Université de Montréal

# Identification de nouvelles séquences capables de se replier en différents motifs d'ARN connus.

par

Véronique Bourdeau

Département de Biochimie

Faculté de Médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae doctor (PhD.)
en Biochimie

Octobre 2000

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

# Identification de nouvelles séquences capables de se replier en différents motifs d'ARN connus

présentée par :

Véronique Bourdeau

a été évaluée par un jury composé des personnes suivantes :

Bernd F. Lang, pres.-r.
Robert Cedergren, dir. de rech (décédé)
Lea Brakier-Gingras, co-directrice
Stephen Michnick, membre du jury
William McClain, exa. externe

Thèse acceptée le : 26/10/00

# Sommaire

L'ARN peut à la fois conserver l'information génétique et catalyser des réactions biochimiques. Cette observation est à la base d'une hypothèse selon laquelle le développement de la vie aurait passé par un « Monde à ARN ». Dans ce monde, toutes les activités nécessaires pour encoder, répliquer et maintenir l'information génétique auraient été effectuées par des molécules d'ARN. Un ARN possède trois niveaux de structures : la structure primaire ou séquence, la structure secondaire qui se compose principalement des agencements divers de régions simple brin et double brin, et la structure tertiaire permettant le rapprochement de différentes régions de structure secondaire. Il est possible de décrire pour plusieurs ARN des petits motifs structuraux caractéristiques de l'activité associée à une région donnée de la molécule. Les ARN et leurs motifs exhibent un fort aspect de dégénérescence entre les niveaux de structures. Effectivement, pour une structure secondaire donnée, plusieurs structures primaires pourront exister. Ce concept se trouve parfaitement illustré par les centaines de séquences d'ARN de transfert (ARNt) connues formant une structure secondaire conservée dite en feuilles de trèfle.

La présente thèse rapporte des travaux effectués sur différents aspects relatifs aux motifs structuraux d'ARN. Dans un premier temps, une approche informatisée a été utilisée pour identifier, parmi les séquences naturelles disponibles, des régions particulières pouvant adopter la même structure que certains motifs d'ARN déjà connus (Chapitre 1, 2 et 3). Deux conclusions majeures sont ressorties de ces recherches : (i) la distribution des motifs d'ARN correspond aux prédictions du modèle de la dérive aléatoire et (ii) plusieurs motifs ont été identifiés qui permettent de développer des hypothèses intéressantes sur leur fonction potentielle (exemple détaillé au Chapitre 2) ou qui démontre la conservation de certains motifs dans un contexte particulier (e. i. : les ribozymes en tête-de-marteau dans l'ADN répétitif; Chapitre 3).

Les Chapitres 4 et 5 comportent des expériences faites sur un motif particulier d'ARN : l'ARNt. Le Chapitre 4 présente l'isolement de séquences pouvant produire des ARNt actifs chez *Escherichia coli* mais possédant une structure secondaire atypique. Cette activité démontre la possibilité de trouver de nouvelles structures primaire et secondaire capables de former une structure tertiaire fonctionnelle (Chapitre 4). La caractérisation de la maturation de deux des ARNt atypiques isolés a également permis d'identifier une activité de clivage présente dans le cytoplasme de *E. coli* qui agit spécifiquement à l'intérieur de leur séquence (Chapitre 5). Cette découverte apporte des indices sur les moyens de défense possibles des cellules contre l'apparition et la conservation d'ARNt atypiques.

Enfin, ces travaux démontrent que les séquences naturelles présentent un fort aspect de dégénérescence des trois niveaux de structures des ARN. L'importance de cette dégénérescence est discutée dans le contexte du Monde à ARN où elle devait représenter une propriété utile à la naissance de la vie.

# TABLE DES MATIÈRES

**Sommaire** iii

**Table des matières** v

**Liste des Tableaux** xi

**Liste des Figures** xii

**Liste des Sigles et Abréviations :**

Introduction et Discussion xiv

Chapitres 1 à 5 xv

**Dédicace** xviii

**Introduction :** 1

1. Au début... 2
2. L'hypothèse du Monde à ARN 3
   2.1 ARN versus ADN 4
   2.2 ARN versus protéines 6
   2.3 ARN ou plus simple ? 8
3. Aspects structuraux des ARN 9
   3.1 Structure primaire 9
   3.2 Structure secondaire 9
   3.3 Structure tertiaire 13
   3.4 Interdépendance des niveaux de structure 14
   3.5 Définition d'un motif d'ARN 15
4. Dégénérescence dans la structure des ARN 18
   4.1 Description 18
   4.2 Implications de la dégénérescence 19

5. Une famille d'ARN particulière : les ARN de transfert 21

5.1 Fonction 22

5.2 Structure 22

5.3 ARNt atypiques 27

5.4 Importance dans le Monde à ARN 28

6. Questions abordées dans cette thèse 29

6.1 Distribution et fonction des motifs d'ARN dans les séquences naturelles 29

6.2 Des ARNt atypiques peuvent-ils être fonctionnels dans le cytoplasme de *Escherichia coli* ? 30

6.3 Pourquoi les ARNt atypiques ne s'expriment que faiblement chez *Escherichia coli* ? 30

**Résultats :** 32

Chapitre 1 33

Article : The distribution of RNA motifs in natural sequences.

Note 35

Contenu 36

I. Abstract 37

II. Introduction 37

III. Materials and Methods 39

a) The search

b) The analysis

IV. Results 41

a) Protein-binding motifs

b) Chemically and catalytically active motifs

c) Small molecule-binding RNA motifs

d) Phylogenetic analysis

V. Discussion 50

a) Origin of the motifs

b) Evolutionary dynamics of the motifs

c) The activity and utility of motifs

d) The advantage of fortuitous targets

e) Conclusion

VI. Acknowledgements 54

VII. References 55

VIII. Tables 59

IX. Figures 65

Chapitre 2 75

Article : Does HIV tat protein also regulate genes of other viruses present in HIV infection?

Note 77

Contenu 78

I. Does HIV tat protein also regulate genes of other viruses present in HIV infection? 79

II. Acknowledgements 81

III. References 82

IV. Figure 83

Chapitre 3 85

Article: Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank

Note 87

Contenu 88

I. Abstract 89

II. Introduction 89

III. Results 91

a) Searching for self-cleaving RNA motifs of the hammerhead type in the GenBank

b) The frequency of mutated versions of the hammerhead self-cleaving RNAs

c) Some intriguing hammerhead motifs that might have functional significance

IV. Discussion 97

V. Methods 99

VI. Acknowledgments 102

VII. References 102

VIII. Tables 107

IX. Figures 113

Chapitre 4 117

Article : Amber suppression in *Escherichia coli* by unusual mitochondria-like transfer RNAs.

Note 119

Contenu 120

I. Abstract 121

II. Introduction 121

III. Materials and Methods 122

a) Strains

b) Construction of the Combinatorial Library

c) tRNA Levels and Aminoacyl-tRNA Levels

d) Identification of the Charged Amino Acid

IV. Results 125

a) Experimental Design

b) Clones Selection and Characterization

c) Sequence Analysis

V. Discussion 128

VI. Acknowledgements 130

VII. References 130

VIII. Tables 132

IX. Figures 136

Chapitre 5 144

Article : Cleavage of mitochondria-like tRNA Expressed in *Escherichia coli*.

Note 146

Contenu 147

I. Abstract 148

II. Introduction 148

III. Material and Methods 150

a) Strains

b) Plasmids

c) *E. coli* S100 extracts

d) Processing of *in vitro* transcribed tRNAs

e) Northern blots

f) Measurements of $Lac^+$ reversion frequency

IV. Results 153

V. Discussion 159

VI. Acknowledgements 162

VII. References 162

VIII. Figures 165

**Discussion** 180

1. Distribution des motifs d'ARN dans GenBank (Chapitre 1 et 3) 181

2. Valeur des motifs potentiels trouvés (Chapitre 1,2 et 3) 182

3. Importance de l'isolement d'ARNt atypiques actifs (Chapitre 4) 183

4. Significations possibles d'une activité de clivage des ARNt atypiques (Chapitre 5) 184

5. Implications générales des travaux de cette thèse 186

5.1 Observation générale 186

5.2 Avantages et désavantages de la dégénérescence des structures des ARN 186

*i)* *Utilité de la dégénérescence* 186

*ii)* *Inconvénient de la dégénérescence* 187

6. Conclusion 189

**Références** 191

**Appendices**

I. Lettre de R.C. Gallo xix

II. Site Internet des résultats du Chapitre 1 xxi

III. Site Internet des résultats du Chapitre 3 xxviii

**Remerciements** xxxvi

## Liste des Tableaux

Introduction

Tableau 1. Résumé des différents types de motifs. 17

Chapitre 1

Table 1. Frequency of RNA motifs in GenBank. 59

Table 2. The organismal distribution of RNA motifs in GenBank. 61

Table 3. Distribution of RNA motifs among GenBank features. 63

Chapitre 3

Table I. Known hammerhead motifs identified in our search. 107

Table II. Distribution of hammerhead and hammerhead-like motifs in the different sections of the GenBank: mutants of the single stranded regions. 109

Table III. Distribution of hammerhead and hammerhead-like motifs in the different sections of the GenBank: mutants of the Helix II and III. 111

Chapitre 4

Table 1. ß-Galactosidase activities and gene sequences of suppressor tRNAs. 132

Table 2. Levels of expression and aminoacylation. 134

# LISTE DES FIGURES

Introduction

Figure 1. Éléments des différents niveaux de structure d'un ARN. 11

Figure 2 Schémas des structures secondaire et tertiaire des ARNt. 25

Chapitre 1

Figure 1. The protein-binding motifs. 65

Figure 2. The organismal distribution of RNA motifs in GenBank. 67

Figure 3. Chemically and catalytically active motifs. 69

Figure 4. Small molecule-binding RNA motifs. 71

Figure 5. The tRNA motif. 73

Chapitre 2

Figure 1. Tat-biding motifs. 83

Chapitre 3

Figure 1. Structures and descriptors of the hammerhead self-cleaving ribozyme motifs. 113

Figure 2. Putative hammerhead motifs. 115

Chapitre 4

Figure 1. The base-pairing pattern of mitochondrial tRNAs. 136

Figure 2. Modifications to the tRNA$^{Ala}$ gene sequence. 138

Figure 3. Properties of active tRNA clones. 140

Figure 4. Secondary structure of isolated tRNAs. 142

Chapitre 5

Figure 1. Representations of the secondary structure of tRNA$^{Ala}_{su+}$ and the mt-like tRNAs –T7 and –T37. 165

Figure 2. *In vitro* processing of tRNA$^{Ala}_{su+}$, and the mt-like tRNAs –T7 and –T37. 167

Figure 3. *In vitro* processing of internally or 5'-end labeled precursor tRNA transcripts. 169

Figure 4. Characterization of the activity producing the "X" fragment. 171

Figure 5. Effect of the temperature on the formation of the "X" fragment. 174

Figure 6. *In vivo* identification of a fragment equivalent to the "X" fragment. 176

Figure 7. Levels of expression of the mature tRNA, suppression activity and mutator effect. 178

# Liste des Sigles et Abréviations
## (Introduction et Discussion)

| | |
|---|---|
| ⊕ | Région du bras supplémentaire des ARNt |
| AA | Région acceptrice des ARNt |
| AC | Région de l'anticodon des ARNt |
| ADN | Acide désoxyribonucléique |
| ARN | Acide ribonucléique |
| ARNm | ARN messager |
| ARN-p | ARN fait de pyrannose plutôt que de ribose |
| ARNr | ARN ribosomique |
| ARNt | ARN de transfert |
| ARNt$^{Ile}$ | ARN de transfert isoleucine |
| ATP | Adénosine triphosphate |
| désoxyUDP | Désoxyribonucléotide d'uridine diphosphaté |
| Ile | Isoleucine |
| PNA | Acide nucléique relié par liens peptidiques plutôt que par des phosphates |
| Pro | Proline |
| RBE | Élément de liaison à la protéine Rev |
| RRE | Élément de réponse à la protéine Rev |
| snoRNA | Petit ARN nucléolaire |
| snRNA | Petit ARN nucléaire |
| tmRNA ou 10Sa RNA | ARN de transfert et messager |
| Trp | Tryptophane |
| VIH | Virus de l'Immunodéficience Humaine |
| VIS | Virus de l'Immunodéficience Simien |

# LISTE DES SIGLES ET ABRÉVIATIONS

## (CHAPITRES 1 À 5)

| | |
|---|---|
| $\gamma$-$^{32}$P-ATP | Adenosine triphosphate containing a radioactive phosphate ($^{32}$P) in position gamma |
| $\alpha$-$^{32}$P-UTP | Uridine triphosphate containing a radioactive phosphate ($^{32}$P) in position alpha |
| AC | Acceptor region of tRNA (Chapter 4) |
| AIDS | Acquired immunodeficency syndrome |
| Ala | alanine |
| AMP | Adenosine monophosphate |
| AN | Anticodon region of tRNA (Chapter 4) |
| APRT | Adenine phosphoribosyl transferase |
| Asn | Asparagine |
| Asp | Aspartic acid |
| ATP | Adenosine triphosphate |
| Bct | Bacterial sequences |
| cDNA | Complementary DNA |
| CDS | Coding sequence |
| CIAR | Canadian institute for advanced research |
| CIP | Calf intestine phosphatase |
| CpG | DNA sequence of 5'…CG…3' (p is for phosphate) |
| -d | Direct orientation of the motif named before |
| DHFR | Dihydrofolate reductase |
| DiTER | Direction des infrastructure Technologiques d'Enseignement de Recherche |
| DNA | Desoxyribonucleic acid |
| DTT | dithiothreitol |
| EDTA | Ethylenediamine tetraacetate |
| EST or Est | Expressed sequence tags |

| | |
|---|---|
| FAD | Flavin adenine dinucleotide |
| FMN | Flavin adenine mononucleotide |
| GSS | Genome survey sequences |
| HCl | Hydrochloric acid |
| His | Histidine |
| HIV | Human immunodeficiency virus |
| HTG or HTGS | High throughput genome sequencing |
| Inv | Invertebrate sequences |
| LTR | Long terminal repeat |
| Mam | Other mammalian sequences |
| $Mg^{2+}$ | Magnesium ion |
| $MgCl_2$ | Magnesium chloride |
| Misc. | Miscellanies |
| mRNA | Messenger RNA |
| mt | mitochondria |
| NaCl | Sodium chloride |
| NAD | Nicotinamide adenine dinucleotide |
| NADPH | Nicotinamide adenine dinucleotide phosphate |
| NSERC | Natural Science and Engineering Research Council |
| nt | nucleotide |
| NTPs | Four nucleotides in triphosphate form |
| $^{-}OH$ | Hydroxide ions (basic conditions) |
| PAT | Patent sequences |
| PCR | Polymerase chain reaction |
| Phg | Phage sequences |
| PKR | Ribosome-associated protein kinase |
| Pln | Plant sequences (including fungi and algae) |
| PMSF | phenylmethylsulfanylfluoride |
| PNK | Polynucleotide kinase |
| Pre_RNA | Precursor RNA |

| | |
|---|---|
| Pri | Primate sequences |
| Prim_trnascript | Primary transcript |
| -r | Reverse orientation of the motifs named before |
| RBE | Rev-binding element |
| Rna | Structural RNA sequences |
| RNA | Ribonucleic acid |
| Rod | Rodent sequences |
| RRE | Rev responsive element |
| rRNA | Ribosomal RNA |
| scRNA | Small cytoplasmic RNA |
| SIV | Simian immunodeficiency virus |
| snoRNA | Small nucleolar RNA |
| snRNA | Small nuclear RNA |
| STS | Sequence tagged sites |
| SYN | Synthetic and chimeric sequences |
| TAR | *trans*-activation response |
| TBE | Tat-binding element |
| T° | temperature |
| tRNA | Transfer RNA |
| $\text{tRNA}_{\text{su+}}^{\text{Ala}}$ | Alanine suppressor transfer RNA |
| Tyr | Tyrosine |
| UNA | Unanotated sequences |
| UNAM | Universidad Nacional Autónoma de México |
| UTP | Uridine triphosphate |
| UTR | Untranslated region |
| Vrl | Viral sequences |
| Vrt | Other vertebrate sequences |
| X-Gal | 5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside |

À Mes Parents :

Pierre et Thérèse Bourdeau

# INTRODUCTION

## 1. Au début...

Un clin d'œil! Seulement l'instant d'un clin d'œil.

L'existence de l'homme moderne (*homo sapiens sapiens*) ne représente en effet presque rien en comparaison de l'âge de l'Univers : 15,8 milliards d'années. À partir de presque rien, une explosion, un " big bang ", un Univers a vu le jour. Dans son expansion, une multitude d'événements se sont enchaînés, dépendants les uns des autres et comme liés par de subtiles chaînes : des lois agissant sur la matière, l'espace et le temps. Depuis le début de cet Univers, il aura fallu plus de 11 milliards d'années pour qu'une petite planète se forme dans un système solaire propice au développement de la vie. Sur ce petit grain de sable dans l'espace du désert céleste, sur cette planète qui est la nôtre, 3,8 milliards d'années de plus se sont écoulées avant que la vie ne se manifeste. Et nous, comme humains, n'existons que depuis les 200 000 dernières années (Wood et Collard, 1999)!

Quelle perception pouvons-nous avoir de la vie avec notre courte expérience? Nous élargissons nos connaissances et nos conceptions de la vie avec chaque petit élément d'information que nous pouvons recueillir ou proposer. Les connaissances que nous avons à ce jour sont incomplètes mais fascinantes. La vie a trouvé le moyen de s'installer et d'abonder sous la forme de milliers et milliards de micro-organismes en l'espace de 3,2 à 3,8 milliards d'années, dans un environnement bien différent de celui que nous connaissons : une soupe pré-biotique (Mojzsis *et al.*, 1999). L'interaction de plusieurs atomes et de différentes molécules aurait permis la formation de molécules organiques capables de se répliquer et de remplir diverses fonctions. L'ensemble de ces molécules fonctionnelles se serait éventuellement assemblé en une plus grande structure : la première cellule. Puisque tous les organismes vivants connus ont pour unité de base la cellule, tous les organismes modernes descendraient de cette première cellule, soit le dernier ancêtre commun (Orgel, 1998).

L'idée générale de l'origine de la vie est acceptée par la majorité des scientifiques oeuvrant dans ce domaine mais les détails en sont encore plus que mystérieux. Comment les éléments nécessaires à la formation de la première cellule se sont-ils créés? Quels éléments ont été les plus importants? Le sont-ils encore? Comment se sont-ils tous assemblés?

## 2. L'hypothèse du Monde à ARN

À partir de la prémisse que l'hypothèse la plus simple est aussi la plus probable, il nous faut considérer des moyens modestes au début de la vie. Les cellules telles que nous les connaissons de nos jours sont beaucoup trop complexes pour avoir été présentes au commencement, en particulier les cellules eucaryotes qui possèdent de nombreux compartiments : noyau, réticulum endoplasmique, Golgi, lysosome, etc. Même les cellules les plus simples sont constituées de trois polymères fonctionnels principaux : les acides désoxyribonucléiques ou ADN (polymères de désoxyribonucléotides), les acides ribonucléiques ou ARN (polymères de ribonucléotides) et les protéines (polymères d'acides aminés). L'ADN représente le matériel de choix pour la conservation de l'information génétique. Sa transmission dans les cellules filles est essentielle au maintien de toute espèce. L'ARN est impliqué surtout dans la traduction de l'information qui est encodée dans l'ADN pour produire différentes protéines. Dans ce processus, plusieurs molécules d'ARN sont nécessaires : des ARN messagers (ARNm), des ARN de transfert (ARNt), des ARN ribosomiques (ARNr) qui s'associent à de nombreuses protéines pour former les ribosomes, site de la traduction. Ainsi, l'ARN peut remplir des tâches fonctionnelles seul ou en conjugaison avec des protéines. Cependant, l'ARN possède aussi le rôle de maintien de l'information génétique, en particulier dans certains virus. Enfin, les protéines possèdent des propriétés chimiques et physiques très différentes de sorte qu'elles peuvent servir à la catalyse d'une multitude de réactions et remplir des fonctions structurales diverses.

En tant que détectives devant le mystère de l'origine de la vie, nous devons considérer comme hypothèses les plus probables celles qui n'impliquent qu'une seule substance fonctionnelle pour remplir toutes les activités requises. Les faits à observer pour identifier cette substance sont peu nombreux. Au niveau de ce qui est possible chimiquement, la synthèse de différents composés organiques simples a été démontrée dans des conditions soupçonnées d'avoir prévalu au début de la vie sur terre. Dans plusieurs expériences, différents mélanges de méthane, d'ammoniaque, d'hydrogène, de cyanure d'hydrogène, d'azote et/ou de formaldéhyde ont été soumis à des décharges électriques, pour imiter des éclairs. Les résultats de ces expériences ont permis de démontrer la possibilité de former certains acides aminés, quelques bases d'acides nucléiques et de sucres (précurseurs des acides nucléiques) (revue par Orgel, 1998). Les organismes les plus simples, possiblement les plus anciens, pouvant encore être étudiés sont constitués à la fois d'ADN, d'ARN et de protéines. Ces trois polymères sont donc tous des suspects à considérer sérieusement. Les spéculations vont ainsi à savoir qui des ADN, des ARN ou des protéines furent les premiers à exister et à supporter la vie?

## 2.1 ARN versus ADN

L'ADN et l'ARN se différencient chimiquement par un groupement hydroxyle (–OH) de plus sur les riboses de l'ARN, là où les désoxyriboses de l'ADN ne contiennent qu'un hydrogène (–H). Trois des bases entrant dans la composition des nucléotides sont communes (adénine, cytosine et guanine) mais l'une diffère d'un polymère à l'autre (ADN : thymine, ARN : uridine). Ces petites différences semblent pourtant bien importantes par rapport à la grande divergence de fonctions trouvée dans la cellule pour ces deux polymères.

Il est possible que ces aspects chimiques aient influencé la détermination naturelle du plus apte de ces deux polymères à soutenir le début de la vie. Toutefois, il est plus facile d'observer les évidences biochimiques qui indiquent clairement que l'apparition de l'ARN a précédé celle de l'ADN (revue par Gilbert et de Souza, 1999). Le premier indice se situe au niveau de la synthèse des nucléotides. Les ribonucléotides sont produits d'abord dans la cellule et servent de précurseurs aux désoxyribonucléotides. Les nucléotides ayant comme base la thymine, bien que retrouvés seulement dans l'ADN, dérivent aussi de la formation préalable de désoxyUDP. Un second indice découle de la dépendance du processus de réplication de l'ADN à la présence d'amorces d'ARN. De façon générale, les enzymes copiant l'ADN sont incapables d'initier *de nuovo* la synthèse de la nouvelle copie d'ADN. Elles ne peuvent qu'allonger le nouveau brin d'ADN à partir d'un fragment préexistant déjà apparié sur l'ADN, une amorce faite d'ARN. Enfin, les extrémités des chromosomes linéaires d'ADN sont construites avec l'aide d'une enzyme, la télomérase, qui utilise une matrice d'ARN pour allonger l'extrémité 3' de l'ADN (revue par O'Reilly *et al.*, 1999).

À un niveau plus fonctionnel, il y a beaucoup plus d'enzymes constituées d'ARN, ou ribozymes, connues que d'enzymes composées d'ADN, ou DNAzymes. Il existe un grand nombre de ribozymes naturels et plusieurs autres ont été développés au laboratoire alors que les DNAzymes identifiées ont été obtenues uniquement suite à des expériences d'évolution *in vitro*, ou SELEX (revue par Joyce et Orgel, 1999; Joyce, 1999). Plus encore, les ribozymes impliquent parfois certains groupements hydroxyles 2'–OH du ribose soit pour leur structure (Egli *et al.*, 1993), soit pour la catalyse (exemples avec le ribozyme en tête-de-marteau : Perreault *et al.*, 1990). Puisque l'ADN ne contient pas ce groupement, un équivalent ADN des ces ribozymes ne serait pas fonctionnel.

Dans l'ensemble, un suspect semble plus sérieux que l'autre. Entre l'ADN et l'ARN, le dernier a probablement précédé le premier!

## 2.2 ARN versus protéines

Lorsqu'il s'agit de discriminer entre les ARN et les protéines comme premier polymère à la source de la vie, c'est le dilemme de l'œuf ou de la poule! Il faut des protéines pour synthétiser de l'ARN, précurseurs et polymères, alors que sans ARN, il est difficile d'imaginer un processus de traduction et une matrice pouvant produire de nouvelles protéines.

Du point de vue chimique, les protéines possèdent une plus grande diversité. Elles sont composées de 20 acides aminés alors que les ARN se composent de seulement quatre acides nucléiques. Les protéines forment des repliements tridimensionnels qui permettent d'exposer les chaînes latérales des acides aminés et les rendent ainsi disponibles à interagir entre eux, avec d'autres protéines ou avec un substrat. Les ARN ont plutôt tendance à former des structures où les bases s'apparient pour former des doubles hélices. Dans ces hélices les groupements fonctionnels des bases sont difficiles d'accès et le squelette des sucres et des phosphates se trouve majoritairement exposé au milieu. De ce point de vue, les protéines semblent plus adéquates pour remplir des rôles fonctionnels et structuraux plus divers que ne le semblent les ARN. Faut-il ajouter que les protéines aujourd'hui remplissent effectivement une plus grande variété de fonctions et ce de façon très efficace ?

Le domaine de l'enzymologie permet cependant de regarder la question du potentiel chimique d'un polymère face à son ligand d'une autre façon. Plutôt que d'analyser les propriétés du polymère seul, le dilemme peut se concentrer sur le potentiel à lier les ligands (Gilbert et de Souza, 1999). Sous cet angle, les ARN autant que les protéines ont la possibilité de positionner des groupements capables de former des ponts hydrogène ou des interactions non-polaires avec tout ligand nonobstant avec quel niveau de facilité.

Un aspect évolutif vient cependant appuyer l'hypothèse de l'ARN comme premier polymère. White (1976) a observé que les protéines utilisent fréquemment des cofacteurs dont la majorité sont des nucléotides ou des bases hétérocycliques qui peuvent être dérivées des nucléotides. Son hypothèse st que ces cofacteurs constituent des reliques des anciennes enzymes d'ARN avant que le Monde protéique ne supplante celui à ARN.

Le point le plus décisif sur le dilemme entre les ARN et les protéines se situe à savoir lequel des deux polymères a le plus grand potentiel pour s'auto-répliquer. À ce niveau, nul ne peut exclure le grand avantage des ARN : la complémentarité du pairage des nucléotides offre un système simple de réplication fidèle. Ayant ainsi beaucoup de difficultés à envisager comment les protéines pourraient s'auto-répliquer, différents scientifiques ont proposé séparément la même hypothèse de l'évolution de la vie basée sur la catalyse et la réplication de l'ARN (Crick, 1968; Orgel, 1968; Woese, 1967). Cette idée fut historiquement oubliée pendant un certain temps jusqu'à la découverte de l'activité de ribozymes au début des années '80 (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982) qui stimula de vives discussions sur l'origine de la vie (Gray et Cedergren, 1993; Lewin, 1986; Pace et Marsh, 1985; Sharp, 1985). Bien que dans les détails plusieurs propositions diffèrent, une même expression les regroupe maintenant sous le terme de Monde à ARN ou " RNA World " avancé pour la première fois par Gilbert (1986). Depuis, beaucoup de chercheurs tentent d'étendre l'éventail des réactions pouvant être catalysées par l'ARN dans le but de confirmer la probabilité que ce Monde à ARN ait existé. Une grande aide est survenue avec l'avènement des techniques de sélection *in vitro* (SELEX) visant à identifier des ARN capables d'effectuer une activité particulière (Ellington et Szostak, 1990). Avec l'application technique, la possibilité de catalyser un système d'auto-réplication par un ARN semble même possible (Ekland et Bartel, 1996). L'hypothèse d'un Monde à ARN précurseur du Monde ADN / ARN / protéines est maintenant généralement acceptée.

## 2.3 ARN ou plus simple?

L'hypothèse du Monde à ARN n'explique pourtant pas tout de l'origine de la vie. Entre autres, de nombreuses questions restent sans réponse quant à la synthèse pré-biotique des nucléotides (particulièrement le D-énantiomère) et leur assemblage en longs polymères avant que des ARN fonctionnels ne catalysent ces événements. Les doutes sont sérieux et divisent la communauté des chercheurs intéressés au problème de l'origine de la vie en deux groupes (revue par Joyce et Orgel, 1999). L'un, optimiste, croit que les chimistes exagèrent les difficultés et que d'heureuses découvertes établiront des voies raisonnables. L'autre est plus pessimiste et considère la nécessité d'un miracle pour permettre l'apparition des polymères d'oligonucléotides sur une terre primitive. Ces préoccupations ont soulevé la venue d'hypothèses suggérant l'existence d'un système génétique alternatif ayant précédé le Monde à ARN. Quel système? Comment aura-t-il donné naissance ensuite au Monde à ARN? Les réponses sont encore spéculatives mais non sans intérêt.

De nombreuses études peuvent être trouvées sur les propriétés de différents analogues des acides nucléiques (revue par Eschnmoser, 1999; et par Joyce et Orgel, 1999). Parmi ces analogues, deux sont particulièrement étudiés. Le premier est un analogue où le ribose des nucléotides est remplacé par un pyrannose pour constituer un polymère formant aussi des appariements Watson-Crick, l'ARN-p. Dans le second, le squelette ribose-phosphate est remplacé par un squelette fait de liens amides. Ce dernier est nommé PNA pour “ peptide nucleic acid ” et il peut également former des hélices doubles pouvant être complémentaire à des séquences d'ARN ou d'ADN. Si de tels systèmes génétiques ont vraiment existé, il serait donc possible qu'une transition se soit produite par complémentarité des bases vers le Monde à ARN.

## 3. Aspects structuraux des ARN

Qu'il ait été le premier ou non, le Monde à ARN est fortement soupçonné d'avoir existé à un moment dans l'évolution de la vie sur terre. L'étude des ARN, des structures qu'ils forment et des activités qu'ils possèdent est donc cruciale pour notre compréhension des possibilités et limites de ce polymère proprement dit vital!

### 3.1 Structure primaire

La structure primaire d'un polymère d'ARN ou séquence représente la succession ordonnée de ses nucléotides, de 5' en 3'. Par simplicité, seulement l'identité des bases azotées de chaque nucléotide est nommée par les lettres A, C, G ou U (Figure 1 A). La structure primaire d'un ARN dépend directement de la région de l'ADN qui l'encode. Cependant, elle peut être modifiée par des événements de maturation, d'édition ou de modification des nucléotides qui sont généralement post-transcriptionnels. Il est intéressant de noter que l'ensemble de toutes les séquences théoriquement possibles n'est pas représenté dans le monde vivant (Miramontes *et al.*, 1995).

### 3.2 Structure secondaire

Les éléments déterminants de la structure secondaire d'un ARN sont principalement des interactions hydrophobes de superposition des bases et des interactions polaires par des ponts hydrogène. Les ponts hydrogène entre les bases sont particulièrement importants. Ils permettent la formation de paires de bases A–U (deux ponts hydrogène) et G–C (trois ponts hydrogène) similaires à ce qui se retrouvent dans l'ADN. Toutefois, contrairement à l'ADN, un autre appariement de bases est aussi très fréquent et qualifié de paire " wobble " G–U (deux ponts hydrogène). Ces derniers ne déforment pas l'hélice mais semblent y introduire tout de même une plus grande fragilité pouvant être exploitée par des protéines (exemple de la paire de bases G–U dans les ARNt$^{Ala}$ qui constitue un site de reconnaissance pour l'alanine ARNt-synthétase; McClain *et al.*, 1988; Beuning *et al.*, 1997; Chang *et al.*, 1999; Strazewski *et al.*, 1999)

Les autres appariements de bases sont aussi retrouvés dans les structures secondaires mais ils ne peuvent se contenir dans une double hélice sans créer des distorsions. Pour cela, ils sont qualifiés de mis-appariements ou " mismatchs ". Pour l'ADN et pour l'ARN, la combinaison des interactions d'appariement et de superposition des bases amène la formation d'une double hélice, dans le cas de l'ARN une hélice de forme A. Les ARN forment des structures secondaires plus diverses que l'ADN pour lequel la double hélice est la constante. La combinaison de régions en double hélice et de régions où des nucléotides restent en simple brin amène la formation de différentes structures secondaires caractéristiques dont certaines sont décrites ci-après (revue par Burkard *et al.*, 1999b) et schématisées à la Figure 1 B.

**Structure tige-boucle**. Lorsqu'une hélice est formée par un même brin d'ARN qui se replie sur lui-même en laissant quelques nucléotides non-appariés, le terme de tige-boucle ou boucle en épingle à cheveux (" hairpin loop ") est utilisé. Cette structure se retrouve fréquemment dans les ARN. Les ARNt contiennent chacun trois structures tige-boucle dont une possède la boucle qui interagit directement avec les nucléotides des codons des ARNm. Les boucles à quatre nucléotides ou tétranucléotides sont les plus fréquentes dans l'ARN et en particulier les boucles GNRA et UNCG (Woese *et al.*, 1990). La boucle d'une structure tige-boucle peut techniquement contenir n'importe quel nombre de nucléotides non-appariés. Toutefois, les larges boucles risquent souvent de former parmi leurs nucléotides des appariements Watson-Crick, G–U ou même des mis-appariements. Dans ce cas, d'autres termes s'appliquent pour décrire la structure obtenue (revue par Burkard *et al.*, 1999b).

**Figure 1.** Éléments des différents niveaux de structure d'un ARN.

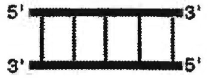
A) Exemple d'une structure primaire d'un ARN ou séquence. B) Schémas de diverses structures. C) Schémas de quelques structures tertiaires pouvant se former dans un ARN.

Adapté de Burkard *et al.* (1999a).

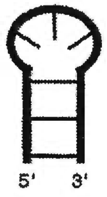
## A) Structure Primaire

5' ... GCUCCCUUAGCAUGGGAGAGUCU ... 3'

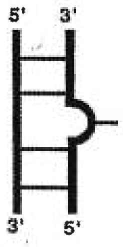
## B) Structures Secondaires



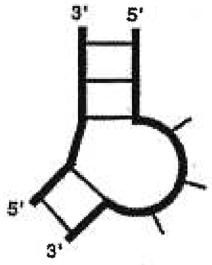
A-form double helix



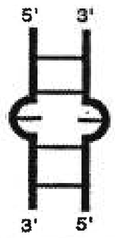
hairpin loop



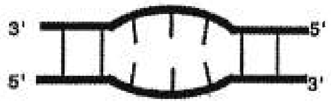
single nucleotide bulge



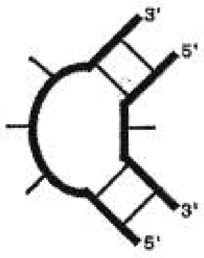
three nucleotide bulge



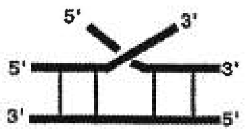
mismatch pair
or, symmetric internal
loop of 2 nucleotides



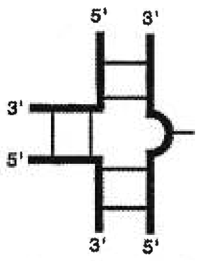
symmetric internal loop



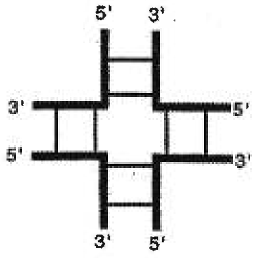
asymmetric internal loop



two-stem junction

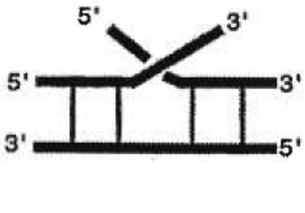


three-stem junction

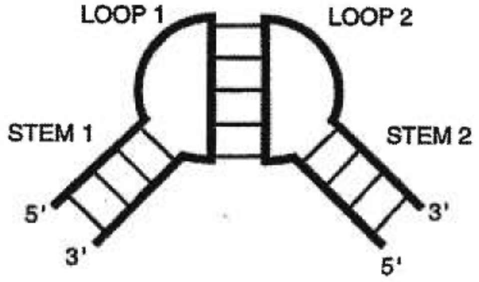


four-stem junction

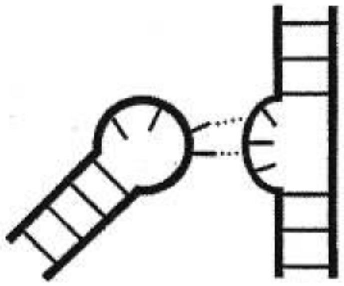
# C) Structures Tertiaires



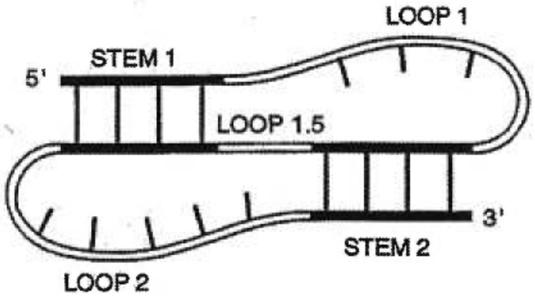
coaxial stack



Kissing hairpins



Hairpin loop - bulge contact



Pseudoknot

**Boucle interne.** Une boucle interne se forme lorsqu'une double hélice d'ARN est interrompue de chaque côté par quelques nucléotides qui ne peuvent s'apparier en A–U, C–G ou G–U. Elle peut être symétrique si chaque côté de l'hélice contient le même nombre de nucléotides non-appariés (parfois nommée mis-appariement pour un ou deux nucléotides de chaque côté de l'hélice) ou asymétrique. Les boucles internes causent des changements dans la structure générale de l'ARN ce qui lui donne la possibilité d'adopter des formes plus diverses et plus flexibles. Elles permettent également de rendre disponibles des nucléotides pouvant participer à des interactions tertiaires avec la même macromolécule d'ARN ou des interactions avec d'autres molécules : substrat, ligand, cofacteur, autre sous-unité, etc. Les protéines liant l'ARN prennent souvent avantage de la distorsion du sillon majeur d'un ARN par la présence de mis-appariements ou d'une boucle interne (Gait et Karn, 1993). Même après une sélection *in vitro*, tous les ARN isolés pour leur capacité d'être liés par la protéine Rev (une protéine du Virus de l'Immunodéficience Humaine – VIH), ou récemment par une protéine à doigts de zinc dérivée de la protéine Zif268, possèdent des boucles internes ou des mis-appariements (Giver *et al.*, 1993a; Giver *et al.*, 1993b; Blancafort *et al.*, 1999) démontrant l'importance de ces structures. Finalement, une boucle interne procure la possibilité de réduire la stabilité de repliement de l'hélice d'ARN où elle se trouve, ce qui peut être avantageux si un réarrangement structural est nécessaire à la fonction de l'ARN.

