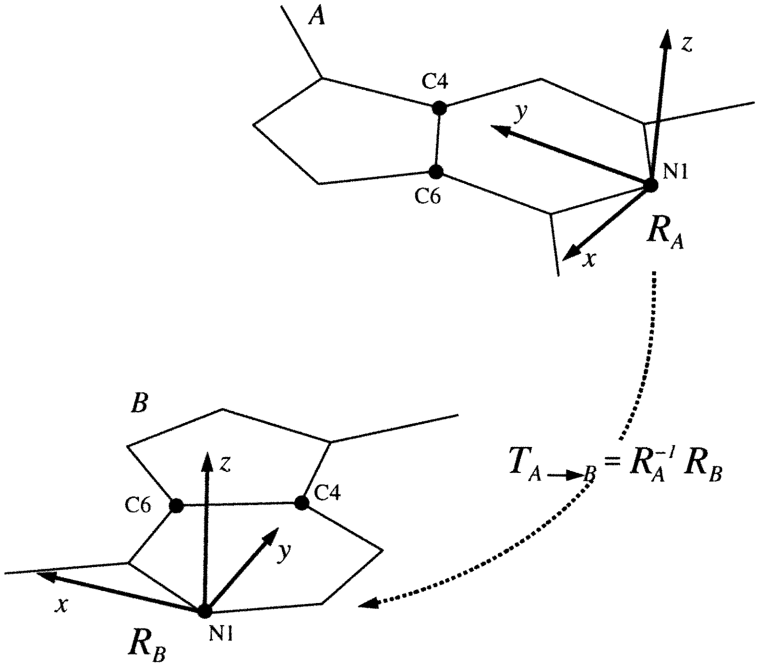
Here, we consider a *RNA three-dimensional structure* as the assembly of its constituent nucleotides in three-dimensional space. We introduce a RNA *conformational search space* defined by molecular contacts (or constraints). The molecular contacts are used in operators that position and orient the nucleotides in three-dimensional space.

A *molecular contact* is formed between two nucleotides, A and B, if they are connected through a phosphodiester bond or if they share a hydrogen bond between their nitrogen bases. The combination of all molecular contacts constitutes the *contact graph* of the RNA. It is self evident from the definition of a molecular contact that in all known RNA three-dimensional structures, every nucleotides make at least one molecular contact with another one. Thus, all RNAs contain at least one *path* of molecular contacts that connects all its constituent nucleotides, which does not contain any cycle, a *spanning tree* of the *nucleotide contact graph*.

In the following, we first present how contact graphs define the conformational search space of RNAs, to position and orient all nucleotides in three-dimensions. Then, a database of spatial relations based on molecular contacts, as observed among pairs of nucleotides in known structures, is introduced.

RNA Conformational Search Space Defined by Molecular Contacts. The premise to use molecular contacts in defining the conformational space of RNAs relies on the fact that molecular contacts contain all the information critical to the global fold of the molecule. Consider the best characterized case of an RNA double-helix. The spatial relation between two bases involved in a Watson-Crick pairing can be used, in conjunction with a canonical base stacking geometry, as a good approximation to position and orient double-helical strands in three-dimensions.

The spatial information is encoded by homogeneous transformation matrices *(6)*. The *local referential* of a nucleotide, A, can be represented by an homogeneous transformation matrix, R_A. R_A is determined by the coordinates of three atoms in A from which three right handed unary orthogonal vectors can be derived. The Cartesian coordinates of the first selected atom, for instance, can be chosen as the origin of the residue (see Figure 1). The *spatial relation* between two nucleotides, A and B, is an homogeneous coordinate transformation matrix, $T_{A\rightarrow B} = R_A^{-1}R_B$. In this way, the spatial relations between any pair of nucleotides forming molecular contacts in the known three-dimensional structures can be extracted and used as building blocks of RNA three-dimensional structures.



An observed contact between nucleotides A and B can be reproduced between any pair of nucleotides, let's say A' and B', by applying the homogeneous transformation matrix $R_{B'}^{-1}T_{A\rightarrow B}R_{A'}$ to the atomic coordinates of B' to position and orient B' with respect to A' as observed between nucleotides A and B; or symmetrically, by applying the homogeneous transformation matrix $R_{A'}^{-1}T_{A\rightarrow B}^{-1}R_{B'}$ to the atomic coordinates of A'. The final result of this manipulation is, in either case, that the new observed spatial relation between A' and B' is exactly the same as that observed between A and B, thus reproducing the same molecular contact in the newly built model.

A *transformational set* is a set of homogeneous transformation matrices associated with a molecular contact type defined by the nature of the nucleotides in contact. For instance, there are four types of RNA bases determining ten different types of pairs by considering that symmetric pairs are likely to share the same types of molecular contacts, and two main types of molecular contacts: paired and connected. Connected nucleotides can be either stacked or not. In practice, we consider only two types of bases (purines and pyrimidines) for contact defined by a phosphodiester bond. This partition of the different molecular contacts gives a possibility of $(10 + 6 = 16)$ different transformational sets.

The number of spanning trees, pairs of nucleotide contacts and homogeneous transformation matrices associated with a contact graph determine the conformational search space *size* of a RNA. The number of homogeneous transformation matrices associated with a molecular contact type is given by the number of occurrences observed in all available RNA three-dimensional structures in the Protein DataBank (PDB) *(7)*, Nucleic acids DataBase (NDB) *(8)* and other personally communicated structures.

The transformations were extracted and classified among the 16 different types of contacts. Those sets were then sorted in such a way that any subset composed of the first n elements represents the most efficient sampling of the addressed space. This property is achieved by selecting, as the first element of the set, the one that minimizes the sum of its distances with all other elements. This element is then considered the most "common" example. The next elements are those that maximizes their distances with all previously included elements. This sorting method supposes the existence of a distance metric to evaluate the difference between two homogeneous transformation matrices. The simplest metric is to sum the squares of the differences between the corresponding matrix elements, the Euclidean distance metric.

Starting from the contact graph, an efficient way to build three-dimensional models is to first determine a reference nucleotide that will be placed arbitrarily in three-dimensional space. From that, a spanning tree of the contact graph is expanded, determining which contacts will be used in the building procedure. For each molecular contact appearing in the spanning tree, the corresponding transformational set is used to systematically search the conformational space for valid three-dimensional models. Molecular contacts represented by edges that are not considered in the selected spanning tree are replaced in the simulation by geometrical constraints to guarantee their satisfaction in the final three-dimensional models.

The computer program MC-SYM is currently used to perform this search. Since the number of spanning trees of a fully connected graph composed of N vertices is N^{N-2}, the problem of selecting the one that is the most likely to generate complete structures is still open. The fuzzy logic approach presented here was developed in part to deal with the inaccuracy introduced by the approximation made while selecting a specific spanning tree.

The Sequence-Structure Relation. Structural hypotheses are derived from available structural data and are distinguished by their patterns of base pairing and stacking. For each structural hypothesis and sequence, a three-dimensional modeling simulation is performed, for instance using the MC-SYM program. In fact, any three-dimensional scheme determining if a three-dimensional model can be built for a given sequence and constraint graph is acceptable. The sequence-structure relation is established by associating the sequences to their consistent structural hypotheses; a link is created if and only if a three-dimensional model can be built. An MC-SYM input script describes the constraint graph and a sequence. By changing the latter, one can easily verify if a different sequence is compatible with the constraint graph.

Terminology and Notation of the Uncertainty Principle

In the history of mathematics, *uncertainty* was approached in the XVIIth century by Pascal and Fermat who introduced the notion of probability. However, probabilities do not allow one to process subjective beliefs nor imprecise or vague knowledge, such as in computer modeling of three-dimensional structure. Subjectivity and imprecision were only considered from 1965, when Zadeh, known for his work in systems theory, introduced the notion of *fuzzy set*. The concept of fuzziness introduces partial membership to classes, admitting intermediary situations between no and full membership. Zadeh's *theory of possibility*, introduced in 1977, constitutes a framework allowing for the representation of such *uncertain* concepts of non-probabilistic nature *(9)*. The concept of fuzzy set allows one to consider imprecision and uncertainty in a single formalism and to quantitatively measure the preference of one hypothesis versus another. Note, however, that Bayesian probabilities could have been used instead.

Consider a finite reference set, X. Events can be defined by subsets of X to which can be assigned coefficients between 0 and 1 evaluating their possibility to occur. In order to define these coefficients, a measure of possibility is introduced, Π, which is a function defined over the power set of X (the set of all subsets composed of the elements of X), $\mathcal{P}(X)$, the parts of X which take their values in $[0, 1]$, such that:

$$\Pi(\emptyset) = 0, \quad \Pi(X) = 1, \tag{1}$$

$$\begin{aligned}&\forall A_1 \in \mathcal{P}(X), \quad A_2 \in \mathcal{P}(X), \ldots \\ &\Pi(\cup_{i=1,2,\ldots} A_i) = \max_{i=1,2,\ldots} \Pi(A_i),\end{aligned} \tag{2}$$

where $\emptyset$ is the empty set and max indicates the maximum value of all values. The possibility associated to the empty set is zero. The possibility of X is one. The possiblity of a series of events (union) is the maximum possibility among these events.

The functions of *belief* concern a quantification of credibility attached to the events. Shafer's *theory of evidence* considers a finite universe of reference, X, upon which are determined belief coefficients obtained by distributing a global mass of belief equal to 1 among all possible events *(10)*. A mass, m, can be defined as follows:

$$m : \mathcal{P}(X) \longrightarrow [0, 1]$$

such that

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \in \mathcal{P}(X)} m(A) = 1.$$

For each set $A \in \mathcal{P}(X)$, the value $m(A)$ represents the degree with which a group of observers believe in the realization of an event from the elements of A. This value, $m(A)$, involves only a single set, the set A, and does not involve any other information for the subsets of A. If there exists additional evidence which confirms the realization of the same event in a subset of A, $B \subset A$, it must be expressed by another value, $m(B)$. Every non empty part A of X, for which $m(A) \neq 0$, is called a *focal element* corresponding to an event believed by the observers. The *belief measure* of such a part A of X is defined by considering all the focal elements implying A:

$$Bel(A) = \sum_{B|B\subseteq A} m(B),$$

that is, the belief of a part A of X is defined by the sum of all parts, B such that A contains B. The *plausibility measure* of A is defined by taking all focal elements related to A:

$$Pl(A) = \sum_{B|B\cap A\neq\emptyset} m(B),$$

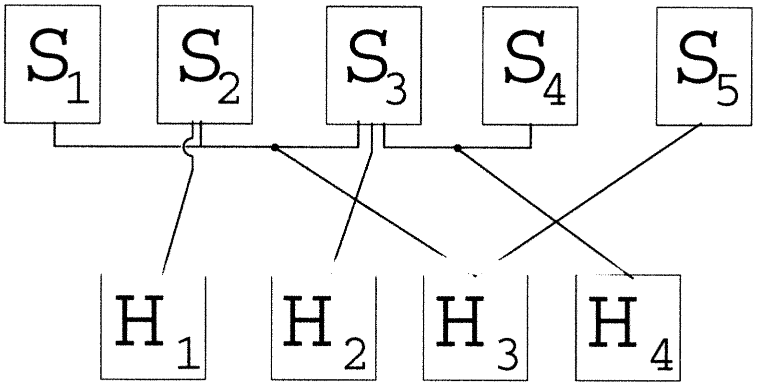
that is, the plausibility of a part A of X is defined by the sum of all parts, B such that B intersects with, or contains any element of, A. The above measures verifies the following relations:

$$Pl(A) = 1 - Bel(\bar{A}) \quad \text{and} \quad Bel(A) \leq Pl(A).$$

where $\bar{A}$ indicates the complement of A in X.

The range $[Bel(A), Pl(A)]$ embeds the imprecise probability, $P(A)$, for any part A of X. A particular case of the mass m is remarkable: consider that all focal elements are singletons of X, that is, beliefs only concern elementary events. Then, every part A of X is such that $Bel(A) = Pl(A)$ and this common value is equal to the probability, $P(A)$.

Calculating Possibilities. Consider the sequence-structure relation in Figure 2. A uniform probability of $\frac{1}{5}$ is assigned to each sequence. Note that the uniform distribution is not a requirement of the mathematical model. It was assumed that all sequences could adopt the same conformation. The possibility for each structural hypothesis to contain the conformation was computed using Zadeh's theory of possibility.



Consider X as the set containing all conformations generated by MC-SYM for all sequence variants. From the sequence-structure relation, R, the focal elements are

$$S_{1,5}, S_2, S_3 \text{ and } S_4,$$

where $S_{1,5} = S_1 \cap S_5$ and S_1, S_2, S_3, S_4, S_5 are subsets of X which contain the three-dimensional conformations associated with the active structure.

From R we have:

$$\begin{aligned} S_{1,5} &= E_{H_3} \\ S_2 &= E_{H_1} \cup E_{H_3} \\ S_3 &= E_{H_2} \cup E_{H_3} \cup E_{H_4} \\ S_4 &= E_{H_4} \end{aligned}$$

where $E_{H_i}, i = 1, 2, 3, 4$, represents the set of conformations that satisfy hypothesis H_i.

A belief coefficient of possibility to contain the conformation is assigned to each focal element:

$$\begin{aligned} m(S_{1,5}) &= \frac{2}{5} \\ m(S_2) &= m(S_3) = m(S_4) = \frac{1}{5}. \end{aligned}$$

The basic probabilities (masses) were assigned by considering that any of the available sequences could adopt the active conformation. The possibility distribution, π, is then

$$\forall x \in X \;\; \pi(x) = 1,$$

which is equivalent in the case of the probability distribution, p, to

$$\forall x \in X \;\; p(x) = \frac{1}{|X|}.$$

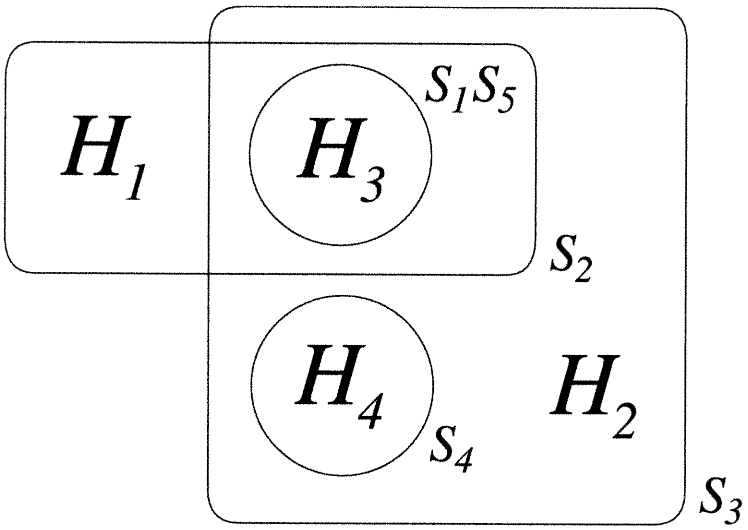
However, given the biological supposition that the active conformation should be found among the structures common to all sequences, the belief coefficients were assigned according to how many sequences are compatible with the hypothesis, that is, for $S_{1,5}$, the belief coefficient of sequences S_1, and S_5,

$$m(S_{1,5}) = \frac{2}{5},$$

based on the fact that two sequences in the set of five sequences were found compatible with a particular subset of the structural hypothesis. Figure 3 shows the focal elements $S_{1,5}, S_2, S_3$andS_4. It is now possible to define the intervals of probabilities (possibilities) for each part of the conformational space.

This situation allows to deduce the intervals of probabilities of each structural hypothesis, that is, the belief measures:

$$\begin{aligned} Bel(E_{H_1}) &= Bel(E_{H_2}) = 0 \\ Bel(E_{H_3}) & \sum_{B|B\subseteq E_{H_3}} m(B) = m(S_{1,5}) = \frac{2}{5} \\ Bel(E_{H_4}) &= \sum_{B|B\subseteq E_{H_4}} m(B) = m(S_4) = \frac{1}{5} \end{aligned}$$



and the plausibility measures:

$$\begin{aligned}
Pl(E_{H_1}) &= \sum_{B|B\cap E_{H_1}\neq\emptyset} m(B) \\
&= m(S_2) = \frac{1}{5} \\
Pl(E_{H_2}) &= \sum_{B|B\cap E_{H_2}\neq\emptyset} m(B) \\
&= m(S_3) = \frac{1}{5} \\
Pl(E_{H_3}) &= \sum_{B|B\cap E_{H_3}\neq\emptyset} m(B) \\
&= m(S_{1,5} + m(S_2) + m(S_3) = \frac{4}{5} \\
Pl(E_{H_4}) &= \sum_{B|B\cap E_{H_4}\neq\emptyset} m(B) \\
&= m(S_3) + m(S_4) = \frac{2}{5}
\end{aligned}$$

which are summarized in Table I. According to the belief and plausibility coefficients calculated for all the hypotheses, H_3, with an imprecise probability over the range $[Bel(E_{H_3}), Pl(E_{H_3})]$, is the one that seems the most likely to be shared by all variant sequences. This does not necessarily indicate, without any doubt, that the actual three dimensional structure will be found in those of H_3. It simply indicates that among the current evaluated hypotheses H_3 is the one that best reflects the combined results of the modeling experiments. With that information in hand, the models generated under H_3 should be carefully examined and used in the design of future laboratory experiments, either to confirm the hypothesis or produce new structural data and hypotheses.

Table I. Belief and Plausibility Measures for H_1, H_2, H_3 and H_5. $Bel(A)$ is the belief value. $Pl(A)$ is the plausibility value.