**Boucle protubérante.** Un ou plusieurs nucléotides non-appariés d'un seul côté d'une double hélice constituent une boucle protubérante ou “ bulge loop ”. Selon les propriétés de superposition de chaque base, une seule pyrimidine protubérante aura tendance à se positionner hors de l'hélice tandis qu'une seule purine protubérante cherchera à s'empiler entre les bases adjacentes dans l'hélice (Burkard *et al.*, 1999b). Une boucle protubérante peut provoquer une courbure de l'hélice dont l'envergure dépendra de la taille de la boucle (Burkard *et al.*, 1999b).

**Jonctions.** Une jonction résulte de la présence de plus de deux hélices reliées en une structure fermée comme par exemple, les trois hélices du ribozyme en tête-de-marteau (" hammerhead ribozyme ") ou les quatre d'un ARNt. Les jonctions sont un facteur marquant dans la détermination des formes que peut adopter un ARN (revue par Burkard *et al.*, 1999b; et par Hermann et Patel, 1999). Des interactions tridimensionnelles importantes s'y produisent comme les superpositions co-axiales (voir ci-après).

### 3.3 Structure tertiaire

L'assemblage de régions de structure secondaire par des interactions de superpositions hydrophobes ou par ponts hydrogène joue un rôle important dans la forme finale qu'adopte un ARN. Parmi ces interactions se retrouvent les superpositions coaxiales (" coaxial stacking ") où les bases aux extrémités de deux hélices se superposent leur permettant de s'aligner l'une sur l'autre, les contacts boucle-boucle (" kissing loops " ou " kissing hairpins ") ou boucle – boucle protubérante et les pseudo-nœuds formés par des nucléotides de la boucle d'une structure tige-boucle s'appariant avec des nucléotides complémentaires localisés à la base de la tige-boucle (Figure 1C) (Burkard *et al.*, 1999b). D'autres interactions comme des appariements triples et quadruples, des motifs de plate-forme (mis-appariements des nucléotides consécutifs d'un brin d'ARN) et des motifs de " zipper " de riboses ont aussi été décrites (revue par Hermann et Patel, 1999).

Évidemment, à cause des fortes charges négatives du squelette de ribose-phosphate de l'ARN et le fait que les structures en hélices exposent le squelette au milieu, les interactions de stabilisation avec des groupements positifs sont grandement favorisées. L'eau joue un rôle important mais les ARN interagissent aussi couramment avec différents cations et en particulier avec le $Mg^{2+}$, rendant ces interactions une partie quasi intégrale de la structure tertiaire d'un ARN (Henkin, 1994; Westhof *et al.*, 1988). Il est possible pour plusieurs ARN d'identifier des sites de liaison spécifiques qui participent soit au bon repliement structural ou même à l'activité catalytique (revue par Feig et Uhlenbeck, 1999; et par Hermann et Patel, 1999).

La description détaillée de tous ces éléments de structure tertiaire sort du cadre de cette thèse mais d'excellentes explications accompagnées d'exemples peuvent être retrouvées dans deux revues récentes : Feig & Uhlenbeck (1999) et Hermann & Patel (1999). Il suffira de mentionner ici qu'elles permettent à des régions distantes de la structure primaire de se rapprocher dans l'espace de façon stable.

La structure tertiaire est la plus fragile des trois niveaux de structure de l'ARN. En effet, des expériences analysant la structure de l'ARN lorsque la température est graduellement élevée ont démontré dans plusieurs cas que la structure tertiaire est la première à se déstabiliser (exemples avec un ribozyme de groupe I : Banerjee *et al.*, 1993; ou avec un ARNt : Crothers *et al.*, 1974; Hilbers *et al.*, 1976).

### 3.4 Interdépendance des niveaux de structure

Il est important de comprendre à travers l'énumération des structures de l'ARN retrouvée ci-dessus, qu'il existe une grande interdépendance entre les niveaux de structure. La structure secondaire est étroitement liée à la structure primaire de l'ARN entre autres à cause de la complémentarité des bases. Ainsi, de faibles changements dans la séquence peuvent entraîner un repliement différent qui changera la stabilité, la flexibilité et/ou l'orientation d'une hélice seule, en interaction avec une autre région de l'ARN ou avec une autre molécule. La disponibilité et l'accessibilité des nucléotides impliqués dans des interactions particulières peuvent alors être compromises. Un exemple pouvant illustrer ce point nous vient des études de repliement de l'ARNr. Plusieurs chercheurs ont observé que les régions initiale (“ leader ”) et d'espacement (“ spacer ”) du transcrit primaire sont nécessaires pour le bon repliement de l'ARNr mature (Balzer et Wagner, 1998; Liiv *et al.*, 1998). Besançon et Wagner (1999) ont même démontré récemment que des interactions importantes surviennent entre la région initiale (" leader ") et l'extrémité 5' du transcrit de l'ARNr 16S permettant à ce dernier d'adopter une conformation adéquate pendant et après sa maturation complète.

Il faut aussi mentionner que la structure secondaire possède également une interdépendance importante avec la structure tertiaire. En effet, certaines interactions tertiaires peuvent favoriser des structures secondaires particulières ce qui augmentera leur stabilité en comparaison à d'autres repliements alternatifs avec des interactions tertiaires moins fortes. Un exemple intéressant est l'effet de la présence de $Mg^{2+}$ qui peut induire le bon repliement de l'ARNt$^{Leu}$ de levure plutôt que la structure secondaire inactive stable qui se formerait en l'absence de $Mg^{2+}$ (Hawkins *et al.*, 1977). De façon similaire avec le domaine P5abc de le ribozyme de l'intron du groupe I de *Tetrahymena thermophila*, l'addition de magnésium induit la formation d'une structure tertiaire contenant de nombreux changements dans les appariements de bases (Wu et Tinoco, 1998). Les auteurs de ce dernier travail ont conclu que le concept du repliement des ARN se produisant d'abord par la formation de la structure secondaire et ensuite par l'organisation des interactions tertiaires n'est pas toujours correct. Enfin, la liaison de protéines peut aussi influencer la structure finale d'un ARN comme la liaison de la protéine S12 qui stimule l'excision d'un intron du bactériophage T4 (Coetzee *et al.*, 1994).

### 3.5 Définition d'un motif d'ARN

Le terme " motif " a été d'abord utilisé pour décrire la structure des protéines. Il est fréquemment employé pour les décrire mais sa signification exacte dépend beaucoup du contexte. Bork et Koonin (1996) ont décrit les quatre différents niveaux de structure pour lesquels l'expression " motif " est utilisée pour décrire des régions de protéines (voir aussi Tableau 1). Le premier, dit petit motif fonctionnel, désigne un très petit nombre d'acides aminés importants pour une activité donnée et qui ont évolué indépendamment du contexte structural environnant, i.e. site de glycosylation. Le second est dit petit motif structural et désigne une région protéique sous un ensemble de contraintes topologiques particulières souvent peu spécifiques (ex : extrémités d'hélices). Le troisième regroupe les motifs fonctionnels et se caractérise par des résidus variables ayant des propriétés semblables et formant des structures caractéristiques comme les régions transmembranaires. Enfin la quatrième catégorie se compose des motifs le plus souvent mentionnés dans la littérature, les motifs conservés. Elle implique la présence d'aspects structuraux et fonctionnels qui permettent de distinguer des protéines spécifiques du reste des protéines existantes ainsi qu'une notion d'ancêtre commun, d'évolution partagée.

De façon semblable, il est possible d'utiliser le terme " motif " pour décrire des éléments de la structure d'un ARN (Tableau 1). Pour l'ARN également, un motif peut être à différents niveaux mais peu de consensus se retrouvent dans la littérature scientifique (Dandekar et Hentze, 1995; Gautheret *et al.*, 1990; Laferrière *et al.*, 1994; Michel et Westhof, 1996). Au niveau de la structure primaire, il s'agirait d'un petit motif de séquence comme, par exemple, des séquences déterminant la demi-vie des ARNm les contenant (" UUAUUUAU " – Caput *et al.*, 1986; ou " AUUUA " – Shaw et Kamen, 1986). Au niveau d'un aspect structural particulier, il est question d'un petit motif de structure. Les mis-appariements G-A constituent un bon exemple de petit motif de structure puisqu'ils sont les mis-appariements les plus communs dans les molécules d'ARN (Gautheret *et al.*, 1994; Hermann et Patel, 1999). Enfin, des motifs structuraux biologiques ou fonctionnels peuvent aussi être décrits pour l'ARN. Nous distinguerons ici, et au Chapitre 1 également, les termes motifs structuraux biologiques et fonctionnels de la façon suivante :

- Un motif représentant une partie d'un ARN dont la structure est importante pour remplir une activité donnée dans un contexte particulier sera dit motif structural biologique. Par exemple, le site de liaison spécifique de la protéine Rev du VIH localisé sur l'ARNm viral et dit RRE (Rev-Responsive

**Tableau 1.** Résumé des différents types de motifs.

| Motifs protéiques : | Exemples : |
|---|---|
| Petit motif fonctionnel | Site de glycosylation |
| Petit motif structural | Extrémité d'une hélice |
| Motif fonctionnel | Région transmembranaire |
| Motif conservé | Site de liaison à l'ADN |

| Motifs d'ARN : | Exemples : |
|---|---|
| Petit motif de structure primaire | Séquence influençant la dégradation des ARNm |
| Petit motif de structure | Mis-appariements G-A |
| Motif structural biologique | Site de liaison à la protéine Tat |
| Motif structural fonctionnel | Ribozyme en tête-de-marteau |

Element; Hadzopoulou-Cladaras *et al.*, 1989; Malim *et al.*, 1989; Hammarskjold *et al.*, 1989; Emerman *et al.*, 1989).

- Un motif pour lequel l'activité biologique a été observée dans plusieurs contextes : *in vitro* et *in vivo* dans différents organismes, sera désigné motif structural fonctionnel. Ce dernier cas convient parfaitement au motif de le ribozyme en tête-de-marteau (revue par Bratty *et al.*, 1993).

## 4. Dégénérescence dans la structure des ARN

### 4.1 Description

Il y a une certaine redondance ou un aspect de dégénérescence associé au repliement des séquences d'ARN en structures secondaires. Waterman a déterminé dès 1978 qu'il était possible de décrire $4^N$ séquences d'ARN de longueur *N* alors que seulement $1,8^N$ structures secondaires différentes peuvent être définies pour ces mêmes molécules (Waterman, 1978). C'est donc dire que l'ensemble de toutes les structures est plus petit que l'ensemble des séquences possibles et plusieurs séquences primaires pourront correspondre à une structure secondaire donnée (Schuster *et al.*, 1994). Il y a de ce fait dégénérescence dans la spécification de la structure secondaire par la structure primaire d'un ARN.

Puisque plusieurs séquences peuvent produire la même structure secondaire, il en va de même pour la formation générale de la structure tertiaire. En effet, tel qu'il peut être observé dans le cas des familles d'ARN fonctionnels (ARNt, ARNr, etc.), plusieurs structures primaires ou séquences peuvent permettre de former une structure tertiaire apte à remplir la même activité. Par exemple, des centaines de séquences différentes existent encodant des ARNt fonctionnant au niveau de la traduction (Sprinzl *et al.*, 1998; http://www.uni-bayreuth.de/departments/biochemie/trna/). Ils doivent tous, pour être fonctionnels, occuper le même espace tridimensionnel général pour ainsi établir des interactions appropriées avec les différents facteurs impliqués dans la traduction et en particulier à l'intérieur des ribosomes. Une autre preuve de la dégénérescence des niveaux de structure se déduit de l'analyse des séquences obtenues suite à des séries de sélection *in vitro* pour une activité de liaison à un substrat donné (Ellington et Szostak, 1990). Les séquences isolées peuvent généralement se classer en groupes partageant une structure secondaire similaire et liant le substrat de la même façon. La structure tertiaire autour du substrat doit forcément être semblable puisque le ligand possède la même structure. Par exemple, Ekland *et al.* (1995) ont isolé, par sélection *in vitro*, des séquences d'ARN pouvant catalyser une réaction de ligation. Ces ARN se regroupent en trois classes de part leur structure secondaire et la spécificité de la région de ligation impliquée dans la liaison. Les auteurs infèrent de ces résultats qu'il existe plusieurs structures d'ARN équivalentes en complexité et en activité.

## 4.2 Implications de la dégénérescence

La première implication de la dégénérescence des niveaux de structure des ARN est la possibilité de retrouver dans différents organismes des ARN possédant différentes séquences tout en partageant une structure très similaire et la même fonction. Cela est effectivement le cas et les ARN identifiés comme responsables d'une même activité constituent tous ensembles une famille. Leur séquence est fréquemment comparée (comparaisons phylogénétiques) pour identifier les régions conservées qui sont susceptibles d'être les plus importantes pour le maintien de la structure ou pour procurer l'activité. Entre autres, il peut être très utile de faire des analyses de covariance des nucléotides situés dans la séquence au niveau des hélices ou des interactions tertiaires soupçonnées. Puisque les structures des ARN présentent une forte dégénérescence, les régions les plus conservées sont les plus susceptibles d'avoir été fixées dans l'évolution pour une raison telle que leur implication dans une fonction particulière :

structurale, de liaison ou catalytique. Une fois que les régions les plus conservées sont identifiées et la structure minimale conservée décrite, il devient plus facile de chercher pour un ARN semblable dans des organismes proches par une recherche de similarité dans les banques de séquences disponibles. En effet, des recherches n'impliquant que la séquence de l'ARN d'intérêt pour retrouver des séquences semblables sont généralement peu efficaces à cause de la dégénérescence possible de la structure primaire d'un ARN fonctionnel vis-à-vis de sa structure secondaire. Des approches impliquant à la fois des aspects de la séquence et de la structure secondaire sont beaucoup plus appropriées (exemple pour les petits ARN nucléolaires : snoRNA, dans Lowe et Eddy, 1999).

Plusieurs groupes ont aussi étudié à fond les aspects théoriques des structures des ARN et de la dégénérescence qu'elles présentent. Fontana *et al.* (1993) puis *Schuster et al.* (1994) ont démontré théoriquement que des structures secondaires typiques peuvent se retrouver dans le proche voisinage de n'importe quelle séquence choisie de façon aléatoire et donc qu'elles sont distribuées partout à travers l'espace des séquences possibles. Plus encore, en mutant seulement 15% des nucléotides d'une séquence donnée, il est possible d'obtenir deux structures aussi éloignées que si elles avaient découlé de deux séquences totalement aléatoires (Fontana *et al.*, 1993; Huynen *et al.*, 1993).

L'étude de la dégénérescence entre la séquence et la structure des ARN amène la notion importante que l'accès à de nouvelles structures par des mutations ponctuelles est beaucoup plus facile qu'anticipé précédemment (Schuster *et al.*, 1994). Il a été démontré que la sélection de nouvelles structures est en fait relativement simple : presque toutes les structures compatibles sont accessibles les unes des autres avec seulement quelques mutations, en moyenne 7,2 mutations, et même les séquences incompatibles peuvent être à une proximité raisonnable d'environ 18 mutations (Schuster *et al.*, 1994). Puisque plusieurs mutations dans la séquence d'un ARN peuvent survenir sans changer la structure secondaire, il s'agit de mutations neutres (Kimura, 1983). Celles-ci permettent une diffusion importante de la séquence ou une exploration de l'ensemble des séquences théoriquement possibles tout en préservant la structure secondaire et la fonction de l'ARN. Ce phénomène donne un accès presque sans limites à de nouvelles structures alternatives qui seront sélectionnées au moment où la nouvelle structure présentera un avantage (Huynen, 1996; Provine, 1986). La dégénérescence des niveaux de structure des ARN implique donc des propriétés bien étudiées théoriquement qui sont importantes d'un point de vue évolutif, en particulier dans le contexte de l'hypothèse du Monde à ARN et de la dérive aléatoire (voir discussion pour un développement plus détaillé de ce point).

## 5. Une famille d'ARN particulière : les ARN de transfert

Les ARNt constituent probablement la famille de petits ARN non-codants (autres qu'ARNm) la plus étudiée jusqu'à maintenant. L'existence des ARNt a été soupçonnée dès 1955 par Francis Crick dans son hypothèse de la présence d'un " adaptateur " pour convertir le langage des acides nucléiques en acides aminés (" *On degenerate templates and the adapter hypothesis* " – première note du "RNA Tie Club"; Watson, 2000). Il aura fallu 10 ans de plus pour que le premier ARNt soit isolé en 1965 (ARNt$^{Ala}$ de levure; Holley *et al.*, 1965). Maintenant, plus de 3000 séquences d'ARNt sont connues (Sprinzl *et al.*, 1998). Leur important rôle dans la cellule, leur structure à la fois petite et complexe ainsi que le fait qu'ils sont fortement soupçonnés d'avoir joué un rôle important dans le Monde à ARN contribuent au grand intérêt que de nombreux chercheurs leur portent.

## 5.1 Fonction

Le rôle principal des ARNt se situe au niveau de la traduction de l'information génétique (Voet et Voet, 1990). Ils permettent de faire correspondre les trois acides nucléiques des codons à l'acide aminé qui doit être inséré dans la protéine naissante selon l'information contenue dans les ARNm. Pour ce faire, ils interagissent directement avec les ARNm au site de traduction : le ribosome. Différents ARNt correspondent aux différents acides aminés et la correspondance est contrôlée par les différentes aminoacyl-ARNt-synthétases puisque celles-ci fixent à l'extrémité 3' des ARNt qu'elles reconnaissent un acide aminé correspondant. Cependant, les ARNt remplissent aussi d'autres fonctions. Ils participent entre autres à la régulation transcriptionelle de plusieurs ARNm encodant des protéines de la voie de synthèse de plusieurs acides aminés (Henkin, 1994) et de la voie de synthèse des dérivés de la porphyrine (Schon *et al.*, 1986). Ils peuvent aussi servir de donneurs d'acides aminés pour la formation d'aminoacylphosphatidylglycérol, de glycyl-lipopolysaccharides (Littauer et Inouye, 1973) ou de certaines protéines nécessitant un transfert terminal (Leibowitz et Soffer, 1969). Certains ARNt servent également d'amorce à la transcription réverse du génome de rétrovirus particulièrement les ARNt$^{\text{Trp}}$ et ARNt$^{\text{Pro}}$ (Dahlberg, 1980; Peters *et al.*, 1977; Waters *et al.*, 1975). Finalement, quelques ARNt$^{\text{Ile}}$ semblent pouvoir agir comme facteurs de transcription (ARNt$^{\text{Ile}}$ dans le ver à soie; Dunstan *et al.*, 1994) ou comme modulateurs du transport des acides aminés possédant une chaîne latérale avec un embranchement (Quay et Oxender, 1980).

## 5.2 Structure

Les ARNt sont de petits ARN d'environ 80 nucléotides (Voet et Voet, 1990). Ils sont transcrits *in vivo* sous forme de précurseurs et un procédé de maturation impliquant de nombreuses enzymes leur permet d'atteindre leur structure fonctionnelle finale. Le processus de maturation diffère dans ses détails s'il survient chez les procaryotes, les eucaryotes ou dans les organelles mais il implique globalement les mêmes étapes. La maturation d'un ARNt implique la coupure des nucléotides superflus en 5' et en 3', la modification de certains nucléotides et dans certains cas, l'excision d'un intron, l'addition de la séquence " CCA " à l'extrémité 3' et l'édition de certains nucléotides (Deutscher, 1995; Martin, 1995; Mazzara et McClain, 1980). La structure primaire finale est donc relativement différente du transcrit synthétisé initialement.

La structure secondaire que les ARNt adoptent est souvent dite canonique puisqu'elle est partagée par la majorité des ARNt. L'expression " en feuilles de trèfle " est généralement utilisée pour décrire son apparence parce que la représentation de la structure secondaire fait effectivement penser à un trèfle à trois feuilles (Figure 2 A). Cinq régions peuvent y être identifiées : la tige acceptrice (" AA ") à l'extrémité de laquelle un acide aminé peut être fixé (tige du trèfle), trois tiges-boucles (les trois feuilles du trèfle) qui sont nommées région D (parce que la boucle contient souvent une base modifiée dite dihydrouridine), région de l'anticodon (" AC ") et région T (parfois aussi TΨC parce que cette succession de nucléotides, dont deux modifiés, se retrouve fréquemment dans la boucle), et enfin une région variable ou bras supplémentaire (⊕) qui fait parfois penser à la quatrième feuille du trèfle. La région de l'anticodon contient les trois nucléotides qui interagissent avec les ARNm (cercles pleins; Figure 2 A). Voici les caractéristiques conservées de la structure des ARNt canoniques (Dirheimar *et al.*, 1995; Steinberg et Cedergren, 1994) :

i) sept paires de bases dans l'hélice acceptrice (" AA ").

ii) deux nucléotides entre les hélices des régions acceptrice et D, constituant le Connecteur 1. Le premier de ces nucléotides est souvent un A ou un U.

iii) trois ou quatre paires de bases dans l'hélice de la région D.

iv) une longueur variable de la boucle de la région D. Les dinucléotides GG ou GA sont fréquemment trouvés à l'intérieur de la boucle.

v) cinq paires de bases Watson-Crick dans l'hélice de l'anticodon (“ AC ”) plus une autre paire de bases non Watson-Crick près de la jonction à quatre hélices (cette dernière paire de bases est plutôt considérée par certains auteurs comme des nucléotides entre les hélices).

vi) sept nucléotides dans la boucle de la région de l'anticodon. Le deuxième nucléotide de la boucle est généralement un U et le sixième un A ou un G.

vii) une longueur variable pour la région du bras supplémentaire constituant le Connecteur 2 (⊕).

viii) cinq paires de bases dans l'hélice de la région T. La paire de bases près de la boucle est habituellement une paire G-C.

ix) sept nucléotides dans la boucle de la région T. La séquence de cette boucle est assez conservée: T✈CNANN, où ✈ représente une pseudouridine.

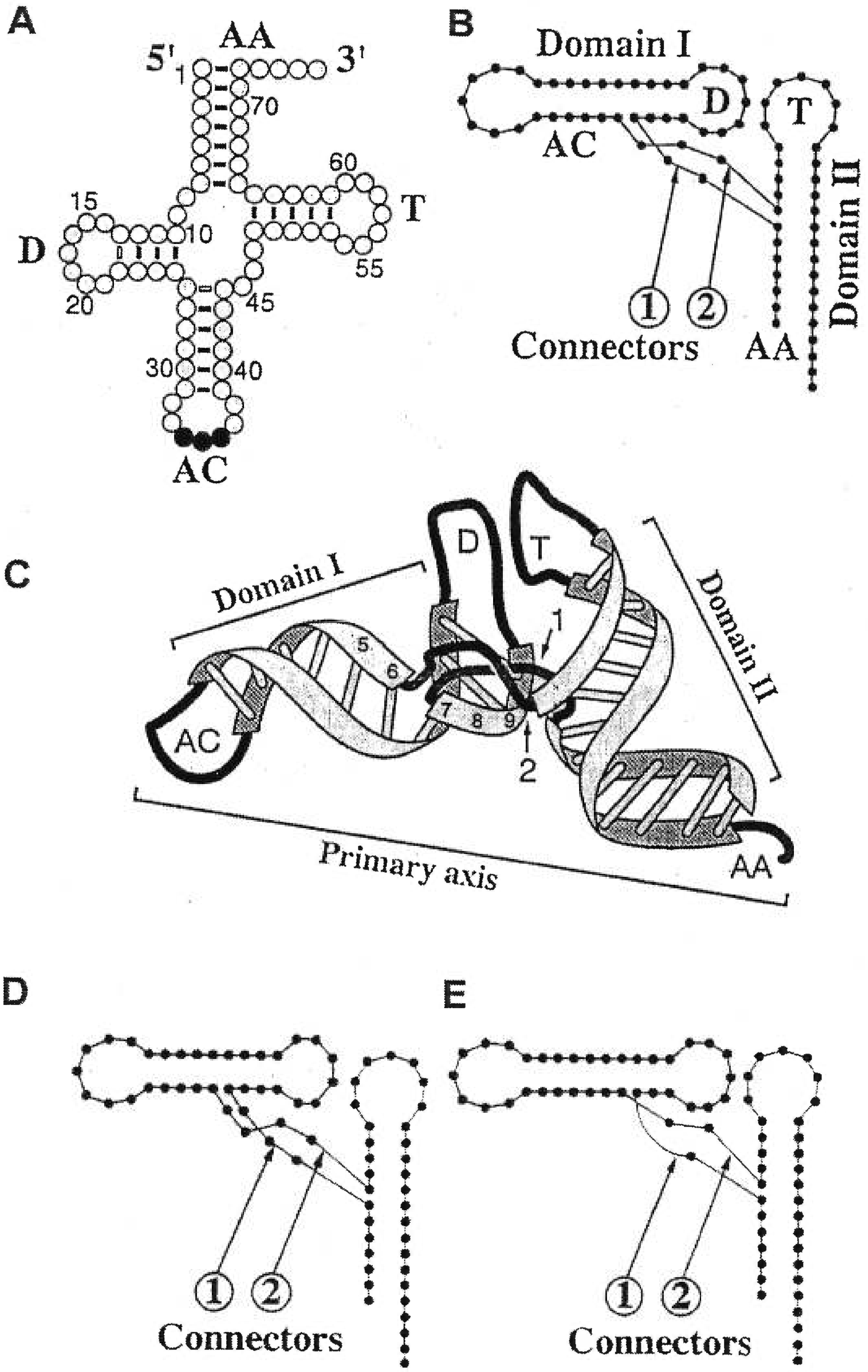
x) aucun nucléotide entre les hélices T et acceptrice.

La structure tertiaire des ARNt est décrite comme un “ L ” (Figure 2 B et C). La première partie du “ L ”, ou Domaine I, se forme par la superposition des hélices des régions de l'anticodon et D. La deuxième partie, le Domaine II, est constituée de la superposition des hélices des régions acceptrice et T. Les Domaines I et II sont reliés par les Connecteurs mentionnés dans les caractéristiques de la structure secondaire. De nombreuses interactions tertiaires maintiennent la structure tridimensionnelle des ARNt en particulier dans le coin du “ L ” ou interagissent les boucles D et T. Parmi les structures tertiaires nommées à la section 3.3, plusieurs ont été identifiées d'abord dans un ARNt (Dirheimar *et al.*, 1995; Hermann et Patel, 1999). Dans le contexte de cette thèse, nous n'entrerons pas plus dans les détails des interactions tertiaires. Il suffit simplement de comprendre que l'ensemble de la structure

**Figure 2.** Schémas des structures secondaire et tertiaire des ARNt.

A) Structure secondaire en “ feuilles de trèfle ” d’un ARNt canonique. B) Schéma en deux dimensions de la structure tertiaire d’un ARNt canonique. C) Schéma en trois dimensions de la structure tertiaire d’un ARNt canonique. D) Schéma en deux dimensions de la structure tertiaire d’un ARNt atypique de type-5 (cinq paires de bases dans l’hélice de l’anticodon). E) Schéma en deux dimensions de la structure tertiaire d’un ARNt atypique de type-7 (sept paires de bases dans l’hélice de l’anticodon).

Adapté de Steinberg et Cedergren (1994).


A
AA
5'
1
3'
70
60
T
15
10
D
55
20
45
30
40
AC
B
Domain I
D
T
AC
Domain II
1
2
Connectors
AA
C
Domain I
D
T
1
5
6
7
8
9
2
AC
Domain II
AA
Primary axis
D
1
2
Connectors
E
1
2
Connectors

permet de positionner l'axe primaire c'est-à-dire les deux régions fonctionnelles des ARNt : l'anticodon et l'extrémité 3' avec son acide aminé.

### 5.3 ARNt atypiques

La structure secondaire en feuilles de trèfle est presque universelle parmi les ARNt. Elle est la règle parmi les ARNt fonctionnant dans les systèmes de traduction cytoplasmique. La situation est cependant très différente parmi les ARNt des mitochondries. Ces derniers présentent une gamme variée de traits singuliers dont des substitutions de bases généralement conservées, des mis-appariements, des hélices plus longues et même la perte des régions D ou T (Dirheimar *et al.*, 1995). Pour certains de ces ARNt atypiques, des repliements secondaires optimisant les appariements des bases ont été développés avec cinq, sept, huit, neuf et dix paires de bases dans l'hélice de l'anticodon (la normale est de six : cinq Watson-Crick et une non Watson-Crick; Steinberg et Cedergren, 1994; Steinberg *et al.*, 1994; Steinberg *et al.*, 1997; Yokogawa *et al.*, 1991). Deux exemples sont schématisés à la Figure 2 (D et E). Dans le Chapitre 4 et 5, ces structures sont qualifiées de Type-5, -7, -8, -9 et -10 selon le nombre de paires de bases dans l'hélice de l'anticodon (nomenclature retrouvée dans Steinberg *et al.*, 1997). Une correspondance intéressante est ressortie de l'observation de ces structures. La longueur de l'hélice de l'anticodon présente une corrélation inverse avec celles des Connecteurs et de l'hélice de la région D (Steinberg et Cedergren, 1994; Steinberg *et al.*, 1997). C'est donc dire que même si la longueur des hélices D et de l'anticodon changent, la constante devrait être le maintien de la dimension du Domaine tridimensionnel I (Figure 2 D et E). Ainsi, la somme des paires de bases ou couches de nucléotides superposés du Domaine formé de l'anticodon et de la région D doit rester de 12. Les changements survenant dans les Connecteurs suivent ceux des hélices de sorte que si la jonction entre les hélices se situe plus près du coin du " L ", ils sont plus courts (Figure 2 E) et inversement si la jonction des hélices est loin (cas avec 5 paires de bases dans l'hélice de l'anticodon, Figure 2 D), ils ont alors besoin d'être plus longs. Somme toute, la structure tertiaire globale conserve les dimensions standards et positionne adéquatement les sites fonctionnels : l'anticodon et le –CCA avec l'acide aminé.

## 5.4 Importance des ARNt dans le Monde à ARN

Selon l'hypothèse du Monde à ARN, à un moment donné de l'évolution de la vie toutes les activités fonctionnelles et de réplication étaient effectuées par des molécules d'ARN (Section 2; Gilbert, 1986). Dans un tel environnement, il n'y a donc pas de pression évolutive pour développer un " adaptateur " de la traduction comme les ARNt ! Weiner et Maizels (1987) ont donc rapidement proposé que les ARNt auraient évolué comme éléments du système de réplication. Étant donnée la structure tertiaire des ARNt en deux domaines ou moitiés, ils ont suggéré que la moitié formée par le Domaine II aurait évolué d'abord comme marqueur génomique (" tag ") à l'extrémité 3' des génomes d'ARN simple brin. La deuxième moitié des ARNt, soit le Domaine I, aurait évolué séparément quand le système de réplication du Monde à ARN serait devenu plus complexe ou avec la venue de la synthèse des protéines lors de la transition au Monde des complexes ribonucléoprotéiques (Monde RNP; Maizels et Weiner, 1999; Schimmel *et al.*, 1993; Schimmel et Ribas de Pouplana, 1995).

# 6. Questions abordées dans cette thèse

## 6.1 Distribution et fonction des motifs d'ARN dans les séquences naturelles

Les études théoriques ont démontré une distribution aléatoire des motifs de structure secondaire dans l'ensemble de toutes les séquences artificielles possibles (Section 4.2). Il demeure donc important de déterminer la distribution des motifs structuraux biologiques ou fonctionnels. En effet, l'évolution aurait-elle pu limiter la présence de certains motifs? Et à l'opposé, peut-on trouver un motif surexprimé dans certaines séquences naturelles?

L'étude des séquences naturelles amène aussi la possibilité d'une fonctionnalité des motifs d'ARN qui peuvent y être identifiés. Est-il possible d'identifier des motifs d'ARN connus dans de nouveaux contextes où ils seraient également actifs? L'identification de la présence d'un motif donné pourrait-elle permettre d'expliquer des observations biologiques encore inexpliquées au point de vue moléculaire?

Avec toutes ces questions en tête, nous avons choisi dans le laboratoire du Dr Cedergren de procéder à une recherche de différents motifs d'ARN dans une vaste base de données de séquences naturelles : GenBank. Notre approche était basée aussi sur l'utilisation d'un programme développé dans notre laboratoire par Gautheret *et al.* (1990) puis Laferrière *et al.* (1994), qui permet de rechercher des motifs d'ARN dont les structures primaire, secondaire et une partie de la tertiaire sont décrites.

Le Chapitre 1 présente la recherche de dix différents motifs d'ARN regroupés en trois catégories : motifs d'ARN liés par des protéines, motifs actifs chimiquement ou au niveau catalytique et motifs d'ARN liant de petites molécules. Une hypothèse intéressante sur l'activité d'un motif identifié dans cette recherche est présentée au Chapitre 2. Enfin, le Chapitre 3 présente une vaste recherche de ribozymes en tête-de-marteau. Trois motifs de départ représentant les trois arrangements possibles de ribozymes en tête-de-marteau agissant en *cis* (centre catalytique et cible dans la même molécule) ainsi que leurs mutants ponctuels sont utilisés.

### 6.2 Des ARNt atypiques peuvent-ils être fonctionnels dans le cytoplasme de *Escherichia coli* ?

Le système de traduction mitochondrial est caractérisé par la présence d'ARNt atypiques qui ne se retrouvent pas normalement dans les systèmes de traduction cytoplasmiques (Section 5.3). Cette observation soulève plusieurs questions. Pourquoi n'ont-ils pas été éliminés des mitochondries? À l'inverse, pourquoi sont-ils absents du cytoplasme? Leur activité est-elle restreinte à l'environnement du système de traduction mitochondrial ou peuvent-ils tout de même être fonctionnels dans un cytoplasme? Est-il possible d'isoler de nouvelles séquences d'ARNt atypiques qui seraient fonctionnelles dans un contexte cytoplasmique?

Dans le but de récolter des informations pouvant aider à résoudre ces questions, nous avons décidé d'évaluer l'activité possible d'une banque d'ARNt atypiques, semblables aux ARNt atypiques de mitochondries, dans le cytoplasme d'*E. coli* (Chapitre 4). L'expression et la sélection, parmi plus de 100 000 séquences différentes, a permis d'isoler de nouveaux ARNt ayant une structure primaire nouvelle, une structure secondaire atypique mais une structure tertiaire similaire à un ARNt normal puisque permettant une activité de traduction.

### 6.3 Pourquoi les ARNt atypiques ne s'expriment que faiblement chez *Escherichia coli* ?

Les nouveaux ARNt atypiques isolés au cours du Chapitre 4 ne présentent qu'une expression faible en comparaison avec l'ARNt contrôle. Ces résultats soulèvent des questions sur l'expression, la stabilité et la dégradation de ces ARNt dans le cytoplasme de *E. coli*. Les ARNt atypiques seraient-ils mal exprimés? Ou sont-ils plus lents à compléter leur maturation? Y a-t-il une dégradation plus rapide des ARNt atypiques matures?

Des expériences supplémentaires de caractérisation de deux des ARNt atypiques actifs isolés précédemment sont présentées au Chapitre 5. Elles montrent l'identification d'un clivage spécifique des ARNt atypiques chez *E. coli*.

# RÉSULTATS

# CHAPITRE 1

## The distribution of RNA motifs in natural sequences

*Nucleic Acids Research* **27**: 4457-4467 (1999)

# The distribution of RNA motifs in natural sequences

Véronique BOURDEAU, Gerardo FERBEYRE, Marie PAGEAU,
Bruno PAQUIN* and Robert CEDERGREN

Département de Biochimie, Université de Montréal,
C.P. 6128, succursale Centre-Ville,
Montréal, QC, Canada, H3C 3J7

*To whom correspondence should be addressed. Tel: +1 514 343 6111 ext. 1938; Fax: +1 514 343 2210; Email: paquinb@magellan.umontreal.ca.

Present address:
Gerardo Ferbeyre, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724.

Dedicated to the late Robert Cedergren.

Received July 16, 1999; Revised and Accepted September 28, 1999

## NOTE

Contribution de chaque auteurs:

V. Bourdeau : a participé à toutes les étapes des recherches, de la construction des pages sur Internet et de l'écriture

G. Ferbeyre : a développé des hypothèses sur des motifs putatifs intéressants et participé à l'écriture

M. Pageau : a automatisé les recherches, développé les programmes SITE et écrit les pages sur Internet des résultats

B. Paquin : a vérifié si une distribution phylogénétique pouvait être trouvée et dirigé la fin du projet

R. Cedergren : a dirigé le projet

CONTENU

I. Abstract

II. Introduction

III. Materials and Methods

- a) The search
- b) The analysis

IV. Results

- a) Protein-binding motifs
- b) Chemically and catalytically active motifs
- c) Small molecule-binding RNA motifs
- d) Phylogenetic analysis

V. Discussion

- a) Origin of the motifs
- b) Evolutionary dynamics of the motifs
- c) The activity and utility of motifs
- d) Taking advantage of fortuitous targets
- e) Conclusion

VI. Acknowledgements

VII. References

VIII. Tables

IX. Figures

## I. ABSTRACT

Functional analysis of genome sequences has largely ignored RNA genes and their structures. We introduce here the notion of 'ribonomics' to describe the search for the distribution of and eventually the determination of the physiological roles of these RNA structures found in the sequence databases. The utility of this approach is illustrated here by the identification in the GenBank database of RNA motifs having known binding or chemical activity. The frequency of these motifs indicates that most have originated from evolutionary drift and are selectively neutral. On the other hand, their distribution among species and their location within genes suggest that the destiny of these motifs may be more elaborate. For example, the hammerhead motif has a skewed organismal presence, is phylogenetically stable and recent work on a schistosome version confirms its *in vivo* biological activity. The under-representation of the valine-binding motif and the Rev-binding element in GenBank hints at a detrimental effect on cell growth or viability. Data on the presence and the location of these motifs may provide critical guidance in the design of experiments directed towards the understanding and the manipulation of RNA complexes and activities *in vivo.*

## II. INTRODUCTION

The realization that conserved amino acid motifs in proteins can often be related to function has greatly aided the evaluation of unidentified open reading frames in sequence databases. Detailed comparison of protein sequence, structure and function has now provided motif databases (1-4), which greatly facilitates the task of inferring function for otherwise uncharacterized coding sequences. As sequences have accumulated, so has the number of recognizable motifs, thereby guaranteeing that an everincreasing role will be played by functional inference or *in silico* analysis of sequence motifs.