A	$Bel(A)$	$Pl(A)$
E_{H1}	0	$\frac{1}{5}$
E_{H2}	0	$\frac{1}{5}$
E_{H3}	$\frac{2}{5}$	$\frac{4}{5}$
E_{H5}	$\frac{1}{5}$	$\frac{2}{5}$

Application to the Leadzyme

The Pb^{2+} cleavage of a specific ribophosphodiester bond in yeast $tRNA^{Phe}$ is the classical model of metal-assisted RNA catalysis. *In vitro* selection experiments have identified $tRNA^{Phe}$ variants a derivative of which, named the leadzyme, is very active in cleavage by Pb^{2+} *(11)*. The leadzyme consists of an RNA duplex with an asymmetric internal loop of six nucleotides *(12)*. Cleavage of the leadzyme domain produces two fragments: one with a terminal 5'-hydroxyl group, and the other with a 3'-phosphomonoester presumably generated via a 2',3'-cyclic phosphodiester intermediate. The two-step reaction mechanism of the leadzyme is reminiscent of protein ribonucleases and distinguishes it from other ribozymes, such as the hammerhead, the hairpin and the hepatitis δ domains, which produce 2'-3'-cyclic phosphates *(13, 14, 15, 16)*. The detailed three-dimensional structure was a requisite to the understanding of the particularities of this reaction.

Modeling of the leadzyme was initiated with a series of structural hypotheses derived from the primary and secondary structures. A list of active analogous sequences was previously isolated by *in vitro* selection experiments *(11)*. The program MC-SYM was used to establish the sequence-structure relation, $R \subseteq S \times H$ by generating conformational libraries for the wild-type sequence and all sequence analogs. The fuzzy logic model was applied to these libraries to identify the most plausible hypothesis that was then experimentally evaluated. Activity data of leadzyme variants that incorporated modified nucleotides into the catalytic core *(17)* led to a new structural hypothesis and a second round of computer modeling. The final model is consistent with all available structural data and provided insight into the catalytic reaction of this ribozyme. The details about the active conformation and the three-dimensional modeling of the leadzyme are reported in a manuscript in preparation, available from the authors.

Conclusion

A mathematical model based on fuzzy logic was developed for the selection of structural hypotheses that are more likely to contain active conformations consistent with a series of analogous sequences. The application of this model is especially useful when multiple-sequence data are available, that however do not reveal sufficient structural aspects to initiate three-dimensional modeling. The MC-SYM program or any other RNA modeling approach can be used to produce the sequence-structure relation. The fuzzy logic model was incorporated in the iteration of computer modeling, hypothesis formation and experimental work. This protocol was successfully applied to the three-dimensional modeling of the leadzyme. The fuzzy logic model made possible the identification of a structural hypothesis that was used in the design of laboratory experiments which, in turn, generated structural data that produced a final consistent model.

Literature Cited

1. Major, F.; Gautheret, D. In *Encyclopedia of Molecular Biology and Molecular Medicine*; Myers, R.A., Ed.; VCH Publishers Inc.: NY, 1996, Vol. 5; pp 371–388.

2. Brown, J.; Nolan, J.; Haas, E.; Rubio, M.; Major, F.; Pace, N. *Proc. Natl. Acad. Sci.* **1996**, 93, pp. 3001–3006.

3. Gautheret, D.; Koonings, D.; Gutell, R. *J. Mol. Biol.* **1994**, 242, pp. 1–8.

4. Michel, F.; Westhof, E. *J. Mol. Biol.* **1990**, 216, pp. 585–610.

5. Major, F.; Turcotte, M.; Gautheret, D.; Lapalme, G.; Fillion, E.; Cedergren, R. *Science* **1991**, 253, pp. 1255–1260.

6. Paul, R. P. Robot Manipulators: Mathematics, Programming, and Control; MIT Press: Cambridge, MA, 1981.

7. Bernstein, F. C.; Koetzle, T. F.; Williams, G.J. B.; Meyer, E.F. J.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Eur. J. Biochem.***1977**, 80, pp. 319–324.

8. Berman, H.; Olson, W.; Beveridge, D.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A.; Schneider, B. *Biophys. J.* **1992**, 63, pp. 751–759.

9. Zadeh, L. In Fuzzy sets and systems I; North-Holland Publishing Company: Amsterdam, Holland, 1977, pp. 3–28.

10. Shafer, G. A mathematical theory of evidence; Princeton Univ. Press: Princeton, NJ, 1976.

11. Pan, T.; Uhlenbeck, O. *Biochemistry* **1992**, 31, pp. 3887–3895.

12. Pan, T.; Uhlenbeck, O. *Nature* **1992**, 358, pp. 560–563.

13. Buzayan, J.; Gerlach, W.; Bruening, G. *Proc. Natl. Acad. Sci.* **1986**, 83, pp. 8859–8862.

14. Hutchins, C.; Rathjen, P.; Forster, A.; Symons, R. *Nucl. Acids Res.* **1986**, 14, pp. 3627–3640.

15. Forster, A. C.; Symons, R. H. *Cell* **1987**, 49, pp. 211–220.

16. Epstein, L.; Gall, J. *Cell* **1987**, 48, pp. 535–543.

17. Chartrand, P.; Usman, N.; Cedergren, R. *Biochemistry* **1997**, 36, pp. 3145–3150.

CHAPITRE 4

Modeling Active RNA Structures Using the Intersection of Conformational Space: Application to the Lead-Activated Ribozyme

S. Lemieux, P. Chartrand, R. Cedergren et F. Major. *RNA*, **4**:739–49, 1998.

RNA (1998), *4*:739–749. Cambridge University Press. Printed in the USA.

Modeling active RNA structures using the intersection of conformational space: Application to the lead-activated ribozyme

SÉBASTIEN LEMIEUX,[1]* PASCAL CHARTRAND,[2]* ROBERT CEDERGREN,[2] and FRANÇOIS MAJOR[1]

[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec, Canada H3C 3J7
[2] Département de Biochimie, Université de Montréal, Montréal, Québec, Canada H3C 3J7

ABSTRACT

The Pb^{2+} cleavage of a specific phosphodiester bond in yeast $tRNA^{Phe}$ is the classical model of metal-assisted RNA catalysis. In vitro selection experiments have identified a $tRNA^{Phe}$ variant, the leadzyme, that is very active in cleavage by Pb^{2+}. We present here a three-dimensional modeling protocol that was used to propose a structure for this ribozyme, and is based on the computation of the intersection of conformational space of sequence variants and the use of chemical modification data. Sequence and secondary structure data were used in a first round of computer modeling that allowed identification of conformations compatible with all known leadzyme variants. Common conformations were then tested experimentally by evaluating the activity of analogues containing modified nucleotides in the catalytic core. These experiments led to a new structural hypothesis that was tested in a second round of computer modeling. The resulting proposal for the active conformation of the leadzyme is consistent with all known structural data. The final model suggests an in-line SN2 attack mechanism and predicts two Pb^{2+} binding sites. The protocol presented here is generally applicable in modeling RNAs whenever the catalytic or binding activity of structural analogues is known.

Keywords: base triple; catalytic RNA; internal loop; MC-SYM; metalo-nucleotide complex; molecular modeling; noncanonical base pairs

INTRODUCTION

The leadzyme is a catalytic RNA that cleaves a specific ribophosphodiester bond in the presence of Pb^{2+} (Pan & Uhlenbeck, 1992b). It was originally isolated by in vitro selection of molecules undergoing cleavage from partially randomized sequence libraries related to the sequence of yeast $tRNA^{Phe}$ (Pan & Uhlenbeck, 1992a). The secondary structure of the leadzyme, shown in Figure 1, consists of two helical domains sandwiching an asymmetric internal loop of six nucleotides. Cleavage is effected at the phosphodiester bond between C1 and G2 (see Fig. 1A). Rather than producing 2′-3′-cyclic phosphates via a transphosphorylation/cleavage reaction as other catalytic RNA domains (Buzayan et al., 1986; Hutchins et al., 1986; Epstein & Gall, 1987; Forster & Symons, 1987), the leadzyme also performs a second hydrolytic reaction of the cyclic phosphate intermediate reminiscent of protein ribonucleases. Thus, the final products possess either a 5′-hydroxyl group or a 3′-phosphomonoester terminus (Pan & Uhlenbeck, 1992b). Although some NMR structural data are available on the conformation of this RNA domain, its detailed three-dimensional structure, a requisite to the understanding of the details of this reaction, is unknown.

Over the past years, we have been developing a computer-based modeling protocol for RNA consisting of first translating primary, secondary, and tertiary structure data into geometrical constraints, then applying a constraint satisfaction solver, MC-SYM (Major et al., 1991), to generate all-atom three-dimensional models, and finally, refining the preliminary models with molecular mechanics energy minimization (Nilsson & Karplus,

Reprint requests to: François Major, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec, Canada H3C 3J7; e-mail: major@iro.umontreal.ca.

*The first two authors contributed equally to the work.

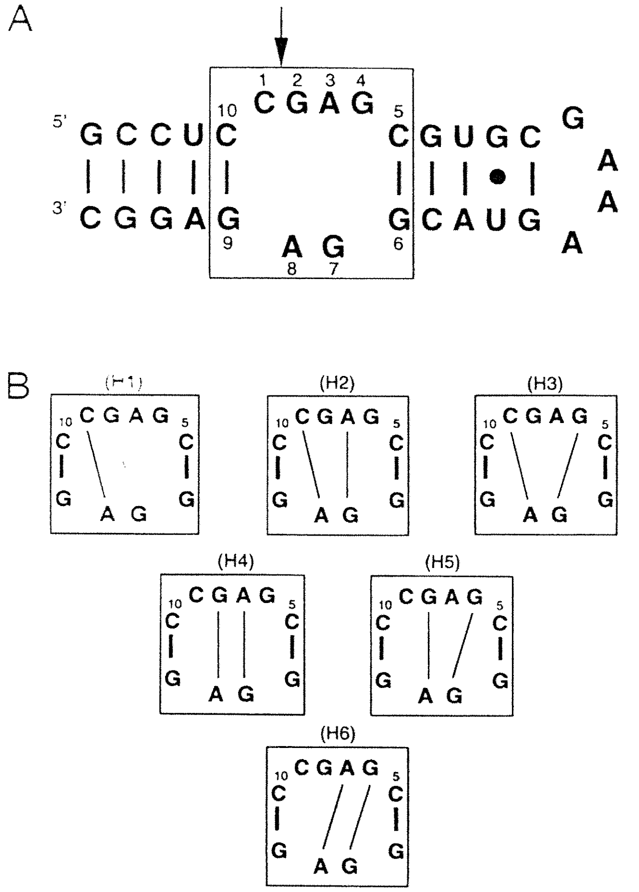


FIGURE 1. A: Primary and secondary structure of the leadzyme. Thick lines indicate Watson–Crick base pairs and bullets indicate non-Watson–Crick base pairs. Arrow indicates the cleavage site between C1 and G2. B: Six structural hypotheses that maximize base pairing in the internal hexaloop. Thin lines indicate non-Watson–Crick base pairs.

1986; Weiner et al., 1986). The small size of the leadzyme rendered it an ideal subject for our procedure. Moreover, the abundance of activity data on structural analogues of the leadzyme provided us with the challenge of including these data in an automatic protocol.

During the modeling of the Rev protein binding site of HIV-1, the ability of a collection of RNA molecules (aptamers) to engage in a common bimolecular interaction was taken to mean that the aptamers share many aspects of their three-dimensional structures (Leclerc et al., 1994). In particular, the geometry of base–base interactions at the binding site should be conserved among aptamers in spite of sequence variation. In the case of the leadzyme, many structural analogues have been isolated and their catalytic activity has been determined. Catalytic activity may be a more powerful reflection of the three-dimensional structure of the molecule than the binding ability of above because analogues have to mimic all conformations along the reaction pathway as well as the energy levels. Therefore, the active conformation of the leadzyme should be found among the conformations common to the leadzyme and its active analogues. We present here the principle of the intersection of conformational space (ICS) of catalytically active molecules, which allows the use of activity data. ICS consists of the generation of a conformational library of each active analogue, and the selection of common conformers from each library. A preliminary three-dimensional model of the leadzyme was produced by ICS, and subsequently refined by experimental data and further modeling (see Table 1). The final structure is consistent with all structural and activity data and suggests a plausible reaction mechanism.

TABLE 1. Overview of the modelling strategy.[a]

Data available	Modelling steps
	Initial models
Primary and secondary structure	1. Formation of six structural hypotheses 2. Generation of conformations for the wild-type sequence using MC-SYM 3. Classification
	Intersection of Conformational Space
In vitro selection	4. Threading of the variant sequences 5. Assignment of probabilities using fuzzy set theory
	Final models
Chemical modification	6. Formation of new structural hypotheses 7. Generation of conformations for the wild-type sequence using MC-SYM 8. Refinement using molecular mechanics energy minimization 9. Insertion of the lead ions 10. Refinement using molecular mechanics energy minimization

[a] Initial models for the wild-type sequence were generated from primary and secondary structure information. Then, the initial models were scanned for common conformations among the wild-type and other active sequences derived from in vitro selection data. Finally, chemical modification experiments were used for verification of the best initial model and generation of a new structural hypothesis. Molecular mechanics energy minimization was then used to obtain the final model.

RESULTS

Initial models

The internal loop of the leadzyme is composed of six nucleotides, four on one strand and two on the other (see Fig. 1A). No firm structural constraint is known with the exception of a pairing between C1 and A8 at pH 6.5 in the absence of lead ions, as suggested by NMR spectroscopy (Legault, 1995; Legault & Pardi, 1994, 1997). To derive the first set of structural hypotheses, a maximum of base pairing and base stacking was assumed because known RNA structures and basic knowledge of RNA thermodynamics suggest that base stacking and base pairing predominate in the stabilization of internal loops (Tinoco et al., 1987; Varani et al., 1989; Wimberly et al., 1993; Cai & Tinoco, 1996). The 4 × 2 asymmetry of the leadzyme internal loop allows for a maximum of two cross-strand base pairs and six ways in which these two base pairs could be distributed (see Fig. 1B). Each of these combinations was evaluated by defining six structural hypotheses: Hypothesis 1 (H1) proposes the pairs C1·A8 and G2·G7. Hypothesis 2 (H2) to hypothesis 6 (H6) involve, respectively, C1·A8 and A3·G7, C1·A8 and G4·G7, G2·A8 and A3·G7, G2·A8 and G4·G9, and A3·A8 and G4·G9.

The conformational library for each hypothesis was generated by the combinatorial assembly of every hydrogen bonding pattern of each base pair. Here, we evaluated only hydrogen bonding involving at least two hydrogen bonds as described by Saenger (1984). Consideration of one hydrogen bond interaction is possible (see below), but would greatly expand the number of structures that would be generated. Thus, a G·G base pair can be constructed using four different hydrogen bonding patterns: GG(III) [Roman numerals refer to the Saenger system of base pair conformations (Saenger, 1984)]: GG(IV), GG(VI), and GG(VII); the A·G base pair by four hydrogen bonding patterns: AG(VIII), AG(IX), AG(X), and AG(XI); the A·A base by three hydrogen bonding patterns: AA(I), AA(II), and AA(V); and the A·C base pair by two hydrogen bonding patterns: AC(XXV) and AC(XXVI).

In addition, conformational patterns involving protonated bases or less than two hydrogen bonds were also implemented for the case of the A·C interaction (see Fig. 2). Arabic numerals refer to this class of interactions, which include: the pseudo-wobble AC(75) (Rould et al., 1989); the major tautomeric forms of the bases AC(71) and AC(78) (Hutchins et al., 1986); and other patterns generated mathematically, AC(68, 69, 70, 72, 73, 74, 76, 77, 79, 80). These latter patterns were evaluated by MC-SYM, even though they have never been observed experimentally. The protonated A·C base pair of type AC(75) suggested by NMR data (Legault & Pardi, 1994) was considered in hypotheses H2 and H3. Using the wild-type sequence, 60 different base pair patterns for H1, H2, and H3, 16 patterns for H4 and H5, and 12 patterns for H6 could be con-

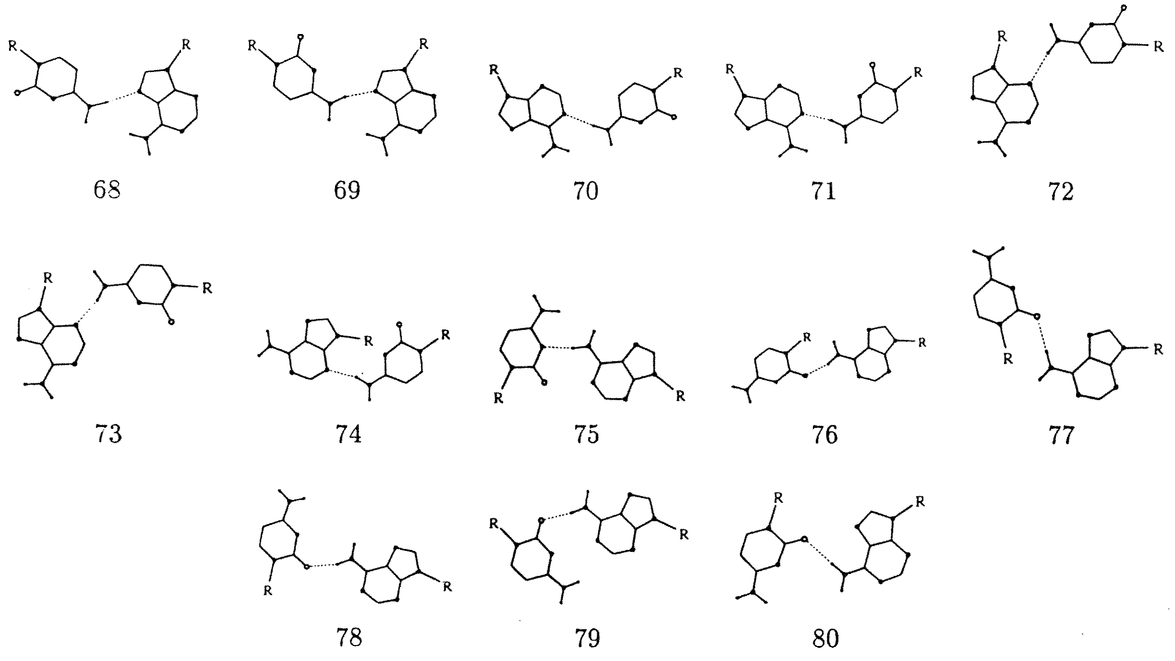


FIGURE 2. A·C base pairing patterns implemented in MC-SYM to test the C1·A8 base pair. All patterns involve a single hydrogen bond, except when protonation occurs, then two hydrogen bonds can be formed.

structed. Additional conformational enrichment was created by consideration of the sugar pucker and glycosyl bond torsion angles. Nucleotides were assigned either the C2′- or C3′-*endo* sugar puckers and either the *anti* or *syn* glycosyl bond torsion angles. Based on existing structures, all nucleotides in double-helical stems including the flanking nucleotides were assigned the A-RNA helix conformation.