RNA remains an enigma in gene sequence research since few systematic attempts have been made to identify RNA coding genes in sequence databases other than those belonging to a few well-known families, such as transfer RNAs (tRNAs) or small nucleolar RNAs (snoRNAs; in particular see ref. 5), as well as other RNAs like Group I and II introns. The degeneracy built into RNA molecules due to their composition being based on only four major nucleotides renders the primary sequence of RNA insufficient, by itself, in defining motifs. Secondary and tertiary structural aspects must therefore be made part of RNA motif definitions. In spite of these complications, evidence is accumulating that RNA motifs will provide the ultimate basis for an understanding of RNA structure and function (6,7).

To introduce our concept of RNA genomics, 'ribonomics', we define and present here the distribution of a number of RNA motifs in the GenBank sequence database (8). Given the lack of consensus on what an RNA motif might be (6,9-11), the use of motif in this work will refer to RNA molecules or parts thereof which have a chemical or ligand-binding activity in a defined context. Our motifs could be called 'functional motifs' because of their demonstrated biological activity under known conditions, but prudence dictates caution: it is unlikely that these motifs would behave similarly in all environments due to factors of conformational variability, accessibility, etc. In light of these complications, we will arbitrarily refer to the motifs under investigation here as 'biological motifs' in order to acknowledge their known activities under certain conditions and to underline their potential to play a biological role in other contexts.

Our search of GenBank employed RNAMOT, a computer search engine developed in our laboratory, which defines primary and secondary structural information within computer readable 'descriptors' (9,10). Previously, features of this strategy have been illustrated in searches to identify tRNAs (9), alternative folding patterns of cytoplasmic tRNAs (12), putative Tat-binding elements (TBEs) in viruses linked to human immunodeficiency infections (13) and a catalytic RNA domain in the repetitive DNA of schistosome (14) and of cricket (Rojas *et al.*, manuscript in preparation). Here, we present extensive searches of the GenBank database for RNA biological motifs implicated in chemical or ligand-binding activities.

## III. MATERIAL AND METHODS

### III. a) The search

The program RNAMOT (9,10), written in C, requires a sequence file in the IUPAC/IUB format and a 'descriptor' defining the motif under investigation. In the course of this study, we have used the release of October 15, 1998 of the sequence data bank: GenBank (NCBI-GenBank flat file release 109.0). Searches were carried out on both strands and all occurrences of motifs involving unidentified bases denoted by N in the database were disregarded. A Power Challenge XL with 32 CPUs IP 19, R4400, 150 MHz processor (3072 Mb) running UNIX IRIX 6.2 was used.

In order to help establish the significance of their presence, frequencies of each motif in the database were compared with frequencies in a random sequence database generated by a uniform pseudo-random number generator (15) with a period length near $2^{121}$. The random sequence databases contained 10 000 sequences of 100 000 nucleotides each; the four nucleotides A, C, G and T were used with equal probabilities. An 'expected' frequency in GenBank (**N**) was calculated from the number of occurrences of each motif in the random databases (**M**) by the following: $\mathbf{N} = (\mathbf{a} \times \mathbf{M})/(10^4 \times 10^5)$, where **a** is the number of nucleotides in GenBank ($2.009 \times 10^9$ in release 109.0).

### III. b) The analysis

Subsequent to the compilation of motif frequency in sequence fragments, the location within the fragment was identified and extracted from the associated documentations by SITE, a suite of programs written in C with Perl, AWK and Bourne Shell scripts. SITE determined the strand sense, and whether the motif was contained partially or wholly within features defined in the documentation of the sequence fragment. From the overall list of some 67 features in GenBank, we have combined and compiled the following subset of features for our classification: 1) mRNAs: sequenced mRNAs or cDNAs, CDS (coding sequences), mat_peptides (maturation peptides) and UTR (untranslated regions); 2) introns; 3) control regions: the CAAT_signals, TATA_signal, enhancers, promoters, -10 and -35_signals; 4) LTR (long terminal repeat); 5) rRNAs (ribosomal RNAs); 6) tRNAs; 7) other RNAs: including pre_RNAs (precursor RNAs), prim_transcript (primary transcript RNAs), guide RNAs, scRNA (small cytoplasmic RNAs) and snRNAs (small nuclear RNAs); 8) satellite & repeat: sequences with satellite DNA features or repeated sequence entries; 9) artificial: patent, synthetic, artificial and oligonucleotide entries not described with features; and finally 10) a miscellaneous category including all other minor features or no description at the location of the result. The occurrence and location files were then converted into the HTML format and linked to GenBank files. The compilation of the results in some cases involved correcting the location assignment of individual sequence entries especially when an mRNA feature is identified, but the motif falls outside identified CDSs, indicating that it is an intron or a UTR.

A possible phylogenetic distribution was also evaluated through the compilation of repeated gene names. For this analysis, the GenBank files EST (expressed sequence tags), GSS (genome survey sequences), HTG (high throughput genome sequencing), PAT (patent sequences), STS (sequence tagged sites), SYN (synthetic and chimeric sequences) and UNA (unannotated sequences) were not considered.

## IV. RESULTS

The lack of a bona fide list of accepted RNA motifs prompted us to make a rather arbitrary selection and definition of motifs. We chose examples of defined individual structures rather than structures derived from the known families of cellular RNAs. Thus with the exception of the known natural occurrences of the hammerhead motif (16) or of the UV-loop motif (17) little or no prior information was available indicating whether these structures would be found in the database. Whenever it was possible, we composed two descriptors which maintained the central core region of the motifs but gave two possible positions for the 5' and 3' ends (see legend of Fig. 1).

### IV. a) Protein-binding motifs

*The Tat-binding element (TBE).* The interaction of the Tat protein with a specific RNA-binding site, the *trans*-activation response region (TAR), is involved in the replication of primate lentiviruses like HIV (human immunodeficiency virus) and SIV (simian immunodeficiency virus) (18,19). Basal transcription from the HIV LTR allows the synthesis of short RNA transcripts, but in the presence of the Tat protein, transcription is enhanced and RNA transcripts are longer (20). The search descriptor for the TBE was defined by the minimal consensus sequence of the TAR element found in different HIV isolates and chemical interference analysis of the binding site (21,22; Fig. 1). In the TBE consensus structure, the two base pairs in the upper and lower stem and the one uridine (U) at the 5' terminus of the internal loop are invariant, since they provide key interactions in the conformation of the TBE when bound to Tat (23).

The RNAMOT search of the GenBank database using the TBE descriptor produced a list of 52 698 occurrences of the motif in its two possible orientations: 26 102 occurrences for the direct orientation and 26 596 for the reverse one (Table 1). After removal of the motifs from the patent and the synthetic or chimeric sequences, 25 518 occurrences of the direct motif and 24 976 of the reverse one are present in the 'natural' GenBank (Table 1). An identical search with the random sequence bank produced 32 320 presences for the direct orientation and 32 887 for the reverse. The frequency of the motif in GenBank is thus on the order of that expected in random sequences or slightly lower. Next, we established the organismal distribution of the motif. From Table 2, it can be seen that the motif is distributed among the organismal classes in roughly the proportion that each class is represented in GenBank (compare 'TBE' versus 'GenBank distribution' columns) and not only in viral sequences where the natural motif has been identified (see also Fig. 2). The distribution of the TBE motif among genetic features was determined and is shown in Table 3. Particularly interesting is the high number of occurrences of the direct motif in LTR features compared with the reverse motif (on the plus strand) and inversely its low representation in the mRNAs of the minus strand whereas the reverse motif has a quite high incidence.

*The Rev-binding element (RBE).* The Rev protein and its binding site, the Rev responsive element (RRE), promote the transport of unspliced transcripts of the HIV genome to the cytoplasm (19,24,25). The primary interaction between Rev and the RRE has been shown to be largely determined by a small, 30 nt region of the RRE, called the Rev-binding element (RBE; 26). Descriptors for the RBE motif were derived from the sequences of RNAs possessing high binding affinity to the Rev protein as isolated by selection from partially randomized sequence pools *in vitro* (see SELEX below; 27). In Figure 1, consensus structures for the four different classes used in our searches are shown. The classes are named according to the non-Watson-Crick interaction bridging two nucleotides in the internal loop (Fig. 1, in bold). Note that the RR+2 class contains a bulge of 2 nt not present in the other classes and that the GGwt class includes the wild-type RBE motif.

The number of occurrences for all the RBE classes in GenBank is: 87 for CA class, which represents 0.31 times what we expected; 122 for RR+2, thus 0.22 times the amount expected; 1059 for AA class with 0.49 times the expected frequency; and 1068 for GGwt class, which is around the expected value (1.01 times; Table 1). The expected number and the observed number of occurrences in GenBank are both quite low except for the GGwt class. The AA class seems to produce twice as many occurrences in the reverse orientation of the motif versus the direct one. From the distribution of the hits in the GenBank files (Table 2; Fig. 2), it is obvious that the frequency of the occurrences of the GGwt class is biased by a huge representation of the motif in the viral sequences because of numerous HIV sequence entries. In fact, if we remove the number of occurrences due to HIV/SIV sequences, the frequency obtained dropped slightly lower than the expected level (Table 1). The distribution among features found in Table 3 shows that the RBE has a relatively high frequency in mRNA and coding sequences.

*The S1-binding motif.* This RNA motif contains a pseudoknot with highly conserved sequence elements in its loops (Fig. 1). The motif binds both the S1 ribosomal protein and the 30S ribosomal subunit from Escherichia coli (28). Such an RNA motif on the 5' UTR of an mRNA might have a regulatory role in translation initiation.

A descriptor of this S1-binding motif was used to search GenBank. Of 135 identified occurrences only two were in the patent or synthetic sequences (Table 1). The remaining 133 represent a frequency slightly higher than what we were expecting (1.38 times the expectation). Surprisingly, the distribution of this motif in GenBank shows a presence higher than expected for a random distribution in the mammalian sequences (especially in primate, Pri, and other mammalian, Mam, files) and the EST, whereas in the bacterial sequences we obtained only half of the expected number (Table 2, Fig. 2). This could mean that the motif has been restricted in bacteria.

### IV. b) Chemically and catalytically active motifs

*The UV-loop motif.* The photoreactive UV-loop motif was adapted from a consensus of similar RNA loop structures found in viroids, 5S rRNA, the sarcin-ricin loop of 28S rRNA and the hairpin ribozyme (17). This internal loop includes a G and a U (Fig. 3, bold) which are covalently cross-linked upon UV radiation (29).

The two descriptors used represent the two orientations (5' or 3') of the motif (Fig. 3). There are 2914 occurrences found in the 'natural' GenBank (1452 occurrences expected, Table 1). The distribution of this motif shows an over-representation in the invertebrates (Table 2, Fig. 2) as well as in rRNA genes (reverse motif; Table 3).

*The hammerhead motif.* The hammerhead ribozyme motif was initially defined as a self-cleaving domain found in plant virusoids and satellite RNAs (reviewed in 16). This motif is composed of three helices surrounding a single-stranded, catalytic core region. Extensive mutagenic analysis has defined the sequence requirements for efficient self-cleavage: changes in the unpaired core region are not tolerated, whereas few sequence restrictions constrain the base paired regions (Fig. 3).

The descriptor used in the GenBank search for the hammerhead motif did not include the base-pairing requirements derived from helices I and III of the consensus hammerhead RNA motif in order that only the catalytic portion of the motif would be found (bold region in Fig. 3). This definition makes it possible to find the substrate portion of the motif at a distant site consistent with a *trans*-cleavage mode *in vivo*, where the cleavage site and the catalytic core could be in different molecules (14). The hammerhead catalytic motif occurs 2788 times in all the GenBank but 85% of these occurrences correspond to artificial sequences, leaving 414 occurrences in the 'natural' GenBank (Table 1), compared to 515 occurrences expected, a ratio of 0.80. The organismal distribution shown in Table 2 and Figure 2 suggests concentrations of hammerhead motifs in both invertebrate and viral sequences, whereas primates seem to be inhospitable to the motif. The motif is also under-represented in the ESTs. The data in Table 3 eloquently support and extend the apparent preference of this motif for repetitive DNA of eukaryotes (14).

*The leadzyme motif.* The leadzyme is a catalytic RNA having the unusual property of being able to cleave a target RNA in the presence of lead, whereas the classical catalytic RNAs require magnesium, manganese or calcium divalent cations (30,31). The original leadzyme was isolated from *in vitro* experiments where partially randomized RNA molecules derived from a tRNA structure were selected for their ability to self-cleave in the presence of lead ion; there are thus no known naturally occurring leadzymes which cleave *in vivo*. [Note that many RNA molecules do show site-specific cleavages in the presence of divalent lead ion *in vitro* (32).] The consensus structure for a catalytically active leadzyme has been determined by extensive chemical and enzymatic characterization (31) and is shown in Figure 3.

Table 1 shows that the frequency of the leadzyme motif is only slightly lower than that expected based on the constraints used in the descriptors (1487/1806 for the direct orientation and 1231/1804 for the reverse one). The motif shows a slight overabundance among the mammals and bacteria (Table 2, Fig. 2), but its presence among the features may be more significant, since an over-representation among mRNA sequences is evidence that the motif is being expressed in RNA and may therefore be involved in some cleavage activity (Table 3).

In contrast to the search for the hammerhead motif above, the entire leadzyme motif, its catalytic and substrate portions, were combined in the descriptor used in this case. If active, leadzyme occurrences in coding sequences would likely be involved in self-cleavage (*cis*-cleavage) of the cited sequence *in vivo*, although transcription of the leadzyme could lead to *trans*-cleavage as well. The fact that the normal intracellular concentration of lead ion would be below the 10-100 µM required for leadzyme activity *in vitro* does not augur well for *in vivo* activity. These identified leadzyme motifs, however, might play a role in lead poisoning.

*The RNA-cleaving DNA enzyme motif.* Santoro and Joyce (33) isolated, by *in vitro* selection, DNA molecules capable of recognizing RNAs by Watson-Crick base pairing and cleaving them. One of them, DNAzyme_8-17, when paired to its target has the consensus structure shown in Figure 3. We decided to use this motif in our search although it is not an RNA motif because single-stranded DNA shares many characteristics with single-stranded RNA (34-36), even though the availability of such a single-stranded DNA motif is unwarranted. Fifty-four occurrences of the DNAzyme_8-17 can be found in the 'natural' GenBank, which represents only 49% of the expected rate (Table 1). Compared to other 'catalytic' RNA motifs, this shows the lowest frequency of occurrences.

### IV. c) Small molecule-binding RNA motifs

*Aptamer motifs.* Aptamers represent a class of RNA molecules that have been isolated and characterized *in vitro* by a technique called SELEX (37). In the first step, a partially or fully randomized pool of RNA molecules is challenged by a potential ligand (37-39). Those RNA molecules bound to the ligand are amplified by PCR after being reverse transcribed into DNA. Cycles of binding and amplification follow until the selected mixture is judged appropriate. Individual molecules are then isolated, sequenced and consensus structures proposed. In the following database searches, we have constructed descriptors based on the consensus structures of the RNA molecules reported in the original publications.

*The neomycin-binding motif B.* Neomycin and other aminoglycosides bind to 16S rRNA in the A site causing misincorporation of amino acids during protein synthesis (40). An oligoribonucleotide, motif A, mimicking the decoding region of 16S rRNA was partially randomized and used by Famulok and Hüttenhoffer (41) for *in vitro* selection. They identified a new group of neomycin-binding RNAs whose consensus structure was named motif B (Fig. 4). This motif was used to screen the GenBank for neomycin-binding sites and a total of 391 occurrences were found among natural sequences, less than half of the 915 expected (Table 1). A particularly high representation of this motif can be noted among the bacterial and vertebrate sequences (Table 2, Fig. 2); however, the expectation that these surplus occurrences might be in rRNA genes finds little support in the present data (Table 3).

*The paromomycin-binding motif.* Paromomycin is another aminoglycoside antibiotic binding to rRNA at the ribosome A site. Recht *et al.* (42) developed a consensus structure of this motif based on the analysis of the critical nucleotides essential for paromomycin (Fig. 4). A total of 831 965 occurrences was found. This number should serve as a warning: the utility of finding a given RNA motif in the database is strictly dependent on the careful definition of the motif descriptor. The paromomycin motif in this case is insufficiently constrained, and the search produces an unmanageable result list. Estimation of the frequency (i.e. in a small database) could avoid a useless search.

*The valine-binding motif.* The valine-binding motif has been exclusively defined by the SELEX technique, and thus there is no evidence that this RNA motif plays any biological role *in vivo* (43). On the other hand, the motif is highly constrained as shown in Figure 4. This motif proves to be very under-represented in GenBank; we found only 24 'natural' occurrences when 98 are expected from the random database (Table 1). These few occurrences are concentrated in the mRNAs (Table 3). Perhaps the clustering of sites in mRNA and the avoidance of this motif signal a functional role for valine.

*The theophylline-binding motif.* RNA ligands have been isolated by Jenison *et al.* (44) following a SELEX for theophylline binding and a counter-SELEX against binding to caffeine. The consensus secondary structure is shown in Figure 4. We use two descriptors that correspond to the two orientations of the theophylline-binding motif found by selection. Putative theophylline-binding motifs were found only 12 times in GenBank. Most of these are in patent and synthetic sequences, leaving only three occurrences of the reverse motif (Table 1). Since the motif contained a lot of constraints, this result is close to what was expected. Because of the low level of incidence, it is not significant to look at the distribution.

*The FMN- and the FAD-binding motifs.* Burgstaller and Famulok (45) performed a SELEX to isolate RNA ligands to flavin adenine mononucleotide (FMN) and flavin adenine dinucleotide (FAD). The consensus motifs are presented in Figure 4. The occurrence of FMN-binding motifs of the direct orientation was comparable to the random distribution (203/177) whereas that of the reverse orientation was slightly higher (255/193) (Table 1). It was found at a frequency nine times higher than expected in vertebrate sequences (Table 2, Fig. 2) and the FAD-binding motif was found more frequently than expected (10/0) (Table 1).

*The ATP-binding motif.* ATP-binding RNAs have been isolated by *in vitro* selection (46). The RNA aptamer consensus obtained recognizes the adenine part of ATP (Fig. 4). The same consensus structure was obtained independently in another experiment that selected RNAs for binding to NAD (45). With two descriptors of the ATP-binding motif, one for each orientation of the motif, we found a total of 7526 'natural' occurrences (Table 1). This is nearly twice that expected in a random situation. Table 2 clearly highlights a high distribution of occurrences in mammalian sequences (primates, rodents and other mammals; see also Fig. 2). A low frequency is observable in invertebrate, plant and bacterial sequences. The ATP-binding motif seems to have no location limitation through the genomes (Table 3).

*The tRNA motif.* We used a general motif for tRNA (Fig. 5) as a positive control for the search. Our scan found 5664 occurrences, which is far in excess over the expected number (none were obtained in the random sequence database we used). Table 2 and Figure 2 show a significantly increased presence of the motif in structural RNA sequences (RNA), invertebrate, plant and bacterial sequences which could come from a bias of known tRNA sequences for these organisms in GenBank. ESTs contain an extremely low frequency of tRNA motifs, as expected. Table 3 ascertains that most tRNAs are in the tRNA feature.

### IV. d) Phylogenetic analysis

We also looked for a possible phylogenetic distribution of these motifs. We were thus searching for multiple occurrences of a particular motif within homologous genes of different species. We found many such examples, but in most cases the distribution of the motifs between closely related species did not match their relationships as inferred from molecular phylogeny. In fact, the evolutionary pattern followed that of the encoded proteins. Two exceptions, however, are the occurrence of the ATP-binding motif in an intron of the adenine phosphoribosyl transferase (APRT) gene of several species of rodents (see below) and the presence of hammerhead motifs in satellite DNAs of several eukaryotes (14,47; Rojas *et al.*, manuscript in preparation).

## V. DISCUSSION

We have reported the results obtained in the building of an RNA motif database. RNA structures have been defined, converted to computer-usable descriptors and used to search the GenBank database. The principal issues rising from these data are the origin of the motifs, the significance of the wide distribution of the RNA motifs in the database and the usefulness of this information.

### V. a) Origin of the motifs

Among the origins for the RNA motifs that can be envisaged are random evolutionary drift of sequences, descent from an ancestral organism (phylogenetic origin) and horizontal transfer between organisms. In the first case, one would expect frequencies of occurrences to reflect the probability of a random formation of motifs and a uniform distribution of motifs throughout natural sequences. Frequencies of motifs in a given organism would be proportional to genome size and the restrictiveness of constraints used in their definition. Motifs would be found indiscriminately among different organismal sources and in transcribed or non-transcribed regions. The distribution of the motifs studied here corresponds very well with these criteria. Moreover, our results confirm the conclusion of Schuster *et al.* (48) that sequences able to fold in the same secondary structure can be found randomly in a space of artificial sequences. They are also consistent with reports showing that a subset of evolved RNAs have a similar distribution of shape elements, as do natural RNAs (49,50). Based on these observations, we favor the random drift origin of the vast majority of RNA motifs in the present version of GenBank. Even if the frequencies of some motifs (RBEs, DNAzyme_8-17, valine-binding motif, UV-loop and the ATP motif) vary significantly from the expected (Table 1), this could reflect an origin by random drift, with strong negative or positive selectivity.

### V. b) Evolutionary dynamics of the motifs

An auxiliary issue to origin is evolutionary flux: can a motif be 'fixed' in a population or is it transitory? Evaluation of this property requires detailed phylogenetic analysis, which is not possible in the current organismally sparse database. The limited analysis we performed confirms quite a wide distribution with few exceptions of conservation such as the ATP motif (see 'ATP in APRT' below) and the hammerheads in satellite DNA. Indeed, in the study of the schistosome catalytic RNA domain, the distribution of the hammerhead motif was determined in related species (14). These data clearly indicate that the hammerhead motif is evolutionarily stable among closely related species of schistosomes. The distribution in conjunction with the biochemical data generated for the motif also shows that the schistosome hammerhead RNA is catalytically active *in vivo*. Thus, even when a motif is generated by a random drift of sequences, this 'evolutionary accident' can be put to profit by the host.

The RNA motifs presented here can be useful or detrimental to the organism and their relatively high (but expected) frequency hints that most of them have little or no effect. Furthermore, it must be kept in mind that distal recognition elements well known *in vivo* have been ignored in our search descriptors. The RBEs, the DNAzyme_8-17 and, in particular, the valine-binding motifs are clearly exceptions. The simplest, obvious explanation for these distributions is that the motifs are somehow detrimental to the host organism.

### V. c) The activity and utility of motifs

The presence of RNA motifs in the database raises important and cogent issues dealing with not only the activity of an RNA motif in a novel context, but also the ability of the organism or an outside entity to take advantage of the motif in that context. It is unlikely that all of the occurrences identified in this work are active since many unfavorable aspects are to be taken into account such as transcription, compartmentalization, alternative folding, co-localization with the interacting protein or molecule and with co-ions, etc. However, because of its intrinsic properties, RNA can accumulate mutations without changing its secondary structure, providing access to new shapes and motifs (48,51). Since RNA structures are dynamic, the presence of a new, putative motif can confer a potential, novel activity to the RNA. Environmental changes can induce alternative folding within the RNA and encourage the formation of the motifs. Thus knowledge of the presence of a motif by itself is useful information because some of these may be bona fide motifs. For example, our recent study of the hammerhead motif found in schistosome repetitive DNA shows that it is expressed and active in schistosomes and may be involved in the regulation of the synaptobrevin-like protein gene via a *trans*-cleavage of the mRNA (14). Other occurrences of the RNA motifs are equally intriguing even if not yet proven. The presence of a neomycin motif B in Giardia could explain its high sensitivity to this antibiotic compared with other eukaryotes (52). In the case of the TBE motif, we have suggested that the presence of putative TBE structures in Kaposi sarcoma associated herpes virus (or HHV8) and hepatitis C virus might be related to the fact that these viral infections are exacerbated by the HIV virus (13). Putative TBEs were also found in viruses like shope fibroma that stimulate HIV replication (53). Here are two particularly interesting occurrences.

*TBE motif in vaccinia virus.* Park *et al.* (54) have reported that HeLa cells express a 'TAR-binding protein' that is a potent inhibitor of the interferon-induced, ribosome-associated protein kinase, PKR, which mediates the antiviral and antiproliferative effects of interferon (55). Vaccinia virus also possesses a similar protein, called E3L whose absence in the replication defective mutant virus can be complemented by the human TAR-binding protein. Park *et al.* (54) have suggested that E3L could provide a means by which the virus could escape the interferon induced antiviral pathway. Finding a TBE motif in the coding region of a subunit of the vaccinia virus RNA polymerase (accession no. VACRNAPSA; position 6000-6034) provides a mechanism: the cellular TAR-binding protein or the viral-encoded counterpart could inhibit PKR by binding to the TBE.

*ATP-binding motif in APRT genes.* Occurrences of the ATP-binding motif have been found in the second intron of the APRT gene in four closely related rodents: Mus pahari (accession no. MPU28721; position 1128-1163), Stochomys longicaudatus (accession no. SLU28723; position 1060-1113), Rattus norvegicus (accession no. RATAPRT; position 1104-1138) and Gerbillus campestris (accession no. GCU28961; position 1193-1227). On phylogenetic trees either based on rodent APRT genes or on other morphological and biochemical analysis (56), the ones with an ATP-binding motif in the intron II cluster in the center of the tree and are phylogenetically linked. In two related rodents, the motif seems to have been lost by a deletion in the intron II of Mus musculus and Mus spicilegus (accession nos M11310 and U28720, respectively). Since the APRT gene is involved in the salvage pathway of adenine synthesis in mammals, the presence of the motif might play a regulatory role.

### V. d) Taking advantage of fortuitous targets

RNA motifs do not have to be used by or be useful to the host cell to provide an important entry point to metabolic manipulation of a cell. As the number of defined small molecule- and protein-binding motifs grows, RNA-based intervention could become the method of choice in inhibiting, stimulating or modulating biological processes. Equally exciting is the possible use of the RNA motif database in the identification of secondary targets, when evaluating drugs for which RNA aptamers have been selected. Since it has been proved that the presence of an RNA aptamer in the mRNA of a given gene inhibit its expression upon binding to the ligand (57), it is quite probable that the fortuitous occurrence of a motif in RNA influences its expression.

### V. e) Conclusion

As the number of unknown sequences accumulates in GenBank databases, recourse to motif search programs will be increasingly useful for *in silico* functional analysis (58). A research strategy based on database searches and experimental approaches was developed. This strategy when coupled to the ability to define new motifs using *in vitro* selection could lead to a virtually limitless source of new concepts in the understanding and use of RNA structures. Using such a strategy, we have already been successful in identifying functional catalytic motifs (14). Likewise, a similar strategy was used by Lowe and Eddy (5) to identify new snoRNA genes in yeast. The occurrences and location files of our searches are available on the web at the URL: http://www.centrcn.umontreal.ca/~bourdeav/Ribonomics.

## VI. ACKNOWLEDGEMENTS

We thank Alice Rae, Nicholas Delihas and Franz Lang for reviewing the manuscript. The authors wish to express their gratitude to NSERC of Canada who financed this project. V.B. holds a doctoral fellowship from NSERC of Canada. The late R.C. was Richard Ivey Scholar of the Canadian Institute for Advanced Research (CIAR) program in Evolutionary Biology. In addition, we thank Bernard Lorazo, Daniel Raymond and André Fourrier of DiTER (Direction des infrastructures Technologiques d'Enseignement et de Recherche) at the Université de Montréal for their assistance.

## VII. REFERENCES

1. Murvai, J., Vlahovicek,K., Barta,E., Szepesvári,C., Acatrinei,C. and Pongor,S. (1999) *Nucleic Acids Res.*, **27**, 257-259.
2. Attwood, T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) *Nucleic Acids Res.*, **27**, 220-225.
3. Henikoff, J.G., Henikoff,S. and Pietrokovski,S. (1999) *Nucleic Acids Res.*, **27**, 226-228.
4. Hofmann, K., Bucher,P., Falquet,L. and Bairoch,A. (1999*) Nucleic Acids Res.*, **27**, 215-219.
5. Lowe, T.M. and Eddy,S.R. (1999) *Science*, **283**, 1168-1171.
6. Michel, F. and Westhof,E. (1996) *Science*, **273**, 1676-1677.
7. Westhof, E., Masquida,B. and Jaeger,L. (1996) *Fold. Des.*, **1**, R78-R88.
8. Benson, D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) *Nucleic Acids Res.*, **26**, 1-7.
9. Gautheret, D., Major,F. and Cedergren,R. (1990) *Comput. Appl. Biosci.*, **6**, 325-331.
10. Laferrière, A., Gautheret,D. and Cedergren,R. (1994*) Comput. Appl. Biosci.*, **10**, 211-212.
11. Dandekar, T. and Hentze,M.W. (1995) *Trends Genet.*, **11**, 45-50.
12. Steinberg, S. and Cedergren,R. (1995) *RNA*, **1**, 886-891.
13. Ferbeyre, G., Bourdeau,V. and Cedergren,R. (1997) *Trends Biochem. Sci.*, **22**, 115-116.
14. Ferbeyre, G., Smith,J.M. and Cedergren,R. (1998) *Mol. Cell. Biol.*, **18**, 3880-3888.
15. L'Écuyer, P. and Andres,T.H. (1997) *Math. Comput. Simulation*, **44**, 99-107.
16. Bratty, J., Chartrand,P., Ferbeyre,G. and Cedergren,R. (1993) *Biochim. Biophys. Acta*, **1216**, 345-359.

17. Burke, J.M. (1996) *Biochem. Soc. Trans.*, **24**, 608-615.

18. Sodroski, J., Rosen,C., Wong-Staal,F., Salahuddin,S.Z., Popovic,M., Arya,S., Gallo,R.C. and Haseltine,W.A. (1985) *Science*, **227**, 171-173.

19. Cullen, B.R. and Greene,W.C. (1989) *Cell*, **58**, 423-426.

20. Karn, J. and Graeble,M.A. (1992) *Trends Genet.*, **8**, 365-368.

21. Weeks, K.M., Ampe,C., Schultz,S.C., Steitz,T.A. and Crothers,D.M. (1990) *Science*, **249**, 1281-1285.

22. Weeks, K.M. and Crothers,D.M. (1991) *Cell*, **66**, 577-588.

23. Puglisi, J.D., Tan,R., Calnan,B.J., Frankel,A.D. and Williamson,J.R. (1992) *Science*, **257**, 76-80.

24. Zapp, M.L. and Green,M.R. (1989) *Nature*, **342**, 714-716.

25. Cullen, B.R. and Malim,M.H. (1991) *Trends Biochem. Sci.*, **16**, 346-350.

26. Tan, R., Chen,L., Buettner,J.A., Hudson,D. and Frankel,A.D. (1993) *Cell*, **73**, 1031-1040.

27. Giver, L., Bartel,D., Zapp,M., Pawul,A., Green,M. and Ellington,A.D. (1993) *Nucleic Acids Res.*, **21**, 5509-5516.

28. Ringquist, S., Jones,T., Snyder,E.E., Gibson,T., Boni,I. and Gold,L. (1995) *Biochemistry*, **34**, 3640-3648.

29. Branch, A.D, Benenfeld,B.J., Baroudy,B.M., Wells,F.V., Gerin,J.L. and Robertson,H.D. (1989) *Science*, **243**, 649-652.

30. Pan, T. and Uhlenbeck,O.C. (1992) *Nature*, **358**, 560-563.

31. Chartrand, P., Usman,N. and Cedergren,R. (1997) *Biochemistry*, **36**, 3145-3150.

32. Ciesiolka, J., Michalowski,D., Wrzesinski,J., Krajewski,J. and Krzyzosiak,W.J. (1998) *J. Mol. Biol.*, **275**, 211-220.

33. Santoro, S.W. and Joyce,G.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 4262-4266.

34. Paquette, J., Nicoghosian,K., Qi,G.R., Beauchemin,N. and Cedergren,R. (1990) *Eur. J. Biochem.*, **189**, 259-265.

35. Breaker, R.R. and Joyce,G.F. (1994) *Chem. Biol.*, **1**, 223-229.

36. Cuenoud, B. and Szostak,J.W. (1995) *Nature*, **375**, 611-614.

37. Tuerk, C. and Gold,L. (1990) *Science*, **249**, 505-510.

38. Joyce, G.F. (1989) *Gene*, **82**, 83-87.

39. Ellington, A.D. and Szostak,J.W. (1990) *Nature*, **346**, 818-822.

40. Davies, J. and Davis,B.D. (1968) *J. Biol. Chem.*, **243**, 3312-3316.

41. Famulok, M. and Huttenhofer,A. (1996) *Biochemistry*, **35**, 4265-4270.

42. Recht, M.I., Fourmy,D., Blanchard,S.C., Dahlquist,K.D. and Puglisi,J.D. (1996) *J. Mol. Biol.*, **262**, 421-436.

43. Majerfeld, I. and Yarus,M. (1994) *Nature Struct. Biol.*, **1**, 287-292.

44. Jenison, R.D., Gill,S.C., Pardi,A. and Polisky,B. (1994) *Science*, **263**, 1425-1429.

45. Burgstaller, P. and Famulok,M. (1994) *Angew Chem. Int. Ed. Engl.*, **33**, 1084-1087.

46. Sassanfar, M. and Szostak,J.W. (1993) *Nature*, **364**, 550-553.

47. Green, B., Pabon-Peña,L., Graham,T.A., Peach,S.E., Coats,S.R. and Epstein,L.M. (1993) *Mol. Biol. Evol.*, **10**, 732-750.

48. Schuster, P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) *Proc. R Soc. Lond. B Biol. Sci.*, **255**, 279-284.

49. Fontana, W., Konings,D.A., Stadler,P.F. and Schuster,P. (1993) *Biopolymers*, **33**, 1389-1404.

50. Reidys, C., Stadler,P.F. and Schuster,P. (1997) *Bull. Math. Biol.*, **59**, 339-397.

51. Huynen, M.A. (1996) *J. Mol. Evol.*, **43**, 165-169.

52. Andrews, B.J., Panitescu,D., Jipa,G.H., Vasile-Bugarin,A.C., Vasiliu,R.P. and Ronnevig,J.R. (1995) *Am. J. Trop. Med. Hyg.*, **52**, 318-321.

53. Tseng, C.K., Hughes,M.A., Hsu,P.L., Mahoney,S., Duvic,M. and Sell,S. (1991) *Am. J. Pathol.*, **138**, 1149-1164.

54. Park, H., Davies,M.V., Langland,J.O., Chang,H.-W., Nam,Y.S., Tartaglia,J., Paoletti,E., Jacobs,B.L., Kaufman,R.J. and Venkatesan,S. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 4713-4717.

55. McMillan, N.A., Chun,R.F., Siderovski,D.P., Galabru,J., Toone,W.M., Samuel,C.E., Mak,T.W., Hovanessian,A.G., Jeang,K.T. and Williams,B.R. (1995) *Virology*, **213**, 413-424.