The data sets corresponding to the six hypotheses were submitted to MC-SYM. Figure 3 shows the order in which individual nucleotides were added to the nascent models in the step-by-step construction procedure. For each hypothesis of Figure 1B, the base pairing pattern and nucleotide conformation combinatorics defined conformational landscapes of more than 5×10^{11} different conformations. Each of these conformations was evaluated for the wild-type sequence. At each step in the construction, the MC-SYM constraint satisfaction solver pruned conformations that violated basic stereochemical rules (see Major et al., 1993). MC-SYM found nearly 15,000 conformations that were compatible with structural hypotheses H2, H3, H4, and H5, but no structures could be constructed for H1 nor H6. To simplify the subsequent structural analysis, the conformers for each hypothesis were classified into 19 classes based on the base pairing patterns (see Table 2). Classes had up to 15 subclasses based on sugar puckers and glycosidic angles.

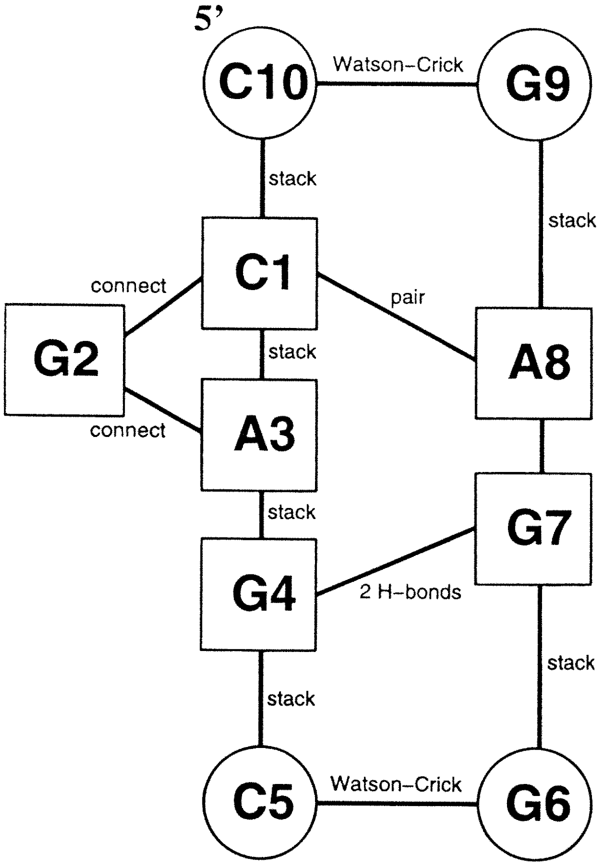


FIGURE 3. Graph of relations to model the leadzyme under H3. Nucleotides are represented by vertices and their spatial relations by edges. Circled nucleotides were assigned C3′-*endo* sugar puckers and *anti* glycosyl torsion (A-RNA helix type). Boxed nucleotides were assigned either the C2′- or C3′-*endo* sugar puckers and either the *anti* or *syn* glycosyl bond torsion angles (sampled). The interpretation of this figure is as follows. The base of C1 stacks on the base of C10. The base of A3 stacks on the base of C1. G2 bulges outside the helix. The base of G4 stacks on the base of A3. The base of C5 stacks on the base of G4. G6 forms a Watson–Crick base pair with C5. G7 forms a base pair involving at least two hydrogen bonds with G4 and stacks on the base of G6. A8 forms a base pair with C1. G9 forms a Watson–Crick base pair with C10 and stacks on the base of A8.

TABLE 2. Summary of MC-SYM simulations for the wild-type leadzyme sequence.[a]

Hypothesis	Class	Subclasses (#)	Structures (%)
H2	AC(69)AG(XI)	8	12.9
	AC(75)AG(XI)	2	6.3
	AC(77)AG(XI)	2	3.8
	AC(78)AG(XI)	4	10.4
	AC(79)AG(XI)	8	4.4
	AC(69)AG(VIII)	9	10.4
	AC(71)AG(VIII)	15	51.7
H3	AC(69)GG(VI)	2	14.0
	AC(71)GG(VI)	2	86.0
H3′	CG(XIX)GG(110)	1	58.3
	CG(XXII)GG(VI)	1	41.7
H4	AG(VIII)AG(VIII)	7	7.2
	AG(VIII)AG(IX)	2	0.4
	AG(IX)AG(VIII)	7	2.4
	AG(IX)AG(XI)	3	3.0
	AG(X)AG(XI)	1	6.8
	AG(XI)AG(XI)	6	80.2
H5	AG(VII)GG(VI)	1	66.7
	AG(IX)GG(VI)	1	33.3

[a] Conformations are defined by hydrogen bonding patterns. Subclasses are based on sugar pucker modes and base orientation at the glycosyl bond. The percentage of models refers to the number of models containing the indicated hydrogen bonds relative to the total number of models in that class.

Computing the ICS of the leadzyme

In theory, sets of conformations should be computed for each active analogue to determine the ICS. In this case, the analogues were identified from in vitro selection (Pan & Uhlenbeck, 1992a), but generally any molecule having the same activity could be considered. A computationally more efficient technique, however, consists of trying to thread the analogue sequences through the conformations found for the wild-type sequence of the leadzyme. The results of this exercise are shown in Figure 4.

The wild-type sequence is the only sequence that generated conformations under H2, H3, H4, and H5 (Fig. 4). The set of hypotheses associated with this molecular modeling experiment is $WT = \{H2, H3, H4, H5\}$ and $m(WT) = 1/9$ (the wild type is one out of nine

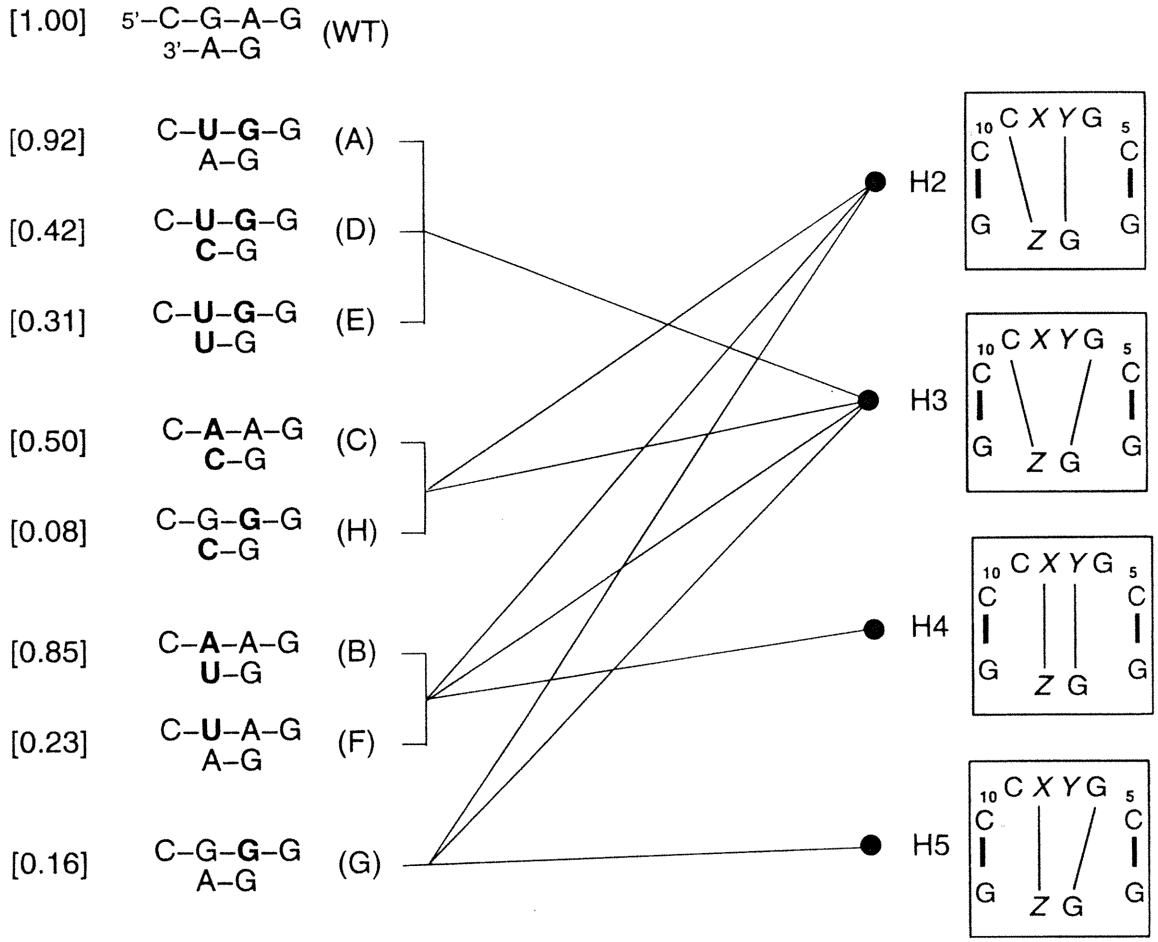


FIGURE 4. Network of folding compatibilities. Numbers in brackets indicate relative activity (Pan et al., 1994). Variant identifiers are in parentheses. Edges connect sequences to the structural hypotheses for which MC-SYM has found consistent conformations. Wild-type sequence can be folded into structural hypotheses H2, H3, H4, and H5. Only H3 accommodates all active analogue sequences.

variants, see Materials and Methods). Variants *A*, *D*, and *E* fold only according to H3, *ADE* = {H3}; variants *B* and *F* fold to H2, H3, and H4, *BF* = {H2, H3, H4}; variants *C* and *H* fold to H2 and H3, *CH* = {H2, H3}; and variant *G* folds to H2, H3, and H5, *G* = {H2, H3, H5}. These folding patterns imply the following basic probabilities: m(*ADE*) = 1/3; m(*BF*) = m(*CH*) = 2/9; and m(*G*) = 1/9, and thus, the likelihood for {H2}, {H3}, {H4}, and {H5} are [0, 2/3], [1/3, 1], [0, 1/3], and [0, 2/9], respectively (see Materials and Methods). These probabilities indicate that H3 has the highest degree of belief, and, thus, is most likely to contain active conformations of the leadzyme. The plausibility of 1 indicates that H3 is the only hypothesis that can accommodate all active variants (see Materials and Methods).

Two major conformational classes are found for H3 (Table 2): 14% feature the AC(69) and GG(VI) pairing patterns and 86% contain the AC(71) and the GG(VI) pairing patterns. The sugar pucker and the glycoside bond rotation further divide these classes into minor subclasses. The RMSD between any pair of models in a given subclass is less than 2.0 Å. Both major classes contain the GG(VI) pairing pattern and allow for the isosteric substitution of C1·A8 by C1·C8 or C1·U8 found among the analogue sequences. Note that the AC(69) and AC(71) base pairing patterns are present among the possible active conformations, but not the AC(75), the protonated pair suggested by the NMR data (Legault & Pardi, 1994). This structure was found only in the conformations compatible with H2.

A new structural hypothesis

During the course of these modeling studies, the activity of different leadzymes having a modification at virtually every functional group in the asymmetric internal loop became available (Chartrand et al., 1997). Functional groups whose modification decreased the catalytic activity of the ribozyme by at least a factor of one order of magnitude when compared to the wild-type leadzyme are shown in Figure 5.

The major inconsistency between the preliminary model and the modified leadzyme data concerned the C1·A8 base pair. Although the functional groups of C1

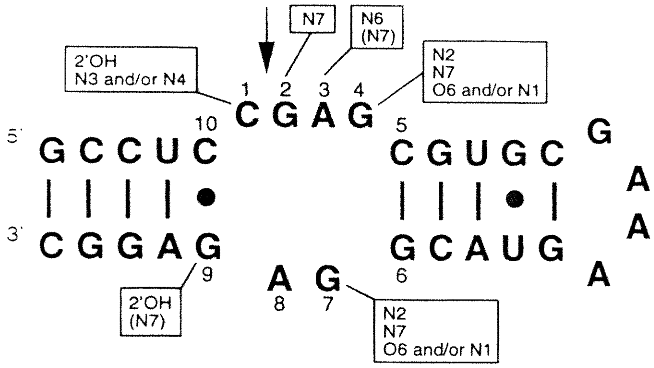


FIGURE 5. Summary of the functional groups found important for the catalytic activity of the leadzyme. Deletion of these functional groups leads to a decrease in activity by more than one order of magnitude compared with the wild-type. Those in parentheses reduce the activity by a factor less than 10×.

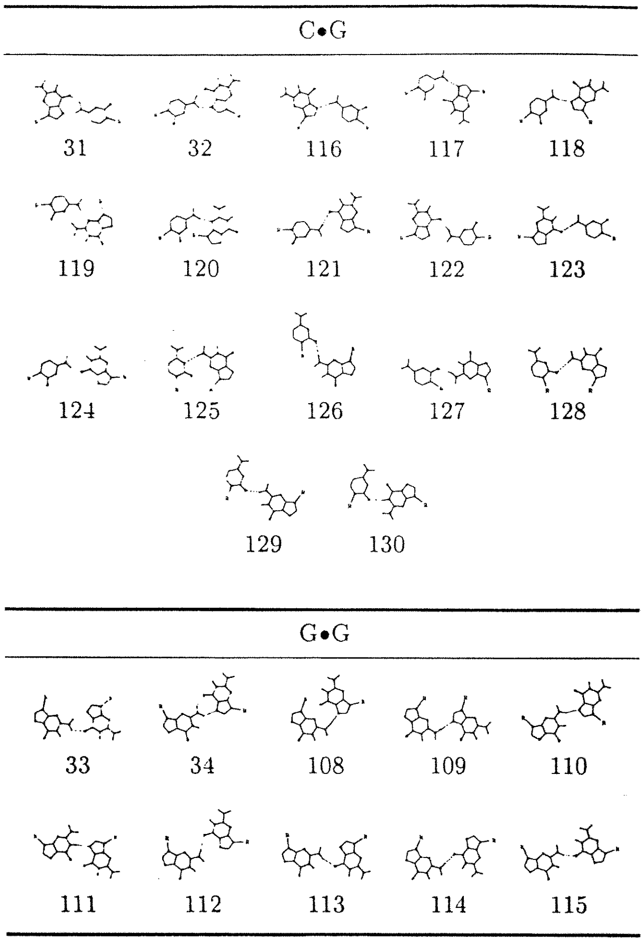


FIGURE 6. C·G and G·G base pairing patterns involving one hydrogen bond, as implemented in MC-SYM to test the C1·G7 and G4·G7 base pairs. The C·G base pairs were assigned canonical Watson–Crick patterns. The G·G base pair was assigned two hydrogen bond pairing patterns only.

were found to be crucial in the modified leadzyme data, A8 was clearly not involved in any interaction (Chartrand et al., 1997). This inconsistency raised the possibility that C1 could be involved in an interaction with G7 because type VI base pairing of G4·G7 does not implicate the functional groups of G7 involved in a Watson–Crick base pair. This possibility gave rise to H3'. a new structural hypothesis not considered in the initial modeling process. H3' is derived from H3, but contains the base triple C1·G7·G4. Given the invariability of C1, G7, and G4, the C1·G7·G4 combination is the only possible triple under H3.

The conformational space of this triple is defined by 19 base pairing patterns for the C1·G7 base pair (Watson–Crick, reverse Watson–Crick, and 17 patterns involving one hydrogen bond; see Fig. 6), and 14 base pairing patterns for the G7·G4 base pair (the three G·G base pairing patterns described in the initial model and 11 patterns involving one hydrogen bond; see Fig. 6). There are 266 theoretical conformations of this base triple, however, only 61 C·G·G triples are sterically sound and, among these, only 11 are possible in the context of the leadzyme hexaloop.

Because stereochemical defects are inherent in models derived from the systematic use of discreet nucleotide conformation sets, molecular mechanics refinement (see Materials and Methods) was necessary to optimize the geometry of the models. Models that were not able to maintain their base pairing patterns during optimization were eliminated from further consideration, leaving only two models that were successfully optimized (see Table 2). One of these two models featured the CG(XIX) and GG(110) base pairing patterns, and was therefore discarded because many of the functional groups found to be important from the modified nucleotide data, i.e., the significant A3:N^6, G7:N^7, G4:N^1, G4:O^6, G4:N^2, and G2:N^7, were not involved in stabilizing interactions and A3:N^6 was positioned far from the active site. On the other hand, the model featuring the triple composed of the CG(XXII) (reverse Watson–Crick) and GG(VI) base pairs satisfied all available structural data.