56. Fieldhouse, D., Yazdani,F. and Golding,G.B. (1997) *Heredity*, **78**, 21-31.

57. Werstuck, G. and Green,M.R. (1998) *Science*, 282, 296-298.

58. Segovia, L. (1998) *Nature Biotechnol.*, **16**, 25.

59. Leclerc, F., Cedergren,R. and Ellington,A.D. (1994) *Nature Struct. Biol.*, **1**, 293-300.

## VIII. TABLES

### Table 1. Frequency of RNA motifs in GenBank.

The number of occurrences found in the GenBank for both orientations of the RNA motifs is shown together (total column) or separately (direct or reverse column). The ‘natural' GenBank refers to the database after removal of entries that are patented sequences (PAT) and synthetic or chimeric sequences (SYN). These values are compared to the results obtained in a random sequence database as indicated in ‘expected frequencies' (see Material and Methods). The ratio is the number of occurrences found in the ‘natural' GenBank over what was expected. *, values excluding HIV and SIV sequences for RBE_GGwt-d.

| | Frequency in GenBank | | Frequency in 'natural' GenBank | Expected frequency | Ratio frequency over expected |
|---|---|---|---|---|---|
| | Total | Direct or reverse | | | |
| FAD | 10 | - | 10 | 0 | - |
| Theophylline-d | 12 | 3 | 0 | 4 | 0 |
| Theophylline-r | | 9 | 3 | 2 | 1.50 |
| Valine-d | 25 | 9 | 9 | 55 | 0.16 |
| Valine-r | | 16 | 15 | 43 | 0.35 |
| DNAzyme_8-17 | 54 | - | 54 | 110 | 0.49 |
| RBE_CA-d | 87 | 35 | 35 | 155 | 0.23 |
| RBE_CA-r | | 52 | 51 | 136 | 0.38 |
| RBE_RR$_{+2}$-d | 122 | 16 | 10 | 47 | 0.21 |
| RBE_RR$_{+2}$-r | | 106 | 10 | 43 | 0.23 |
| S1 | 135 | - | 133 | 96 | 1.38 |
| Neomycin | 402 | - | 391 | 915 | 0.43 |
| FMN-d | 469 | 209 | 203 | 177 | 1.15 |
| FMN-r | | 260 | 255 | 193 | 1.32 |
| RBE_AA-d | 1059 | 369 | 357 | 1019 | 0.35 |
| RBE_AA-r | | 690 | 641 | 1025 | 0.63 |
| RBE_GGwt-d | 1068 | 860<br>(361)* | 792<br>(356)* | 486 | 1.63<br>(0.73)* |
| RBE_GGwt-r | | 208 | 202 | 533 | 0.38 |
| Hammerhead | 2788 | - | 414 | 515 | 0.80 |
| Leadzyme-d | 2808 | 1517 | 1487 | 1806 | 0.82 |
| Leadzyme-r | | 1291 | 1231 | 1804 | 0.68 |
| UV-loop-d | 2956 | 1565 | 1543 | 718 | 2.15 |
| UV-loop-r | | 1391 | 1371 | 734 | 1.87 |
| ATP-d | 7693 | 3438 | 3319 | 1918 | 1.73 |
| ATP-r | | 4255 | 4207 | 2085 | 2.02 |
| tRNA | 5841 | - | 5664 | 0 | - |
| TBE-d | 52 698 | 26 102 | 25 518 | 32 320 | 0.79 |
| TBE-r | | 26 596 | 24 976 | 32 887 | 0.76 |
| Paromomycin-d | 831 965 | 418 554 | 407 954 | 466 255 | 0.87 |
| Paromomycin-r | | 413 411 | 404 188 | 462 539 | 0.87 |

**Table 2. The organismal distribution of RNA motifs in GenBank.**

Percentage of occurrences in the different groups of organisms or subparts of the 'natural' GenBank for the motifs having a total number of occurrences over 100 in Table 1. The GenBank distribution is the relative size of each GenBank group relative to the total size of the 'natural' GenBank. Pri, primate sequences; Rod, rodent sequences; Mam, other mammalian sequences; Vrt, other vertebrate sequences; Inv, invertebrate sequences; Pln, plant sequences (including fungi and algae); Bct, bacterial sequences; Vrl, viral sequences; Phg, phage sequences; Rna, structural RNA sequences; Est, expressed sequence tag sequences; Misc., genome survey, high throughput genomic sequencing, sequence tagged site and unannotated sequences.

| | Pri | Rod | Mam | Vrt | Inv | Pln | Bct | Vrl | Phg | Rna | Est | Misc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 21 | 2 | 5 | 2 | 5 | 7 | 3 | 1 | 0 | 0 | 41 | 14 |
| Neomycin | 11 | 5 | 1 | 8 | 10 | 5 | 15 | 6 | 0 | 0 | 22 | 18 |
| FMN-d | 17 | 3 | <1 | 2 | 7 | 10 | 4 | 1 | 0 | 0 | 31 | 24 |
| FMN-r | 9 | 9 | 1 | 9 | 8 | 7 | 7 | 3 | 0 | 0 | 28 | 17 |
| DNAzyme_8-17 | 11 | 0 | <1 | 2 | 4 | 11 | 24 | 6 | 0 | 0 | 33 | 9 |
| RBE_AA-d | 18 | 6 | 6 | 1 | 8 | 7 | 8 | 3 | 0 | 0 | 22 | 20 |
| RBE_AA-r | 16 | 2 | 1 | <1 | 5 | 4 | 6 | 1 | <1 | 0 | 47 | 17 |
| RBE_GG-d | 5 | 2 | <1 | 1 | 1 | 1 | 4 | 63 | 0 | 0 | 17 | 5 |
| RBE_GG-r | 17 | 6 | 3 | 1 | 4 | 2 | 15 | 6 | 0 | 0 | 29 | 15 |
| Hammerhead | 3 | 1 | 0 | 2 | 20 | 9 | 9 | 26 | <1 | 0 | 21 | 7 |
| Leadzyme-d | 17 | 5 | 2 | 2 | 3 | 3 | 10 | 2 | <1 | 0 | 43 | 13 |
| Leadzyme-r | 14 | 5 | 2 | 1 | 6 | 4 | 11 | 3 | <1 | 0 | 42 | 12 |
| UV-loop-d | 14 | 2 | 1 | 1 | 15 | 7 | 5 | 3 | 0 | <1 | 28 | 24 |
| UV-loop-r | 14 | 2 | 1 | 1 | 17 | 11 | 5 | 2 | <1 | 1 | 21 | 25 |
| ATP-d | 25 | 5 | 2 | 1 | 2 | 3 | 4 | 2 | 0 | <1 | 37 | 20 |
| ATP-r | 26 | 4 | 1 | 1 | 3 | 2 | 4 | 2 | <1 | 0 | 35 | 20 |
| tRNA | 2 | 1 | <1 | 1 | 15 | 30 | 38 | 0 | 1 | 5 | 1 | 5 |
| TBE-d | 15 | 3 | 1 | 1 | 7 | 8 | 7 | 6 | <1 | <1 | 34 | 19 |
| TBE-r | 13 | 4 | 1 | 1 | 7 | 8 | 7 | 3 | <1 | <1 | 37 | 19 |
| Paromomycin-d | 15 | 3 | 1 | 1 | 8 | 8 | 6 | 2 | <1 | <1 | 35 | 20 |
| Paromomycin-r | 15 | 3 | 1 | 1 | 8 | 7 | 6 | 3 | <1 | <1 | 34 | 21 |
| GenBank distribution | 15 | 3 | 1 | 1 | 7 | 7 | 6 | 3 | <1 | <1 | 36 | 21 |

**Table 3. Distribution of RNA motifs among GenBank features.**

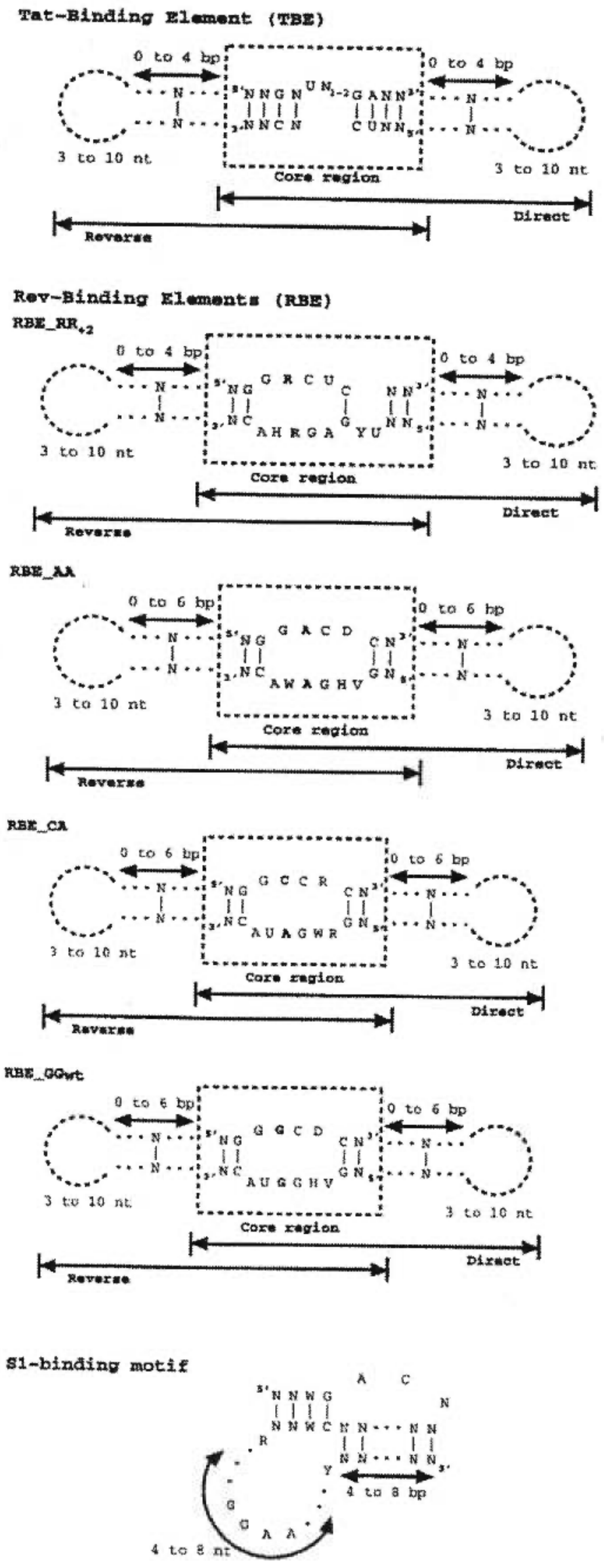
Occurrences of the RNA motifs in different gene regions as annotated in the feature section of the GenBank report of each entry (see Material and Methods). Plus strand, sequence submitted to GenBank; Minus strand, complementary sequence from the sequence submission; Misc., not described. Note that on the minus strand the ‘other RNAs' column includes rRNA and tRNA.

| | Plus Strand | | | | | | | | | | Minus Strand | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mRNA | introns | Control region | LTR | rRNA | tRNA | others RNAs | Satellite & repeat | artificial | Misc. | mRNA | introns | Control region | others RNAs | artificial | Misc. |
| FAD | 3 | | | | | | | | | 4 | | | | | | 3 |
| Theophylline-d | | | | | | | | | 3 | | | | | | | |
| Theophylline-r | 1 | | | | | | | | 6 | | | | | | | 2 |
| Valine-d | 2 | | | | | | | | | 2 | 1 | | | | | 4 |
| Valine-r | 3 | | | | | | | | 1 | 6 | | | | | | 6 |
| DNAzyme 8-17 | 16 | | | 1 | | | | | | 10 | 1 | | | | | 26 |
| RBE_CA-d | 4 | 1 | | | | | | | | 8 | 2 | | | | | 20 |
| RBE_CA-r | 16 | | | | | | | | | 9 | | | | | | 27 |
| RBE_RR$_{+2}$-d | 5 | | | | | | | | 6 | 2 | 1 | | | | | 2 |
| RBE_RR$_{+2}$-r | 3 | | | | | | | | 96 | 2 | | | | | | 5 |
| S1 | 38 | | | | | | | | 2 | 29 | 4 | 2 | | | | 60 |
| Neomycin B | 89 | 8 | 2 | | 5 | | | | 7 | 76 | 14 | | | | 5 | 196 |
| FMN-d | 62 | 2 | | | | | | | 4 | 45 | 4 | 4 | | | 2 | 86 |
| FMN-r | 74 | 4 | | | | | | | 27 | 49 | 7 | 1 | | | 1 | 97 |
| RBE_AA-d | 126 | 3 | | | | | | 1 | 3 | 76 | 12 | 2 | | | 6 | 140 |
| RBE_AA-r | 186 | 5 | 2 | | 2 | | | | 43 | 107 | 16 | | | | | 329 |
| RBE_GG-d | 512 | 7 | | | 2 | | | | 52 | 132 | 5 | | | | 3 | 147 |
| RBE_GG-r | 61 | 2 | | | 2 | | | | 2 | 54 | 3 | 2 | | | 1 | 81 |
| Hammerhead | 115 | 4 | | | 3 | | | 30 | 2344 | 88 | 16 | 1 | | | 19 | 168 |
| Leadzyme-d | 621 | 9 | 1 | | 4 | | 1 | 4 | 10 | 212 | 33 | 6 | | | 14 | 602 |
| Leadzyme-r | 514 | 17 | 1 | | 4 | 1 | 2 | 1 | 17 | 166 | 24 | 2 | | | 33 | 509 |
| UV-loop-d | 292 | 32 | 1 | 2 | 4 | | 1 | 4 | 3 | 441 | 17 | 15 | | | 2 | 751 |
| UV-loop-r | 252 | 27 | 2 | | 35 | | 4 | 3 | 6 | 359 | 31 | 11 | | 1 | 7 | 653 |
| ATP-d | 1155 | 57 | 5 | | 8 | | 3 | 5 | 18 | 874 | 48 | 21 | | 4 | 16 | 2041 |
| ATP-r | 942 | 40 | 2 | | 6 | | | 17 | 39 | 665 | 28 | 13 | | 1 | 51 | 1634 |
| tRNA | 31 | 5 | 13 | | 17 | 3431 | 49 | | 128 | 548 | 4 | 4 | 1 | 1082 | 9 | 519 |
| TBE-d | 6484 | 303 | 82 | 503 | 118 | 24 | 50 | 33 | 226 | 5406 | 382 | 124 | | 25 | 160 | 12 182 |
| TBE-r | 6198 | 294 | 44 | 10 | 133 | 6 | 99 | 41 | 202 | 5352 | 1072 | 111 | | 12 | 285 | 12 737 |

## IX. FIGURES

### Figure 1. The protein-binding motifs.

(Top) Secondary structure of the TBE. (Middle four) Secondary structure of the four Rev-binding elements: RBE_RR+2, RBE_AA, RBE_CA and RBE_GGwt, which were named according to the postulated interaction of the bold nucleotides (59). Note that RBE_GGwt includes the wild-type motif found in HIV and SIV. (Bottom) Structure of the S1-binding motif. Whenever possible, the RNA motifs were given two orientations (direct and reverse) by maintaining the core region and by varying the position of the stem-loop that completes the motif. The letter code represents the following nucleotides: B = C, G or U; D = A, G or U; H = A, C or U; K = G or U; M = A or C; N = A, C, G or U; R = A or G; S = C or G; V = A, C or G; W = A or U; Y = C or U (IUPAC-IUB code). Note that N-N constraints imply Watson-Crick or G-U pairing. T and U are considered the same.


Tat-Binding Element (TBE)

Rev-Binding Elements (RBE)

RBE_RR+2

RBE_AA

RBE_CA

RBE_GGwt

S1-binding motif

**Figure 2. The organismal distribution of RNA motifs in GenBank.**

Graphic representation of the percentages of occurrences of the RNA motifs and the GenBank distribution as shown in Table 2.

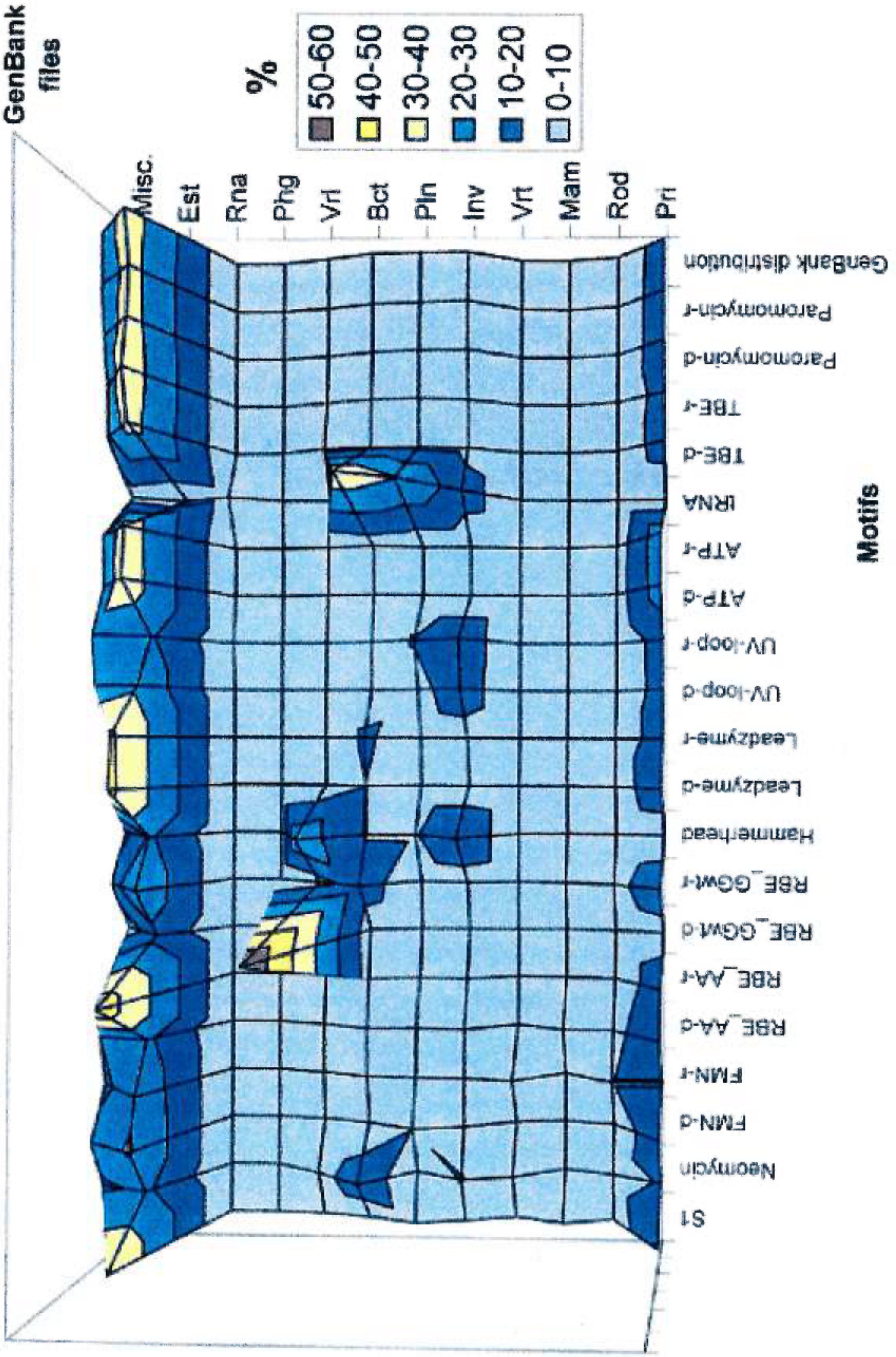
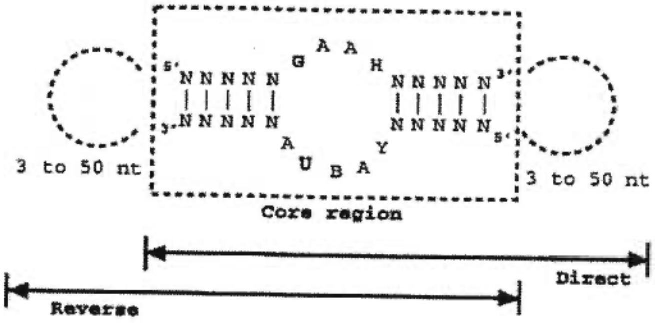

GenBank files
%
50-60
40-50
30-40
20-30
10-20
0-10
Misc.
Est
Rna
Phg
Vrl
Bct
Pln
Inv
Vrt
Mam
Rod
Pri
GenBank distribution
Paromomycin-r
Paromomycin-d
TBE-r
TBE-d
tRNA
ATP-r
ATP-d
UV-loop-r
UV-loop-d
Leadzyme-r
Leadzyme-d
Hammerhead
RBE_GGwt-r
RBE_GGwt-d
RBE_AA-r
RBE_AA-d
FMN-r
FMN-d
Neomycin
S1
Motifs

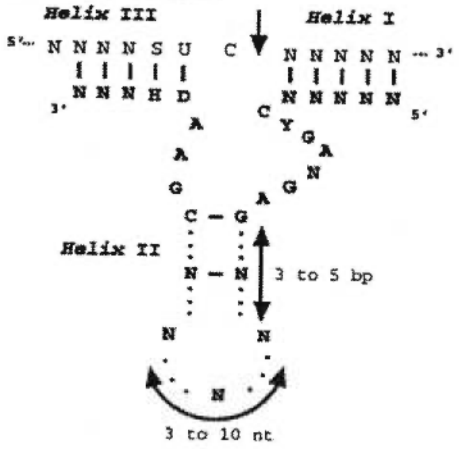
**Figure 3. Chemically and catalytically active motifs.**

(Top) The secondary structure of the UV-loop motif is presented. The nucleotides in bold are involved in cross-linking upon UV irradiation. (Second from top) The hammerhead motif is shown in bold with its target RNA. This motif is composed of three helical regions Helix I, II, and III surrounding a catalytic, non-helical region. In searches with this motif, only the lower bold part of the structure has been encoded in the descriptor to allow distant potential substrates to be found. (Third from top) The leadzyme motif, note that in contrast to the hammerhead motif, both catalytic and substrate portions of this motif were encoded into the descriptor for the searches. (Bottom) The DNAzyme_8-17 motif is presented in bold with its target RNA. As with the hammerhead, only the portion in bold was used in searches. As explained in Figure 1, two orientations were given to the motifs when possible (direct and reverse). The arrows indicate the position of the catalytic cleavage site. The letter code is defined in the legend of Figure 1.

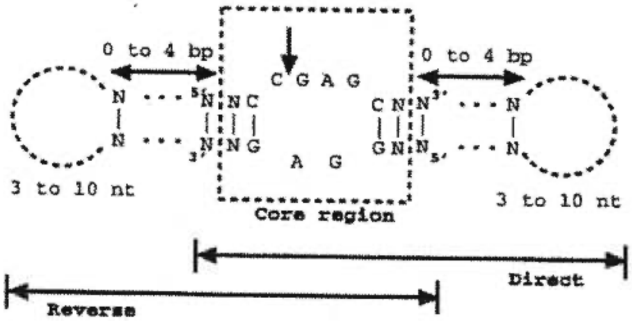
UV-loop motif


5' N N N N N G A A H N N N N N 3'
3' N N N N N A U B A Y N N N N N 5'
3 to 50 nt
3 to 50 nt
Core region
Direct
Reverse

Hammerhead motif



Leadzyme motif



DNAzyme_8-17 motif

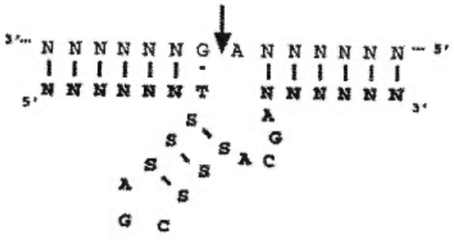
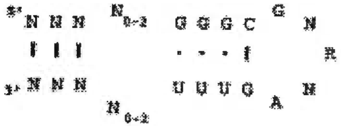

3'... N N N N N N G A N N N N N N ... 5'
5' N N N N N N T N N N N N N 3'
S S S S S S A G C A G C
A G C

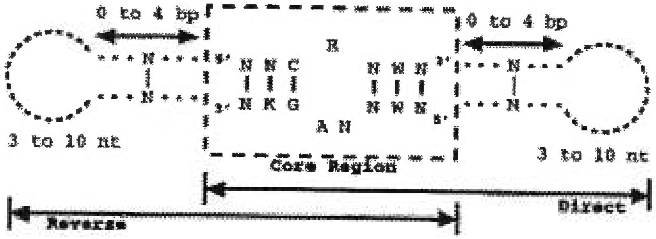
**Figure 4. Small molecule-binding RNA motifs.**

The structure of the neomycin-binding motif B (top left), the paromomycin motif (second left), the FMN- (flavine adenine mononucleotide) binding motif (third left), the flavine adenine (FAD) motif (bottom left), the valine-binding motif (top right), the theophylline-binding motif (middle right) and the ATP-binding motif (bottom right) are presented. In the ATP-binding motif, the asterisks indicate that one mismatch is permitted in the stem. The letter code is defined in the legend of Figure 1 and two orientations were given to the motifs when possible (direct and reverse).

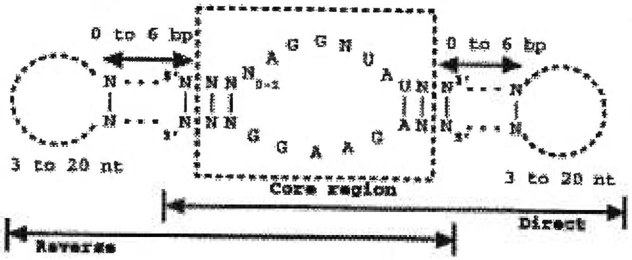
Neomycin-binding motif B



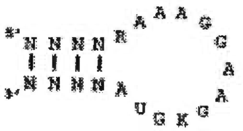
Paromomycin-binding motif



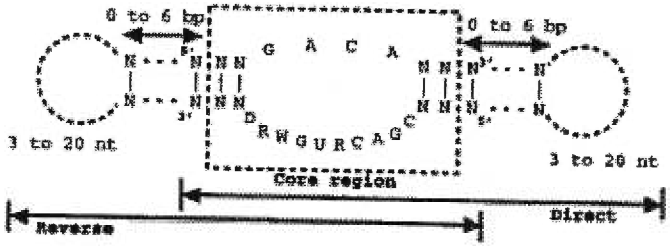
FMN-binding motif



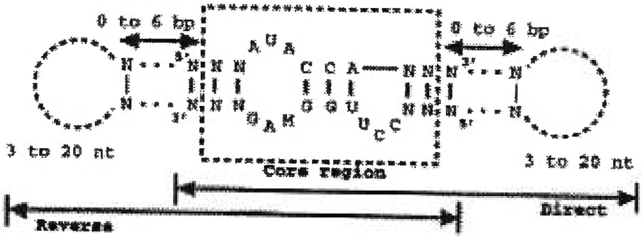
FAD-binding motif



Valine-binding motif



Theophylline-binding motif



ATP-binding motif

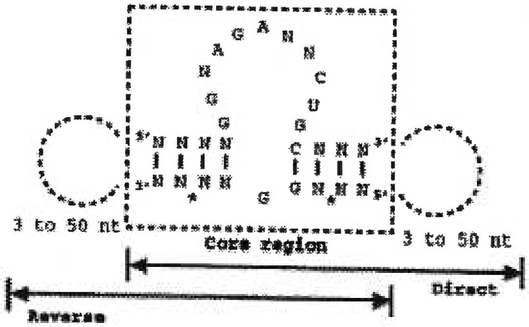
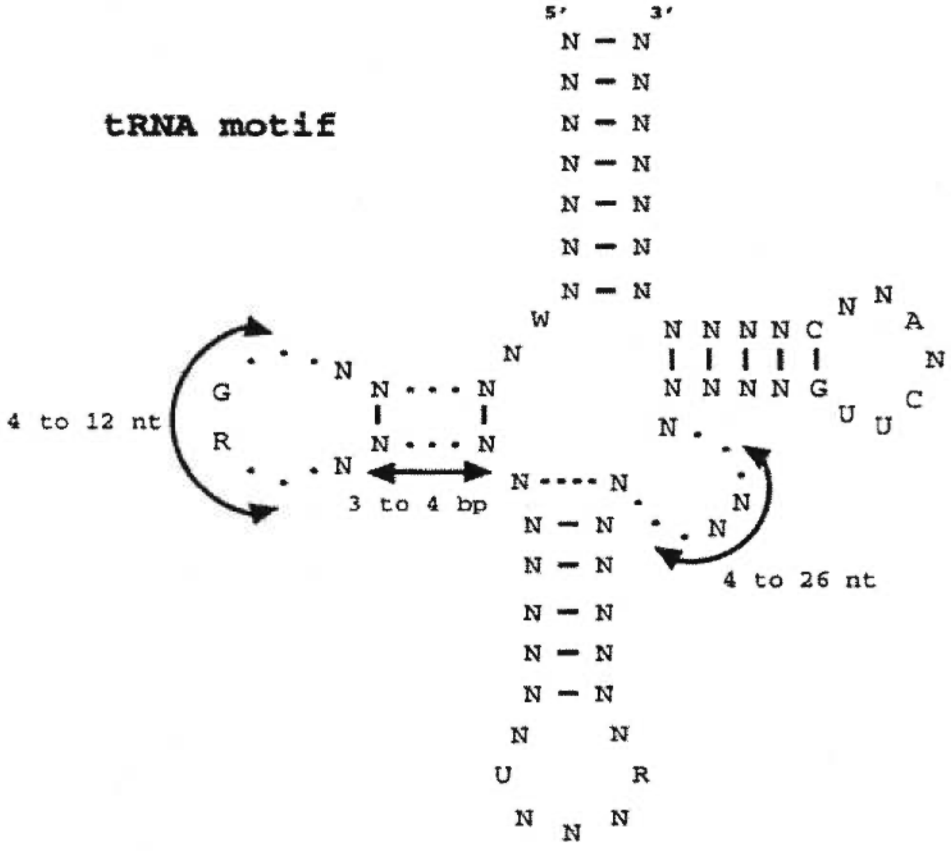


**Figure 5. The tRNA motif.**

Secondary structure of the canonical tRNA motif containing no terminal 'CCA' or introns. The letter code is defined in the legend of Figure 1.



tRNA motif

# CHAPITRE 2

## Does HIV tat protein also regulate genes of other viruses present in HIV infection?

*Trends in Biological Sciences* **22**: 115-116 (1997)

# Does HIV tat protein also regulate genes of other viruses present in HIV infection?

Gerardo FERBEYRE, Véronique BOURDEAU
and Robert CEDERGREN

Département de Biochimie,
Université de Montréal, Montréal, Québec,
Canada H3C 3J7

Corresponding author:
Robert Cedergren, Département de Biochimie, Université de Montréal,
Montréal, Québec, Canada H3C 3J7
Fax: (514) 343-2210
Email: ceder@bch.umontreal.ca

## NOTE

Contribution de chaque auteurs:

| | |
|---|---|
| G. Ferbeyre : | a développé l'hypothèse |
| V. Bourdeau : | a effectué la recherche informatisée |
| R. Cedergren : | a dirigé le projet |

## CONTENU

I. Does HIV tat protein also regulate genes of other viruses present in HIV infection?

II. Acknowledgements

III. References

IV. Figure

### I. Does HIV tat protein also regulate genes of other viruses present in HIV infection?

The availability of database containing gene and protein sequences has engendered great excitement for both experimental and structural biologists. The identification and the use of protein sequence motifs are among the delicacies that have been served to the functionally started genome community, because often information on the functionality can be inferred from a recognized motif contained in a protein sequence.

RNA is another pivotal macromolecule in the cell having the conformational properties of linear polymer, and thus folds into complex, three-dimensional arrangements resembling protein structures in nature, although differing greatly in detail[1]. This complexity renders RNA as functionally discriminatory as proteins, even though they are not generally catalytic. However, RNAs are involved in very specific complexes regulating cellular metabolism and/or viral functions.

Proteins attain great sophistication and variability in their tertiary structures owing to their construction from the 20 amino acid building blocks. RNA charms the biologist by its uncanny use of only four building blocks to form different patterns of intermolecular base pairing. This particularity is at the root of a structural paradigm change in going from proteins to RNA: it is not enough to search databases for a sequence of RNA as in the case of proteins; RNA motifs must be defined by their base-pairing properties as well. For this reason, information about RNA motifs present in sequence databases has remained veiled because we have lacked the appropriate tools to convert them into biological meaning.

We have recently developed a versatile search engine in our laboratory called RNAMOT[2]. This program takes, as input, the database sequences and carries out searches based on a series of criteria defined by the user. Although a sequence string is an option in the search, the power of the algorithm is entrenched in its ability to examine base-pairing patterns in the sequence database as well. Primary sequence and secondary or tertiary structural features can all be made part of a search "descriptor". The program returns all seqquences in the database containing the descriptor. A scoring routine allows the user to evaluate the quality of the match with the descriptor.

This platform has already been used successfully to examine parts of the sequence database for an unusual RNA-like domain[3]. We now report that we have search the GenBank sequence database of April 15, 1996 for a number of RNA structural motifs. One such search, using he characteristics of the tat protein-binding motif found in the HIV virus and formally defined by Weeks and Crothers[4] (Fig. 1a), has permitted a potentially significant observation. We have identified a tat-binding motif in Kaposi's sarcoma-associated herpes virus and in the hepatitis C virus, which exactly matches the descriptor.

In Kaposi's sarcoma-associated herpes virus, the tat-binding element was found in the gene coding for the glycoprotein H (GenBank accession number KSU40377; Fig. 1b)[5]. Although the glycoprotein H gene is found in all other herpes viruses and is considered essential for infectivity[6], the tat-binding domain is present only in this virus, and thus is not a feature of herpes viruses in general.

In the hepatitis C virus, the tat-binding element was found in the gene coding for NS4 protein (GenBank accession number HCVNS34, Fig. 1c). The combinatorics of this motif predict that one of these structures should be found in every 2.16 x $10^6$ nucleotides. In a simulation experiment, one of these motifs was found in 100 random sequences of 10,00 nucleotides. Thus, the probability of finding this motif in the 20,705 nucleotides genome of the herpes virus is 0.089 (searching in both directions), and in the much smaller hepatitis virus genome, 0.00432. Although these probabilities could indicate a chance occurrence, the origin of the motif is irrelevant, as the issue we are raising is whether the tat protein opportunistically recognizes them.

The structures have predicted $\Delta G$ values between - 6.2 and -6.4 kcal $mol^{-1}$ (M. Zuker, unpublished; see also http://www.ibc.ustl.edu/~zuker/rna/frm1.cgi). Although these values are less than the *bona fide* tat structure, their predicted melting point is a respectable 63.2°C and 72.4°C, respectively.

Kaposi's sarcoma is a common disease among AIDS patients, but its exact etiology remains controversial. The discovery of Chang *et al.*[7] of a herpes virus in Kaposi's sarcoma cells from AIDS patients raised the possibility that HIV could facilitate infection by this virus. Significantly, Vogel *et al.*[8] reported that transgenic mice expressing the gene that encodes the tat protein generated dermal lesions similar to Kaposi's sarcoma, although the presence of a herpes virus was not determined. The effect of the tat protein could be exercised via the tat-binding motif that we have found.

The tat protein could facilitate viral infection or activate an otherwise latent virus in AIDS patients by transcriptional modulation of glycoprotein H gene expression[9], or by overriding the interferon-induced inhibition of protein synthesis during infection[10]. Although hepatitis C infection can occur independently of an HIV infection, its frequency is much higher among AIDS patients[11]. Thus, the HIV tat protein could modulate hepatitis C infection as well by mechanisms similar to those described above for Kaposi's sarcoma-associated herpes virus.

Ellington *et al.* have recently speculated on viral use of arginine-binding pockets typified by the tat-binding motif in order to rapidly take command of cellular metabolism upon infection[12]. Although our data do not support this hypothesis, tat-binding motifs are also found in RNAs of chromosomal origin. In any case, our data suggest strongly that the viral communication link in multiviral infections involving HIV could be assured by the tat protein.

## II. Acknowledgments

This work was supported by a grant from Natural Science and Engineering Research Council (NSERC) of Canada. R. Cedergren is fellow of the Evolutionary Biology program of the Canadian Institute for Advanced Research. V. Bourdeau holds a predoctoral fellowship from NSERC.

## III. References

1 Uhlenbeck, O.C. (1995) *RNA* 1, 4-6

2 Laferrière, A., Gauteret, D. and Cedergren, R. (1994) CABIOS 10, 211-212

3 Steinberg, S and Cedergren, R. (1995) RNA 1, 886-891

4 Weeks, K. M. and Crothers, D. M. (1991) Cell 66, 577-588

5 Moore, P. S. e al. (1996) J. Virol. 70, 549-558

6 Gompels, U. A., Craxton, M. A. and Honess, R. W. (1988) J. Virol. 69, 2819-1829

7 Chang, Y. et al. (1994) Science 266, 1865-1869

8 Vogel, J. et al (1988) Nature 335, 606-611

9 Karn, J. and Graeble, M. A. (1992) Trends Genet. 8, 365-368

10 McMillan, N. A. et al. (1995) Virology 213, 413-323

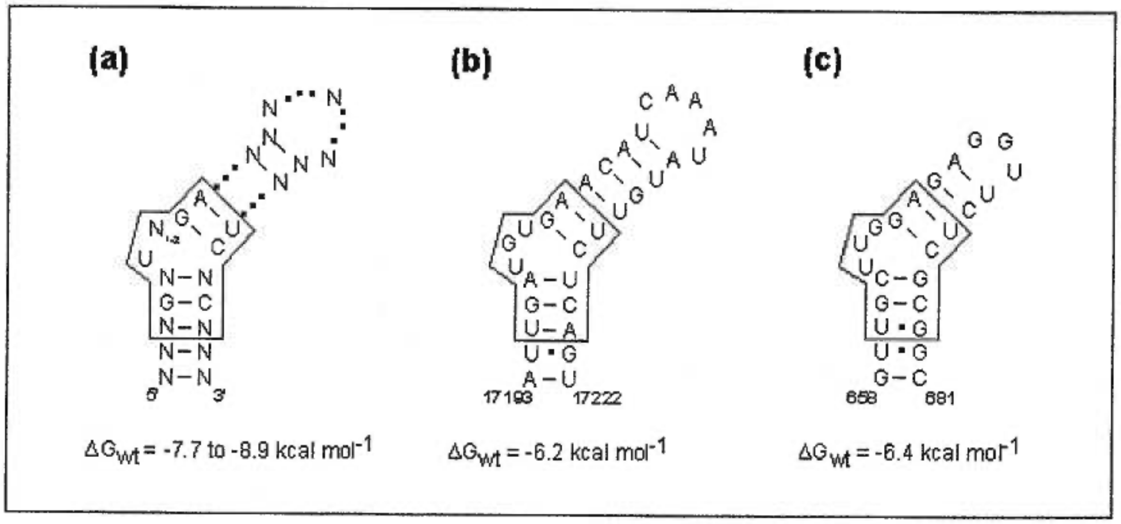
11 Cribier, B. et al (1995) AIDS 9, 1131-1136

12 Ellington, A., Leclerc, F. and Cedergren, R. (1966) Nat. Struct. Biol. 3, 981-984

## IV. Figure

### Figure 1. Tat-binding motifs.

(a) The tat-binding element consensus as defined b Weeks and Crothers[4]. (b) The putative tat-binding element in the glycoprotein H gene of Kaposi's sarcoma-associated herpes virus. (c) The putative tat-binding element in the NS4 gene of hepatitis C virus. All ΔG values are at 37°C.



$\Delta G_{wt}$ = -7.7 to -8.9 kcal mol$^{-1}$

$\Delta G_{wt}$ = -6.2 kcal mol$^{-1}$

$\Delta G_{wt}$ = -6.4 kcal mol$^{-1}$

# CHAPITRE 3

## Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank

*Genome Research* **10**: 1011-1019 (2000)

# Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank.

Gerardo FERBEYRE*[1], Véronique BOURDEAU*[2], Marie PAGEAU[2],
Pedro MIRAMONTES[3] and Robert CEDERGREN†[2]

1) Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA 11724.
2) Département de Biochimie, Université de Montréal, C.P.6128, Succursale Centre-Ville, Montréal, Québec, Canada H3C 3J7.
3) Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, México.

* G.F and V.B contributed equally to this work.

† Deceased.

April 20, 2000

Correspondence to: Dr. Gerardo Ferbeyre
Cold Spring Harbor Laboratory,
1 Bungtown Rd, P.O. Box 100
Cold Spring Harbor, New York 11724, USA.
Tel.: (516) 367 8424
Fax: (516) 367 8454
email: ferbeyre@cshl.org

## NOTE

Contribution de chaque auteurs:

G. Ferbeyre : a proposé le projet, développé des hypothèses sur des motifs putatifs intéressants et participé à l'écriture

V. Bourdeau : a décrit les motifs à chercher, participé à l'analyse des résultats ainsi qu'à l'écriture des pages Internet et de l'article

M. Pageau : a effectué la recherche informatisée et écrit les pages sur Internet des résultats

P. Miramontes : a proposé le projet et développe une autre partie de l'idée.

R. Cedergren : directeur de G.F., V.B. et M.P.

CONTENU

I. Abstract
II. Introduction
III. Results
a) Searching for self-cleaving RNA motifs of the hammerhead type in the GenBank
b) The frequency of mutated versions of the hammerhead self-cleaving RNAs
c) Some intriguing hammerhead motifs that might have functional significance
IV. Discussion
V. Methods
VI. Acknowledgments
VII. References
VIII. Tables
IX. Figures

## I. ABSTRACT

Hammerhead ribozymes were previously found in satellite RNAs from plant viroids and in repetitive DNA from certain species of newts and schistosomes. To find out if this catalytic RNA motif has a wider distribution, we decided to scrutinize the GenBank database for RNAs that contain hammerhead or hammerhead-like motifs. The search shows a widespread distribution of this kind of RNA motif in different sequences suggesting that they might have a more general role in RNA biology. The frequency of the hammerhead motif is half of that expected from a random distribution but this fact comes from the low CpG representation in vertebrate sequences and the bias of the GenBank for those sequences. Intriguing motifs include those found in several families of repetitive sequences, in the satellite RNA from the carrot red leaf luteovirus, in plant viruses like the spinach latent virus and the elm mottle virus, in animal viruses like the hepatitis E virus and the caprine encephalitis virus and in mRNAs like those coding for cytochrome P450 oxidoreductase in the rat and the hamster.

## II. INTRODUCTION

The hammerhead ribozyme was originally discovered as a self-cleaving motif in viroids and satellite RNAs. These RNAs replicate using the rolling circle mechanism, which generates long multimeric replication intermediates. They use the cleavage reaction to resolve the multimeric intermediates into monomeric forms. The region able to self-cleave has three base paired helices (I-III) connected by two conserved single stranded regions and a bulged nucleotide (Forster and Symons 1987; for reviews see Symons 1992; Bratty et al. 1993; Birikh et al. 1997). The hammerhead ribozyme also seems to function in the generation of unit length sequences from multimeric transcripts of repetitive DNA sequences. Two of these RNAs have been characterized: one in several newt species (Epstein and Gall 1987) and the other one in three Schistosome species (Ferbeyre et al. 1998). Among the repetitive sequences of these two organisms, it should be noted that not all contained a *bona fide* hammerhead ribozyme. Indeed, many mutations were also found creating variants of the original motif. Overall, the rather limited distribution of this motif contrast with the simplicity of its secondary structure where only a core of 14 nucleotides is absolutely required for cleavage.

We have recently conducted an extensive research of different RNA motifs in the GeneBank database (Bourdeau et al. 1999). The results showed that most of the motifs were randomly distributed among gene sequences suggesting that most RNA motifs originate by random drift. We now wish to extend these observations to the self-cleaving hammerhead ribozyme and its variants where either an essential nucleotide in the single strand positions is allowed to be random or the identity of a conserved base pair from Helix II and III is changed. We found that most of the hammerhead motifs are apparently under-represented among gene sequences but this comes from the bias of the GenBank for sequences with low CpG representation. We also report the finding of intriguing motifs in several repetitive sequences and mRNAs.

# III. RESULTS

## a) Searching for self-cleaving RNA motifs of the hammerhead type in the GenBank.

The hammerhead ribozyme can be described by three helices separated by three single stranded regions of conserved nucleotides. There are three equivalent conformations of the self-cleaving hammerhead depending on which helix bears the 5' and 3' end of the motif. We named them HH-I, HH-II and HH-III (Fig. 1). The descriptors composed as input for the search program are presented beside each motif and described in the legend of Figure 1 (see also Methods). They were designed to detect any sequence with all the minimal nucleotide requirements to have some catalytic activity and with the possibility to fold like the hammerhead. In this context, it is expected that sequences will be found that combine several non optimum features and be inactive for this reason, i.e., a non GUC cleavage, a C in position 4, short helices and long loops. It is also possible that they contain all the requirements for being catalytically active but the active conformation is inaccessible because the RNA molecule that bears them folds into an alternative secondary structure.

The search for hammerhead self-cleaving motifs through the GenBank database (Benson et al. 1999) was done using the program RNAMOT (Gautheret et al. 1990; Laferrière et al. 1994). The sequences detected with our descriptors are referred as occurrences. The ability of the descriptors to identify the hammerhead motifs already characterized is illustrated in Table I. The program recognizes most of the known plant derived hammerheads (Symons 1997; see also http://callisto.si.usherb.ca/~jpperra/organisms.html; Bussière et al. 1996; Lafontaine et al. 1999) and all those present in satellite

DNA sequences. It is to be noted that there is no known natural incidence of a hammerhead of the HH-II type.

Table II presents the frequencies of occurrences of potential hammerhead motifs in the different sections of the GenBank as well as the expected frequencies calculated from the number of occurrences obtained in a database of random sequences. In general the number of occurrences observed are half of the frequency expected if our motifs were randomly distributed among the sequences of the GenBank. HH-I and HH-II detect twice as many motifs than HH-III because we design the motifs in a way that Helix III has a two base pairs requirement in HH-I and HH-II descriptors versus three base pairs in the HH-III descriptor (see Methods). This increase was predicted by the number of occurrences obtained in the random database.

**b) The frequency of mutated versions of the hammerhead self-cleaving RNAs.**

We also composed descriptors for variants of the hammerhead ribozyme motif. Substitutions were made by replacing, one at a time, each of the essential nucleotides located in the single stranded regions of the ribozyme core by N (bolded in Fig. 1) or by changing the identity of each one of the two conserved base pairs of the hammerhead motif (also bolded in Fig. 1).

Table II presents the data on the distribution of the mutated variants of HH-I, HH-II and HH-III from the single stranded region. It is expected that every mutant will increase the frequency of occurrences by a factor of 4 because we changed the requirements in every position from only one to all four nucleotides except in position 4 where C and U were already allowed and in the cleavage site where only G was originally excluded. Thus, in position 4 we expected to double the frequency and in the cleavage site we expected a 25 % increase. The results are mostly those anticipated based on these calculations. However the mutants of position 12 doubled the expected increase in all the orientations. This effect was not uniformly observed in the different subdivisions of the GenBank. Actually most of the extra occurrences are located in the files containing ESTs and mammalian sequences. These preferences were not observed in the random database where the mutants showed the anticipated increase in their frequency in comparison with the original motif. The number of occurrences obtained in the virus section of the GenBank for the HH-III-8 variant: 722 instead of the 113 expected (HH-III has 3774 expected occurrences and viruses represent 3% of the GenBank). However, a quick analysis of the occurrences obtained with this descriptor revealed that most of them are the same motif repeated in 679 hepatitis C sequences.

Table III presents the frequencies obtained with the mutant hammerhead ribozymes using a different identity for the conserved base pair of Helix II or III (positions 10.1:11.1 and 15.1:16.1). One striking observation is that all the mutants in Helix II (iiNN) have a total of occurrences two to six times higher than expected whereas the mutants in Helix III (iiiNN) have half the expected frequency. One more interesting point is the high number of occurrences obtained with the three orientations of the hammerhead ribozyme having a A:U base pair in Helix II (10.1:11.1) instead of the usual G:C.

The mutants in position 12 and the mutants of the conserved base pair of Helix II have in common that they disrupt the presence of a dinucleotide CpG in the resulting sequence. It is well known that "CpG" is under represented in vertebrate sequences (Karlin and Mrazek 1997). The GenBank is biased for those sequences mainly due to human and rodent entries. In those files, the mutants that disrupt the CpG requirement have a higher frequency. In order to confirm that the overall frequency of the hammerhead motifs containing CpG dinucleotides is half of the expected one due to the low CpG content of vertebrate sequences, we built a new random database in which the frequency of CpG was reduced by half in favor of either AG, CA, CC, CT, GG and TG to simulate the frequencies observed by Karlin and Mrazek (1997; see Methods). In this database, we observed an overall doubling of the original expected frequencies for all the motifs needing a CpG but not for the others (data not shown).

Still, the mutants with a A:U base pair in position 10.1:11.1 of the Helix II have a very high frequency in all three conformations of the motif: two to three times higher than expected even considering the CpG effect discussed above. So far we have no explanation for this intriguing observation.

Finally, we made three more searches by changing the cleavage site from NUH to NHH based on the report of Kore et al. (1998) that such hammerheads were still active. We obtained for these new mutants a number of occurrences corresponding to half what we expected according to the search in an equal A-C-G-T random database. Moreover, like for the previous motifs, the number of occurrences in the GenBank is comparable to the expected frequency according to the search in the reduced for CpG database. All the occurrences found in the GenBank are available in our web site at http://www.centrcn.umontreal.ca/~bourdeav/HH.

### c) Some intriguing hammerhead motifs that might have functional significance.

This section present a sample of motifs considered interesting either because of their location or because their structure is optimal for self-cleavage The hammerhead ribozyme occurs naturally in satellite RNAs, viroids and transcripts from repetitive sequences. The probability of finding an active hammerhead should be higher among these genetic elements. Several potential hammerhead motifs were found in distinct families of repetitive DNA.

Hammerhead ribozymes were found in the satellite DNA from *Dolichopoda schiavazzii* (cricket) using the HH-I descriptor (example in Fig. 2A). Fourteen have a conserved HH-I motif and two have a HH-I-iiGU motif (G:U in position 10.1:11.1 instead of G:C). This ribozyme cleaves after CUA (Rojas et al.; manuscript in preparation). Helix I have the GG:CC base pairs and the internal loop common to the hammerhead motifs in schistosomes (Ferbeyre et al. 1998) and newts (Pabon-Peña et al. 1991). It is noteworthy that among the 20 similar sequences submitted to GenBank, the four sequences not found through the search contained either mismatches in one of the helices or combined two point mutations.

A hammerhead-like motif was detected in the Kpn-13 family of human repetitive DNA using the descriptor HH-I-4 (Fig. 2B). The motif is found in several ESTs containing Kpn-repetitive sequences (also known as L1-repetitive elements) indicating its expression at the RNA level. All the occurrences contain a disabling A at position 4 but one (AA564135) possesses a C. The latter motif is inactivated by a G per A substitution at position 12. Variants of this motif are also found in genomic clones containing Kpn repetitive sequences. Intriguingly, the L1 motif interrupting the dystrophin gene of a muscular dystrophy patient (accession HSU09115) also has a disruption in Helix I. Four additional hammerhead-like motifs were found in the satellite DNA array from the rodent *Microtus chrotorrhinus* (accession MICSATB, position 921-1079, not shown), in the repetitive DNA from the protozoan parasite *Theileria parva* (accession S37077, position 84-223, not shown) with the descriptor HH-I-7 and in mouse repetitive DNA with descriptors for the HH-I-iiUA and HH-III-iiAU motifs (Fig. 2C and D). The first two motifs are predicted to be inactive because they contain A instead of G in position 12.

Viruses are good candidates for using catalytic RNA motifs because they would give them independence from cellular proteins. We have found several new intriguing hammerhead motifs in different viruses (Fig. 2E). Two similar hammerhead ribozyme motifs were found in the 5' UTR of two viruses of the Ilarvirus genus, family of Bromoviridae, which are single stranded positive RNA viruses. One motif is in the spinach latent virus (accession PMOVRNA3, position 252-331) and the other in the Elm mottle virus (accession SLU57048, position 250-329) (Fig. 2E). Both motifs were found using the HH-III descriptor. The region containing the hammerhead is highly conserved among these viruses. The hammerhead motif found with HH-II in a RNA associated to carrot red luteovirus that is also very interesting because satellite RNAs were the first molecules found to contain hammerhead ribozymes (Fig. 2F). This motif is predicted to cleave after AUA. Mammalian viruses also contain potential hammerhead ribozymes and two of them found with HH-II are illustrated in Figure 2G and H, one in the hepatitis E virus and the other in the caprine encephalitis virus.

Two hammerhead motifs in human mRNAs are also presented in Figure 2I and J. Self cleaving motifs in mRNA might regulate gene expression by promoting RNA decay. The genes coding for the interferon-induced DAP1 and the neuroleukin gene possess potentially active hammerhead motifs found with HH-III that are predicted to cleave after UUC and CUC respectively. Perhaps even more remarkable are the conserved hammerhead motifs found in the genes coding for NADPH-cytochrome P450 oxidoreductase both in the rat and the hamster (Fig. 2K and L). All together, the motifs presented here suggest that the hammerhead ribozyme might have functions other than those previously suggested for satellite RNA and transcripts for repetitive sequences.

## IV. DISCUSSION

We have used the search engine RNAMOT to scrutinize the GenBank for potential self-cleaving hammerhead ribozyme motifs. Our search extends earlier efforts to find a subset of potential hammerheads in *E. coli* sequences (Ruffner et al 1990). As this motif has relatively few structural constraints, we designed an extensive set of descriptors for both the wild type motif and variants of its essential nucleotides. The results show a wide distribution of potential hammerhead-like motifs in all regions of the GenBank with a higher frequency for the variants that do not require the presence of a CpG dinucleotide in the final sequence of the motifs. This CpG dinucleotide in positions 11.1 and 12 is not absolutely required for self cleavage since other base pairs are acceptable in positions 10.1:11.1. We conclude that the reduction we observed in the frequency of most hammerhead motifs in this search is fortuitous.

We expect that the great majority of the motifs found here are inactive because we designed descriptors which include mutations or non-optimal features of the hammerhead self-cleaving motif (Ruffner et al. 1990). However our results illustrate the possibility that natural sequences might end up forming self-cleaving motifs by random drift. In other words, it would be sufficient to mutate one or two residues to activate the potential hammerhead ribozymes described here. This is not only true for the hammerhead ribozyme motif since other RNA motifs can be found randomly in natural sequences (Fontana et al. 1993; Reidys et al. 1997; Bourdeau et al. 1999).

The use of variants of the hammerhead ribozyme was stimulated by previous work that showed that satellite DNA encoding hammerhead ribozymes is enriched with mutated variants of the motif (Zhang and Epstein 1996; Ferbeyre et al. 1998). The ribozyme motif found in the cricket satellite DNA follows this rule since 14 of the 20 sequences deposited until now in the GenBank contains an active motif. Other mutant hammerheads were found in different families of repetitive DNA using descriptors for hammerhead-like motifs, raising the possibility that other members of these families, not yet sequenced, contain the active motifs. The occurrence of hammerhead ribozymes in transcripts of repetitive DNA from different species suggests a functional role for the self-cleavage reaction in the propagation and/or the metabolism of these transcripts. We have previously proposed that self-cleavage might limit the expansion of repetitive sequences through the genome by retrotransposition (Ferbeyre et al. 1998). This model predicts that recent insertions of these elements will contain disabling mutations in the hammerhead motif. The family of L1 repetitive elements for example contains mutated versions of the hammerhead and members of this family still retrotranspose in humans, sometimes causing genetic diseases (Holmes et al. 1994). Another intriguing possibility is that viroids and satellite RNAs originated from transcripts of repetitive sequences when these transcripts parasitize a viral replication machinery. Subsequently they might jump from one organism to another using the virus as a vector and as a result their distribution will cross phylogenetic barriers.

A large number of ESTs and mRNAs were found here to possess hammerhead-like motifs. In order to test any role of the hammerhead motifs identified in this work, we need a combination of biochemical and genetic analysis. Our group has finished the characterization of hammerhead motifs in repetitive DNA of Schistosome (Ferbeyre et al. 1998) and the Cricket (Rojas et al., in preparation). All the occurrences we found in the GenBank are available in our web site (URL http://www.centrcn.umontreal.ca/~bourdeav/HH) for those interested in finding where "hammers" can cut.

## V. METHODS

The pattern searching for RNA secondary structures was carried out by RNAMOT (Gautheret et al. 1990; Laferrière et al. 1994). The inputs for this program are nucleotide sequences and a descriptor file defining the structural motif to be searched. RNAMOT reports all the occurrences of the motif as well as its positions along the sequence. Two of the three helices defining the hammerhead self-cleaving motif are closed by loops. The remaining helix connects the motif to the rest of the RNA molecule. As a result there are three ways of defining a self-cleaving hammerhead ribozyme motif. We have built descriptors for these three different orientations of the motif taking into account the following constraints (Fig. 1):

1) Three nucleotides in Helix I. Helix I has no specific nucleotide requirements although the hammerhead motif found in the newt and in Schistosome possess a conserved GG:CC base pairing, three nucleotides downstream of the cleavage site as well as an internal loop further downstream (Pabon-Peña et al. 1991; Ferbeyre et al. 1998).

2) The conserved sequence CYGANGA. This sequence is part of the catalytic core of the ribozyme and is entirely conserved with the exception of position 7. In the latter although all nucleotides are accepted the preferred ones are U then G or A and finally C. More recently position 4 was reported to accept also U so we have included this feature in our search (Ambros and Flores 1998)

3) Three nucleotides in Helix II. There is a strong preference for a R:Y base pair in positions 10.1:11.1 but the pair G:C confers the better activity and was the only one allowed in our original descriptors.

4) The conserved sequence GAA is absolutely required for catalysis. In the X-ray model of the hammerhead, nucleotides G12 and A13 form two reverse Hoogsteen G-A base pairs with nucleotides A9 and G8 respectively while A14 form a non Watson Crick base pair with N7 (Scott et al. 1995).

5) Helix III requires an A:U base pair which is also of non Watson Crick type and a minimum of one more pair in two of the orientations (HH-I and HH-II). When the helix is open as in HH-III, two more pairs are required.

6) The cleavage site was defined as NUH (H is any nucleotide but G). However, natural ribozymes contain GUC, GUA, AUA and AUC because they allow the highest reaction rates (Shimayama et al. 1995; Ferbeyre et al. 1998).

7) The loops closing the helices were allowed to have from 0 to 100 nucleotides.

Sixty-three additional mutants were also included in the study. These were derived from the original motifs shown in Figure 1 by changing either one base in the conserved single stranded regions for an N (any nucleotide; 30 mutants), the identity of one of the constrained base pair (positions 10.1:11.1 and 15.1:16.1; 30 mutants) or by changing the cleavage site from NUH to NHH (three more motifs; Kore et al. 1998).

The search was carried on in the release of July 15, 1998, of the GenBank sequence database (NCBI-GenBank flat file release 108.0). Searches were carried out on both strands and all occurrences of motifs involving unidentified bases denoted by N in the database were disregarded. A Power Challenge XL with 32 CPUs IP 19, R4400, 150 MHz processor (3072 Mbytes) running UNIX IRIX 6.2 was used.

In order to help to establish the significance of their presence, frequencies of each motif in the database were compared with frequencies in a random sequence database generated by a uniform pseudo-random number generator (L'Écuyer and Andres 1997) with a period length near $2^{121}$. The random sequence databases contained 1000 sequences of 100,000 nucleotides each; the four nucleotides A, C, G and T were used with equal probabilities. An "expected" frequency *N* in GenBank was calculated from the number *M* of occurrences of each motif in the random databases as follows: $N = (a \times M) / (10^4 \times 10^5)$, where *a* is the number of nucleotides in GenBank ($1.797 \times 10^9$ in the release 108.0).

The random database reduced in CpG dinucleotides was generated using the same procedure but each time a CpG dinucleotide was created a second generator (evolving in parallel) would enter in function to decide if yes or no (50% frequency) the dinucleotide would be changed. If a change had to take place a third generator (also evolving in parallel) would be able to choose among six replacing dinucleotides: AG, CA, CC, CT, GG or TG (choices made according to the dinucleotide frequencies reported by Karlin and Mrazek 1997). The "expected" frequency was evaluated as before.

## VI. ACKNOWLEDGMENTS

The authors wish to express their gratitude to Bruno Paquin for valuable comments, and to NSERC of Canada which financed this project. V.B. holds a doctoral fellowship from NSERC of Canada. The late R.C. was Richard Ivey Scholar of the Canadian Institute for Advanced Research (CIAR) program in Evolutionary Biology. We acknowledge previous efforts from Dr. Daniel Gautheret to search for hammerhead sequences with RNAMOT in our laboratory. In addition, we thank Bernard Lorazo, Daniel Raymond and André Fourrier of the DITER (Direction des infrastructures technologiques d'enseignement et de recherche) at the Université de Montréal for their assistance. P.M. wishes to thanks the hospitality of the Université de Montréal and the Institute of Physics, UNAM.

## VII. REFERENCES

Ambros, S., and R. Flores. 1998. *In vitro* and *in vivo* self-cleavage of a viroid RNA with a mutation in the hammerhead catalytic pocket. *Nucleic Acids Res.* **26:** 1877--1883.

Benson, D.A., M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp, and D.L. Wheeler. 1999. GenBank. *Nucleic Acids Res.* **27:** 12--17.

Birikh, K.R., P.A. Heaton, and F. Eckstein. 1997. The structure, function and application of the hammerhead ribozyme. *Eur. J. Biochem.* **245:** 1--16.

Bourdeau, V., G. Ferbeyre, M. Pageau, B. Paquin, and R. Cedergren. 1999. The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.* **27**: 4457--4467.

Bratty, J., P. Chartrand, G. Ferbeyre, and R. Cedergren. 1993. The hammerhead RNA domain, a model ribozyme. *Biochim. Biophys. Acta* **1216:** 345--359.

Bussière, F., D. Lafontaine, and J.-P. Perreault. 1996. Compilation and analysis of viroid and viroid-like RNA sequences. *Nucleic Acids Res.* **24:** 1793--1798.

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

Ferbeyre, G., J.M. Smith, and R. Cedergren. 1998. Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol. Cell. Biol.* **18:** 3880--3888.

Fontana, W., D.A. Konings, P.F. Stadler, and P. Schuster. 1993. Statistics of RNA secondary structures. *Biopolymers* **33:** 1389--1404.

Forster, A.C., and R.H. Symons. 1987. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49:** 211--220.

Gautheret, D., F. Major, and R. Cedergren. 1990. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.* **6:** 325--331.

Hertel, K.J., A. Pardi, O.C. Uhlenbeck, M. Koizumi, E. Ohtsuka, S. Uesugi, R. Cedergren, F. Eckstein, W.L. Gerlach, R. Hodgson, and R.H. Symons. 1992. Numbering system for the hammerhead. *Nucleic Acids Res.* **20:** 3252.

Holmes, S.E., B.A. Dombroski, C.M. Krebs, C.D. Boehm, H.H. Jr Kazazian. 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.* **7:** 143--148.

Karlin, S., and J. Mrazek. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* **94:** 10227--10232.

Kore, A.R., N.K. Vaish, U. Kutzke, and F. Eckstein. 1998. Sequence specificity of the hammerhead ribozyme revisited; the NHH rule. *Nucleic Acids Res.* **26:** 4116--4120.

L'Écuyer, P., and T.H. Andres. 1997. A random number generator based on the combination of four LCGs. *Math. Comput.Simulation* **44:** 99--107.

Laferrière, A., D. Gautheret, and R. Cedergren. 1994. An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.* **10:** 211--212.

Lafontaine, D.A., P. Deschênes, F. Bussière, V. Poisson, and J.-P. Perreault. 1999. The viroid and viroid-like RNA database. *Nucleic Acids Res.* **27:** 186--187.

Pabon-Peña, L.M., Y. Zhang, and L.M. Epstein. 1991. Newt satellite 2 transcripts self-cleave by using an extended hammerhead structure. *Mol. Cell. Biol.* **11:** 6109--6115.

Reidys, C., P.F. Stadler, and P. Schuster. 1997. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull. Math. Biol.* **59:** 339--397.

Ruffner, D.E., G.D. Stormo, and O.C. Uhlenbeck. 1990. Sequence requirements of the hammerhead RNA self-cleavage reaction. *Biochemistry* **29:** 10695--10702.

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

Shimayama T., S. Nishikawa, and K. Taira. 1995. Generality of the NUX rule: kinetic analysis of the results of systematic mutations in the trinucleotide at the cleavage site of hammerhead ribozymes. *Biochemistry* **34:** 3649--3654.

Symons, R. 1992 . Small catalytic RNAs. *Annu. Rev. Biochem.* **61:** 641--671.

Symons, R. 1997. Plant pathogenic RNAs and RNA catalysis. *Nucleic Acid Res.* **25:** 2683--2689.

Zhang, Y., and L.M. Epstein. 1996. Cloning and characterization of extended hammerheads from a diverse set of caudate amphibians. *Gene* **172:** 183--190.

## VIII. TABLES

Table I.	Known hammerhead motifs identified in our search.

| HH-I | HH-III |
|---|---|
| - Avocado sunblotch viroid<br>- Tobacco Ringspot virus satellite RNA<br>- Ambyostoma talpoideum satellite 2<br>- Cryptobranchus alleganiensis satellite 2<br>- Cyrrops pyrrhogaster satellite 2<br>- Eurycea longicauda satellite 2<br>- Plethodon glutinosus satellite 2 | - Schistosoma mansoni DNA for repeated sequences<br>- Barley yellow virus satellite RNA<br>- Chrysanthemum chlorotic mottle viroid<br>- Cherry small circular viroid-like RNA<br>- Lucerne transient streak virus RNA 2<br>- Peach latent mosaic viroid<br>- Subterranean clover mottle virus satellite RNA |

**Table II. Distribution of hammerhead and hammerhead-like motifs in the different sections of the GenBank: mutants of the single stranded regions.**

The motifs are named as explained in the Methods and in Figure 1. The "expected" number of occurrences was obtained by searching the different motifs in a database of 1000 random sequences of 100,000 nucleotides (equal representations of A, C, G and T) and correcting the frequency to the relative size of the GenBank. The "total over expected" shows the ratio of occurrences obtained in the GenBank versus the expected ones according to the search done in a random database with the size of the GenBank.

pri: Primate sequence entries (from the two GenBank files); rod: Rodent sequence entries; mam: Other mammalian sequence entries; vrt: Other vertebrate sequence entries; inv: Invertebrate sequence entries; pln: Plant sequence entries (including fungi and algae); bct: Bacterial sequence entries; rna: Structural RNA sequence entries; vrl: Viral sequence entries; phg: Phage sequence entries; syn: Synthetic and chimeric sequence entries; una: Unannotated sequence entries; est: EST (expressed sequence tag) (from 23 GenBank files); pat: Patent sequence entries; sts: STS (sequence tagged site) sequence entries; gss: GSS (genome survey sequence) sequence entries; htg: HTGS (high throughput genomic sequencing) sequence entries.

| | pri | rod | mam | vrt | inv | pln | bct | rna | vrl | phg | syn | una | est | pat | sts | gss | htg | total | total over<br>expected | expected<br>(A=C=G=T) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HH-I | 51 | 21 | 0 | 14 | 176 | 119 | 130 | 0 | 37 | 3 | 5 | 0 | 186 | 45 | 4 | 17 | 108 | 916 | 0.56 | 1635 |
| HH-I-3 | 182 | 72 | 12 | 45 | 997 | 560 | 553 | 2 | 159 | 16 | 6 | 0 | 702 | 125 | 10 | 119 | 486 | 4046 | 0.57 | 7081 |
| HH-I-4 | 270 | 29 | 0 | 32 | 365 | 228 | 226 | 10 | 69 | 3 | 8 | 1 | 358 | 55 | 4 | 33 | 272 | 1963 | 0.55 | 3576 |
| HH-I-5 | 310 | 70 | 14 | 97 | 506 | 400 | 309 | 2 | 77 | 10 | 7 | 1 | 641 | 60 | 13 | 92 | 382 | 2991 | 0.46 | 6488 |
| HH-I-6 | 197 | 60 | 17 | 37 | 516 | 312 | 399 | 3 | 107 | 5 | 8 | 1 | 700 | 93 | 11 | 67 | 288 | 2821 | 0.40 | 6973 |
| HH-I-8 | 171 | 44 | 11 | 22 | 790 | 407 | 455 | 10 | 138 | 11 | 5 | 0 | 521 | 86 | 10 | 61 | 420 | 3162 | 0.46 | 6883 |
| HH-I-9 | 247 | 56 | 12 | 29 | 508 | 260 | 454 | 0 | 107 | 9 | 7 | 3 | 536 | 99 | 9 | 65 | 344 | 2745 | 0.38 | 7314 |
| HH-I-12 | 946 | 224 | 41 | 103 | 754 | 613 | 454 | 0 | 180 | 13 | 8 | 2 | 1856 | 137 | 48 | 263 | 794 | 6436 | 0.94 | 6865 |
| HH-I-13 | 209 | 76 | 16 | 41 | 428 | 308 | 462 | 5 | 93 | 10 | 11 | 0 | 739 | 143 | 12 | 66 | 226 | 2845 | 0.42 | 6775 |
| HH-I-14 | 317 | 94 | 30 | 36 | 404 | 377 | 487 | 7 | 104 | 5 | 30 | 5 | 737 | 102 | 8 | 77 | 327 | 3147 | 0.50 | 6362 |
| HH-I-17 | 65 | 25 | 0 | 18 | 201 | 144 | 162 | 0 | 50 | 4 | 11 | 0 | 225 | 53 | 4 | 19 | 122 | 1103 | 0.54 | 2031 |
| HH-II | 83 | 33 | 1 | 5 | 175 | 126 | 141 | 1 | 80 | 1 | 1 | 3 | 435 | 30 | 4 | 8 | 91 | 1218 | 0.54 | 2246 |
| HH-II-3 | 180 | 78 | 8 | 29 | 994 | 645 | 715 | 3 | 204 | 14 | 25 | 23 | 1022 | 102 | 11 | 82 | 502 | 4637 | 0.66 | 7063 |
| HH-II-4 | 133 | 56 | 7 | 11 | 387 | 261 | 278 | 1 | 100 | 1 | 2 | 5 | 658 | 43 | 7 | 39 | 188 | 2177 | 0.51 | 4277 |
| HH-II-5 | 244 | 82 | 16 | 37 | 573 | 477 | 372 | 7 | 141 | 10 | 4 | 4 | 910 | 70 | 12 | 71 | 362 | 3392 | 0.39 | 8788 |
| HH-II-6 | 234 | 81 | 23 | 17 | 522 | 348 | 977 | 38 | 153 | 7 | 7 | 9 | 856 | 74 | 8 | 59 | 311 | 3724 | 0.49 | 7674 |
| HH-II-8 | 209 | 64 | 11 | 22 | 921 | 385 | 519 | 16 | 130 | 3 | 10 | 4 | 1265 | 53 | 7 | 65 | 452 | 4136 | 0.54 | 7710 |
| HH-II-9 | 255 | 58 | 8 | 24 | 457 | 319 | 540 | 6 | 140 | 9 | 11 | 5 | 884 | 82 | 10 | 51 | 281 | 3140 | 0.39 | 7979 |
| HH-II-12 | 1290 | 253 | 61 | 60 | 879 | 716 | 534 | 1 | 209 | 14 | 18 | 3 | 2491 | 125 | 56 | 273 | 1027 | 8010 | 0.97 | 8285 |
| HH-II-13 | 214 | 91 | 45 | 26 | 421 | 326 | 534 | 7 | 162 | 8 | 23 | 3 | 1031 | 71 | 7 | 53 | 228 | 3250 | 0.44 | 7404 |
| HH-II-14 | 281 | 81 | 15 | 37 | 414 | 524 | 493 | 6 | 142 | 7 | 2 | 4 | 1038 | 87 | 9 | 67 | 256 | 3463 | 0.43 | 8069 |
| HH-II-17 | 100 | 35 | 1 | 7 | 220 | 152 | 173 | 1 | 97 | 1 | 1 | 4 | 504 | 36 | 7 | 16 | 108 | 1463 | 0.53 | 2786 |
| HH-III | 42 | 6 | 0 | 2 | 93 | 67 | 65 | 0 | 21 | 1 | 2 | 0 | 96 | 57 | 3 | 8 | 64 | 527 | 0.55 | 952 |
| HH-III-3 | 96 | 32 | 8 | 18 | 625 | 305 | 245 | 0 | 74 | 6 | 3 | 0 | 310 | 72 | 5 | 74 | 291 | 2164 | 0.64 | 3397 |
| HH-III-4 | 295 | 12 | 1 | 21 | 192 | 128 | 98 | 0 | 36 | 3 | 5 | 0 | 204 | 63 | 3 | 23 | 229 | 1313 | 0.76 | 1725 |
| HH-III-5 | 144 | 38 | 16 | 35 | 301 | 205 | 231 | 0 | 34 | 2 | 2 | 0 | 388 | 76 | 11 | 32 | 219 | 1734 | 0.48 | 3612 |
| HH-III-6 | 134 | 29 | 5 | 18 | 290 | 183 | 197 | 9 | 58 | 2 | 3 | 0 | 426 | 66 | 10 | 33 | 180 | 1643 | 0.46 | 3540 |
| HH-III-8 | 92 | 26 | 4 | 6 | 528 | 203 | 260 | 0 | 722 | 1 | 2 | 0 | 294 | 154 | 4 | 29 | 255 | 2580 | 0.68 | 3774 |
| HH-III-9 | 109 | 22 | 0 | 12 | 239 | 143 | 196 | 0 | 87 | 3 | 4 | 0 | 462 | 63 | 7 | 38 | 164 | 1549 | 0.40 | 3918 |
| HH-III-12 | 658 | 97 | 40 | 32 | 369 | 318 | 217 | 0 | 98 | 3 | 6 | 0 | 1071 | 135 | 20 | 121 | 477 | 3662 | 1.03 | 3540 |
| HH-III-13 | 119 | 29 | 8 | 14 | 226 | 176 | 246 | 3 | 53 | 6 | 2 | 0 | 345 | 70 | 6 | 28 | 142 | 1473 | 0.41 | 3576 |
| HH-III-14 | 133 | 33 | 10 | 11 | 242 | 184 | 195 | 1 | 65 | 2 | 21 | 0 | 408 | 86 | 5 | 39 | 146 | 1581 | 0.45 | 3504 |
| HH-III-17 | 49 | 6 | 0 | 3 | 123 | 81 | 86 | 0 | 23 | 2 | 2 | 0 | 154 | 67 | 4 | 11 | 74 | 685 | 0.49 | 1402 |
| GenBank<br>relative size | 0.14 | 0.03 | 0.01 | 0.01 | 0.07 | 0.07 | 0.07 | <0.01 | 0.03 | <0.01 | <0.01 | <0.01 | 0.37 | 0.02 | 0.01 | 0.06 | 0.10 | 1.00 | | |

**Table III. Distribution of hammerhead and hammerhead-like motifs in the different sections of the GenBank: mutants of the Helix II and III.**

See legend of Table II.