DISCUSSION

A conformation consistent with catalytic activity

The conformation proposed by this procedure, as shown in Figure 7, satisfies many criteria that could be expected of an active conformation. In particular, the atom C1:O$^{2'}$, the presumed nucleophile, is aligned with the G2:P and G2:O$^{5'}$, as required of an in-line cleavage mechanism. A potential Pb^{2+} binding site is found adjacent to the cleavage site and is bounded by the base triple on one side and the backbone atoms of C1 and G2 on the other (see Fig. 7A). The Pb^{2+} cofactor is presumed to be bound in proximity to the catalytic site to promote the nucleophilic character of the C1:$O^{2'}$. In this position, the electronegative binding pocket is defined by A3:N^6, A3:N^7, G2:P, C1:$O^{2'}$, G4:O^6, and G4:N^7, all of which have been shown to be important for catalytic activity and could contribute to the affinity of the cation at this site.

The in-line mechanism suggested by this structure, although likely, differs from the adjacent mechanism suggested for the yeast $tRNA^{Phe}$, which involves pseudorotation at the scissile phosphate (Brown et al., 1985). Current experimental data do not permit a distinction between these two mechanisms. The model also suggests that the catalytic lead cation could be involved in the stabilization of the transition-state trigonal bipyramid conformation of the phosphate group. The distance between the Pb-O^1P is less than 4 Å compared with the much longer distance observed in the lead binding pocket of $tRNA^{Phe}$ (Brown et al., 1985). Our model correctly predicts the specificity of cleavage because the C1:$O^{2'}$H is the only 2′-OH within the catalytic pocket (see Fig. 7B).

A second lead binding site?

The loss in free energy of transition-state stabilization, ΔG, due to the deletion of G7:N^2, is 3.2 kcal/mol, a value similar to that observed at G12:N^2 in the magnesium ion binding pocket of the hammerhead ribozyme (Scott et al., 1995). In the leadzyme model, G7:N^2 is involved in only one hydrogen bond with C1:N^3 and could therefore be available for ion binding. Also, modifications in this region reduced the cooperative binding of lead cations, whereas modifications in the catalytic site had no effect (Chartrand et al., 1997).

We thus propose a second cationic binding pocket involving G7:N^2, G9:$O^{2'}$, and C1:N^4. This binding could explain the finding that conversion of the G9–C10 base pair into a C9–G10 base pair or removing the G9:$O^{2'}$H led to an important loss of activity. A standard Watson–Crick base pair positions the G9:$O^{2'}$H toward the adenine base in position A3, whereas a reverse Watson–Crick base pair, as shown in Figure 7B, positions it below the plane of the G9 base, increasing its exposure to the Pb^{2+}, in the presumed binding pocket. This region contains several phosphodiester bonds that would be good potential coordination sites for the cation as well (Fig. 7).

To permit energy minimization of the metalo-nucleotide complex, Pb^{2+} ions were parameterized as indicated in Materials and Methods. The energy-minimized model is shown in Figure 7C. During minimization, the guanine base at position 2 moved slightly back toward the outside of the catalytic core, creating a better pocket for Pb^{2+} than in the initial model. However, the most striking aspect of this optimized structure is how the phosphodiester chain folds back on itself in the G9, A8, and G7 region. This folded-back conformation is due to a reverse Watson–Crick at the G9–C10 pair and the bulged-out nucleotide A8. Undoubtedly, this chain reversal would normally be unlikely due to the close juxtaposition of the charged phosphates in the looped back region; however, in the leadzyme structure, these charges could be shielded very effectively by the Pb^{2+} in the second binding pocket. The finding that an equimolar amount of neodymium, Nd^{3+}, and Pb^{2+} increases the catalytic activity of the leadzyme (Ohmichi & Sugimoto, 1997) could be due to the fact that only large ions provide sufficient shielding of the phosphate ions at this site. In the context of our two metal-ion binding site model, Nd^{3+} would bind in the structural site and Pb^{2+} in the catalytic site, because Nd^{3+} is inactive alone. Higher Nd^{3+}/Pb^{2+} ratios reduce the activity because Nd^{3+} ions may compete for the catalytic binding site.

No data are currently available on the functional groups involved in the formation of the catalytic site of the second step of the leadzyme reaction, the cyclic phosphate hydrolysis. However, according to our model, the propensity of the leadzyme to catalyze the hydrolysis of the cyclic phosphate could be due to: (1) the fact that the 5′ leaving group is an unpaired nucleotide and thus less likely to be in position for a reversal of the reaction; (2) the availability of water due to the state of hydration of the duodecavalent Pb^{2+} is much greater than Mg^{2+}, the usual metallic catalyst of ribozyme reaction; and, (3) the active conformation could exist long enough for a second chemical step.

Applicability of the ICS method

Structural modeling using the ICS approach, like any modeling project, is an attempt to reconcile all data into a three-dimensional form. This technique should not be confused with structural determination, because modeling necessarily depends on the quantity and, above all, quality of the data used during the process. In this sense, computer modeling with qualitative data can be thought of as a low-resolution data transformation process. Nevertheless, these models, even those of low resolution, offer the experimentalist a composite view of a variety of experimental results in an easy to understand form, thereby facilitating the design of more definitive experiments (Cedergren & Major, 1998).

In the present case, additional assumptions such as that concerning the conformation of the helical regions (the A-RNA helix conformation) and maximizing base pairing in the internal loop region of the molecule have been used. Basic knowledge of RNA thermodynamics suggests that base stacking and base pairing predominate in the stabilization of internal loops (Tinoco et al., 1987; Varani et al., 1989; Wimberly et al., 1993; Cai & Tinoco, 1996). Also, it is assumed that the nucleotide conformation at each position is present in the conformational set used in the preliminary model. However, energy minimizations used to construct refined models

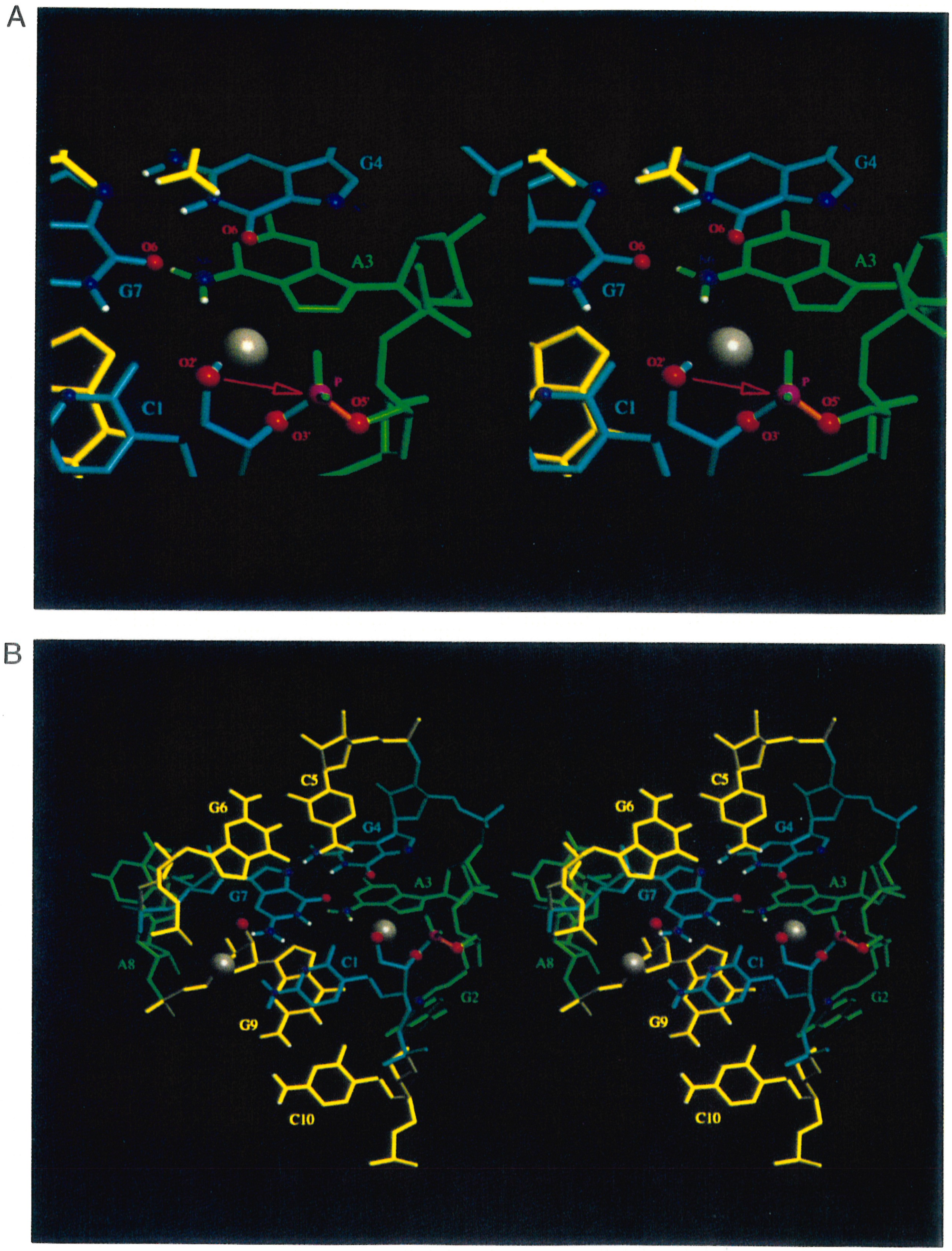


FIGURE 7. (*Figure continues on facing page.*)

limit the effect of this assumption. These types of assumptions are more practical than fundamental because they are used primarily to speed up the algorithm, and in some cases could be done away with entirely. It is important to understand that the ICS technique that we developed is in itself not dependent on these assumptions and could be applied to cases where no such assumptions were made.

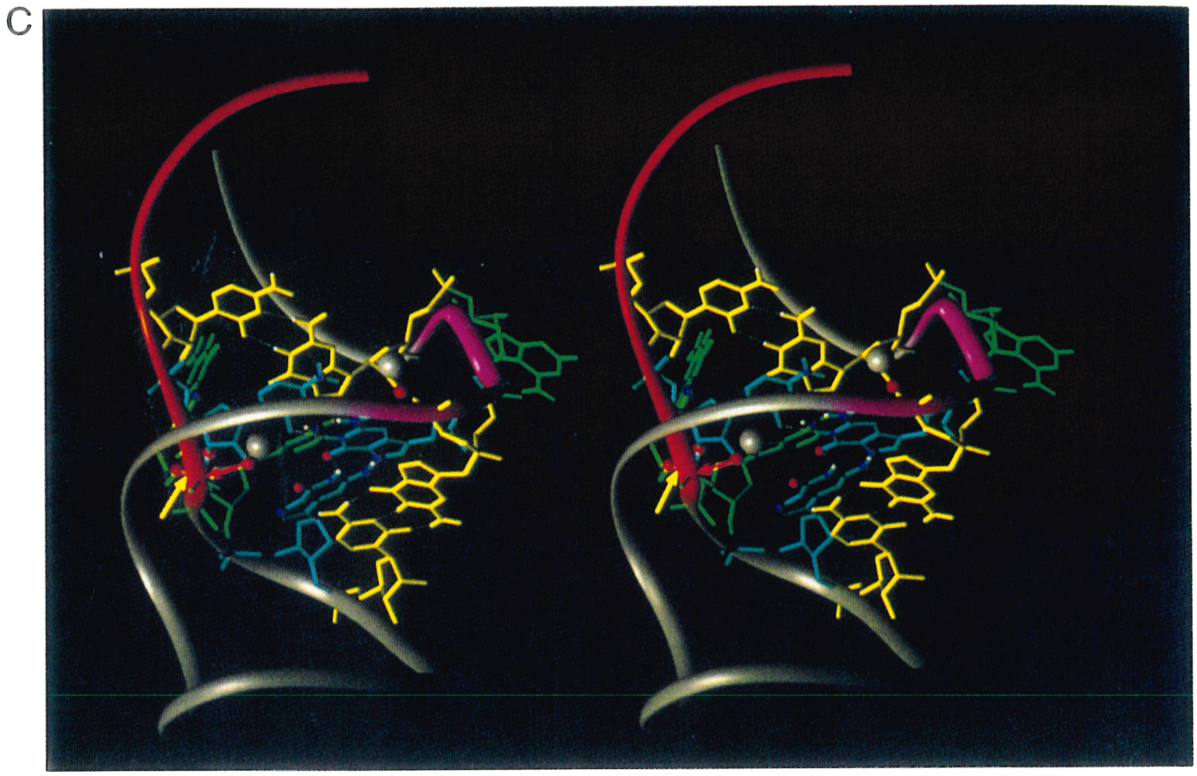


FIGURE 7. The proposed model. Nucleotides of the C·G·G base triple are shown in cyan. The C1:$O^{2'}$H, the P between C1 and G2 and the G2:$O^{5'}$ are aligned, suggesting the in-line attack indicated by the red arrow. The P-$O^{5'}$ scissile bond is shown in orange. The lead ions are shown in gray. The flanking G9·C10 and C5-G6 base pairs are shown in yellow. G2, A3, and A8 are shown in green. Atoms determined to be important for the catalytic activity are drawn with CPK surfaces using blue for nitrogen, red for oxygen, magenta for phosphate, and white for hydrogen atoms. Atoms with smaller CPK surfaces were not tested individually with modification experiments (see Fig. 4). **A:** In-line attack. A metal binding pocket can be observed whose internal surface is made up of many functional groups that were found important in modification experiments. **B:** Stereo view of the active site. **C:** Stereo view of the leadzyme complex with lead. The backbone ribbon of the nucleotides from the leadzyme 5′ end to C1 is shown in red. Near the second lead ion binding pocket, the phosphodiester chain folds back on itself as indicated by the magenta ribbon of nt 6–9. The backbone ribbon of all other nucleotides is shown in gray. Hydrogen bonds are shown as white dashed lines.

A model being dependent on available data is subject to change, as more structural data become available. A key element of the present work is that we were able to propose a single conformation that responds to a certain number of criteria expected of a catalytic molecule and is consistent with all available data on the leadzyme except for the proposed AC base pair, as is the modified nucleoside data as well. We believe that the ICS method should produce a truer picture of the active conformation of the leadzyme as the number of analogues increases. Even more precise structural data would be necessary to establish the nature of the structural interrelationships among the functional groups found important in the chemical modification data that are not involved directly in interactions.

MATERIALS AND METHODS

Formalism for the ICS approach

If we assume that each active sequence variant has an equal probability of adopting any given conformation, then probabilities can be assigned to structural hypotheses, X, in the following way: $\mathrm{m}(X) = 0$, if none of the variants are compatible with the models in X, and $\mathrm{m}(X) = 1$, if all variants are compatible with at least one model of the hypothesis in X. The likelihood of a single structural hypothesis, $\{x\}$, is given by an interval of probabilities, $[\mathrm{Bel}(\{x\}), \mathrm{Pl}(\{x\})]$, where $\mathrm{Bel}(\{x\}) = \Sigma_{Y \subseteq \{x\}} \mathrm{m}(Y)$ and $\mathrm{Pl}(\{x\}) = \Sigma_{Y \cap \{x\}} \mathrm{m}(Y)$ (Zadeh, 1983). The belief of x, $\mathrm{Bel}(\{x\})$, corresponds to the sum of the basic probabilities for the sets that contain exactly x, equal to $\{x\}$, and the plausibility of x, $\mathrm{Pl}(\{x\})$, corresponds to the sum of the basic probabilities of the sets that include x (Major et al., 1998).

Molecular mechanics energy minimization

MC-SYM structures were refined using molecular mechanics calculations performed by the molecular simulation program sander, from the Amber 4.1 suite of programs (Pearlman et al., 1995) using the Amber 94 forcefield. All 1–4 electrostatic interactions were reduced by a 1.2 factor as suggested for the 94 Amber forcefield. A distance-dependent dielectric model, $\epsilon = 4R_{ij}$, for the Coulombic representation of electrostatic interactions was used, as suggested by Weiner et al.

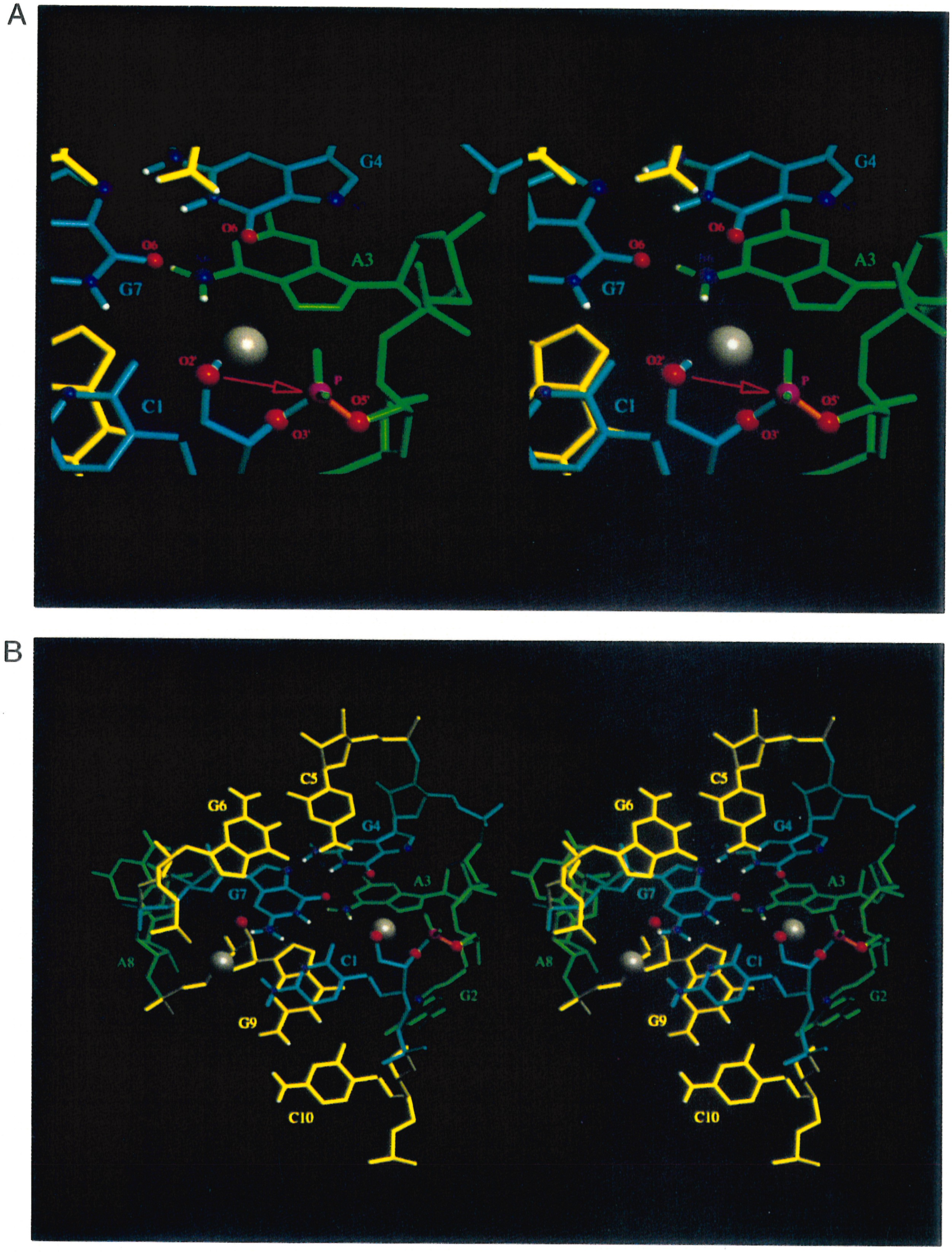


FIGURE 7. (*Figure continues on facing page.*)

limit the effect of this assumption. These types of assumptions are more practical than fundamental because they are used primarily to speed up the algorithm, and in some cases could be done away with entirely. It is important to understand that the ICS technique that we developed is in itself not dependent on these assumptions and could be applied to cases where no such assumptions were made.

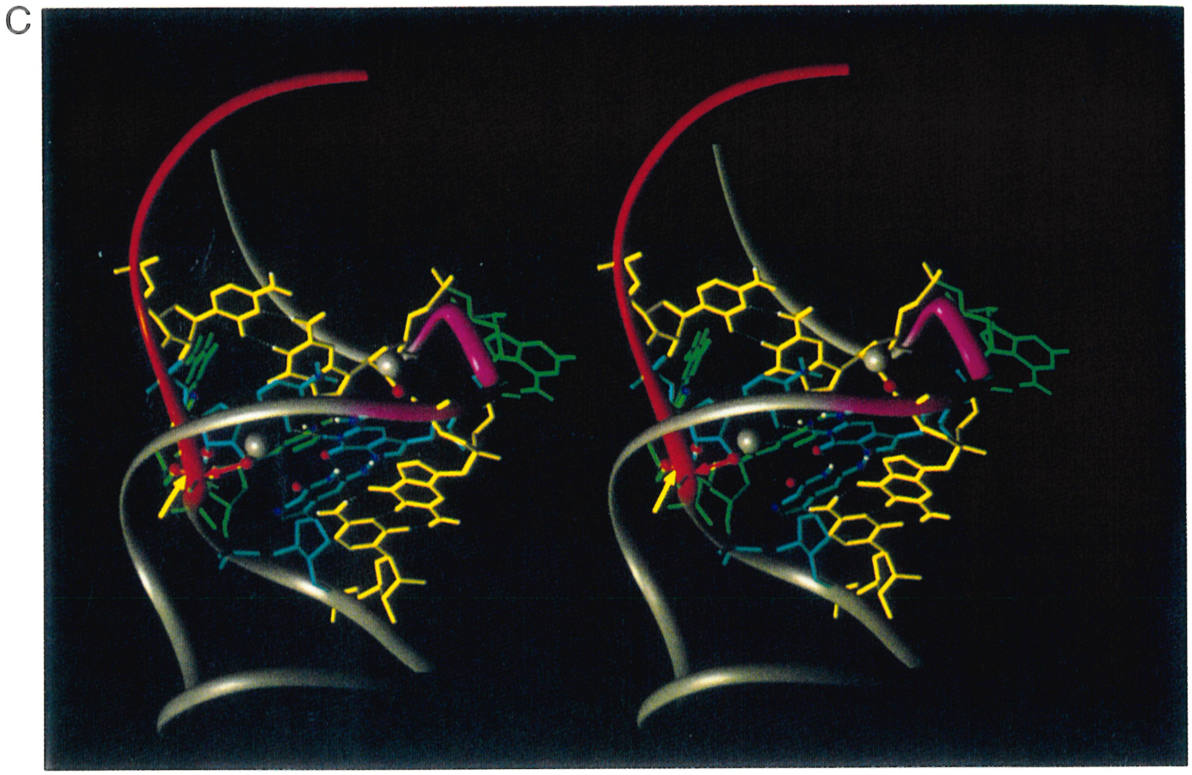


FIGURE 7. The proposed model. Nucleotides of the C·G·G base triple are shown in cyan. The C1:$O^{2'}$H, the P between C1 and G2 and the G2:$O^{5'}$ are aligned, suggesting the in-line attack indicated by the red arrow. The P-$O^{5'}$ scissile bond is shown in orange. The lead ions are shown in gray. The flanking G9·C10 and C5-G6 base pairs are shown in yellow. G2, A3, and A8 are shown in green. Atoms determined to be important for the catalytic activity are drawn with CPK surfaces using blue for nitrogen, red for oxygen, magenta for phosphate, and white for hydrogen atoms. Atoms with smaller CPK surfaces were not tested individually with modification experiments (see Fig. 4). **A:** In-line attack. A metal binding pocket can be observed whose internal surface is made up of many functional groups that were found important in modification experiments. **B:** Stereo view of the active site. **C:** Stereo view of the leadzyme complex with lead. The backbone ribbon of the nucleotides from the leadzyme 5′ end to C1 is shown in red. Near the second lead ion binding pocket, the phosphodiester chain folds back on itself as indicated by the magenta ribbon of nt 6–9. The backbone ribbon of all other nucleotides is shown in gray. Hydrogen bonds are shown as white dashed lines.

A model being dependent on available data is subject to change, as more structural data become available. A key element of the present work is that we were able to propose a single conformation that responds to a certain number of criteria expected of a catalytic molecule and is consistent with all available data on the leadzyme except for the proposed AC base pair, as is the modified nucleoside data as well. We believe that the ICS method should produce a truer picture of the active conformation of the leadzyme as the number of analogues increases. Even more precise structural data would be necessary to establish the nature of the structural interrelationships among the functional groups found important in the chemical modification data that are not involved directly in interactions.

MATERIALS AND METHODS

Formalism for the ICS approach

If we assume that each active sequence variant has an equal probability of adopting any given conformation, then probabilities can be assigned to structural hypotheses, X, in the following way: $\mathrm{m}(X) = 0$, if none of the variants are compatible with the models in X, and $\mathrm{m}(X) = 1$, if all variants are compatible with at least one model of the hypothesis in X. The likelihood of a single structural hypothesis, $\{x\}$, is given by an interval of probabilities, $[\mathrm{Bel}(\{x\}), \mathrm{Pl}(\{x\})]$, where $\mathrm{Bel}(\{x\}) = \Sigma_{Y \subseteq \{x\}} \mathrm{m}(Y)$ and $\mathrm{Pl}(\{x\}) = \Sigma_{Y \cap \{x\}} \mathrm{m}(Y)$ (Zadeh, 1983). The belief of x, $\mathrm{Bel}(\{x\})$, corresponds to the sum of the basic probabilities for the sets that contain exactly x, equal to $\{x\}$, and the plausibility of x, $\mathrm{Pl}(\{x\})$, corresponds to the sum of the basic probabilities of the sets that include x (Major et al., 1998).

Molecular mechanics energy minimization

MC-SYM structures were refined using molecular mechanics calculations performed by the molecular simulation program sander, from the Amber 4.1 suite of programs (Pearlman et al., 1995) using the Amber 94 forcefield. All 1–4 electrostatic interactions were reduced by a 1.2 factor as suggested for the 94 Amber forcefield. A distance-dependent dielectric model, $\epsilon = 4R_{ij}$, for the Coulombic representation of electrostatic interactions was used, as suggested by Weiner et al.

(1984; 1986). As a first step, energy minimization has been performed using the steepest descent for 50 steps, then the conjugate gradient method was applied until the maximum derivative was less than 0.01 kcal/mol/Å. The alignment of the C1:O$^{2'}$ with G2:P and G2:O$^{5'}$, and the base pairs assigned by MC-SYM were restrained during the first 5,000 steps. No cutoff was used during the minimization.

Parameterization of the lead ion

The charge parameter was fixed at 2.0. The van der Waals parameters were initially assigned a basic radius of 1.4 Å, corresponding to the Pauling radii. The potential well depth parameter, *e*, was fixed at 0.15 kcal/mol, corresponding to a slightly higher potential than that of the phosphate atom.

Molecular mechanics energy minimization with lead ions

Two Pb^{2+} were added to the leadzyme structure. The first was placed between atoms G4:O^6 and C1:C$^{1'}$. The second was inserted in the region bounded by the positions of atoms G6:P, G7:N^2, C1:N^4, and G9:O$^{2'}$. Both sites correspond to a high electronegativity surface (not shown). Molecular mechanics calculations were performed as described above except that the Amber 94 forcefield was modified to include the Pb^{2-} parameters. For the first 5,000 steps, the cations were restrained to their starting positions. The conjugate gradient method was applied until the maximum derivative was less than 0.1 kcal/mol/Å. Then, the restraints were removed and the minimization was continued until the maximum derivative was less than 0.01 kcal/mol/Å. The calculated empirical potential energy of the minimized lead structure was −168.6 kcal/mol. The atomic coordinates of the minimized model can be found at: http://www-lbit.iro.umontreal.ca/structures/leadzyme.pdb.

GAAA loop

The GAAA tetranucleotide loop of the leadzyme was modeled from the structural features of the solution structure of the GNRA class of tetranucleotide loops (Heus & Pardi, 1991). The MC-SYM script encoding the structural constraints defined for the GAAA loop describes a search tree of 28,946 nodes from which only eight models were consistent. The maximum RMSD (excluding H atoms) among the eight models was approximately 0.9 Å. One of the GAAA models found by MC-SYM was appended to complete the leadzyme structure.

ACKNOWLEDGMENTS

We thank Dr. Marielle Foucrault for preliminary work on the initial model, Majid Ftouhi for discussions on the mathematical formalism of the ICS technique, Stanislaw Oldziej for his help with the parameterization of the lead ion, Dr. Daniel Gautheret for providing a program to compute the non-Watson–Crick base pairing geometries, and the referees for useful and constructive comments. P.C. was supported by a predoctoral fellowship from the Medical Research Council (MRC) of Canada. R.C. is Richard Ivey Fellow of the Canadian Institute for Advanced Research programme in evolutionary biology. F.M. is a fellow of the MRC of Canada. This work was supported by the MRC of Canada.

Received August 12, 1997; returned for revision October 2, 1997; revised manuscript received April 10, 1998

REFERENCES

Brown R, Dewan J, Klug A. 1985. Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of yeast phenylalanine tRNA. *Biochemistry 24*:4785–4801.

Buzayan J, Gerlach W, Bruening G. 1986. Satellite tobacco ringspot virus RNA: A subset of the RNA sequence is sufficient for autolytic processing. *Proc Natl Acad Sci USA 83*:8859–8862.

Cai Z, Tinoco IJ. 1996. Solution structure of loop a from the hairpin ribozyme from tobacco ringspot virus satellite. *Biochemistry 35*: 6026–6036.

Cedergren R, Major F. 1998. Modeling the tertiary structure of RNA. In: Simons RW, Grunberg-Manago M, eds. *RNA structure and function.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 37–75.

Chartrand P, Usman N, Cedergren R. 1997. The effect of structural modifications on the activity of the leadzyme. *Biochemistry 36*: 3145–3150.

Epstein L, Gall J. 1987. Self-cleaving transcripts of satellite DNA from the newt. *Cell 48*:535–543.

Forster AC, Symons RH. 1987. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell 49*:211–220.

Heus H, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science 253*:191–194.

Hutchins C, Rathjen P, Forster A, Symons R. 1986. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. *Nucleic Acids Res 14*:3627–3640.

Leclerc F, Cedergren R, Ellington A. 1994. A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nature Struct Biol 1*:293–300.

Legault P. 1995. Structural studies of ribozymes by heteronuclear NMR spectroscopy [thesis]. Boulder, Colorado: University of Colorado.

Legault P, Pardi A. 1994. In situ probing of adenine protonation in RNA by 13C NMR. *J Am Chem Soc 116*:8390–8391.

Legault P, Pardi A. 1997. Unusual dynamics and pK_a shift at the active site of a lead-dependent ribozyme. *J Am Chem Soc 119*: 6621–6628.

Major F, Gautheret D, Cedergren R. 1993. Reproducing the three-dimensional structure of a transfer RNA molecule from structural constraints. *Proc Natl Acad Sci USA 90*:9408–9412.

Major F, Lemieux S, Ftouhi A. 1998. Computer RNA three-dimensional modeling from low-resolution data and multiple-sequence information. In: Leontis NB, SantaLucia J Jr, eds. *Molecular modeling of nucleic acids.* Washington DC: American Chemical Society. ACS Symposium Series 682. pp 394–404.

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

Nilsson L, Karplus M. 1986. Empirical energy functions for energy minimization and dynamics of nucleic acids. *J Comp Chem 7*:591–616.

Ohmichi T, Sugimoto N. 1997. Role of Nd^{3+} and Pb^{2+} on the {RNA} cleavage reaction by a small ribozyme. *Biochemistry 36*:3514–3521.

Pan T, Dichtl B, Uhlenbeck O. 1994. Properties of an in vitro selected Pb^{2+} cleavage motif. *Biochemistry 33*:9561–9565.

Pan T, Uhlenbeck O. 1992a. In vitro selection of RNAs that undergo autolytic cleavage with Pb^{2+}. *Biochemistry 31*:3887–3895.

Pan T, Uhlenbeck O. 1992b. A small metalloribozyme with a two-step mechanism. *Nature 358*:560–563.

Pearlman D, Case D, Caldwell J, Ross W, Cheatham T, Ferguson D, Seibel G, Singh U, Weiner P, Kollman P. 1995. *Amber 4.1.* San Francisco, California: University of California.

Rould M, Perona J, Soll D, Steitz T. 1989. Structure of *E. coli* glutaminyl-tRNA synthetase complexed with tRNA^{Gln} and ATP at 2.8 Å resolution. *Science 246*:1135–1142.

Saenger W. 1984. *Principles of nucleic acid structure*. New York: Springer-Verlag.

Scott W, Finch J, Klug A. 1995. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell 81*:991–1002.

Tinoco IJ, Davis P, Hardin C, Puglisi J, Walker G, Wyatt J. 1987. RNA structure from A to Z. *Cold Spring Harbor Symp Quant Biol 52*:135–146.

Varani G, Wimberly B, Tinoco IJ. 1989. Conformation and dynamics of an RNA internal loop. *Biochemistry 28*:7760–7772.

Weiner S, Kollman P, Nguyen D, Case D. 1986. An all atom force field for simulations of proteins and nucleic acids. *J Comput Chem 7*:230–252.

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. 1984. A new forcefield for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc 106*:765–784.

Wimberly B, Varani G, Tinoco IJ. 1993. The conformation of loop E of eukaryotic ribosomal RNA. *Biochemistry 32*:1078–1087.

Zadeh LA. 1983. Commonsense knowledge representation based on fuzzy logic. *Computer 16*:61–65.

CHAPITRE 5

Function of hexameric RNA in packaging of bacteriophage ϕ29 DNA In Vitro

F. Zhang, S. Lemieux, X. Wu, D. St-Arnaud, C.T. McMurray, F. Major et D. Anderson. *Molecular Cell*, **2**: 141–7, 1998.

Molecular Cell, Vol. 2, 141–147, July, 1998, Copyright ©1998 by Cell Press

Function of Hexameric RNA in Packaging of Bacteriophage ϕ29 DNA In Vitro

Feng Zhang,* Sébastien Lemieux,‡
Xiling Wu,§ Daniel St.-Arnaud,‡
Cynthia T. McMurray,§ François Major,‡
and Dwight Anderson†‖
*Department of Genetics and Cell Biology
†Departments of Microbiology and Oral Science
University of Minnesota
Minneapolis, Minnesota 55455
‡Departement d'Informatique et de Recherche
Opérationnelle
Université de Montréal
Montréal, Quebec
Canada H3C 3J7
§Department of Pharmacology
Mayo Clinic and Foundation
Rochester, Minnesota 55905

Summary

A cyclic hexamer of the 120-base prohead RNA (pRNA) is needed for efficient in vitro packaging of the *B. subtilis* bacteriophage ϕ29 genome. This capacity of pRNA to form higher multimers by intermolecular base pairing of identical subunits represents a new RNA structural motif. Dimers of pRNA are likely intermediates in formation of the cyclic hexamer. A three-dimensional model of the pRNA hexamer is presented.