|  | pri | rod | mam | vrt | inv | pln | bct | rna | vrl | phg | syn | una | est | pat | sts | gss | htg | total | total over expected | expected (A=C=G=T) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HH-I | 51 | 21 | 0 | 14 | 176 | 119 | 130 | 0 | 37 | 3 | 5 | 0 | 186 | 45 | 4 | 17 | 108 | 916 | 0.56 | 1635 |
| HH-I-iiAU | 569 | 100 | 30 | 36 | 822 | 545 | 426 | 3 | 58 | 7 | 8 | 0 | 1113 | 93 | 32 | 341 | 544 | 4727 | 2.56 | 1851 |
| HH-I-iiCG | 343 | 59 | 12 | 20 | 203 | 117 | 512 | 11 | 24 | 2 | 1 | 11 | 669 | 69 | 15 | 66 | 221 | 2355 | 1.27 | 1851 |
| HH-I-iiGU | 452 | 65 | 9 | 38 | 361 | 227 | 203 | 0 | 73 | 3 | 49 | 2 | 794 | 58 | 15 | 93 | 332 | 2774 | 1.77 | 1564 |
| HH-I-iiUA | 606 | 124 | 21 | 82 | 406 | 364 | 260 | 2 | 165 | 10 | 2 | 0 | 941 | 43 | 15 | 175 | 436 | 3652 | 2.03 | 1797 |
| HH-I-iiUG | 455 | 69 | 25 | 38 | 354 | 268 | 287 | 4 | 57 | 6 | 24 | 2 | 664 | 47 | 9 | 92 | 268 | 2669 | 1.43 | 1869 |
| HH-I-iiiCG | 51 | 19 | 5 | 6 | 91 | 63 | 163 | 0 | 43 | 2 | 2 | 0 | 355 | 77 | 1 | 10 | 27 | 915 | 0.48 | 1887 |
| HH-I-iiiGC | 121 | 53 | 13 | 13 | 103 | 116 | 189 | 1 | 48 | 1 | 6 | 0 | 494 | 60 | 2 | 28 | 52 | 1300 | 0.74 | 1761 |
| HH-I-iiiGU | 77 | 31 | 18 | 12 | 98 | 116 | 130 | 0 | 21 | 0 | 0 | 1 | 284 | 44 | 1 | 17 | 64 | 914 | 0.58 | 1564 |
| HH-I-iiiUA | 52 | 23 | 2 | 8 | 118 | 89 | 79 | 2 | 13 | 0 | 8 | 0 | 177 | 28 | 2 | 18 | 38 | 657 | 0.39 | 1707 |
| HH-I-iiiUG | 36 | 19 | 4 | 6 | 100 | 63 | 155 | 6 | 21 | 2 | 152 | 0 | 154 | 73 | 1 | 19 | 90 | 901 | 0.50 | 1797 |
| HH-II | 83 | 33 | 1 | 5 | 175 | 126 | 141 | 1 | 80 | 1 | 1 | 3 | 435 | 30 | 4 | 8 | 91 | 1218 | 0.54 | 2246 |
| HH-II-iiAU | 719 | 117 | 40 | 39 | 860 | 572 | 453 | 0 | 118 | 13 | 2 | 0 | 1021 | 87 | 24 | 190 | 663 | 4918 | 2.76 | 1779 |
| HH-II-iiCG | 404 | 97 | 21 | 25 | 183 | 113 | 148 | 1 | 50 | 1 | 7 | 0 | 856 | 38 | 20 | 83 | 284 | 2331 | 1.34 | 1743 |
| HH-II-iiGU | 560 | 96 | 36 | 39 | 352 | 270 | 237 | 1 | 192 | 5 | 1 | 0 | 783 | 35 | 15 | 124 | 367 | 3113 | 1.77 | 1761 |
| HH-II-iiUA | 668 | 126 | 133 | 39 | 465 | 382 | 308 | 4 | 85 | 10 | 22 | 1 | 843 | 66 | 21 | 248 | 462 | 3883 | 2.12 | 1833 |
| HH-II-iiUG | 694 | 123 | 36 | 24 | 335 | 239 | 260 | 1 | 71 | 0 | 5 | 0 | 749 | 67 | 17 | 212 | 319 | 3152 | 1.81 | 1743 |
| HH-II-iiiCG | 65 | 28 | 3 | 14 | 96 | 45 | 215 | 2 | 15 | 4 | 1 | 0 | 159 | 10 | 1 | 20 | 49 | 727 | 0.40 | 1833 |
| HH-II-iiiGC | 136 | 67 | 16 | 21 | 92 | 103 | 232 | 0 | 20 | 7 | 1 | 2 | 498 | 51 | 2 | 19 | 72 | 1339 | 0.71 | 1887 |
| HH-II-iiiGU | 69 | 42 | 3 | 12 | 101 | 88 | 182 | 0 | 33 | 6 | 2 | 0 | 365 | 30 | 1 | 19 | 65 | 1018 | 0.48 | 2103 |
| HH-II-iiiUA | 60 | 16 | 3 | 11 | 107 | 78 | 102 | 1 | 27 | 0 | 0 | 0 | 136 | 8 | 1 | 21 | 65 | 636 | 0.35 | 1833 |
| HH-II-iiiUG | 50 | 22 | 4 | 27 | 85 | 77 | 129 | 0 | 39 | 0 | 0 | 0 | 154 | 23 | 0 | 17 | 63 | 690 | 0.30 | 2282 |
| HH-III | 42 | 6 | 0 | 2 | 93 | 67 | 65 | 0 | 21 | 1 | 2 | 0 | 96 | 57 | 3 | 8 | 64 | 527 | 0.55 | 952 |
| HH-III-iiAU | 360 | 59 | 17 | 11 | 482 | 324 | 241 | 0 | 35 | 7 | 11 | 1 | 723 | 31 | 6 | 354 | 328 | 2990 | 5.05 | 593 |
| HH-III-iiCG | 203 | 37 | 4 | 9 | 98 | 58 | 101 | 0 | 9 | 0 | 0 | 0 | 469 | 8 | 6 | 42 | 132 | 1176 | 1.17 | 1006 |
| HH-III-iiGU | 251 | 29 | 3 | 19 | 194 | 126 | 121 | 0 | 43 | 4 | 0 | 0 | 440 | 40 | 13 | 52 | 217 | 1552 | 1.73 | 899 |
| HH-III-iiUA | 333 | 48 | 10 | 23 | 246 | 224 | 153 | 13 | 39 | 1 | 0 | 0 | 486 | 22 | 5 | 70 | 257 | 1930 | 2.24 | 863 |
| HH-III-iiUG | 285 | 56 | 11 | 31 | 212 | 139 | 113 | 0 | 29 | 0 | 1 | 0 | 399 | 27 | 8 | 63 | 148 | 1522 | 2.02 | 755 |
| HH-III-iiiCG | 23 | 9 | 2 | 3 | 45 | 22 | 108 | 0 | 8 | 1 | 0 | 0 | 158 | 52 | 1 | 7 | 19 | 458 | 0.50 | 917 |
| HH-III-iiiGC | 78 | 14 | 24 | 15 | 38 | 66 | 100 | 0 | 12 | 2 | 1 | 0 | 173 | 27 | 2 | 18 | 38 | 608 | 0.57 | 1060 |
| HH-III-iiiGU | 52 | 18 | 10 | 13 | 47 | 51 | 83 | 1 | 18 | 1 | 11 | 1 | 235 | 7 | 4 | 9 | 43 | 604 | 0.53 | 1150 |
| HH-III-iiiUA | 43 | 15 | 7 | 3 | 72 | 64 | 137 | 2 | 8 | 3 | 304 | 0 | 98 | 133 | 3 | 18 | 126 | 1036 | 1.56 | 665 |
| HH-III-iiiUG | 26 | 17 | 6 | 5 | 40 | 33 | 82 | 3 | 17 | 1 | 152 | 0 | 97 | 64 | 4 | 8 | 58 | 613 | 0.67 | 917 |
| GenBank |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| relative size | 0.14 | 0.03 | 0.01 | 0.01 | 0.07 | 0.07 | 0.07 | <0.01 | 0.03 | <0.01 | <0.01 | <0.01 | 0.37 | 0.02 | 0.01 | 0.06 | 0.10 | 1.00 |  |  |

## IX. FIGURES

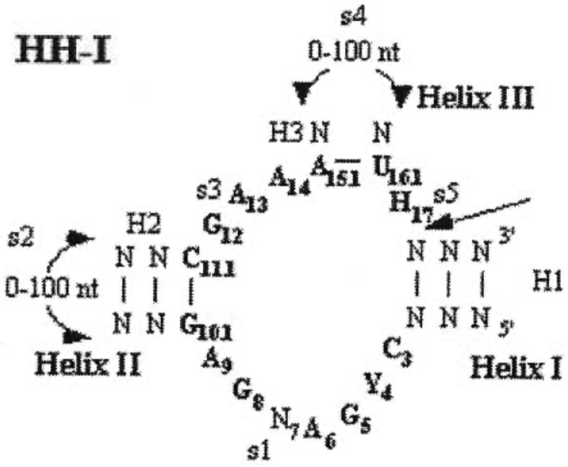
**Figure 1. Structures and descriptors of the hammerhead self-cleaving ribozyme motifs.**

The three descriptors: HH-I, HH-II and HH-III, are defined by which helix is at the 5' end and named according to the helix number (Hertel et al. 1992). Each descriptor is composed of single stranded (s) and double stranded (H) regions. The regions are first named in order from 5' to 3' and then specified for their length (minimum:maximum), number of mismatches (in the case of H only) and presence of specific nucleotides. For example, HH-I consist of the following features: H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1 where H1 is an helix of a fixed length of three base pairs with no mismatches and no specific nucleotides; H2 is also of three base pairs with no mismatches but with a starting G-C base pair; H3 is an helix of two base pairs beginning with an A-U base pair; s1 is a single stranded region of 7 nucleotides exactly with a specific sequence; s2 vary between 0 and 100 undetermined nucleotides; and so on. The hammerhead-like motifs are the same as the three shown but with an "N" replacing one of the nucleotides in bold or with a different identity of one of the base pairs in bold. These motifs are named according to the original motif and the position of the mutation, e.g., HH-I-3 motif is as HH-I but with an N instead of a C at position 3 thus HH-I-3 descriptor has a modified s1 as follow: s1 7:7 NYGANGA, similarly with HH-I-iiAU which is a HH-I motif with a A:U base pair in the Helix II instead of a G:C thus the descriptor HH-I-iiAU has this particularity: H2 3:3 0 ANN:NNU. The cleaving site is after $H_{17}$.

(H= A, C or U, N= A, C, G or U, Y= C or U). See Methods for the basis of the sequence requirements.

Descriptors:

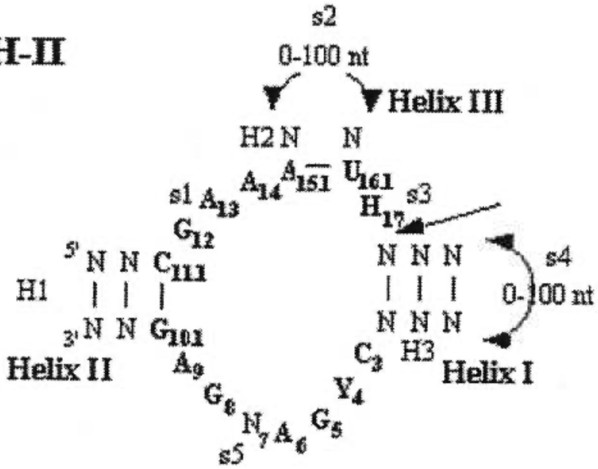
HH-I



H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

H1 3:3 0
H2 3:3 0 GNN:NNC
H3 2:2 0 AN:NU
s1 7:7 CYGANGA
s2 0:100
s3 3:3 GAA
s4 0:100
s5 1:1 H

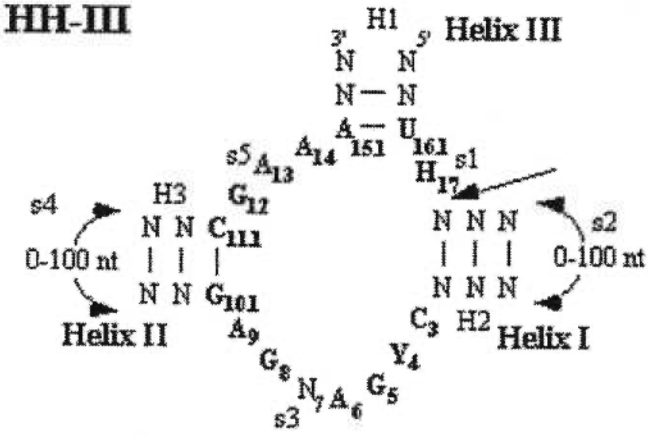
HH-II



H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

H1 3:3 0 NNC:GNN
H2 2:2 0 AN:NU
H3 3:3 0
s1 3:3 GAA
s2 0:100
s3 1:1 H
s4 0:100
s5 7:7 CYGANGA

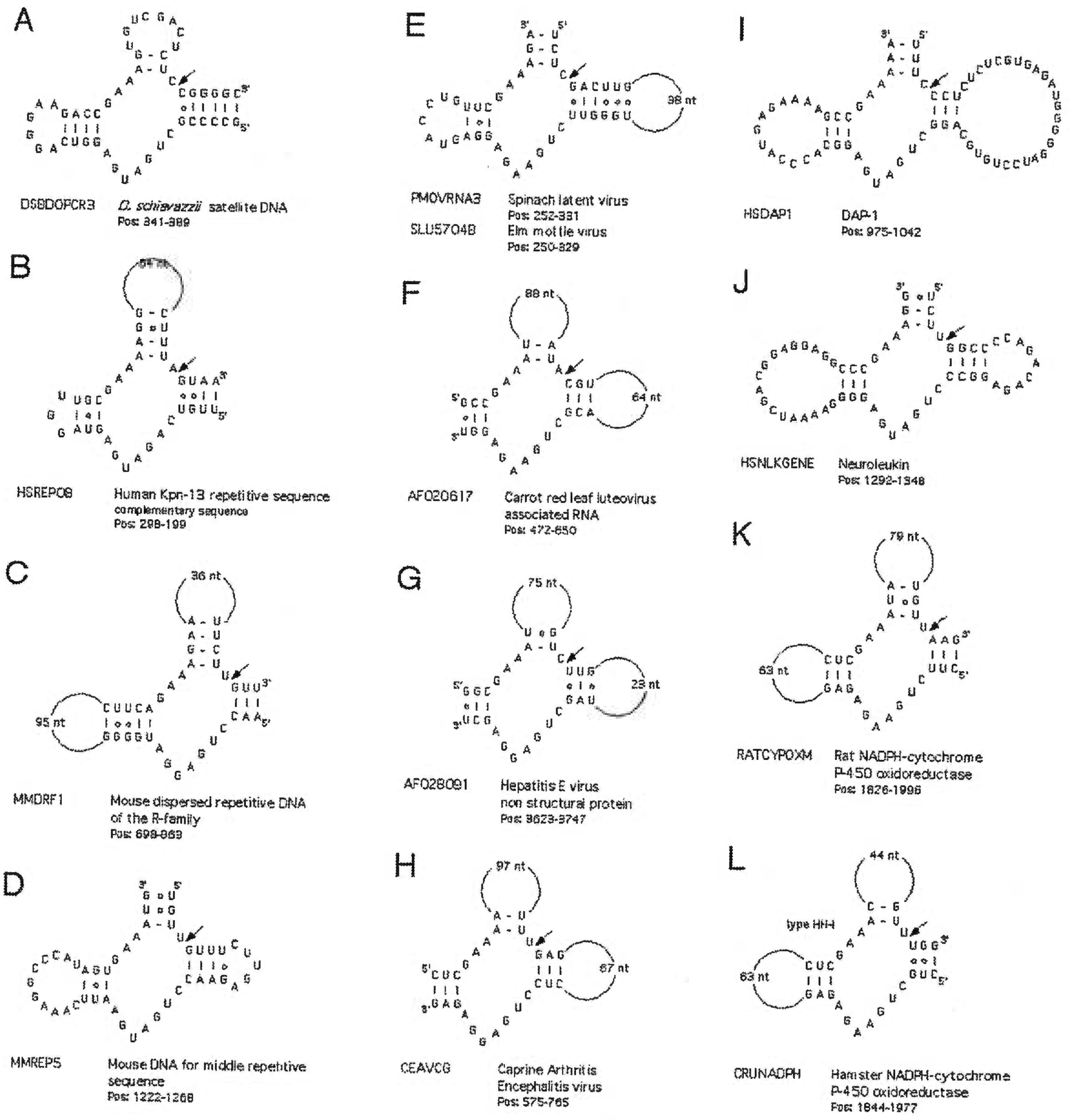
HH-III



H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

H1 3:3 0 NNU:ANN
H2 3:3 0
H3 3:3 0 GNN:NNC
s1 1:1 H
s2 0:100
s3 7:7 CYGANGA
s4 0:100
s5 3:3 GAA

**Figure 2.** **Putative hammerhead motifs.**



A DSBDOPCR3 *D. schiavazzii* satellite DNA Pos: 341-389

B HSREP08 Human Kpn-13 repetitive sequence complementary sequence Pos: 298-199

C MMDRF1 Mouse dispersed repetitive DNA of the R-family Pos: 698-869

D MMREP5 Mouse DNA for middle repetitive sequence Pos: 1222-1268

E PMOVRNA3 Spinach latent virus Pos: 252-331
SLU57048 Elm mottle virus Pos: 250-329

F AF020617 Carrot red leaf luteovirus associated RNA Pos: 472-650

G AF028091 Hepatitis E virus non structural protein Pos: 3623-3747

H CEAVCG Caprine Arthritis Encephalitis virus Pos: 575-765

I HSDAP1 DAP-1 Pos: 975-1042

J HSNLKGENE Neuroleukin Pos: 1292-1348

K RATCYPOXM Rat NADPH-cytochrome P-450 oxidoreductase Pos: 1826-1996

L CRUNADPH Hamster NADPH-cytochrome P-450 oxidoreductase Pos: 1844-1977

# CHAPITRE 4

## Amber suppression in *Escherichia coli* by unusual mitochondria-like transfer RNAs

*Proceedings of the National Academy of Sciences of the United States of America* **95**: 1375-1380 (1998)

Biochemistry

# Amber suppression in *Escherichia coli* by unusual mitochondria-like transfer RNAs

Véronique BOURDEAU*, Sergey V. STEINBERG*, Gerardo FERBEYRE,
Rémi EMOND, Nicolas CERMAKIAN, and Robert CEDERGREN†

Département de Biochimie, Université de Montréal
Montréal, PQ, Canada H3C 3J7

* V.B. and S.V.S. contributed equally to this work.

† To whom correspondence should be addressed at: Université de Montréal, Département de biochimie, C.P. 6128, succursale Centre-Ville, Montréal, PQ H3C 3J7, Canada. Email: ceder@bch.umontreal.ca.

Keywords: suppressor, mitochondrial, translation, evolution.

Abbreviations: mt, mitochondria; $tRNA_{su+}^{Ala}$, alanine suppressor transfer RNA; X-Gal, 5-bromo–4–chloro–3–indolyl–ß–D-galactopyranoside; DHFR, dihydrofolate reductase.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF003201-AF003233).

This paper was submitted directly (Track II) to the Proceedings Office.
Edited by Paul R. Schimmel, The Scripps Research Institute, La Jolla, CA, and approved December 4, 1997 (received for review September 23, 1997)

## NOTE

Contribution de chaque auteurs:

| | |
|---|---|
| V. Bourdeau : | a participé à toutes les expériences |
| S.V. Steinberg : | a développé l'hypothèse et analysé les ARNt obtenus |
| G. Ferbeyre : | a élaboré l'approche expérimentale |
| R. Emond : | a participé au criblage de la banque d'ARNt |
| N. Cermakian : | a participé à l'évaluation de l'activité des ARNt |
| R. Cedergren : | a dirigé le projet |

CONTENU

I. Abstract

II. Introduction

III. Materials and Methods

a) Strains

b) Construction of the Combinatorial Library

c) tRNA Levels and Aminoacyl-tRNA Levels

d) Identification of the Charged Amino Acid

IV. Results

a) Experimental Design

b) Clones Selection and Characterization

c) Sequence Analysis

V. Discussion

VI. Acknowledgements

VII. References

VIII. Tables

IX. Figures

## I. ABSTRACT

The "cloverleaf" base-pairing pattern was established as the structural paradigm of active tRNA species some 30 years ago. Nevertheless, this pattern does not accommodate the folding of certain mitochondrial tRNAs. For these recalcitrant tRNAs, we have proposed structures having from 5 to 10 base pairs in the anticodon stem rather than the canonical 6. The absence of these types of tRNAs in cytoplasmic translation systems, however, raises the possibility that they may not be *bona fide* alternate folding patterns for active tRNA molecules. For this reason, we have designed new tRNA genes based on our model of unusual mitochondrial tRNAs, having 7, 8, 9, and 10 base pairs in the anticodon stem with other modifications to the D-stem and connector regions. We show here that these synthetic genes produce tRNAs that actively suppress amber codons *in vivo*.

## II. INTRODUCTION

The vast majority of tRNAs encoded in the genomes of all cell types, chloroplasts, mitochondria (mt), and viruses fold into the standard cloverleaf secondary structural pattern. Some mt tRNA sequences are unusual, however, in that they do not fit this pattern. Structures for these tRNAs have been proposed based on optimal base pairing patterns and the hypothesis that the distance between the anticodon and acceptor stem must be identical in all tRNAs, so that they can extend themselves in the same way between the messenger RNA and the site of aminoacyl transfer on the ribosome (1, 2). These patterns involve a double zipper principle, where the 5-10 base pairs in the anticodon stem vary inversely with the number of base pairs in the D-stem and the length of the connector regions (Fig. 1). These changes are compensatory and result in three-dimensional structures for tRNAs, which are virtually superimposable on the normal three-dimensional "L" structure in spite of their unusual secondary structure. Other unusual features of mt translation including abbreviated ribosomal RNAs (3-5) and nonuniversal genetic codes (6-8) etc. raise the issue that these tRNAs may be simply another anomaly of mt rather than representatives of alternate folding patterns, because they are not found in cytoplasmic translation. We show here, to the contrary, that tRNAs incorporating these unusual features actively suppress amber mutations in *Escherichia coli*, and covariation analysis of their sequences supports the patterns that we have assigned to the unusual mt tRNAs.

## III. MATERIALS AND METHODS

### III. a) Strains

Three *E. coli* strains have been used: Top10 (F$^-$ *mrcA* Δ(*mrr-hdsRMS-mcrBC*) ϕ80*lac*ZΔM15 Δ*lac*X74 *deoR recA1 araD*139 Δ(*ara-leu*)7697 *galU galK rpsL endA1 nupG*) from Invitrogen; XAC-1 (F’ *lac*I$_{373}$*lac*Z$_{\mu 118\ \mathrm{am}}$ *pro*B+/F$^-$ (*lac-proB*)$_{\mathrm{XIII}}$ *nalA rif arg*E$_{\mathrm{am}}$ *ara*) (9); and XAC/A16 (Δ*lacproB nalA rif arg*E$_{\mathrm{am}}$/F’ *lac*I$_{\mathrm{q}}$ amber-Z fusion *proB*) containing the pDa3am plasmid (10).

### III. b) Construction of the Combinatorial Library

The template oligonucleotide coding for the combinatorial tRNA library and two flanking primers with restriction enzyme sites for EcoRI and PstI (sequences on request) were synthesized by General Synthesis and Diagnostics (Toronto). These were then PCR-amplified for 5 min at 95°C, followed by 30 cycles of: 30 s at 94°C, 30 s at 42°C, and 30 s at 72°C, with Vent DNA polymerase (2 units)/100 pmol of each primer/20 mM of dNTPs/100 ng of the template. All enzymes were from New England Biolabs. The double-stranded DNA obtained was then digested with PstI and EcoRI, purified on a Sephadex G-50 column, and cloned into the pGFIB-I plasmid, using a 3:1 insert-to-plasmid ratio and T4 DNA ligase. The plasmid had been predigested with the same enzymes and dephosphorylated with calf intestine alkaline phosphatase by using a 3:1 insert:plasmid ratio and T4 DNA ligase. Electroporation in the TOP10 strain yielded over 500,000 colonies (four times the sequence complexity of the library). Plasmid DNA from 12 randomly selected clones was isolated, and the sequences of the encoded tRNA genes confirmed the randomized nature of the expected positions. Plasmid DNA from this library was prepared by extracting the DNA from all 5 x $10^5$ colonies by using the alkaline lysis protocol (11). This preparation was used to transform cells of the XAC-1 strain. Cells were plated on LB medium with 50 µg/ml ampicillin and 20 µg/ml 5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside (X-Gal) and left to grow overnight at 37°C. Blue colonies developed during a 24-h incubation at 4°C. Plasmids from the blue colonies were isolated and used to retransform XAC-1 to ensure that the phenotype was dependent on the presence of the plasmid. All other protocols unless otherwise mentioned were carried out according to Sambrook et al. (11).

### III. c) tRNA Levels and Aminoacyl-tRNA Levels

The tRNA and aminoacyl-tRNA levels were measured by using an adaptation of previous protocols (12, 13). Cells from a fresh transformation of XAC-1 with isolated plasmids were grown in 10-ml cultures by using 2 x YT broth containing 50 µg/ml ampicillin. During log phase, cells were harvested by centrifugation at 4°C. All subsequent steps were carried out at 4°C. The pellet was resuspended in 0.5 ml of 0.3 M sodium acetate (pH 5.2) and 1 mM EDTA and extracted with an equal volume of phenol equilibrated with 0.3 M sodium acetate (pH 5.2). After vortexing for 1 min and 10 min on ice, the mixture was centrifuged to isolate the aqueous layer, which was subjected to precipitation on dry ice with 2 vol of ethanol. The pellet from centrifugation was washed with a 0.5-ml vol of 70% ethanol/10 mM sodium acetate (pH 5.2) and resuspended in 40 µl of 10 mM sodium acetate (pH 5.2)/1 mM EDTA. The concentration of the RNA isolations is determined according to the absorbance of a 1:1000 dilution. Samples of 2-4 µg of each RNA preparation were distributed in two tubes, one of which was maintained at 4°C, whereas 1.5 µl of 0.5 M Tris (pH 9) was added to the other and incubated at 37°C for 25 min. To each sample, 1.5 µl of loading buffer containing 0.1 M sodium acetate (pH 5.2), 8 M urea, 0.05% bromophenol blue, and 0.05% xylene cyanol was added to the samples, and they were loaded on a 6.5% polyacrylamide gel containing 8 M urea, 0.1 M sodium acetate (pH 5.2) and run overnight at 300 V at 4°C. After electrophoresis, RNA was transferred by electroblotting the portion of the gel around the xylene cyanol on a nylon membrane (Hybond-N, Amersham, Buckinghampshire, UK) and hybridized with a probe corresponding to the anticodon stem-loop of the tRNA library and with a probe for 5S RNA (positions 34-53 in the *E. coli* 5S sequence).

Aminoacylation levels were obtained from the quantity of aminoacyl-tRNA and uncharged tRNA in lanes that had not been treated with Tris by scanning autoradiograms of the gels and evaluating the intensity with the Scion Image PC program (Beta 1 release, 1997). tRNA levels were calculated by scanning the autoradiograms and normalizing to the quantity of 5S RNA in the same lane.

### III. d) Identification of the Charged Amino Acid

The identification of the amino acid charged on the tRNAs from clone T7 and T59 was accomplished by their introduction into strain XAC/A16 carrying the plasmid pDa3am, which contains an amber codon at the third position of the dihydrofolate reductase (DHFR) gene. DHFR was isolated on a methotrexate resin (Sigma) according to the protocol of McClain and Foss (10) and Normanly et al. (14) and subsequently eluted with folic acid. Separation of the folic acid from DHFR was done with a Bio-Rad Econopac Q column with a gradient from 0 to 1 M KCl. The fractions with DHFR were pooled and desalted on Centricon-30 (Amicon). DHFR was microsequenced in the laboratory of C. Lazure (Institut de recherche clinique de Montréal) after SDS/15% PAGE and transfer to a PVDF membrane (Amersham).

# IV. RESULTS

## IV. a) Experimental Design

Although the activity of these unusual structures in protein synthesis was the main purpose of this work, we also wished to evaluate the structural inferences that have been made to rationalize how the unusual tRNAs could be fitted to the standard three-dimensional structure (1, 2, 15). Thus, we adopted the strategy to incorporate the unusual features into the background of a completely different, non mitochondrial tRNA as a stringent test of our knowledge. Because the conversion of a normal tRNA molecule to the mt type requires modification of the D-stem and connector regions, a tRNA must be selected whose identity determinants are elsewhere so that amino acid charging ability and specificity would not be jeopardized. The *E. coli* $tRNA_{UGC}^{Ala}$ gene was chosen for modification because its main identity element, the G3-U70 base pair, is located in the acceptor stem (16-18). Fig. 2 shows the modifications of the *E. coli* $tRNA_{UGC}^{Ala}$ gene to incorporate the following mt characteristics. 1) The G26·A44 base pair was changed to G26·C44 in the designed gene. Because the geometry of this interaction may be a factor in arresting base pairing in the anticodon stem, rendering it a Watson-Crick interaction could favor extension of the stem (19). 2) G10 was deleted to eliminate the 10·25 base pair and to favor the formation of a 25·45 pair. Position 25 was represented in the designed gene as either an A, a common nucleotide at the last position of the anticodon stem, or a C to encourage propagation of the stem. 3) T8 was deleted from the gene sequence to shorten the connector region between the acceptor stem and the D-stem. This change plus deletion of G10 guarantees that tRNAs from this gene could not have a standard cloverleaf secondary structure (1, 15). 4) Positions 13, 14, 15, 21, and 22 of the D-domain were randomized to permit all combinations of nucleotides at these positions, which could be sensitive to both secondary and tertiary interactions. 5) The unpredictability of interactions at strategic positions of the variable loop prompted the additional randomization of positions 46, 47, and 48. 6) Finally, the TGC anticodon of tRNA$^{Ala}$ gene was converted to CTA to permit reading of the UAG amber codon. Judicious placement of this codon in a reporter gene allows simple evaluation of tRNA *in vivo* activity by the well known amber suppression test. Moreover, amber mutations in both the argE and the *lacI/lacZ* genes of the XAC-1 strain of *E. coli* prompted its use in this test: suppression would render the strain prototrophic for arginine and blue in the presence of X-Gal (9).

### IV. b) Clones Selection and Characterization

The tRNA gene library was prepared by PCR amplification of synthetic DNA encoding the tRNA gene with eight randomized (4 and 5 above) and one binary position (2 above) producing a library having a sequence complexity of 1.3 x 10$^5$. The amplification product was cloned into the pGFIB-I plasmid (9) and then used to transform the XAC-1 strain (see Materials and Methods). Plasmids from blue colonies were isolated and utilized to retransform XAC-1 to ensure that the suppressor phenotype was not a result of a host mutation. Fig. 3A shows the plate growth of several isolated clones in the presence and absence of X-Gal and in the absence of an arginine growth supplement. Generally, darker colored colonies also had better growth characteristics in the medium lacking arginine, suggesting a certain correlation between the suppression levels of the two amber codons.

Thirty-three plasmids demonstrated reproducible suppression of the amber mutations, and their characterization included the determination of the ß-galactosidase activity of cellular extracts and the sequence of the plasmid-borne tRNA genes (Table 1). Suppression varied between 0.3 and 40% of the tRNA$_{su+}^{Ala}$ control. Certain tRNA genes contained deletions (T59 and T39) or insertions of nucleotides (T28, T42, etc.) not present in the original gene.

tRNAs encoded in these plasmids were further characterized by the determination of their *in vivo* aminoacylation level and their level of expression. Fig. 3B is a representative autoradiogram of a polyacrylamide gel run at acid pH, demonstrating the amino acid charging level of the unusual tRNAs and the control $tRNA_{su+}^{Ala}$. Analysis of these data showed that aminoacylation of all tRNAs was between 70 and 90 ± 5% (Table 2). Although generally the charging level of $tRNAs_{u+}^{Ala}$ was greater than the variant tRNAs, these differences were within the experimental error. The steady state level for all tRNAs was also calculated by using the amount of 5S RNA as an internal standard. The levels of the variant tRNAs were found to be consistently lower than that of the control tRNA (Table 2). The specific activity of these tRNAs (that is the level of activity divided by quantity of aminoacyl-tRNA) is as much as three times the value of the control. This value must be taken with some caution, however, because suppression is saturable due to the negative effect of excessive read-through of stop codons. On the other hand, the gel mobility of the charged and noncharged tRNAs is consistent with the predicted length, indicating that nuclease processing of the precursor tRNA was not disrupted by the unusual structure of the tRNAs.

The identity of the amino acid inserted at the site of the amber codon was determined by using the pDa3am plasmid, which contains an amber mutation at the third codon of the encoded DHFR gene. The XAC/A16 strain containing the pDa3am plasmid was transformed with the plasmid bearing either the T7 or the T59 tRNA clone. These tRNA clones were chosen because they were associated with high ß-galactosidase activity and represented vastly different structural groups (see Table 1). DHFR was isolated on a methotrexate resin, and microsequencing of the protein from both transformants showed the presence of alanine at the position of the nonsense codon.

### IV. c) Sequence Analysis

In all, 22 of the 33 tRNA gene sequences had 7 base pairs in the anticodon stem and one or two connector nucleotides (Type 7-1 and 7-2, Fig. 4). For the remaining 11 tRNAs, the optimal base pairing produces structures with 8, 9, and 10 base pairs in the anticodon stem (Fig. 4). All tRNAs fold into patterns that have been proposed previously for mt tRNAs (1, 2). None of the 33 tRNA sequences could be folded into the common cloverleaf pattern.

The top of Fig. 4 focuses on the analysis of nucleotide covariation in the positions of the D-stem that were randomized in the original library. For Type 7-1 tRNAs, the third and fourth pairs of the D-stem, positions 13-22 and 14-21, are Watson-Crick pairs in 12 and 10 cases of 13, respectively. In all nine Type 7-2 tRNAs, the three D-stem base pairs are Watson-Crick base pairs. Because an insufficient number of unusual mt tRNA sequences is available in the sequence database to establish the base pairing pattern we have proposed, our data help to validate the base pairing pattern of at least the Type 7 tRNAs. Moreover, the fact that randomized positions in the synthetic library produced highly base paired D-stems testifies to the importance of this stem in active tRNA species.

## V. DISCUSSION

Because mt may be more tolerant to slower and/or less precise protein synthesis, we had anticipated that the mt-like tRNAs may not be active in cytoplasmic translation. Our suggestion that eukaryotes and archaea are protected from the alternate, mt-type, folding pattern of certain tRNAs by dimethylation of a particular guanosine (19) also raises the spector that these tRNAs would be detrimental. In contrast, our data show that these tRNAs are efficient suppressors (Table 2), although it is true that high speed and/or high precision protein synthesis in response to a normal codon has not been demonstrated. Only the state level of these tRNAs seems deficient, possibly reflecting instability or poor processing. On the other hand, efficient suppressors might only exist at low levels, because otherwise they would interfere with normal termination of translation. Also, the acid gel experiment shown in Fig. 3B demonstrates that aminoacylation and maturation of the 5'- and 3'-termini are normal.

The fact that these mt-like tRNAs function well as suppressor tRNAs in the cytoplasmic protein synthesis begs the question of why they are not normally found in cytoplasmic systems. Of course, our data address only the effect of the presence of these tRNAs during a few generations; the true evolutionary issue deals with the long term effect on the survival of the cell. It might be possible, however, that the conformational space of the unusual tRNAs was never explored during the evolution of the cytoplasmic protein synthesis system. We think not because: (i) the origin of mt is rooted in the $\alpha$-purple bacteria, and the unusual tRNAs must be considered as derived from normal structures rather than remnants of primordial molecules; (ii) the simple mechanism for the production of the unusual tRNAs has been proposed, which involves only slippage of base pairing in the D-stem area (1); (iii) the distribution of these tRNAs among distantly related mt clearly demonstrates several independent origins for this class of tRNAs (15). Therefore, this structure must have originated several times during the evolution of cytoplasmic translation, but then it must have been rejected systematically.

Our ability to prepare active tRNAs having unusual folding motifs in light of previous attempts (refs. 22 and 23 and V. Bourdeau et al., unpublished results) could be because of our gene design strategy and the use of randomized positions in the unusual gene sequence such that many sequence variants could be screened simultaneously. In any case, these experiments demonstrate that the classic cloverleaf pattern of tRNA is not a necessary condition for tRNA activity even in cytoplasmic protein synthesis and that the self-compensating structures that we have proposed for the unusual mt tRNAs more faithfully represent the structure/function paradigm of active tRNA species.

## VI. ACKNOWLEDGEMENTS

We thank Dr. C. Lazure of the Clinical Research Institute of Montréal for microsequencing of the DHFR protein, Dr. J. Abelson and Dr. W. McClain for the gift of strains and plasmids, and Dr. W. McClain for valuable comments on this work. This work was supported by a grant from the Natural Science and Engineering Research Council of Canada (NSERC), a NSERC scholarship to V. Bourdeau, and a Fonds pour la Formation de Chercheurs et l'Aide à la Recerche (Québec) scholarship to N. Cermakian. R.C. is the Richard Ivey Fellow of the Canadian Institute of Advanced Research.

## VII. REFERENCES

1. Steinberg, S. & Cedergren, R. (1994) *Nat. Struct. Biol.* **1**, 507-510.
2. Steinberg, S., Gautheret, D. & Cedergren, R. (1994) *J. Mol. Biol.* **236**, 982-989.
3. Eperon, I.-C., Anderson, S. & Nierlich, D. P. (1980) *Nature (London)* **286**, 460-467.
4. Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. (1981) *Cell* **26**, 167-180.
5. de la Cruz, V. F., Lake, J. A., Simpson, A. M. & Simpson, L. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1401-1405.
6. Barrell, B. G., Bankier, A. T. & Drouin, J. (1979) Nature (London) **282**, 189-194.
7. Bonitz, S. G., Berlani, R., Coruzzi, G., Li, M., Macino, G., Nobrega, F. G., Nobrega, M. P., Thalenfeld, B. E. & Tzagoloff, A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3167-3170.
8. Jukes, T. H. & Osawa, S. (1993) *Comp. Biochem. Physiol. [B]* **106**, 489-494.

9. Normanly, J., Masson, J. M., Kleina, L. G., Abelson, J. & Miller, J. H. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6548-6552.

10. McClain, W. H. & Foss, K. (1988) *J. Mol. Biol.* **202**, 697-709.

11. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning, a Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd. Ed..

12. Vashney, U., Lee, C.-P. & RajBandary, U. L. (1991) *J. Biol. Chem.* **266**, 24712-24718.

13. Gabriel, K., Schneider, J. & McClain, W. H. (1996) *Science* **271**, 195-197.

14. Normanly, J., Kleina, L. G., Masson, J. M., Abelson, J. & Miller, J.-H. (1990) *J. Mol. Biol.* **213**, 719-726.

15. Steinberg, S., Leclerc, F. & Cedergren, R. (1997) *J. Mol. Biol.* **266**, 269-282.

16. Prather, N. E., Murgola, E. J. & Mim, B. H. (1984) *J. Mol. Biol.* **172**, 177-184.

17. McClain, W. H. & Foss, K. (1988) *Science* **240**, 793-796.

18. Hou, Y. M. & Schimmel, P. (1988) *Nature (London)* **333**, 140-145.

19. Steinberg, S. & Cedergren, R. (1995) *RNA* **1**, 886-891.

20. Masson, J. M. & Miller, J. H. (1986) *Gene* **47**, 179-183.

21. Miller, J. H. (1972*) Experiments in Molecular Genetics* (Cold Spring Harbor Lab. Press, Plainview, NY).

22. Kumazawa, Y., Schwartzbach, C. J., Liao, H. X., Mizumoto, K., Kaziro, Y., Miura, K., Watanabe, K. & Spremulli, L. L. (1991) *Biochim. Biophys. Acta* **1090**, 167-172.