Introduction

RNA pseudoknots are intramolecular tertiary interactions involving hairpin or interior loops that create quasi-continuous double-helical stem regions. Pseudoknots facilitate specific folding of RNA molecules, confer recognition in reactions catalyzed or mediated by RNA, and provide sites for interaction with proteins (Wyatt and Tinoco, 1993). For example, biological function in translation, including frameshifting during the translation of retroviral mRNAs and autoregulation of translation of the phage T4 gene 32 mRNA by gene 32 protein, requires pseudoknots. Consensus structures of RNA ligands to HIV reverse transcriptase and nerve growth factor generated by systematic evolution of ligands by exponential enrichment are pseudoknotted (Gold et al., 1993). Additionally, "kissing" interaction between loop regions of nonidentical molecules is a structural motif contributing to regulation of replication, conjugation, and gene expression; for example, RNA I interacts with the nascent RNA II transcript to regulate plasmid ColE1 replication (Tomizawa, 1993).

Intermolecular pseudoknots that link identical copies of RNA to form higher multimers have not been described previously. Here, we demonstrate the formation of hexamers by intermolecular base pairing of 120-base prohead RNA (pRNA) encoded by the *Bacillus subtilis* bacteriophage ϕ29. pRNA is essential for in vitro packaging of the ϕ29 DNA–gene product 3 (DNA–gp3) complex (Guo et al., 1987a). Multiple copies of pRNA are bound to the portal vertex (head–tail connector) of the precursor capsid (prohead), the site of DNA packaging (Guo et al., 1987b; Reid et al., 1994a; Trottier and Guo, 1997). pRNA is thought to mediate docking of supercoiled DNA–gp3 with the prohead, orient packaging, and unite with gene product 16 to form the DNA translocating ATPase (Grimes and Anderson, 1989b, 1990, 1997; Anderson and Reilly, 1993). The 120-base form of the 174-base pRNA transcript forms dimers and hexamers that are detected by mutagenesis, native polyacrylamide gel electrophoresis, and analytical ultracentrifugation. The capacity of mutant RNA(s) to form cyclic hexamers is a requirement in the reconstitution of RNA-free proheads for efficient DNA–gp3 packaging in a defined in vitro system.

Results and Discussion

Requirement for Intermolecular Base Pairing of pRNA in DNA Packaging

The secondary structure of the 120-base form of pRNA has a high helical content and three loops (Bailey et al., 1990; Figure 1a). pRNAs of ϕ29 and related phages have limited sequence similarity yet fold into a conserved secondary structure that is important for function (Bailey et al., 1990; Figure 1b). Within the folded ϕ29 pRNA molecule, the CE loop and D loop structures (Bailey et al., 1990) contain regions G and G′, respectively, which are complementary (45–48 AACC and 82–85 GGUU) and are capable of intramolecular pseudoknot formation (Reid et al., 1994c) or intermolecular self-association. To test whether interaction of the G and G′ regions is required in DNA–gp3 packaging, mutant pRNAs with altered bases at positions 45–48 AACC to GCGA (mutant F6), 82–85 GGUU to UCGC (mutant F7), or the double mutant 45–48 AACC to GCGA and 82–85 GGUU to UCGC (mutant F6/F7) (Figure 2a) were combined with RNA-free proheads and the DNA packaging activities measured in vitro (Reid et al., 1994c; Figure 2b). Base changes in F6 and F7 abrogate the ability for intramolecular base pairings between the CE and D loop regions. Base changes in F6/F7 are compensatory mutations that restore the potential for intramolecular base pairings and pseudoknot formation. The G and G′ regions are indeed critical for function, since neither mutant pRNA F6 nor F7 can mediate DNA–gp3 packaging (Figure 2b, lanes 4 and 5). However, the F6/F7 double mutant that restores the potential for pseudoknot formation is as active as wild-type pRNA in DNA–gp3 packaging (Figure 2b, lane 6). These data confirm that interactions between the G and G′ loop regions are essential for DNA–gp3 packaging but do not distinguish whether the interactions are intra- or intermolecular.

Efficient DNA–gp3 packaging by a mixture of RNAs F6 and F7 provided the crucial evidence for intermolecular interactions of the G and G′ regions (Figure 2b, lane

‖To whom correspondence should be addressed.

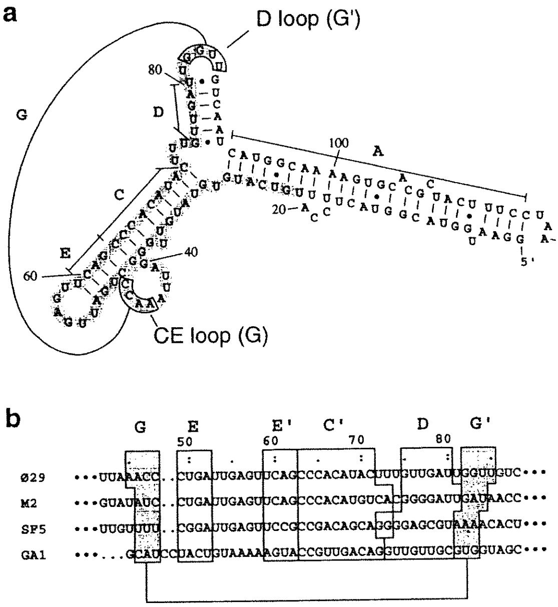

a
D loop (G')
CE loop (G)
G
D
C
E
A
80
100
20
40
60
120
5'
b
G E E' C' D G'
50 60 70 80
Ø29 •••UUAAACC..CUGAUUGAGUUCAGCCCACAUACUUUGUUGAUUGGUUGUC•••
M2 •••GUAUAUC..CUGAUUGAGUUCAGCCCACAUGUCACGGGGAUUGAUAACC•••
SF5 •••UUGUUUU..CGGAUUGAGUUCCGCCGACAGCAGGGGAGCGUAAAACACU•••
GA1 •••...GCAUCCUACUGUAAAAAGUACCGUUGACAGGUUGUUGCGUGGUAGC•••

Figure 1. Secondary Structure and Sequence Alignment of pRNAs

(a) Secondary structure of the 120-base form of wild-type pRNA (Bailey et al., 1990; Reid et al., 1994c). The line shows the proposed tertiary interaction. Helices are designated A, C, D, E, and G. The prohead binding domain (Reid et al., 1994a) is marked by shading.

(b) Phylogenetic sequence alignment of ϕ29 and related phage pRNAs that includes the sequence of domain I for each RNA (Bailey et al., 1990; Reid et al., 1994c). Helical regions are boxed and designated by capital letters. The tertiary interaction is shaded and designated by G and G'.

7). Indeed, F6-F7 multimers were observed by native polyacrylamide gel electrophoresis at room temperature (Figure 2c, lane 4). The individual pRNAs F6 and F7 migrated faster than the F6+F7 mixture, the double mutant F6/F7, and the wild-type pRNA (Figure 2c, lanes 1–5). In mixtures of pRNAs F6 and F7 at 2:1 or 1:2 (Figure 2c, lanes 6 and 7, respectively) about 2/3 of the RNA ran at the slower position. Thus, the multimeric slower form in the F6+F7 mixture contained equal amounts of RNAs F6 and F7. Further, the wild-type and F6/F7 RNAs showed a single multimer at room temperature. In Tris-borate-EDTA gels, all of the pRNAs ran at the faster position, demonstrating that formation of pRNA multimers required Mg^{2+} (data not shown).

Function of pRNA Multimers in DNA Packaging

The base pairing necessary for multimerization of pRNA F6/F7 and the mixture F6+F7 is rationalized in Table 1. Both the wild-type and the F6/F7 pRNAs can form cyclic (closed) multimers with any number of subunits, whereas F6+F7 can only form cyclic multimers with an even number of subunits because noncanonical pairings prevent multimerization of odd numbers of subunits. To define the stoichiometry of the F6+F7 mixture that is required for efficient DNA–gp3 packaging, RNA-free proheads were reconstituted with different ratios of the pRNAs. In these and all reconstitution experiments to follow, 12 copies of pRNA were used per prohead. Proheads incubated first with F6 (six copies) followed by F7 (six copies), or vice versa, were fully active in DNA–gp3 packaging (Figure 2d, lanes 4 and 5). Mixing F6 and F7 in a 1:1 ratio was optimal for biological activity, while an excess of either F6 or F7 resulted in a proportional decrease in DNA–gp3 packaging (data not shown). Also, a 1:1 mixture of F6 and F7 was the most effective competitor of prohead binding of wild-type pRNA (data not shown). Thus, F6 and F7 interact to form an oligomer with an even number of subunits.

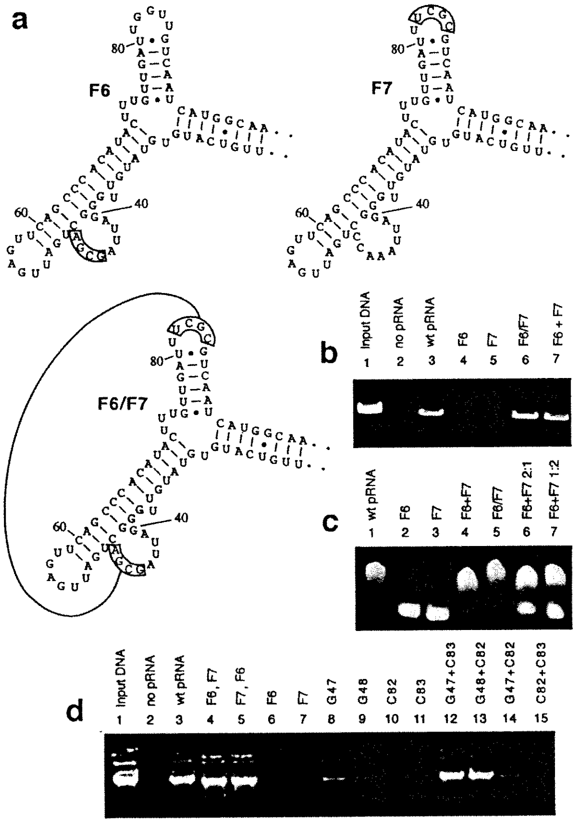


Figure 2. Mutagenesis, Gel Electrophoresis, and DNA Packaging to Test and Characterize the G Helix Interaction

(a) Sequence changes of mutants F6, F7, and F6/F7 are boxed, and the line shows the tertiary interaction.

(b) In vitro DNA–gp3 packaging activity of RNA-free proheads reconstituted with wild-type or mutant pRNAs. Lane 1 shows the DNA–gp3 added to each reaction (Input DNA). Lanes 2–7 show DNA–gp3 packaged by proheads reconstituted with the designated pRNA(s). Unpackaged DNA was digested with DNase; then, the packaged, DNase-resistant DNA–gp3 was extracted from filled heads and quantified following agarose gel electrophoresis.

(c) Native polyacrylamide gel electrophoresis of wild-type (wt) or mutant pRNA(s) at room temperature. The lanes contained 0.9 μg total of the designated pRNA(s).

(d) In vitro DNA–gp3 packaging activity of proheads reconstituted with wild-type or mutant pRNAs. Lane 1 shows the DNA–gp3 added to each reaction. Lanes 2–15 show DNA–gp3 packaged by proheads reconstituted with the designated pRNA(s). F6,F7 and F7,F6 refer to F6 followed by F7 or vice versa.

Analysis of the DNA–gp3 packaging activities of mutant pRNAs defined the base pairs required for multimerization. Bases in either the G or G′ regions (see Figure 1a) were altered as follows: A45 was changed to G or C (mutants G45 and C45); A46 to G or C (mutants G46 and C46); C47 to G, A, or U (mutants G47, A47, and

Table 1. Capabilities of pRNAs to Form Dimers and Cyclic Multimers

pRNA	Pseudoknot Residues	Dimer (Y/N)	Cyclic Multimers	(Y/N)
Wild-type	CC-GG			
	CC-GG	Y		
	CC-GG		3	Y
	CC-GG		4	Y
	CC-GG		5	Y
	CC-GG		6	Y
F6	GA-GG			
	GA-GG	N		
F7	CC-CU			
	CC-CU	N		
F6/F7	GA-CU			
	GA-CU	Y		
	GA-CU		3	Y
	GA-CU		4	Y
	GA-CU		5	Y
	GA-CU		6	Y
F6+F7	GA-GG			
	CC-CU	Y		
	GA-GG		3	N
	CC-CU		4	Y
	GA-GG		5	N
	CC-CU		6	Y

Pseudoknot residues are 47,48—83,82, and changes from wild-type are underlined. F6 and F7 cannot form dimers because of noncanonical pairings and mismatches, respectively. F6/F7 behaves as wild-type pRNA, while F6+F7 can form cyclic multimers with an even number of subunits. F6 and F7 were inactive in DNA–gp3 packaging, while F6/F7 and F6+F7 had the full activity of wild-type pRNA (Figure 2b).

U47); C48 to G, A, or U (mutants G48, A48, and U48); G82 to A, U, or C (mutants A82, U82, and C82); G83 to A or C (mutants A83 and C83); U84 to G (mutant G84); and U85 to G (mutant G85). All pRNAs changed at positions 45, 46, 84, or 85 were active in DNA–gp3 packaging (data not shown). Mutants G47, G48, C82, and C83 had little or no DNA–gp3 packaging activity (Figure 2d, lanes 8–11), demonstrating that base pairing between residues 47–83 and 48–82 was necessary and sufficient for biological activity. The requirement for these two base pair interactions in DNA packaging was also demonstrated by in vitro selection of pRNA aptamers for prohead binding (Zhang and Anderson, 1998).

Mutant complementations demonstrated that dimers are likely intermediates in oligomerization of pRNA (Figure 2d, lanes 12–15). Mutant pairs G47+C83 and G48+C82 that can form dimers and multimers with an even number of subunits gave efficient DNA–gp3 packaging. The mutant pair G47+C82 can form dimers, but the dimers cannot multimerize, and the packaging activity was only l0% that of the wild-type. The mutant pair C82+C83 cannot dimerize, and DNA–gp3 packaging was not detected.

All pRNAs that could not form dimers because of base pairing mismatches were inactive in DNA–gp3 packaging (Table 2A). When a noncanonical base pairing was needed for dimerization (Table 2B) or dimers could not interact to form higher multimers (Table 2C), DNA–gp3 packaging activities were about 10% that of wild-type. Dimer interactions that include a noncanonical base pairing had DNA–gp3 packaging activities ranging from 20% to 60% that of wild-type (Table 2D), whereas canonical base pairing yielded full DNA–gp3 packaging (Table 2E). Since dimers interact to form an oligomer with an even number of subunits, three-way complementations were performed to determine the precise number of interacting subunits (Table 2F). The three mutant pRNAs in each complementation could form dimers, and the dimers could interact to form a cyclic hexamer. DNA–gp3 packaging in these complementations approached or equaled that of wild-type pRNA. The complementations suggest that the pRNA multimers contain six molecules, because multimers of 4 or 7 molecules would require noncanonical base pairing, and multimers of 5 or 8 molecules are excluded by double mismatches. Thus, pRNA dimers can support inefficient DNA packaging, but linkage of dimers to form a hexamer is needed for full packaging activity.

Direct Demonstration of the pRNA Hexamer in Solution

Physical proof of hexamer formation was obtained by analytical ultracentrifugation (Figure 3). At equilibrium, each pRNA species forms a concentration gradient that depends on molecular mass (Figure 3a). For the wild-type, the hexamer was the major form in solution (Figure 3b; each monomer has a mass of 40 kDa). The hexamer was observed at 4k rpm over a 13-fold concentration range, and a dimer intermediate was observed at higher speeds or at low concentration. Trimers, tetramers, or pentamers were not detected as major forms in solution. Additionally, no monomer was detected (at 12k rpm) in the wild-type, and the best fit to the data yielded a dimer with a slightly higher molecular mass than predicted due to contribution from the hexamer (that influences the steeper part of the concentration gradient). These

Table 2. Requirements for pRNA Multimerization in DNA Packaging

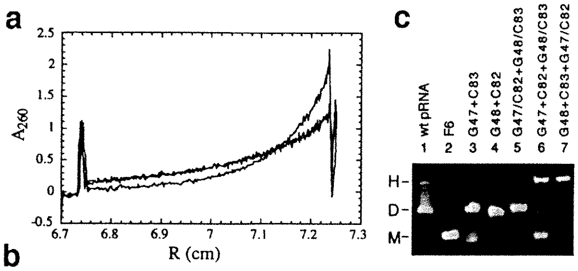
pRNA(s)	Pseudoknot Residues	DNA–gp3 Packaging (%)
(A) C82	CC-GC CC-GC	Not detected
C83	CC-CG CC-CG	
(B) G47	GC-GG GC-GG	10
G48	GC-GG CG-GG	10
(C) C82+G48/C83	CC-GC CG-CG	5
C83+G47/C82	CC-CG GC-GC	15
(D) G48/C83+G47	CG-CG GC-GG	20
G47/C82+G48	GC-GC CG-GG	60
(E) G47+C83	GC-GG CC-CG	100
G48+C82	CG-GG CC-GC	100
(F) G47/A83+U47/C82+ G48/C83	GC-AG UC-GC CG-CG	100
G48/A83+U48/C83+ G47/C82	CG-GA CU-CG GC-GC	100

Pseudoknot residues are 47,48—83,82, and changes from wild-type are underlined. The DNA-gp3 packaging data, expressed as % of wild-type, are from a representative experiment. (A), Dimers cannot form; (B), Dimerization includes a noncanonical base pairing; (C), Dimers cannot interact to form higher multimers; (D), Dimer interactions to form higher multimers include noncanonical base pairing; (E), Dimers interact to form an oligomer with an even number of subunits; (F), Three-way complementations produce cyclic hexamers by interaction of three dimers and exclude oligomers of 4, 5, 7, or 8 subunits. Other mutants and combinations giving similar results for categories (A)–(F) were: (A) G48/C83, G47/C82, U47/C82, and U48/C83; (B) G47/A83 and G48/A82; (C) G47+C82, G48+C83, G47/A83+U47/C82, and G48/A82+U48/C83; (D) U47/C82+G48/C83, G48/C83+G47/A83, U48/C83+G47/C82, and G47/C82+G48/A82; and (F) G47+C82+G48/C83 and G48+C83+G47/C82.

results suggest that the dimer was the only major intermediate in formation of the hexamer.

In contrast to the wild-type, mutants F6 and F7, which could not support DNA–gp3 packaging (Figures 2b and 2d), were monomers in solution (Figure 3b). Absence of the hexamer in F6 and F7 was evident from the concentration gradients that lacked steep slopes at the bottom of the cell where the higher molecular weight forms were in equilibrium (Figure 3a). In contrast to F6 and F7, the F6/F7 double mutant and the F6+F7 mixture yielded forms that were similar to wild-type, predominantly hexamers with dimer intermediates (Figure 3b). As with wild-type, curve fitting did not exclude the presence of trimer, tetramer, and pentamer; however, if present, they were not major intermediates.

The sedimentation data aided interpretation of native polyacrylamide gel electrophoresis of pRNA at 4°C (Figure 3c). F6 (lane 2) marked the position of monomers, and the mutant combinations G47+C83 and G48+C82 (Table 2E; lanes 3 and 4) and G47/C82+G48/C83 (lane 5) produced dimers. Two three-way complementations that package DNA–gp3 efficiently (Table 2; lanes 6 and 7) showed monomers and a form larger than dimers (hexamers) that was also observed in wild-type. Thus, hexamers are likely formed from dimers, consistent with the DNA–gp3 packaging results (Table 2; Figures 2b and 2d). The amount of hexamer observed for biologically active pRNAs varies; for example, G47-C83 (Figure 3c, lane 3) provides less stability than the wild-type C47-G83, and G48-C82 (Figure 3c, lane 6) more stability than the wild-type C48-G82.



b

pRNA	concentration (μM)	4,000 rpm ^{a} (kD)	8,000 rpm ^{a} (kD)	12,000 rpm ^{a} (kD)
wt	0.08	226 ± 46	89 ± 13	92 ± 4
	0.4	227 ± 16	238 ± 17 / 80 ± 6	104 ± 2
	1	221 ± 13	234 ± 33 / 77 ± 4	107 ± 1
F6	0.4	-- ^{b}	52 ± 2	49 ± 1
F7	0.4	-- ^{b}	55 ± 2	51± 1
F6/F7	0.4	222 ± 19	252 ± 21 / 83 ± 4	96 ± 1
F6+F7	0.4	238 ± 22	225 ± 37 / 82 ± 4	91 ± 1

Figure 3. Molecular Mass of the Wild-Type and Mutant pRNAs in Solution Determined by Analytical Ultracentrifugation

(a) A representative concentration gradient formed from the wild-type (steep slope at the bottom of the cell) and the F6 and F7 mutants (superimposed curves) at 8,000 rpm, 4°C, and a sample concentration of 0.4 μM. A_{260} is the absorbance at 260 nm; R (cm) is the distance in centimeters from the solution/air meniscus toward the bottom of the cell.

(b) Results of the best fit to the data. The 4k, 8k, and 12k rpm speeds represent optimal equilibrium conditions for the hexamer, dimer and monomer, respectively; superscript (*a*), each concentration gradient was fit to both a single and a double component model. The best fit to the data is listed. In all cases, the data fit well to a single component or a double component but not both; superscript (*b*) (---) indicates that the data could not be fit. The F6 + F7 mixture contained equal concentrations (0.2 μM) of each RNA to obtain a final concentration of 0.4 μM.

(c) Native polyacrylamide gel electrophoresis of wild-type (wt) or mutant pRNA(s) at 4°C. The lanes contained 0.45 μg total of the designated pRNA(s). The fast, intermediate, and slow forms are hypothesized to be monomers (M), dimers (D), and hexamers (H), respectively.

Three-Dimensional Modeling of the pRNA Hexamer

Three-dimensional models of the pRNA hexamer were built by use of the RNA computer modeling program MC-SYM (Major et al., 1991). The structural constraints to define an MC-SYM script for the pRNA hexamer were

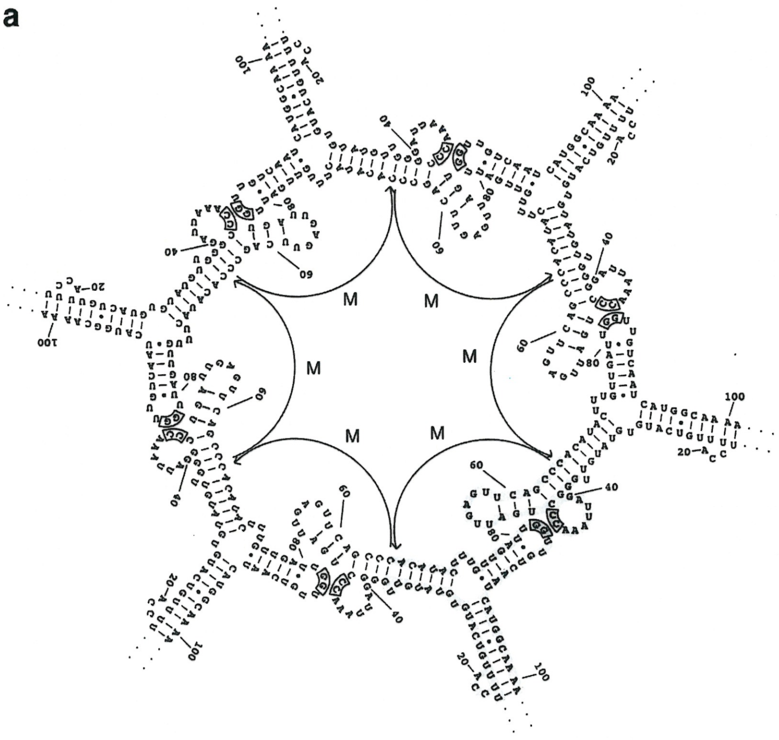
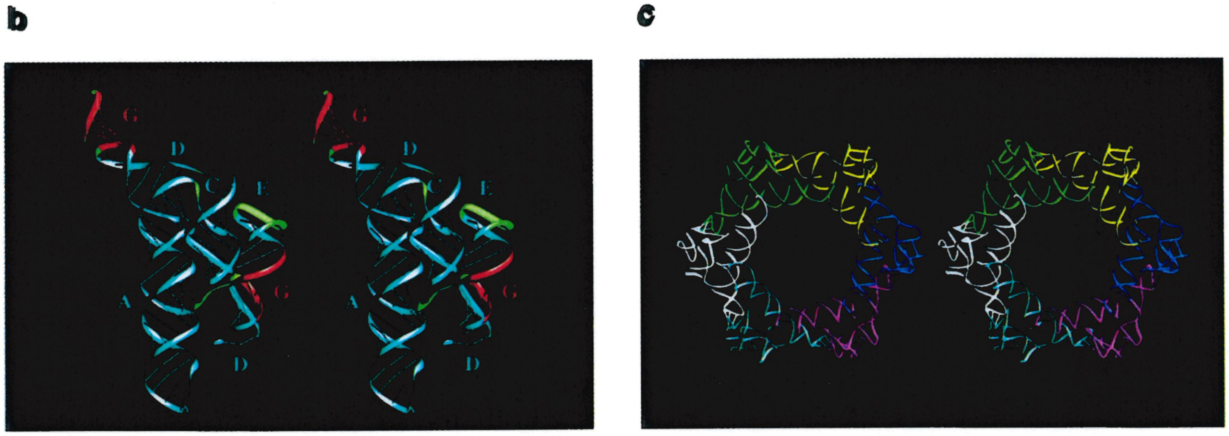


Figure 4. Secondary Structure and Three-Dimensional Model of the pRNA Hexamer

(a) Secondary structure of the pRNA hexamer. The nucleotides of a construction subunit are shown in the shaded area. The transformation matrices, computed between the G37 bases, are indicated by the arrows. Watson-Crick base pairs are indicated by lines, and the wobble base pairs by dots. The residues required for intermolecular base pairing are boxed. Only residues 16–101, 5′ to 3′, of each 120-base molecule are shown. (M) designates the transformation matrix that links two consecutive monomers.

(b) Stereo view of the three-dimensional structure of the pRNA monomer. The intermolecular G stems are represented in red. The upper G stem connects monomers i and i+1. The lower right G stem connects monomers i-1 and i. All other stems are drawn in cyan. The incomplete lower right D stem is part of the monomer i-1. Single-stranded and unconstrained regions are shown in green. The Watson-Crick and wobble stem base pairs are indicated by small cylinders.

(c) Stereo view of the pRNA hexamer. A different color is assigned to each monomer. The four intermolecular Watson-Crick base pairs are represented by small cylinders. The images were generated by the program Ribbons, version 2.45, by Mike Carson at the University of Alabama at Birmingham, Center for Molecular Crystallography.



derived from the secondary structure shown in Figure 1a. A dimer was composed of two monomers linked by the G stem, which contains four Watson-Crick base pairs. A subunit was defined by sectioning the dimer at the G37:C65 base pair (Figure 4a). The conformational space of one subunit was explored by MC-SYM. The standard A-RNA type helix was assumed for all stems, and the GU base pairs were assigned the wobble hydrogen-bonding pattern. The nucleotides in the loop regions were assigned any type of conformation (C3′-endo and C2′-endo pucker modes, and SYN and ANTI torsion around the glycosyl bond), and the number of assignments was reduced to limit the number of loop conformations. For each subunit model found, the transformation matrix, M, connecting the complementary parts of the monomers, was computed and iteratively applied six times. A cycle closure constraint, $M^6 \cong I$, was introduced, and 1125 hexamer models composed of identical monomer conformations were derived. The root mean square deviation between any pair of models was within 2 Å, indicating that all models had a similar global shape. The three-dimensional structure (Figure 4b) suggests parallel side-by-side positioning of stems A, C, and E. The intermolecular stem G is coaxial and stacked on stem E. Stem D is positioned at the bottom of stem A, forming an angle of approximately 60° with stems A, C, and E. This angle induces the formation of a six-monomer cycle (Figure 4c). One side of the E loop is completely exposed to the solvent, and this loop serves a critical role in prohead binding (Reid et al., 1994b, 1994c; Zhang and Anderson, 1998). Mutational analysis (Reid et al., 1994c) supports a prediction of stabilization of the monomer by interaction of residues U72–U74 and U29 with the E loop and by noncanonical pairing of U72 with U29. Other interactions may occur between G40–A44 and either stem A or D, including U81.

Multimeric pRNA within the Dynamic DNA Packaging Machine

The multiple copies of pRNA bound to the prohead portal vertex (head–tail connector) can be removed from the prohead and reattached, with concomitant loss and restoration of DNA packaging activity (Guo et al., 1987b). The connector has 12-fold (Carazo et al., 1986) or 13-fold (Dube et al., 1993; Tsuprun et al., 1994) symmetry. Our three-dimensional reconstruction from electron micrographs of individual connectors appears as a nearly hemispherical shell with a diameter of 140–150 Å, a 35 Å diameter hole at the pole, and 13 protruding subunits near the equator (Tsuprun et al., 1994). The three-dimensional model of the pRNA hexamer (Figure 4c) has a diameter similar to that of the connector, and maintenance of these dimensions in binding would result in superposition of the cyclic pRNA hexamer on the connector to form a double-ring structure. This complex is the crux of the DNA packaging machine.

The ϕ29 DNA–gp3 packaging substrate matures in a separate pathway that involves lariat formation initiated by the terminal protein gp3, binding of multiple copies of the ATPase gp16 at the lariat loop junction, and supercoiling of the lariat loop (Grimes and Anderson, 1997). The gp16, an RNA-dependent ATPase, is hypothesized to recognize pRNA to link the prohead and DNA. Our model of packaging is that the supercoiled DNA wraps around the outside of the connector–pRNA complex, the left end of the DNA is freed to enter the capsid, and the connector rotates with the aid of ATP hydrolysis to translocate the DNA. Approximately six copies of gp16 are added in packaging, and the binding of gp16 by the prohead shields pRNA from RNase (Grimes and Anderson, 1990). Both the connector and gp16 possess RNA recognition motifs, and both have a role in assembling the consummate DNA translocating ATPase that includes pRNA. The challenge is to understand the assembly-regulated configuration of pRNA and proteins that constitutes the dynamic DNA translocating machine.

Weak interactions in the pRNA hexamer may provide conformational lability needed for interactions with the connector and gp16. Thus, the intermolecular base pairing between pRNA monomers may share characteristics of 3D domain swapping in proteins (Bennett et al., 1995). In domain swapping, one domain of a monomeric protein is replaced by the same domain from an identical protein chain, forming either an intertwined dimer or a higher oligomer. In this way, a single pRNA sequence could form distinct structures that would be close in energy and able to rapidly interconvert. Domain swapping may permit interconversion between a dimer with two closed loop–loop interfaces and the cyclic hexamer during discrete steps of DNA packaging. A high resolution crystal structure of a 26-base RNA pseudoknot that forms a dimer by domain swapping has been obtained (S. E. Lietzke et al., submitted). The structure illustrates how RNA can form higher-order multimers like the pRNA hexamer.

Self-cleaving viroid and newt RNAs may only be active as dimers (Forster et al., 1988). To our knowledge, biologically relevant higher multimers of RNA have not been physically and genetically characterized previously. Multimerization of identical molecules of pRNA by intermolecular base pairing to form a discrete structure essential in viral DNA packaging expands the vista of RNA structure and function.

Experimental Procedures

DNA–gp3 Packaging In Vitro

RNA-free proheads were purified from *Escherichia coli* strain HMS174(pAR7-8-8.5-10) that expresses the prohead structural proteins from the cloned genes (Guo et al., 1991). Wild-type and mutant pRNAs were produced by in vitro T7 transcription from pRT71 or its derived plasmid templates (Reid et al., 1994b), and proheads were reconstituted by the addition of twelve copies of pRNA in 25 mM Tris–HCl (pH 7.8), 5 mM $MgCl_2$ and 50 mM NaCl (0.5× TMS). In vitro DNA–gp3 packaging was done as described (Grimes and Anderson, 1989a), except that the 20 μl reaction mixture of proheads with pRNA, DNA–gp3, the ATPase gp16, and 0.5 mM ATP (Sigma) was in 0.5× TMS.

Gel Electrophoresis of pRNA

Native polyacrylamide gel electrophoresis of pRNA (Chang and Tinoco, 1994) was modified as follows. RNA in 10 μl of incubation buffer (10 mM $MgCl_2$, 0.1 mM EDTA, 10 mM Tris–HCl [pH 7.0]) was heated at 80°C for 30 s, cooled immediately in ice, and equilibrated at 4°C or room temperature for 30 min. The sample was mixed with 2.5 μl of 40% (w/v) sucrose, loaded into the gel, and run at 2.5 W at 4°C or room temperature. The electrophoresis buffer contained 10 mM $MgCl_2$, 0.1 mM EDTA, and 100 mM Tris–HEPES (pH 7.8). The nondenaturing 6% gel was prepared using this buffer with a 19:1 weight ratio of acrylamide to *N*, *N'*-methylenebisacrylamide.

Oligonucleotide-Directed Mutagenesis

Plasmid pRT71 (Reid et al., 1994b, 1994c) was used for oligonucleotide-directed mutagenesis (Deng and Nickoloff, 1992) with components of the Transformer™ Site-Directed Mutagenesis Kit and the Trans Oligo™ ScaI/StuI (Clontech). Mutagenesis was performed using the manufacturer's instructions, with an additional round of ScaI (or StuI) digestion and transformation of DH5α cells to further enrich for the desired mutations. All first-round mutagenic primers, and the Trans and Switch Oligonucleotides, were purchased from National Bioscience. For the first round of mutation, Trans Oligo ScaI/StuI GTGACTGGTGAGGCCTCAACCAAGTC was used as the selection primer to change the unique ScaI site to a StuI site on pRT71.

For the second round of selection, the Switch Oligo StuI/ScaI GTGACTGGTGAGTACTCAACCAAGTC was used to change the StuI site back to ScaI on pRT71. The oligonucleotides used to randomize C47, C48, G82, and G83 were GGGGATTAAA(A, G, T)CCTGATTGAG; GGATTAAAC(A, G, T)CTGATTGAG; TTTGTTGATT(A, C, T)GTTGTCAATC; and TTGTTGATTG(A, C, T)TTGTCAATCA, respectively. All second-round mutagenic primers were from Genosys: mutants G47/C82 and G47/A83 were made using oligonucleotide GTTGGGGATTAAAGCCTGATTGAGTTC and mutant plasmids C82 and A83, respectively; mutant U47/C82 was made using mutant C82 plasmid and the oligonucleotide TTGGGGATTAAATCCTGATTGAGTTC; mutants G48/C83 and G48/A82 were made using the oligonucleotide GGATTAAACGCTGATTGAG and mutant plasmids C83 and A82, respectively; and mutant U48/C83 was made using mutant C83 plasmid and the oligonucleotide TGGGGATTAAACTCTGATTGAGTTCA.