23. Hou, Y. M. & Schimmel, P. (1992) *Biochemistry* **31**, 4157-4160

## VIII. TABLES

**Table 1. ß-Galactosidase activities and gene sequences of suppressor tRNAs.**

ß -Galactosidase activity was determined using the method of Miller (21) in the presence of 50 µg/ml ampicillin and the use of chloroform and SDS. The activity is presented as a percentage of the activity of the control $tRNA_{su+}^{Ala}$. The same plasmid without a tRNA insert gave 0.075% activity. tRNA genes were sequenced using the Sanger method by the Organellar Genome Megasequencing Project Laboratory of the Université de Montréal under the direction of G. Burger. Sequences were grouped depending on the predicted number of base pairs in the anticodon stem. Type 7-1 and 7-2 tRNAs differ in the number of nucleotides (1 or 2) in the Connector 1. * indicates the presence of a bulged nucleotide in the anticodon stem (15). Nucleotides in the randomized positions are in bold; and those unexpected from the original design are underlined. Noncanonical base pairs are in italics. These sequence data have been submitted to the GenBank database under accession numbers AF003201-AF003233. AC refers to the aminoacceptor stem; D, to the D-stem; AN, to the anticodon stem; and T, to the T stem.

| Clone | % activity | AC | | D | | | | D | AN | | AN | | T | | T | AC | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Type 7-1 | | | | | | | | |
| T19 | 8 | GGGGCTA | A | CTC | GTC | TGG | GA | GAG | AGCCTGC | TTCTAAC | GCAGGCG | TGT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T24 | 2 | GGGGCTA | A | CTC | ATC | TGG | GA | GAG | AGCCTGC | TTCTAAC | GCAGGCG | CCT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T51 | 10 | GGGGCTA | A | CTC | AAC | TGG | GA | GAG | AGCCTGC | TTCTAAC | GCAGGCG | GTC | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T42 | 2 | GGGGCTA | A | CTGT | AGC | TGG | G | ACAG | AGCCTGC | TTCTAAC | GCAGGCG | TTT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T44 | 2 | GGGGCTA | A | CTCT | GC | TGG | G | AGAG | AGCCTGC | TTCTAAC | GCAGGCG | TAA | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T49 | 6 | GGGGCTA | A | CTCT | GC | TGG | G | AGAG | CGCCTGC | TTCTAAC | GCAGGCG | GCA | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T40 | 5 | GGGGCTA | A | CTCC | GC | TGG | G | GGAG | AGCCTGC | TTCTAAC | GCAGGCG | TTT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T46 | 10 | GGGGCTA | A | CTCC | AC | TGG | G | GAAG | CGCCTGC | TTCTAAC | GCAGGCG | GCG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T23 | 0.3 | GGGGCTA | A | CTTAC | C | TGG | | GTAAG | AGCCTGC | TTCTAAC | GCAGGCG | TAT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T52 | 0.9 | GGGGCTA | A | CTCAT | C | TGG | | GTGAG | AGCCTGC | TTCTAAC | GCAGGCG | TGT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T30 | 7 | GGGGCTA | A | CTTCC | - | TGG | | GGAAG | CGCCTGC | TTCTAAC | GCAGGCG | TGT | TGCGG | TTCGATT | CCGCA | TAGCTCC | ACCA |
| T15 | 0.8 | GGGGCTA | A | CTCTT | C | TGG | G | GAGAG | AGCCTGC | TTCTAAC | GCAGGCG | TTC | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T28 | 0.3 | GGGGCTA | A | CTTAC | C | TGG | GG | GTGAG | AGCCTGC | TTCTAAC | GCAGGCG | TCG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| | | | | | | | | | Type 7-2 | | | | | | | | |
| T16 | 3 | GGGGCTA | AC | TTGGC | | TGG | | GCCAG | AGCCTGC | TTCTAAC | GCAGGCG | TCT | TGCGG | TTCGATC | CCGCT | TAGCTCC | ACCA |
| T29 | 6 | GGGGCTA | AC | TTGTC | | TGG | | GACAG | AGCCTGC | TTCTAAC | GCAGGCG | TAG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T38 | 5 | GGGGCTA | AC | TTGTC | | TGG | | GACAG | AGCCTGC | TTCTAAC | GCAGGCG | TTCG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T43 | 23 | GGGGCTA | AC | TTGAC | | TGG | | GTCAG | AGCCTGC | TTCTAAC | GCAGGCG | GTG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T47 | 1 | GGGGCTA | AC | TTAGC | | TGG | | GCTAG | AGCCTGC | TTCTAAC | GCAGGCG | AAG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T48 | 2 | GGGGCTA | AC | TTTAC | | TGG | | GTAAG | CGCCTGC | TTCTAAC | GCAGGCG | GTG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T50 | 3 | GGGGCTA | AC | TTCTC | | TGG | | GAGAG | AGCCTGC | TTCTAAC | GCAGGCG | TTT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T59 | 40 | GGGGCTA | AC | TTGTC | | TGG | | GACAG | AGCCTGC | TTCTAAC | GCAGGC-T | AG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T31 | 2 | GGGGCTA | AC | TTCTC | | TGG | | GAGAG | CGCCTGC | TTCTAAC | GCAGGCG | TTG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| | | | | | | | | | Type 8 | | | | | | | | |
| T8 | 11 | GGGGCTA | A | CTC | CAC | TGG | GT | GAG | ACGCCTGC | TTCTAAC | GCAGGCGC | GC | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T25 | 4 | GGGGCTA | A | CTC | ACC | TGG | GC | GAG | GCGCCTGC | TTCTAAC | GCAGGCGT | AT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T12* | 5 | GGGGCTA | A | CTC | ACC | TGG | GC | GAG | GCG-CTGC | TTCTAAC | GCAGGCGT | AT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T1 | 3 | GGGGCTA | -C | TT-GC | | TGG | | GCTA | GCGCCTGC | TTCTAAC | GCATGCGC | CG | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T7 | 11 | GGGGCTA | A | CTCC | GC | TGG | G | GGAG | GCGCCTGC | TTCTAAC | GCAGGCGT | TT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T18 | 3 | GGGGCTA | A | CTTG | AC | TGG | G | CAAG | GCGCCTGC | TTCTAAC | GCAGGCGA | TT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T41* | 2 | GGGGCTA | A | CTCT | AC | TGG | G | AGAG | CGC-CTGC | TTCTAAC | GCAGGGCG | ATT | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| | | | | | | | | | Type 9 | | | | | | | | |
| T39 | 7 | GGGGCTA | A | CTC | A-C | TGG | - | GAG | AGCGCCTGC | TTCTAAC | GCAGGCGTT | T | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T22 | 0.7 | GGGGCTA | AC | TGC | AC | TGG | | GCG | AGCGCCTGC | TTCTAAC | GCAGGCGTT | T | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| T45 | 5 | GGGGCTA | AC | TCC | GC | TGG | | GGA | AGCGCCTGC | TTCTAAC | GCAGGCGTT | C | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |
| | | | | | | | | | Type 10 | | | | | | | | |
| T37 | 2 | GGGGCTA | AC | TAC | GC | TGG | | GTA | AGGCGCCTGC | TTCTAAC | GCAGGCGTTA | | TGCGG | TTCGATC | CCGCA | TAGCTCC | ACCA |

**Table 2. Levels of expression and aminoacylation.**

* Structural type defined in Fig. 4.

† Obtained by averaging ß-galactosidase activity from three independent cellular extracts (21).

‡ Determined by scanning the first lane of each clone in the autoradiogram of acid polyacrylamide gels (Fig. 3B). These values are the average of three independent experiments.

§ Obtained from the total tRNA (tRNA + aminoacylated tRNA) in each of the two lanes of the clones. The quantity of tRNA in each band was first normalized to the quantity of 5S RNA in the same lane (internal standard) and expressed as a percentage of the $tRNA_{su+}^{Ala}$ control (Fig. 3B). These values were obtained from three experiments.

| Clone | Type[*] | % of activity[†] | % of aminoacylation[‡] | % of expression[§] |
|---|---|---|---|---|
| Ala | 6 | 100 | 91 ± 4 | 100 |
| 46 | 7-1 | 10 ± 0.7 | 86 ± 10 | 5 ± 3 |
| 49 | 7-1 | 6 ± 0.3 | 83 ± 13 | 13 ± 12 |
| 43 | 7-2 | 23 ± 0.7 | 96 ± 6 | 6 ± 2 |
| 25 | 8 | 4 ± 0.6 | 86 ± 14 | 9 ± 7 |
| 7 | 8 | 11 ± 3.8 | 66 ± 24 | 12 ± 3 |
| 39 | 9 | 7 ± 0.1 | 90 ± 10 | 15 ± 8 |
| 37 | 10 | 2 ± 0.2 | 84 ± 14 | 22 ± 15 |

# IX. FIGURES

**Figure 1. The base-pairing pattern of mitochondrial tRNAs.**

This representation inspired by the three-dimensional L structure of the tRNA shows schematically how changes in the number of base pairs in the anticodon stem and those of the D-stem and the number of nucleotides in the connector regions can be compensatory (1-5). Each black or white circle represents a nucleotide, and pairing is indicated by a bar between positions. Alternate structures are produced by attaching the connector regions to the two nucleotides indicated in the boxes from 5 to 10. These numbers also represent the number of base pairs in the structure resulting from the attachment. AN refers to the anticodon loop and adjacent stem; D, to the D-loop and -stem; and T, to the T-loop and -stem.

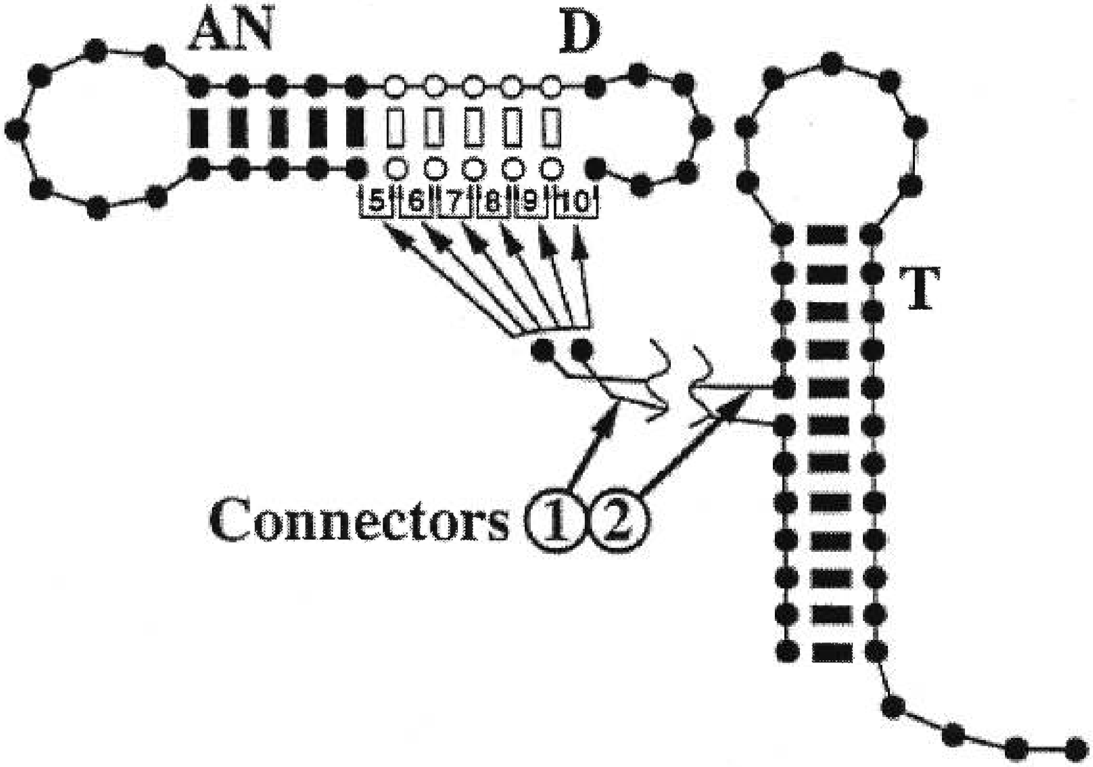

AN
D
T
5
6
7
8
9
10
Connectors ① ②

**Figure 2. Modifications to the tRNA$^{Ala}$ gene sequence.**

The base pair G26·A44 was converted to the Watson-Crick pair G26·C44. T8 and G10 were deleted. Nucleotide 25 was encoded by a mixture of A and C. Positions 13-15 and 21-22 of the D-domain as well as positions 46-48 of the variable loop were randomized, and the anticodon TGC was replaced by CTA. Ts are used here, because the gene sequence is given.

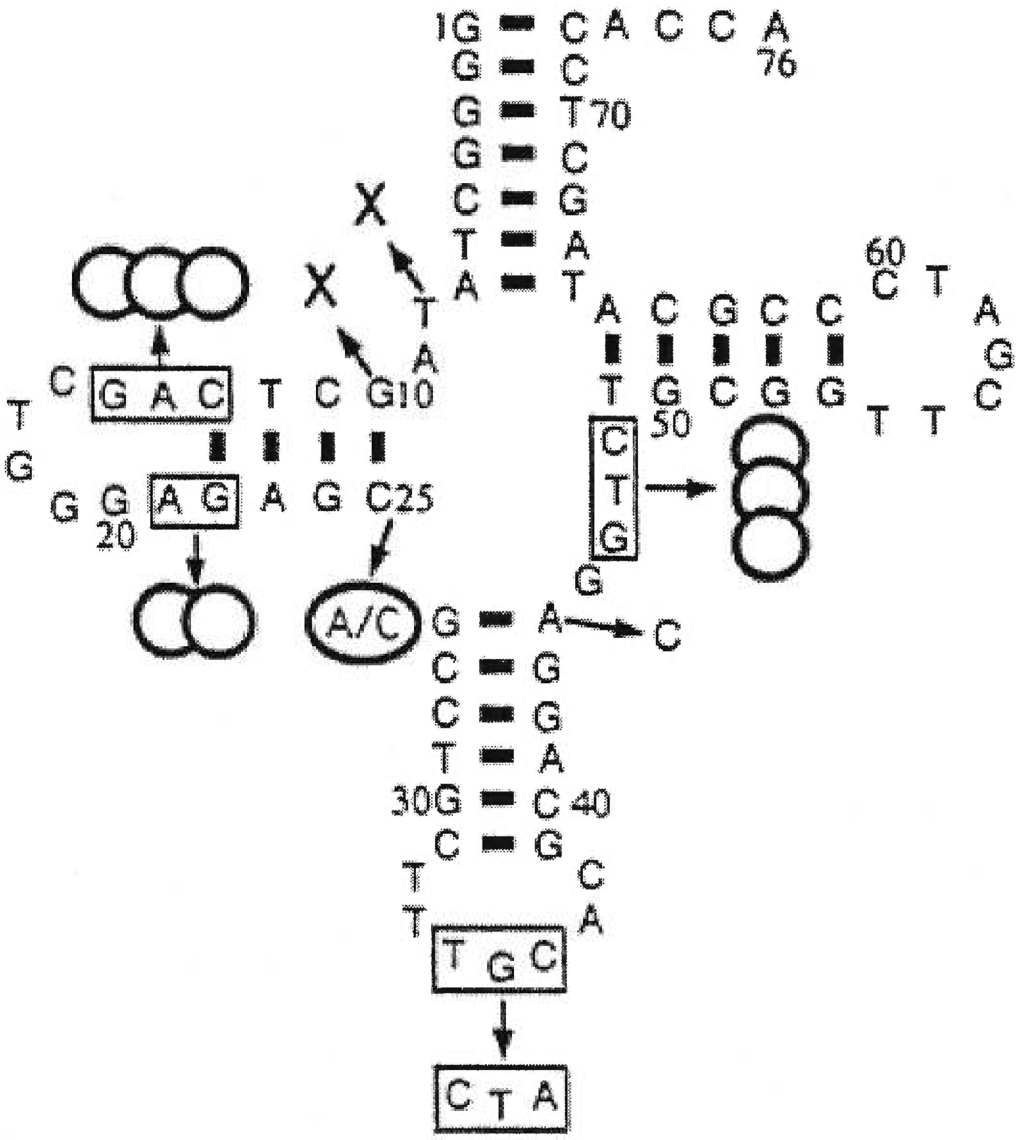

1G
76
T70
60
G10
50
20
C25
A/C
30G
C 40
T G C
C T A
C T G
G A C
A G
X
X

**Fig. 3. Properties of active tRNA clones.**

**(A)** Colony growth of different tRNA clones in XAC-1. Newly transformed colonies were replated on M9 minimal medium with ampicillin (50 µg/ml), without arginine, and with or without X-Gal (20 µg/ml). Ala represents the plasmid containing the $tRNA_{su+}^{Ala}$ gene; Phe, the $tRNA^{Phe}$ suppressor gene; and "–" the plasmid with no insert. The top image is the clone index; the middle shows colony growth in the presence of X-Gal with no added arginine; and the bottom shows growth in the absence of X-Gal and arginine. **(B)** Aminoacylation levels of various tRNA clones. Total RNA was isolated from different clones. One sample of each was treated with Tris (pH 9) and then analyzed with a second, untreated sample on a 6.5% polyacrylamide gel at pH 5.2 (see Materials and Methods). Lanes containing samples which were treated with Tris are indicated by "+," and those which were untreated are indicated by "–" Northern blot hybridization was carried out with a probe representing the anticodon part of the tRNA library and with a 5S probe. The percentage of aminoacylation was calculated by determining the intensity of the bands representing the aminoacylated and nonaminoacylated species. The levels of expression were determined in the same way and normalized to the level of the 5S RNA in each lane.

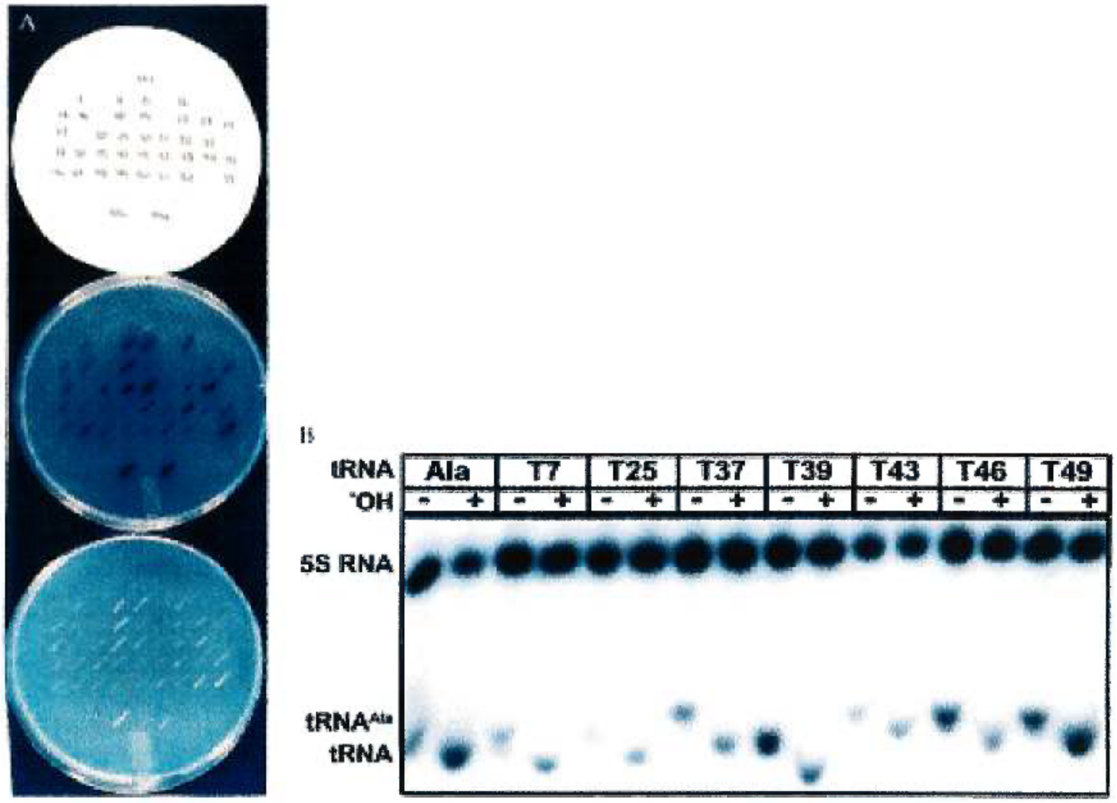
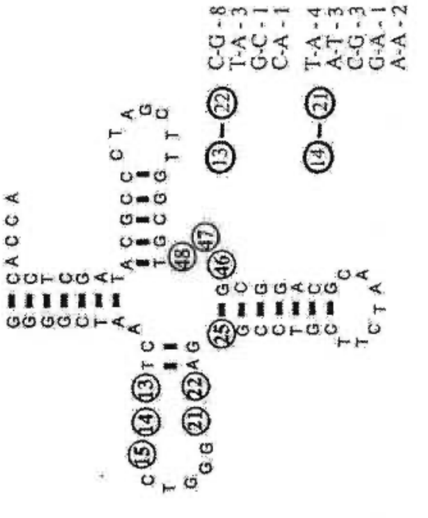
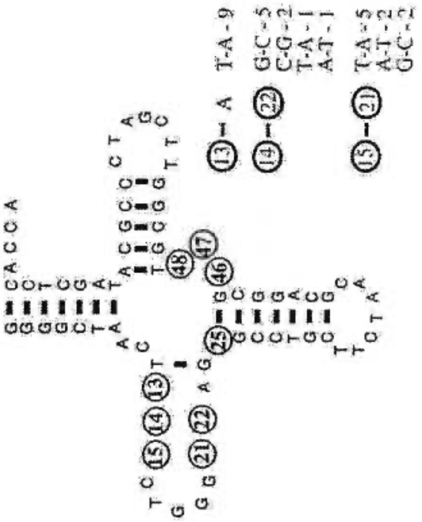

A
B
tRNA
Ala
T7
T25
T37
T39
T43
T46
T49
·OH
- + - + - + - + - + - + - + - +
5S RNA
tRNA^Ala
tRNA

**Fig. 4. Secondary structure of isolated tRNAs.**

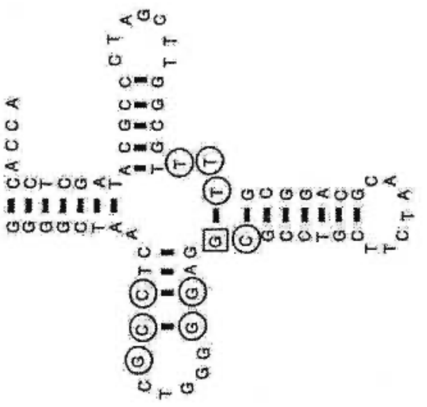
The optimal secondary structure patterns are shown for the type 7-1, 7-2, 8, 9, and 10 tRNAs, which were isolated in these experiments. Types 7-1 and 7-2 are shown with the covariation data from the D-stem of these tRNAs. None of these sequences can be folded into the canonical cloverleaf structure.



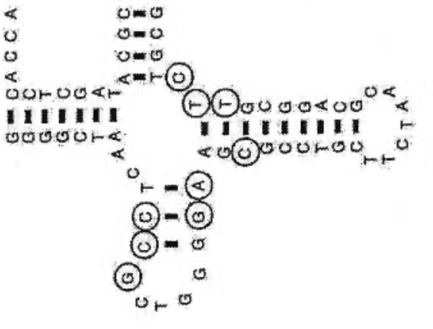
Type 7-1



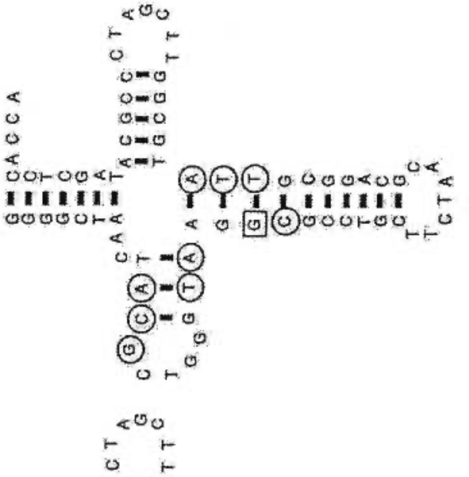
Type 7-2



Type 8
**T7**



Type 9
**T45**



Type 10
**T37**

# CHAPITRE 5

## Cleavage of mitochondria-like tRNAs Expressed in *Escherichia coli.*

*Unpublished*

# Cleavage of mitochondria-like tRNAs expressed in *Escherichia coli.*

Véronique BOURDEAU[1†], Theodore SMIRLIS[2], Robert CEDERGREN[1*], Bruno PAQUIN[1] and Nicholas DELIHAS[2]

1. Département de Biochimie, Université de Montréal,
Montréal, PQ, Canada H3C 3J7
2. Department of Molecular Genetics and Microbiology, School of Medicine,
SUNY, Stony Brook, NY, USA 11794-5222

† To whom correspondence should be addressed at: Université de Montréal, Département de biochimie, C.P. 6128, succursale Centre-Ville, Montréal, PQ H3C 3J7, Canada. Email: bourdeav@magellan.umontreal.ca.

*Deceased.

## NOTE

Contribution de chaque auteurs:

V. Bourdeau : a élaboré et effectué la majorité des expériences et a écrit l'article

T. Smirlis : a effectué des expériences du taux de mutation

R. Cedergren : directeur de V.B.

B. Paquin : a supervisé le projet et est superviseur de V.B.

N. Delihas : a supervisé le projet et est directeur de T.S.

CONTENU

I. Abstract

II. Introduction

III. Materials and Methods

a) Strains

b) Plasmids

c) *E. coli* S100 extracts

d) Processing of *in vitro* transcribed tRNAs

e) Northern blots

f) Measurements of Lac$^+$ reversion frequency

IV. Results

V. Discussion

VI. Acknowledgements

VII. References

VIII. Figures

## I. ABSTRACT

Mitochondrial (mt) transfer RNAs (tRNAs) often harbor many unusual structural features causing their secondary structure to differ from the conventional cloverleaf. Despite this fact, tRNAs with such irregularities, termed mt-like tRNAs, were found to be active in *Escherichia coli* as suppressors of mutated reporter genes. However, their steady-state level of expression and activity were low compared to a control $\text{tRNA}^{\text{Ala}}_{\text{su+}}$. To investigate the reason for their low expression, the processing of the *in vitro* transcripts of two mt-like tRNAs was evaluated using *E. coli* S100 extract. We found a defective 3' processing of the mt-like tRNA transcripts as well as a novel cleavage product. The activity responsible for this cleavage is sensitive to heat and phenol extraction and is stimulated by a low concentration of EDTA. No such cleavage product was obtained with a transcript of $\text{tRNA}^{\text{Ala}}_{\text{su+}}$. A similar fragment can be identified in total RNA extracted from growing cells indicating that this efficient cleavage activity also operates *in vivo*. Finally, we show that expression of mt-like tRNAs tends to increase the cellular mutation rate suggesting that the cleavage activity we have discovered might protect the bacteria from atypical tRNAs.

## II. INTRODUCTION

The secondary structure of transfer RNAs (tRNAs) is generally represented by a cloverleaf. However, not all tRNAs can adopt this pattern. In particular, mitochondrial (mt) tRNAs often possess one or more "odd" features including base substitution at conserved positions, mismatches, extended helices or lack of either the D arm or the T arm (1,2). Different foldings for these tRNAs have been proposed where structural compensations help to maintain the overall normal tertiary L-shape. These include cases where the combination of shorter connectors and a shorter D region would compensate for an extended anticodon stem, thus preserving a tertiary structure similar to that of the normal tRNAs (3-5). Another type of structural compensation consists of maintaining the length of the helical domain formed by short anticodon and D stems through the stacking of an extra bulge nucleotide or the intercalation of a nucleotide from the connectors (6). Thus, the tertiary compensations allow the irregular tRNAs to be capable of interacting with both the messenger RNA and the aminoacyl transfer site in the ribosome. In accordance with the first type of compensation, mt-like tRNAs with a longer anticodon stem and a shorter D region have recently been isolated for their ability to suppress a stop codon located within a reporter gene (7). Their activity in the cytoplasm of *Escherichia coli* proves the relevance of the structural compensation and the capacity of atypical tRNAs to be functional even where they are not naturally found. The efficiency of suppression of the mt-like tRNAs varied between 0.2% and 40% of that of the control $tRNA^{Ala}_{su+}$ while their steady-state level of expression reached only 5% to 22% of the expression obtained with the $tRNA^{Ala}_{su+}$ (7).

Many cases of mutant tRNAs with impaired expression level have been reported. Often, accumulation of precursor tRNA would be visible by Northern blot indicating the presence of a processing defect (example in ref. 8). However, this was not the case of the mt-like tRNAs expressed in *E. coli* where accumulation of such a precursor was not detected (7). In another case, a low level of accumulation of a mutated initiator tRNA *in vivo* was shown to correlate with the instability of both the precursor and the mature tRNA when incubated in *E. coli* S100 extract (9).

We thus choose to use the approach of *in vitro* processing in *E. coli* S100 extract to investigate the fate of the transcripts of two mt-like tRNAs. Our study shows an impaired 3' processing of the mt-like precursor tRNA transcripts as well as the appearance and accumulation of a new fragment resulting from a cleavage of the precursor tRNAs within the mt-like tRNA sequence. Interestingly, expresion of mt-like tRNAs in *E. coli* also leads to the appearance of a cleavage product that could be the result of the same activity we found *in vitro*. Finally, we reported that even if the mt-like tRNAs are expressed at low levels in *E. coli*, their suppression level can be similar to what is obtained with equivalent levels of $tRNA^{Ala}_{su+}$. Conversely, we also present evidence that a slightly higher mutation frequency is observed in cells containing mt-like tRNAs.

## III. MATERIAL AND METHODS

### a) Strains

Two *E. coli* strains were used: XAC-1 (F' $lacI_{373}lacZ_{u118}$ am *proB*+/F⁻ $\Delta$(*lac-proB*)$_{XIII}$ *nalA rif* $argE_{am}$ *ara*) (10) for cloning and *in vivo* experiments and the MRE 600 strain lacking RNaseI (*rna*) (11) for the preparation of the S100 extract.

### b) Plasmids

The $tRNA^{Ala}_{su+}$ as well as two clones isolated from the original library: T7 and T37 (see ref. 7), were subcloned by PCR into pBlueScript $SK^{+}$ (Stratagene) that possesses both T3 and T7 RNA polymerase promoters (PCR protocol is in ref. 7; the other cloning protocols were from ref. 12). The isolated clones were sequenced to confirm the presence of the correct tRNA gene in front of the T3 RNA polymerase promoter. The resulting plasmids were named pBS-Ala, pBS-T7 and pBS-T37.

### c) *E. coli* S100 extracts

Preparation of S100 extracts from *E. coli* MRE 600 was performed as described in RajBhandary & Ghosh (13) with the following adaptations: cells were first diluted in PBS containing 1 mM phenylmethylsulfonylfluoride (PMSF) and 5 mM dithiothreitol (DTT), incubated with lysozyme for 30 minutes on ice, lysed, sonicated and centrifuged 2h at 25 000 rpm (SW28; ~100,000 x g). We used for the column a Q-sepharose resin (Pharmacia).

One mM PMSF and 5 mM DTT were added to all solutions. The dialysis buffer contained 50% glycerol instead of 10% glycerol / 15% polyethylene glycol.

**d) Processing of *in vitro* transcribed tRNAs**

Plasmids pBS-Ala, pBS-T7 or pBS-T37 were digested with *Hind*III (New England Biolabs – NEB), purified on agarose gel with QIAgen columns and then used as a template for *in vitro* transcription (100 ng per reactions). The incubation mixture (50 µL total) consisted of 40 mM Tris-HCl (pH 8.0), 25 mM NaCl, 8 mM $MgCl_2$, 2 mM spermidine-(HCl), 5 mM DTT, 0.4 mM each of ATP, CTP and GTP, 0.2 mM of UTP, 15 µCi of $\alpha$-$^{32}$P-UTP and 1 µL of T3 RNA polymerase (50 units, GIBCO). After two hours at 37°C, reactions were extracted with equal volumes of phenol/chloroform and ethanol precipitated. Products were resolved on 7% polyacrylamide gels containing 8 M urea. Transcripts were located by autoradiography and eluted by incubating the gel pieces at 37°C overnight in 0.5 ml of 0.5 M ammonium acetate, 0.01 M magnesium acetate, 1 mM EDTA and 0.1% SDS. Labeled RNA was recovered by precipitation with two volumes of ethanol in the presence of glycogen carrier (20 µg/ml) and dissolved in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA. Aliquots containing ~200,000 cpm of radioactivity were diluted to 50 µL with transcription buffer. Processing reactions started by the addition of either 1.2 µL of water or of *E. coli* S100 extract (~5 µg total protein) and incubated at 37°C. Aliquots were removed at various times (0, 5, 15, 30 and 60 minutes) and extracted with phenol/chloroform in the presence of carrier tRNA (50 µg). The extracted RNA was precipitated with ethanol in the presence of glycogen and analyzed on a 7% polyacrylamide/8 M urea gel. When pre-treatments were involved, the extract was first diluted 1:2 in water and incubated at the desired temperature for 15 minutes or diluted 1:10 and extracted with an equal volume of phenol/chloroform. The volume equivalent to 1.2 µL of the non-diluted extract was then added to the reaction.

Alternatively, to lable transcripts at the 5'-end, *in vitro* transcription was carried out with 0.4 mM each of all four NTPs. The full-length transcript was located by ethidium bromide staining of the acrylamide gel, cut and eluted as above. The RNA obtained was dephosphorylated with CIP (Calf intestine phosphatase – NEB) in the presence of RNase inhibitor and phosphorylated with PNK (T4 Polynucleotide kinase – NEB) in the presence of $\gamma$-$^{32}$P-ATP (30µCi at >5000mCi/mmol).

### e) Northern blots

Unlabelled *in vitro* transcripts of tRNA and mt-like tRNAs were incubated in S100 extracts and aliquots were run on polyacrylamide gel as described above. Similarly, total RNA from cultured cells was extracted and fractionated on a 6.5% polyacrylamide/ 8 M urea/ 0.1 M sodium acetate (pH 5.2) gel (7,14,15). The RNA was transferred by electroblotting onto a nylon membrane (Hybond-N, Amersham, Buckinghampshire, UK). Hybridization with probes complementary to the anticodon stem-loop, the 5' or 3'-end of the primary transcripts and control 5S RNA (positions 34–53 in the *E. coli* 5S sequence) was carried on as described in McClain et al. (15) (probe sequences available upon request).

### f) Measurements of Lac$^{+}$ reversion frequency

Strains CC101-CC106 (graciously provided by J.H. Miller, 16) were transformed by heat shock with pGFIB-tRNA$^{Ala}_{su+}$, pSUF-tRNA$^{Ala}_{su+}$, pGFIB-T7 or pGFIB-T37. Glycerol stocks from fresh colonies of the strains alone or transformed with different plasmids were prepared and kept at –70°C. LB cultures supplemented with 60 µg/ml ampicillin (for pGFIB-tRNA$^{Ala}_{su+}$, pGFIB-T7 and pGFIB-T37) or 25 µg/ml chloramphenicol (for pSUF-tRNA$^{Ala}_{su+}$) were inoculated from the glycerol stocks and left to grow overnight at 37°C with agitation. Cell number was evaluated by optical density at 600 nm. Dilutions of the cultures were plated on rich LB medium, with antibiotic when needed, to mesure the number of colony forming cells (or living cells). After an overnight growth at 37°C, the number of colonies was counted. In parallel, ~$10^9$ cells were plated on lactose minimal medium with 5-bromo-4-chloro-3-indolyl-ß-D-galactoside (X-gal), and antibiotic when required, and incubated for 48 hours at 37°C. The number of revertant $Lac^+$ colonies was then counted and the frequency of reversion was reported over $10^8$ living cells: [number of $Lac^+$ x 500 / number of colonies on LB x $10^9$] x $10^8$. The final frequency of reversion (shown in Fig. 7) was calculated from the results of 11 separate cultures by the method of the median (17) but in the cases were the median was ≤ 0.5, we used the method of the proportion of cultures without mutants (17).

## IV. RESULTS

In order to evaluate the reason for the low expression level of mt-like tRNAs in *E. coli*, the kinetics of the processing and the stability of mt-like tRNAs obtained *in vitro* were investigated. We chose to work with two mt-like tRNAs isolated previously (7): the first, tRNA-T7, has eight base pairs in the anticodon stem and the second, tRNA-T37, has a stem of ten base pairs (Fig. 1). We used $tRNA^{Ala}_{su+}$ with the normal six base pairs in the anticodon stem as a control (Fig. 1). These tRNAs were cloned downstream of a T3 RNA polymerase promoter so that they could be transcribed *in vitro* (see Material and Methods). Purified transcripts were incubated at 37°C for different times in the presence of *E. coli* S100 extract. The latter contains the necessary enzymatic activities to obtain mature tRNAs from normal *E. coli* precursor tRNA transcripts (9,13,18). The products were then separated by electrophoresis on a 7% polyacrylamide/8 M urea gel. The bands appearing after 5 minutes incubation and migrating close to the tRNA precursors represent a progressive exonucleolytic 3' end processing (Fig. 2). A band migrating in the position of the mature form of the tRNA became visible between 15 and 30 minutes (Fig. 2). Two bands were distinguished that represent either the last steps of the 3' processing (sometimes dependent on previous RNaseP cleavage) or the removal of AMP from the 3'-CCA terminus (example in ref. 19; reviewed in ref. 20). The identity of all the bands was confirmed by probing Northern blots of unlabeled processing profiles with oligonucleotides corresponding to either the 5' or the 3' region of the full transcript (data not shown).

In contrast to the processing profile of $tRNA^{Ala}_{su+}$, the 3' processing of the mt-like tRNAs seemed to be impaired since some of the intermediates of this process accumulated with time (Fig. 2). Even more striking was the presence of an additional band (marked by a star in Fig. 2) in both processing profiles of the T7 and T37 mt-like tRNA transcripts. This band, or the "X" fragment, has an intermediate size migrating between that of the 3' processing products of the full-length transcript and the mature mt-like tRNA.

We performed a Southern blot analysis of plasmid DNA using the "X" fragment as a probe to determine which parts of the primary transcript were present in the "X" fragment. The plasmid containing the mt-like tRNA-T7 gene was digested with restriction enzymes in order to generate fragments containing either the whole tRNA gene, the 5' half or the 3' half of its primary transcript. We found that the "X" fragment hybridized to the 5' half of the primary transcript (data not shown) suggesting that the fragment contained unprocessed 5' sequences, but had lost a part of the 3' region. In confirmation, the "X" fragment was visualized by probing a Northern blot of an unlabeled processing profile with an oligonucleotide complementary to the 5' extremity of the full-length transcript (not shown). Then, we investigated whether the first nucleotide of the *in vitro* transcription was present in the "X" fragment by comparing the processing bands obtained with transcripts labeled internally (transcription with $\alpha$-$^{32}$P-UTP) or at the 5'-terminus (unlabeled transcript dephosphorylated and rephosphorylated in 5' with a radioactive phosphate - $\gamma$-$^{32}$P-ATP; see Material and Methods). As shown in Figure 3, the bands corresponding to the 3' processing of the $tRNA^{Ala}_{su+}$ were both obtained when the transcripts were labeled internally or at the 5'-end, but only the pre-tRNA transcripts labeled internally displayed the mature tRNA band. In the case of the mt-like tRNAs-T7 (Fig. 3), the 3' processing as well as the "X" fragment were visible with the transcripts from both labeling reactions. This indicated that the first nucleotide of the *in vitro* transcript was still a part of the "X" fragment and that any change(s) that diminished the size of the band occurred on the 3' side.