Sedimentation Equilibrium Measurements

Sedimentation equilibrium measurements were performed on a Beckman Optima XL-A equipped with a UV-Vis detection system (Beckman Instruments, Palo Alto, CA). Samples were diluted into a buffer containing 10 mM Tris–HCl (pH 7.0), 10 mM $MgCl_2$, and 0.1 mM EDTA to obtain a final concentration between 0.4 and 1.0 μM. The identical sample buffer was used as a reference in all experiments. Double sector cells (30.0 mm height with 12.0 mm length) were used in all experiments. Generally, 180 μl of sample and 195

μl buffer were added into the sample and reference sector of each cell, respectively. At 4.0°C, equilibrium was achieved at 24 hr when the rotor speed was between 4k and 12k rpm. The partial specific volume (ν_o) of the pRNA is 0.588 ml/gm (Meselson et al., 1957; Sueoka et al, 1959). Equilibrium of the sample is reached when the force on the molecule is equal to the opposing force due to diffusion, described by equation 1,

$$C_r = C_{fe} [(\omega^2/2RT)(M(1 - \nu\rho))(r^2 - F^2)],$$

where C_r is the concentration of the solute at radius r, C_f is the concentration at reference distance F, ω is the angular velocity, R is the gas constant, 8.314×10^7 erg $mol^{-1}K^{-1}$, T is the temperature in Kelvin, ν is the partial specific volume, ρ is the solution density, and M is the molecular mass (van Holde, 1985). Because all forms were present in the sample, speed and concentration were varied to observe a particular form. The optimum speed may be calculated if the partial specific volume and targeted molecular mass is known. Since the monomer pRNA has a molecular mass of 40 kDa, the optimum equilibrium speed for pure hexamer, dimer, and monomer pRNA is 4k, 8k, or 12k rpm, respectively. Since the hexamer diffuses more slowly than the monomer, smaller force (4k rpm) is required to offset hexamer diffusion. Under these conditions, low molecular weight material will not sediment, while the hexamer or other high molecular weight material will form a gradient across the sector-shaped cell. The monomer will form a gradient at higher speeds (12k rpm). Under these conditions, high molecular weight complexes will pellet to the bottom of the cell. The molecular mass of the pRNA form was calculated from the best fit of the data to equation 1 using the nonlinear Levenberg-Marquardt fitting routine (Press et al., 1992) by KaleidaGraph (Synergy Software, Reading, PA). All gradients were fit using either a double or a single component model. The best fit to the data is reported.

Acknowledgments

This work was supported by National Institutes of Health research grants GM39931 and DE03606 to D. A. and by the Medical Research Council (MRC) of Canada. F. M. is a fellow of the MRC. We thank C. Peterson, S. Grimes, P. Jardine, and R. Chen for technical assistance and discussions.

Received March 5, 1998; revised May 13, 1998.

References

Anderson, D., and Reilly, B.E. (1993). Bacteriophage ϕ29 morphogenesis. In *Bacillus subtilis* and Other Gram Positive Bacteria: Physiology, Biochemistry, and Molecular Genetics, J.A. Hoch, R. Losick, and A.L. Sonenshein, eds. (Washington, D.C.: ASM Publications), pp. 859–867.

Bailey, S., Wichitwechkarn, J., Johnson, D., Reilly, B.E., and Anderson, D. (1990). Phylogenetic analysis and secondary structure of the *Bacillus subtilis* bacteriophage RNA required for DNA packaging. J. Biol. Chem. *265*, 22365–22370.

Bennett, M.J., Schlunegger, M.P., and Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. Prot. Sci. *4*, 2455–2468.

Carazo, J.M., Donate, L.E., Herranz, L., Secilla, J.P., and Carrascosa, J.L. (1986). Three-dimensional reconstruction of the connector of bacteriophage ϕ29 at 1.8 nm resolution. J. Mol. Biol. *192*, 853–867.

Chang, K.Y., and Tinoco, I. (1994). Characterization of a "kissing" hairpin complex derived from the human immunodeficiency virus genome. Proc. Natl. Acad. Sci. USA *91*, 8705–8709.

Deng, W.P., and Nickoloff, J.A. (1992). Site-directed mutagenesis of virtually any plasmid by eliminating a unique site. Anal. Biochem. *200*, 81–88.

Dube, P., Tavares, P., Lurz, R., and van Heel, M. (1993). The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. EMBO J. *12*, 1303–1309.

Forster, A.C., Davies, C., Sheldon, C.C., Jeffries, A.C., and Symons, R.H. (1988). Self-cleaving viroid and newt RNAs may only be active as dimers. Nature *334*, 265–267.

Gold, L., Allen, P., Binkley, J., Brown, D., Schneider, D., Eddy, S.R., Tuerk, C., Green, L., MacDougal, S., and Tasset, D. (1993). RNA: The shape of things to come. In The RNA World, R.F. Gesteland and J.F. Atkins, eds. (Plainview, NY: Cold Spring Harbor Laboratory Press), pp. 497–509.

Grimes, S., and Anderson, D. (1989a). *In vitro* packaging of bacteriophage ϕ29 DNA restriction fragments and the role of the terminal protein gp3. J. Mol. Biol. *209*, 91–100.

Grimes, S., and Anderson, D. (1989b). Cleaving the prohead RNA of bacteriophage ϕ29 alters the in vitro packaging of restriction fragments of DNA–gp3. J. Mol. Biol. *209*, 101–108.

Grimes, S., and Anderson, D. (1990). RNA dependence of the bacteriophage ϕ29 DNA packaging ATPase. J. Mol. Biol. *215*, 559–566.

Grimes, S., and Anderson, D. (1997). The bacteriophage ϕ29 packaging proteins supercoil the DNA ends. J. Mol. Biol. *266*, 901–914.

Guo, P., Erickson, S., and Anderson, D. (1987a). A small viral RNA is required for in vitro packaging of bacteriophage ϕ29 DNA. Science *236*, 690–694.

Guò, P., Bailey, S., Bodley, J.W., and Anderson, D. (1987b). Characterization of the small RNA of the bacteriophage ϕ29 DNA packaging machine. Nucleic Acids Res. *15*, 7081–7090.

Guo, P., Erickson, S., Xu, W., Olson, N., Baker, T., and Anderson, D. (1991). Regulation of phage ϕ29 prohead shape and size by the portal vertex. Virology *183*, 366–373.

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

Meselson, M., Stahl, F.W., and Vinograd, J. (1957). Equilibrium sedimentation of macromolecules in density gradients. Proc. Natl. Acad. Sci. USA *43*, 581–588.

Press, W.H., Teukolsky, S.A., Vettering, W.T., and Flannery, B.P. (1992). In Numerical Recipes in C. (New York, NY: Cambridge University Press), pp. 656–706.

Reid, R.J.D., Bodley, J.W., and Anderson, D. (1994a). Characterization of the prohead-pRNA interaction of bacteriophage ϕ29. J. Biol. Chem. 269, 5157–5162.

Reid, R.J.D., Bodley, J.W., and Anderson, D. (1994b). Identification of bacteriophage ϕ29 prohead RNA domains necessary for in vitro DNA–gp3 packaging. J. Biol. Chem. *269*, 9084–9089.

Reid, R.J.D., Zhang, F., Benson, S., and Anderson, D. (1994c). Probing the structure of bacteriophage ϕ29 prohead RNA with specific mutations. J. Biol. Chem. *269*, 18656–18661.

Sueoka, N., Marmur, J., and Doty, P. (1959). Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature *183*, 1429–1431.

Tomizawa, J. (1993). Evolution of functional structures of RNA. In The RNA World, R.F. Gesteland and J.F. Atkins, eds. (Plainview, NY: Cold Spring Harbor Laboratory Press), pp. 419–445.

Trottier, M., and Guo, P. (1997). Approaches to determine stoichiometry of viral assembly components. J. Virology *71*, 487–494.

Tsuprun, V., Anderson, D., and Egelman, E. (1994). The bacteriophage ϕ29 head-tail connector shows 13-fold symmetry in both hexagonally-packed arrays and as single particles. Biophys. J. *66*, 2139–2150.

van Holde, E. (1985). Physical Biochemistry (Engelwood Cliffs, NJ: Prentice-Hall), pp. 110–129.

Wyatt, J.R., and Tinoco, I. (1993). RNA structural elements and RNA function. In The RNA World, R.F. Gesteland and J.F. Atkins, eds. (Plainview, NY: Cold Spring Harbor Laboratory Press), pp. 465–496.

Zhang, F., and Anderson, D. (1998). *In Vitro* selection of bacteriophage ϕ29 prohead RNA aptamers for prohead binding. J. Biol. Chem. *273*, 2947–2953.

CHAPITRE 6

Conclusion

L'ensemble du travail présenté dans ce mémoire vient étendre significativement la méthode MC-SYM, intialement présentée dans [13], au niveau de trois aspects importants, soit le traitement de multiples séquences par l'utilisation de la logique floue, la définition de l'espace des conformations et l'intégration d'une contrainte de formation d'un multimère.

Le chapitre 3 met en place un formalisme basé sur la logique floue permettant d'intégrer dans la modélisation 3-D d'un ARN l'information de séquences multiples. La popularité croissante de la méthode de sélection *in vitro* laisse croire que ce type de formalisation permettra d'accélérer l'obtention de modèles 3-D pour de petites molécules d'ARN. L'étude de sites spécifiques des ARN ribosomaux est aussi un défi intéressant pour l'application de cette méthodologie, puisqu'un grand nombre de séquences sont connues, et que l'élucidation de la structure 3-D de ces sites rendrait possible le développement de nouvelles classes d'agent antimicrobiaux. De plus, l'approche présentée permet aussi d'obtenir des modèles 3-D pour chacune des séquences observées. Ce type de résultat peut être mis à profit lors de l'étude de la fonction de la molécule modélisée. Par contre, puisque l'approche utilise un grand nombre de recherches discrètes indépendantes, les temps de calcul nécessaires deviennent rapidement inacceptables lors du traitement de plus grandes séquences.

Dans les dix dernières années, les techniques de modélisation ont dû s'adapter à l'amélioration et à l'arrivée de plusieurs méthodes expérimentales permettant d'obtenir de l'information structurelle sur les ARN. Les années à venir laissent encore entrevoir de grands changements, l'arrivée de la cryo-microscopie électronique (*cryo-EM*) et l'amélioration continuelle des techniques de crystallographie n'en sont que des exemples. Ceci suggère la mise au point de systèmes de modélisation flexibles dans lesquels il est facile d'implanter l'utilisation de nouveaux types d'information. C'est ce qui a été fait au chapitre 2 et le chapitre 5 présente une démonstration de cette versatilité par l'addition dans l'engin de recherche d'un nouveau type de contrainte.

6.1 Développement futurs

L'utilisation de recherches heuristiques sur l'espace des conformations permettrait de réduire grandement les temps de calcul nécessaires à l'obtention de modèles préliminaires. L'implantation de ces heuristiques exigerait la mise au point d'une fonction d'évaluation permettant de quantifier la qualité d'un modèle. La tâche n'est pas simple puisqu'une telle fonction ne devrait pas pénaliser les incohérences locales des structures construites par l'exploration des relations de bases azotées. Toute approche basée sur le calcul d'une fonction d'énergie réaliste serait donc inutilisable ou nécessiterait le raffinement préalable des modèles, ce qui s'avère trop coûteux.

La mise au point d'une variante heuristique de MC-SYM permettrait entre autre l'application de la méthode à la modélisation de très gros ARN (les ARN ribosomaux par exemple) ou la possibilité d'obtenir rapidement des modèles approximatifs pour un grand nombre d'hypothèses structurelles. Les résultats obtenus devraient aussi être moins sensibles au choix de la taille d'échantillonnage des relations binaires puisqu'il serait possible d'utiliser de grands échantillons de manière systématique.

L'utilisation de la méthodologie présentée au chapitre 2 et utilisée aux chapitres suivants nécessite encore un grand nombre d'interventions du modélisateur, faisant appel à ses intuitions sur la structure. Il serait très intéressant de rationaliser ces intuitions et de permettre une automatisation complète du processus de modélisation 3-D. En plus d'accélerer l'obtention du modèle 3-D, ceci permettrait de rendre complètement objective la démarche de modélisation.

L'introduction des graphes de relations entre bases azotées suggère l'utilisation de plusieurs méthodes dérivées de l'analyse et de l'optimisation de réseaux pour indiquer au modélisateur le moyen idéal de construire le modèle. Parallèlement à l'analyse des informations contenues dans ce graphe, beaucoup des intuitions de l'utilisateur viennent par l'utilisation, pendant un certain temps, d'une même base de donnée de relations binaires. Ceci suggère que la mise au point d'outils permettant une analyse automatisée de la base de donnée est une étape importante à la formalisation des intuitions du modélisateur. Dans ce domaine, on cherchera à optimiser l'exploration d'un espace des conformations chimiquement valides en améliorant la sélection d'un échantillon de relations de bases azotées.

Un des points critiques de la modélisation est le fait que, dans plusieurs cas, le modélisateur est dans une situation où l'information disponible n'est pas suffisante à la construction directe d'un modèle. Dans ces cas, on doit proposer une série d'hypothèses structurelles plausibles (voir chapitre 4 pour un exemple de cette approche). Du choix de ces hypothèses dépendra l'exactitude et la précision des résultats obtenus. Il serait très intéressant de formaliser cette approche et d'automatiser le processus de génération d'hypothèses structurelles. Une fois plusieurs modèles obtenus pour chacune des hypothèses, il serait possible d'analyser les différences structurelles entre les classes de modèles et de suggérer les expérimentations permettant d'obtenir rapidement un modèle définitif. Cette approche s'apparente grandement aux algorithmes de construction d'un arbre de décision optimal et devrait fournir une grande aide lors de la collaboration entre modélisateurs et expérimentatlistes.

Le domaine de la modélisation 3-D des ARN est un champ en pleine expansion et dont les résultats sont essentiels à la compréhension des mécanismes moléculaires impliquant ces ARN. La mise en évidence récente du rôle important joué par les ARN dans les mécanismes de réplication de plusieurs virus (HIV, hépatite, mosaïque du tabac, etc.) confirme la pertinence des études structurelles entreprises sur les ARN et plus particulièrement de la modélisation 3-D de ces structures. La mise au point de méthodes et de logiciels de modélisation est un prérequis à l'évolution de l'étude structurelle des ARN puisqu'elle permettra l'obtention rapide et objective de modèle 3-D des ARN d'intérêt.

RÉFÉRENCES

[1] F. Major et D. Gautheret. Computer Modeling of RNA Three-Dimensional Structures. *Encyclopedia of Molecular Biology and Biotechnology*, R.A. Meyers Ed., Wiley-VCH, NY, 1995.

[2] J.D. Watson et F.H.C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribonucleic Acid. *Nature*, **171**:694–967, 1953.

[3] L.M. Blumenthal. *Theory and Application of Distance Geometry*, Chelsea, NY, 1970.

[4] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, NJ, 1976.

[5] M.S. Bazaraa et C.M. Shetty. *Nonlinear Programming: Theory and Algorithms*, Wiley, NY, 1979.

[6] T.R. Cech. RNA splicing: three themes with variations. *Cell*, **34**:713–6, 1983.

[7] W. Saenger. *Principles of Nubleic Acid Structure*, Springer-Verlag, NY, 1984.

[8] M. Zuker et D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, **46**: 591–621, 1984.

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

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

[11] J.-P. Perreault, T.F. Wu, B. Cousineau, K.K. Ogilvie et R. Cedergren. Mixed deoxyribo- and ribo-oligonucleotides with catalytic activity. *Nature*, **344**: 565–7, 1990.

[12] A. Kolinski et J. Skolnick. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *Journal of Molecular Biology*, **221**: 499–531, 1991.

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

[14] F. Major, D. Gautheret et R. Cedergren. Reproducing the three-dimensional structure of a transfer RNA molecule from structural constraints. *Proc. Natl. Acad. Sci. U.S.A.*, **90**:9408–12, 1993.

[15] D. Gautheret, F. Major et R. Cedergren. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *Journal of Molecular Biology*, **229**: 1049–64, 1993.

[16] A. Malhotra, R.K. Tan et S.C. Harvey. Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques. *Biophysical Journal*, **66**: 1777–95, 1994.

[17] C. Gaspin et E. Westhof. An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *Journal of Molecular Biology*, **254**: 163–74, 1995.

[18] M.S. Waterman. *Introduction to computational biology*, Chapman & Hall, NY, 1995.

[19] P. Chartrand, N. Usman et R. Cedergren, Effect of structural modifications on the activity of the leadzyme. *Biochemistry*, **36**: 3145–50, 1997.

[20] M.H. Kolk, H.A. Heus et C.W. Hilbers. The structure of the isolated, central hairpin of the HDV antigenomic ribozyme: novel structural features and similarity of the loop in the ribozyme and free in solution. *EMBO Journal*, **16**: 3685–92, 1997.

[21] F. Major, S. Lemieux et M. Ftouhi. Computer RNA Three-Dimensional Modeling from Low-Resolution Data and Multiple-Sequence Information. *Molecular Modeling and Structural Determination of Nucleic Acids*, N.B. Leontis et J. Santa Lucia Eds., American Chemical Society Books, Washington, D.C., 1997.

[22] F. Mueller et 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**: 524–44, 1997.

[23] T. Pan. Novel and variant ribozymes obtained through in vitro selection. *Current Opinion in Chemical Biology*, **1**:17–25, 1997.

[24] D.A. Benson, M.S. Boguski, D.J. Lipman, J. Ostell et B.F. Ouellette. GenBank. *Nucleic Acids Research*, **26**: 1–7, 1998.

[25] R. Cedergren et F. Major. Modeling the Tertiary Structure of RNA. *RNA Structure and Function*, Cold Spring Harbor Laboratory Press, NY, 1998.

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

[27] S. Lemieux, S. Oldziej et F. Major, Representing and Infering RNA Three-Dimensional Structures, *Encyclopedia of Computational Chemistry*, N.L. Allinger et *al.* Eds., John Wiley & Sons, West Sussex, England, 1998.

[28] S.R. Lynch et I. Tinoco Jr. The structure of the L3 loop from the hepatitis delta virus ribozyme: a syn cytidine. *Nucleic Acids Research*, **26**: 980–7, 1998.

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