Subsequently, we performed experiments to optimize the cleavage reaction. A pre-incubation of the extract at room temperature (24°C), before the incubation with the transcripts, did not affect the activity responsible for the formation of the fragment (Fig. 4A). However, a similar pre-incubation at 65°C or 75°C or a phenol/chloroform extraction of the extract prior to incubation with the *in vitro* transcripts prevented the formation of the "X" fragment as well as the 3' processing event (Fig. 4A). These results indicate that the activity contained a protein component. We also tested the activity for sensitivity to some ions and to ATP. Variation of ionic strength through an increase of the NaCl concentration (from 25 mM to 150 mM), or of the $Mg^{2+}$ concentration (from 8 mM to 25 mM) or even addition of ATP (up to 5 mM) did not markedly affect the formation of the "X" fragment (Fig. 4B). However, addition of 5 mM or 10 mM of EDTA increased significantly the amount of the "X" fragment. A further augmentation of EDTA to 15 mM resulted in slower 3' processing and retardation in the formation of the "X" fragment (Fig. 4B). Thus, the activity responsible for the cleavage of the mt-like tRNA transcripts was either independent of divalent ions concentration or was influenced by structural folding variations that could occur in the tRNA transcripts when EDTA was present. Possibly, EDTA acted to limit the availability of $Mg^{2+}$.

It is known that the formation of the proper tertiary structure of tRNA transcripts requires the presence of $Mg^{2+}$ ions (21,22). Another factor able to influence the formation of stable tertiary interactions is temperature (23,24). In fact, tRNA transcripts, which do not have base modifications, are more readily melted than their fully modified counterparts, especially at low $Mg^{2+}$ concentration (21). To evaluate the instability of the tRNA tertiary structure contributed to the “X” fragment formation, we repeated the experiments at 24°C in the presence of only 3 mM $Mg^{2+}$. At 24°C, the mt-like tRNA should have the possibility to form more stable tertiary interactions than at 37°C. Nonetheless, the “X” fragment was still present under this condition (Fig. 5). The processing seemed overall slower (or more degraded), even with the $tRNA^{Ala}_{su+}$, transcript, probably because 24°C is a sub-optimal temperature for the enzymatic activities of the *E. coli* extract. We also evaluated the possibility of weakening the tertiary structure of $tRNA^{Ala}_{su+}$ transcript in an attempt to obtain a processing profile similar to those of the mt-like tRNAs. For this, we incubated $tRNA^{Ala}_{su+}$ transcripts in S100 extract at 50°C. However, no extra bands were observed after one hour incubation (Fig. 5). Because it was already determined that the formation of the “X” fragment was heat sensitive, we did not incubated at higher temperature. Thus, the cleavage producing the “X” fragment *in vitro* seems to be specific to the mt-like tRNAs.

Since all the previous experiments were conducted *in vitro* using *E. coli* extracts, we decided to investigate if a similar phenomenon could be observed *in vivo*. Northern blot analysis using a specific probe exclusively recognizing the anticodon of suppressor tRNAs showed that mt-like tRNAs were present at a significantly lower level than $tRNA^{Ala}_{su+}$ (Fig. 6, left; see also ref. 7). In contrast, when we used an oligonucleotide complementary to the first 19 nucleotides of the predicted *in vivo* primary transcript (nine nucleotides upstream of the mature mt-like tRNAs and ten nucleotides within the mt-like tRNA sequence), we observed another band (Fig. 6, right). In this case, the observed band is smaller than the mature mt-like tRNA which is consistent with the fact that, in the *in vivo* expression system, the 5’ region is only nine nucleotides long instead of 65 nucleotides. Thus, a fragment containing a short 5’ region but lacking a part of the 3’ half of the tRNA would become smaller than the mature form of the mt-like tRNA. Note that the second probe contained two mismatches to the $tRNA^{Ala}_{su+}$, which explained why the hybridization to the mature forms of this tRNA was not as strong as the one with the mt-like tRNAs (Fig. 6). The finding *in vivo* of a band that could be the equivalent of to the "X" fragment found *in vitro* suggests an explanation for the low steady-state levels of mature mt-like tRNA observed: proper accumulation was prevented by a cleavage of the precursor somewhere within the mt-like tRNA sequence. The cleavage occurs probably in the anticodon region since the probe complementary to that region does not bind to the band corresponding to the *in vivo* fragment.

Considering the cleavage of mt-like tRNA primary transcripts and the important reduction of steady-state levels of mature mt-like tRNAs *in vivo* versus $tRNA^{Ala}_{su+}$, we decided to re-evaluate the activity of the mt-like tRNAs. First, we wanted to know whether the observed activity reported in our previous work (7) corresponded to the activity obtained using similar amount of $tRNA^{Ala}_{su+}$. We thus used three new plasmid vectors that express the $tRNA^{Ala}_{su+}$ at lower levels (kindly provided by Gabriel and McClain, ref. 14) to compare steady-state expression levels and suppression activities. Figure 7A shows the relative steady-state level of expression of the mature tRNAs (the expression of mature $tRNA^{Ala}_{su+}$ in the pGFIB plasmid is set at 100%), and Figure 7B displays the relative suppression efficiency evaluated through ß-galactosidase activity (the activity of the cells containing pGFIB-$tRNA^{Ala}_{su+}$ is again set at 100%). The different expression levels of $tRNA^{Ala}_{su+}$ correlated well with their corresponding suppression activity. Moreover, the suppression system did not appear to be saturated in the range of expression we reached. The steady-state expression levels and suppression activities obtained with the mt-like tRNAs are hence easy to compare with similar levels of $tRNA^{Ala}_{su+}$. Mt-like tRNA-T7 in pGFIB plasmid was present at a level halfway between those of pSUF-$tRNA^{Ala}_{su+}$ and pB2-$tRNA^{Ala}_{su+}$ vectors. Its suppression activity was also between the average of those two clones. This means that the efficiency of the mature mt-like tRNA-T7 was similar to that of the two clones of the mature $tRNA^{Ala}_{su+}$. However, the mt-like tRNA-T37 seemed to have a lower efficiency than $tRNA^{Ala}_{su+}$ because its expression level was intermediate between those of pSUF-$tRNA^{Ala}_{su+}$ and pB2-$tRNA^{Ala}_{su+}$ but the activity corresponded to a vector expressing less $tRNA^{Ala}_{su+}$: pBATS-$tRNA^{Ala}_{su+}$.

The presence of one or more enzymes that degrade mt-like tRNAs suggested that these mt-like tRNAs might be harmful to the normal cell function. We thus investigated whether mt-like tRNAs could increase the mutation rate. We took advantage of the system developed by Cupples & Miller consisting of six *E. coli* strains containing an episomal *lacZ* gene altered at an essential glutamic acid (codon position 461, ref. 16). In each of these strains a specific mutation is needed to restore the right amino acid and revert the Lac$^-$ phenotype to Lac$^+$ (strains CC101: A-T → C-G, strains CC102: G-C → A-T, CC103: G-C → C-G, CC104: G-C → T-A, CC105: A-T → T-A, and CC106: A-T → G-C). These strains have been used to determine the mutagenic specificity of mutagenes and mutator alleles (16,25). We thus transformed pGFIB-tRNA$^{Ala}_{su+}$, pSUF-tRNA$^{Ala}_{su+}$ (since it expressed levels of tRNA$^{Ala}_{su+}$ similar to the steady-state levels of mt-like tRNAs), pGFIB-T7 and pGFIB-T37 in the six strains (kindly provided by J.H. Miller, 16), to evaluate whether the mt-like tRNAs have an altered mutagenic effect relative to the tRNA$^{Ala}_{su+}$. In four of the strains, little or no difference between the mutation rate of the mt-like tRNAs and the tRNA$^{Ala}_{su+}$ were obtained (not shown). Nevertheless, in two of the strains significant differences from the controls were observed. With the expression of mt-like tRNA-T7, we found that the number of Lac$^+$ revertants per $10^8$ living cells was higher than for the tRNA$^{Ala}_{su+}$ expressed from either the pSUF (expressing levels of tRNA$^{Ala}_{su+}$ equivalent to the steady-state ones of mt-like tRNAs) or the pGFIB plasmids (highest expression of tRNA$^{Ala}_{su+}$ then mt-like tRNAs) when a G-C → C-G transversion was required (strain CC103). Similarly, the expression of the mt-like tRNA-T37 caused more A-T → G-C transitions and thus more Lac$^+$ reversion than tRNA$^{Ala}_{su+}$ expressed from the pGFIB plasmid (strain CC106). Thus, the mt-like tRNA-T7 seemed to be a slight mutator with a higher rate of G-C → C-G transversions while the mt-like tRNA-T37 was a mutator stimulating A-T → G-C transitions. Taking into account the short term nature of the experiments described above, performed under controlled laboratory conditions, it seems plausible that mt-like tRNAs can be even more harmful over a longer time lapse and in natural conditions.

## V. DISCUSSION

We report the discovery of the aberrant processing of mt-like tRNAs *in vitro* and *in vivo*. Our *in vitro* results indicate that in the presence of *E. coli* S100 extract, precursors of mt-like tRNAs are processed more slowly at their 3' extremity and, more importantly, that a fragment resulting from a cleavage within the mt-like tRNA sequence accumulates more readily than the mature form (Fig. 2). Since pre-incubation at 65-75°C or phenol extraction inactivated the activity producing this fragment (Fig. 4A), we believe that it was dependent, totally or partially, on a protein component. Moreover, we observed that the activity forming the fragment did not appear to be dependent on the presence of divalent ions since an increase in the accumulation of the fragment was observed upon addition of EDTA. In the same line of logic, increasing the $Mg^{2+}$ concentration did not improve the formation of the "X" fragment (Fig. 4B). Several *E. coli* endoribonucleases do not require divalent cations: RNases I, IV, F and N (26). We can discard RNase I because the S100 extracts were prepared from cells mutated for this enzyme and the remaining RNases are unlikely since they are not known to efficiently cleave within tRNA sequences.

Because of the increased formation of the "X" fragment upon addition of EDTA (indirect reduction of free $Mg^{2+}$ to around 3 mM when 5 mM of EDTA are added), we also suspected that lowering the $Mg^{2+}$ concentration would increase the susceptibility of mt-like tRNAs to cleavage. Since the presence of $Mg^{2+}$ is known to be essential for the proper tertiary folding of *in vitro* transcribed tRNAs (22), the stability of the tertiary structure might constitute the recognition element for the activity observed. Yet, our attempts to better stabilize mt-like tRNAs-T7 and -T37 or to destabilize $tRNA^{Ala}_{su+}$ tertiary interactions through temperature variations did not show any effects in the specificity of the “X” fragment formation from mt-like tRNAs (Fig. 5). This indicates that there is a specific feature in the mt-like tRNAs (sequence or stability of secondary and/or tertiary structure) that is recognized by a cellular activity. Two events might be occurring. First, the transcript of the mt-like tRNAs could fold in an alternative secondary structure promoting the specific cleavage activity. Another explanation deals with the possibility of a defect in the tertiary structure of the mt-like tRNAs, when compared to normal tRNAs, that we could not mimic for the $tRNA^{Ala}_{su+}$ under the chosen conditions. This would be consistent with the recent report of Hayashi and al. (27) suggesting a new type of thermal instability for the bovine mt $tRNA^{Ser}_{UGA}$. The observation is very important especially in regard to the possible roles that this cleavage activity could play in normal bacterial cells. For example, it might have a function in the turnover of cellular tRNAs degrading both misfolded newly synthesized tRNAs (arising as the result of transcription, processing or folding errors) and “old” tRNAs, two processes poorly understood. It is attractive to think that this activity might even be protecting the cells from harmful effects of atypical tRNAs like the ones we have studied here.

We are thus reporting the detection of the activity of an enzyme from *E. coli* that cleaves within mt-like tRNAs during the processing of their precursor. To date, only few enzymes, or tRNAses, have been identified that can efficiently cleave within tRNA sequences. Of great interest are $tRNA^{Lys}$-specific anticodon nuclease (ACNase, gene PrrC) and the colicin E5. ACNase was first identified in a rare strain of *E. coli* and later found in other bacteria. It cleaves cellular $tRNAs^{Lys}$ at the 5’ of the wobble position. It is expressed in a latent form and its activity is triggered by a viral infection creating a bacterial suicide response (28,29; review in ref. 30). On the other hand, colicin E5 is a plasmid encoded toxin expressed by *E. coli* to inhibit the protein synthesis of neighboring bacteria. After binding to a specific receptor and subsequent translocation, it cleaves Tyr, His, Asn and Asp tRNAs at the 3’ of the wobble position occupied by a queunine modification (31). Finally, it is worth mentioning a member of the RNase A family, angiogenin, which cuts tRNAs better than all other members of its family (32). It will be interesting to evaluate how widespread and important is the activity observed here.

The identification of an *in vivo* fragment derived from the mt-like tRNAs and the low steady-state level of their mature forms also prompted us to re-evaluate the actual translation efficiency of the mature form. Our results show that the activity of suppression obtained with the mt-like tRNAs is comparable to the one attained by similar steady-state levels of $tRNA^{Ala}_{su+}$. This suggests that the evolutionary reason for atypical tRNAs of this type to be absent from cytoplasmic translation systems is not their weak activity. Something else must be at stake, possibly misreading. We then evaluated the mutator activity of the mt-like tRNAs compared to the $tRNA^{Ala}_{su+}$ expecting that, as suppressor tRNAs, they would cause at least a certain level of mutations. Our results indeed showed that mt-like tRNA-T7 induced G-C → C-G transversions and mt-like tRNA-T37 stimulated A-T → G-C transitions, respectively, more than the $tRNA^{Ala}_{su+}$. Interestingly, this is not the first report of a mutator effect of tRNAs. Slupska et al. (33) reported the identification of two mutator tRNAs isolated after mutagenesis: *mutA* and *mutC* (25). Their frequencies of reversion toward A-T → T-A and G-C → T-A transversions were significantly higher than the ones reported here. However, both *mutA* and *mutC* encoded structurally normal glycine tRNAs with a different anticodon corresponding to aspartic acid. This mutation represented an important change in the properties of the amino acid to be inserted and the insertions were assumed to be efficient since it occurs via perfect codon-anticodon interaction. In the presented case of the mt-like tRNAs, misreading instead of normal codon-anticodon interactions would most likely be responsible for the mutator activity. Moreover, the mt-like tRNAs bear an alanine amino acid that probably would not frequently cause important structural defects in proteins since it is a small, uncharged amino acid. Further studies will be required to fully characterize the misreading activity of the mt-like tRNA. However, our observation of the mutator effect of mt-like tRNAs makes it tempting to speculate that the efficient cleavage activity we discovered is a defense mechanism against inappropriate tRNAs.

## VI. ACKNOWLEDGMENTS

The authors wish to express their gratitude to Dawn Ashburn for her assistance as well as to Dr L. Brakier-Gingras, Dr. G. Ferbeyre and Dr S.V. Steinberg for valuable comments and the review of the manuscript. We thank Dr. J. Abelson, Dr. W.H. McClain and Dr. J. H. Miller for the gift of strains and plasmids. This work was supported by a grant from the Natural Science and Engineering Research Council of Canada (NSERC) and a NSERC scholarship to V.B. The late R.C. was Richard Ivey Fellow of the Canadian Institute of Advanced Research (CIAR) program in Evolutionary Biology. N.D. is supported by DRAC Award, School of Medicine, SUNY, Stony Brook.

## VII. REFERENCES

1. Dirheimer, G., Keith, G., Dumas, P. and Westhof, E. (1995) "Primary, Secondary, and Tertiary Structures of tRNAs", In Söll, D., and RajBhandary, Y. (eds.), tRNA: Structure, Biosynthesis, and Function. American Society for Microbiology, Washington, DC, pp. 93-126.

2. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. and Steinberg, S. (1998) *Nucl. Acids Res.*, **26,** 148-153.

3. Steinberg, S. and Cedergren, R. (1994) *Nat. Struct. Biol.*, **1,** 507-510.

4. Steinberg, S., Gautheret, D. and Cedergren, R. (1994) *J. Mol. Biol.*, **236,** 982-989.

5. Watanabe, Y., Kawai, G., Yokogawa, T., Hayashi, N., Kumazawa, Y., Ueda, T., Nishikawa, K., Hirao, I., Miura, K. and Watanabe, K. (1994) *Nucl. Acids Res.*, **22,** 5378-5384.

6. Steinberg, S., Leclerc, F. and Cedergren, R. (1997) *J. Mol. Biol.*, **266,** 269-282.

7. Bourdeau, V., Steinberg, S. V., Ferbeyre, G., Émond, R., Cermakian, N. and Cedergren, R. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95,** 1375-1380.

8. McClain, W. H. and Foss, K. (1988) *Science*, **241,** 1804-1807.

9. Ramesh, V., Varshney, U. and RajBhandary, U. L. (1997) *RNA*, **3,** 1220-1232.

10. Normanly, J., Masson, J. M., Kleina, L. G., Abelson, J. and Miller, J. H. (1986) *Proc. Natl. Acad. Sci. U.S.A.*, **83,** 6548-6552.

11. Cammack, K. A. and Wade, H. E. (1965) *Biochem. J.*, **96,** 671-680.

12. Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning, a laboratory manual, 2nd Ed. Ed., Cold Spring Harbor Laboratory Press, Plainview, NY.

13. RajBhandary, U. L. and Ghosh, H. P. (1969) *J. Biol. Chem.*, **244,** 1104-1113.

14. Gabriel, K. and McClain, W. H. (1999) *J. Mol. Biol.*, **290,** 385-389.

15. McClain, W. H., Jou, Y. Y., Bhattacharya, S., Gabriel, K. and Schneider, J. (1999) *J. Mol. Biol.*, **290,** 391-409.

16. Cupples, C. G. and Miller, J. H. (1989) *Proc. Natl. Acad. Sci. U.S.A.*, **86,** 5345-5349.

17. Lea, D. W. and Coulson, C. A. (1949) *J. Genet.*, **49,** 264-285.

18. Schedl, P., Roberts, J. and Primakoff, P. (1976) *Cell*, **8,** 581-594.

19. Li, Z. and Deutscher, M. P. (1994) *J. Biol. Chem.*, **269,** 6064-6071.

20. Deutscher, M. P. (1995) "tRNA Processing Nucleases", In Söll, D., and Rajbhandary, U. L. (eds.), tRNA: Structure, Biosynthesis and Function. Am. Soc. Micro., Washington, D.C., pp. 51-65.

21. Derrick, W. B. and Horowitz, J. (1993) *Nucl. Acids Res.*, **21,** 4948-4953.

22. Maglott, E. J., Deo, S. S., Przykorska, A. and Glick, G. D. (1998) *Biochemistry*, **37,** 16349-16359.

23. Crothers, D. M., Cole, P. E., Hilbers, C. W. and Shulman, R. G. (1974) *J. Mol. Biol.*, **87,** 63-88.

24. Hilbers, C. W., Robillard, G. T., Shulman, R. G., Blake, R. D., Webb, P. K., Fresco, R. and Riesner, D. (1976) *Biochemistry*, **15,** 1874-1882.

25. Michaels, M. L., Cruz, C. and Miller, J. H. (1990) *Proc. Natl. Acad. Sci. U.S.A.*, **87,** 9211-9215.

26. Deutscher, M. P. (1985) *Cell*, **40,** 731-732.

27. Hayashi, I., Kawai, G. and Watanabe, K. (1998) *J. Mol. Biol.*, **284,** 57-69.

28. David, M., Borasio, G. D. and Kaufmann, G. (1982) *Proc. Natl. Acad. Sci. U.S.A.*, **79,** 7097-7101.

29. Levitz, R., Chapman, D., Amitsur, M., Green, R., Snyder, L. and Kaufmann, G. (1990) *EMBO J.*, **9,** 1383-1389.

30. Kaufmann, G. (2000) *Trends Biochem. Sci.*, **25,** 70-74.

31. Ogawa, T., Tomita, K., Ueda, T., Watanabe, K., Uozumi, T. and Masaki, H. (1999) *Science*, **283,** 2097-2100.

32. Saxena, S. K., Rybak, S. M., Davey, R. T. J., Youle, R. J. and Ackerman, E. J. (1992) *J. Biol. Chem.*, **267,** 21982-21986.

33. Slupska, M. M., Baikalov, C., Lloyd, R. and Miller, J. H. (1996) *Proc. Natl. Acad. Sci. U.S.A.*, **93,** 4380-4385.

## VIII. FIGURES

**Figure 1.** Representations of the secondary structure of tRNA$^{Ala}_{su+}$ and the mt-like tRNAs –T7 and –T37.

The three tRNAs used in this study are represented in their cloverleaf secondary structures. The modifications made in the tRNA$^{Ala}_{su+}$ sequence to construct the library of mt-like tRNAs, from which the tRNAs –T7 and –T37 were isolated (7), are indicated. Note that both mt-like tRNAs contained an insertion of a G in position 23.

"X →" stands for a deletion in the library. The randomized positions are circled. The open square corresponds to a partially randomized position allowing either A or C. "C →" indicates the replacement of a nucleotide by a C.

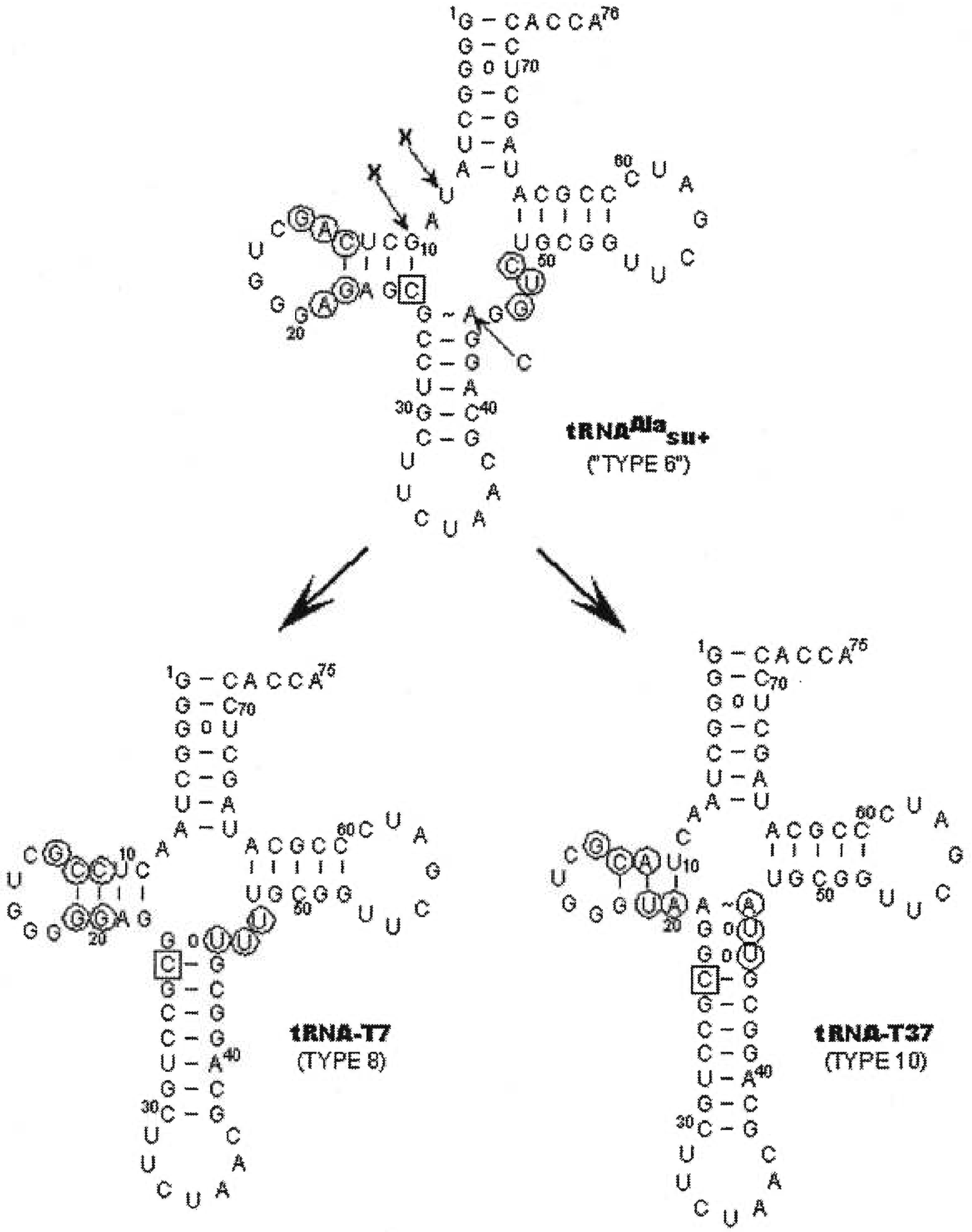

tRNAAla su+
("TYPE 6")
tRNA-T7
(TYPE 8)
tRNA-T37
(TYPE 10)

**Figure 2.** *In vitro* processing of tRNA$^{Ala}_{su+}$ and the mt-like tRNAs –T7 and –T37.

Autoradiography of 7% polyacrylamide/8 M urea gel containing samples of tRNA primary transcripts incubated for various times in the absence or in the presence of S100 extract (details in Material and Methods). The first bands resulting from the 3' processing and the mature tRNA bands are indicated. "*" highlights the novel cleavage product (the "X" fragment).

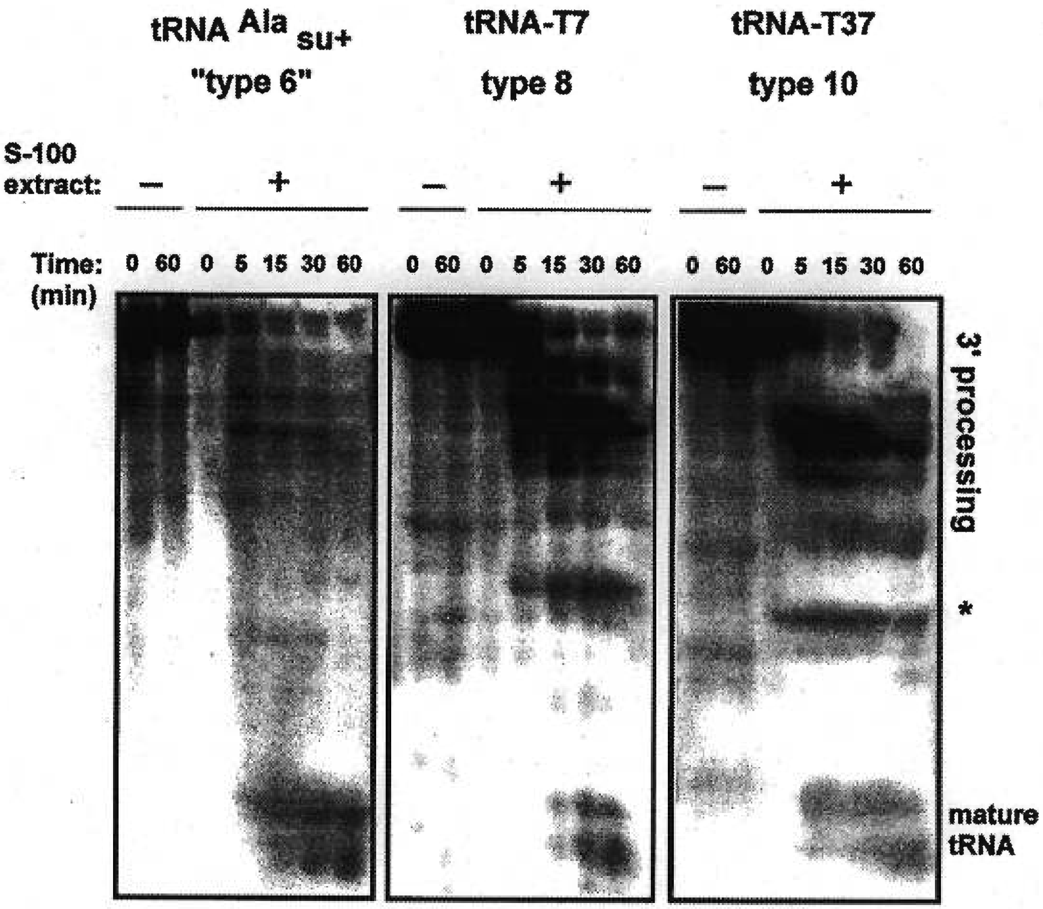

tRNA Ala su+
"type 6"
tRNA-T7
type 8
tRNA-T37
type 10
S-100
extract:
–
+
–
+
–
+
Time:
(min)
0 60 0 5 15 30 60
0 60 0 5 15 30 60
0 60 0 5 15 30 60
3' processing
*
mature
tRNA

**Figure 3.** *In vitro* processing of internally or 5'-end labeled precursor tRNA transcripts.

Transcripts of $tRNA^{Ala}_{su+}$ and of mt-like tRNA-T7 were synthesized in the presence of $\alpha$-$^{32}$P-UTP (lanes 1-2 for $tRNA^{Ala}_{su+}$ and lanes 5-6 for mt-like tRNA-T7) or labeled after transcription with $\gamma$-$^{32}$P-ATP (lanes 3-4 for $tRNA^{Ala}_{su+}$ and lanes 7-8 for mt-like tRNA-T7) (see Material and Methods for details). The primary transcripts were incubated with *E. coli* S100 extract for 30 minutes. Samples were resolved on a 7% polyacrylamide/8 M urea gel. This autoradiography shows bands resulting from the 3' processing, the complete processing (mature tRNA for $tRNA^{Ala}_{su+}$) and a previously unreported cleavage (the "X" fragment), as indicated.

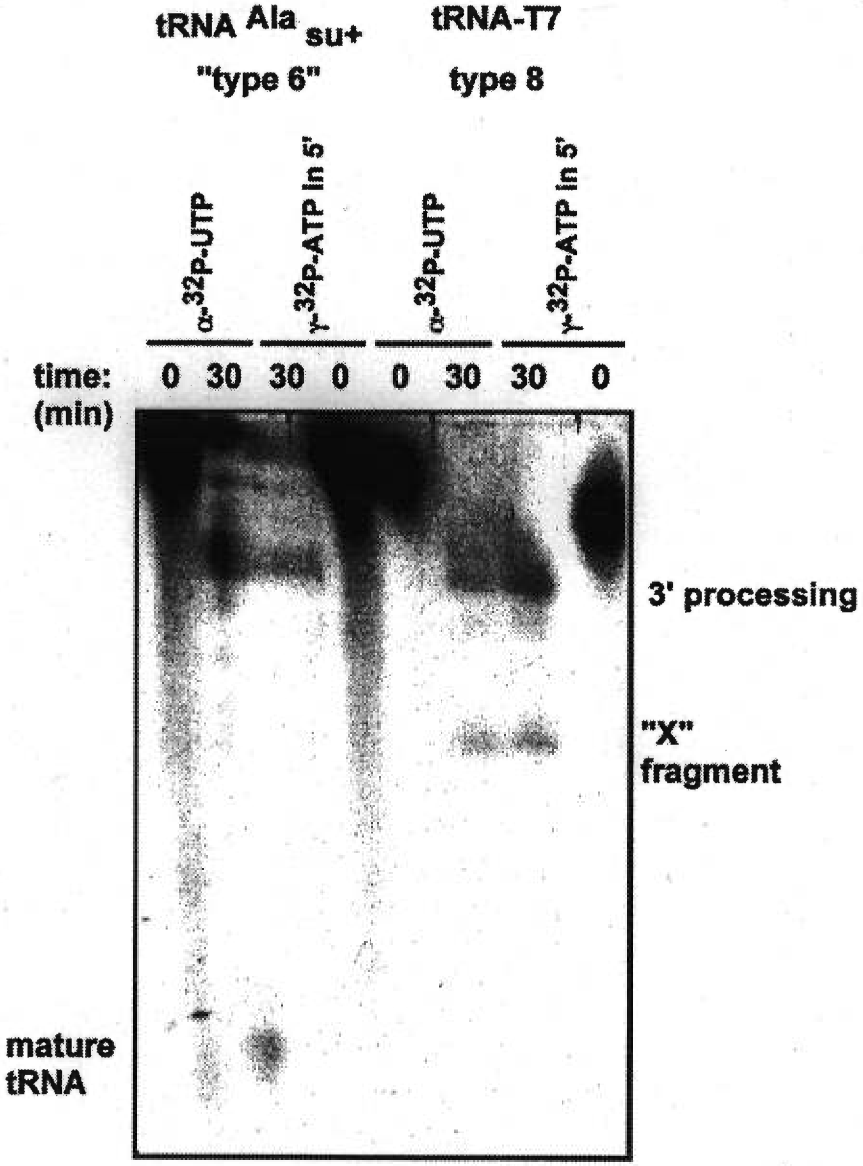
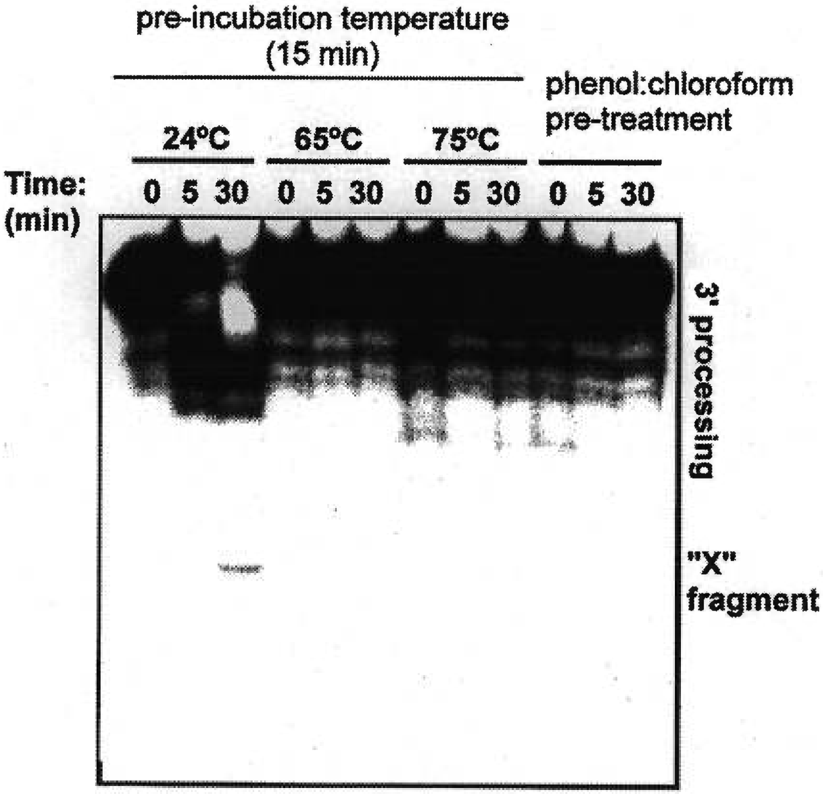

tRNA Ala su+
"type 6"
tRNA-T7
type 8
α-32P-UTP
γ-32P-ATP in 5'
α-32P-UTP
γ-32P-ATP in 5'
time:
(min)
0
30
30
0
0
30
30
0
3' processing
"X"
fragment
mature
tRNA

**Figure 4.** Characterization of the activity producing the "X" fragment.

A) Effect of temperature and phenol/chloroform extraction: *In vitro* processing of mt-like tRNA-T37 primary transcript after pre-incubation of the *E. coli* S100 extract at the indicated temperatures or after a short extraction with phenol/chloroform. B) Buffer conditions: *In vitro* processing of mt-like tRNA-T37 primary transcript in the transcription buffer only (–: lanes 1-3) or in the presence of additional NaCl (lanes 3-6), $MgCl_2$ (lanes 7-9), EDTA (lanes 10 through 18) or ATP (lanes 19-21). In each case, samples were taken before and after 5 or 30 minutes of incubation in *E. coli* S100 extract. The bands resulting from the 3' processing as well as the "X" fragment are indicated.


pre-incubation temperature
(15 min)
phenol:chloroform
pre-treatment
24°C
65°C
75°C
Time:
(min)
0 5 30 0 5 30 0 5 30 0 5 30
3' processing
"X"
fragment

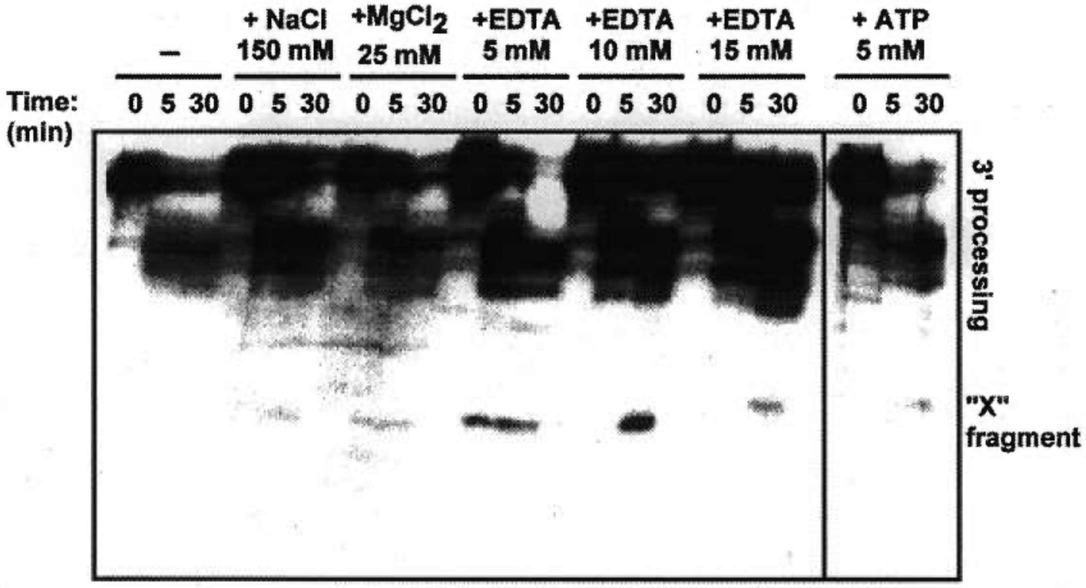

–
+ NaCl 150 mM
+MgCl2 25 mM
+EDTA 5 mM
+EDTA 10 mM
+EDTA 15 mM
+ ATP 5 mM
Time: (min)
0 5 30 0 5 30 0 5 30 0 5 30 0 5 30 0 5 30 0 5 30
3' processing
"X" fragment

**Figure 5.** Effect of the temperature on the formation of the "X" fragment.

Autoradiography of an *in vitro* processing of $\text{tRNA}^{\text{Ala}}_{\text{su+}}$, mt-like tRNA–T7 and –T37 transcripts incubated for the indicated times in the absence or in the presence of *E. coli* S100 extract at three different temperatures: 24°C, 37°C (as previously) and 50°C. The transcription buffer was supplemented with 5 mM EDTA for these experiments.

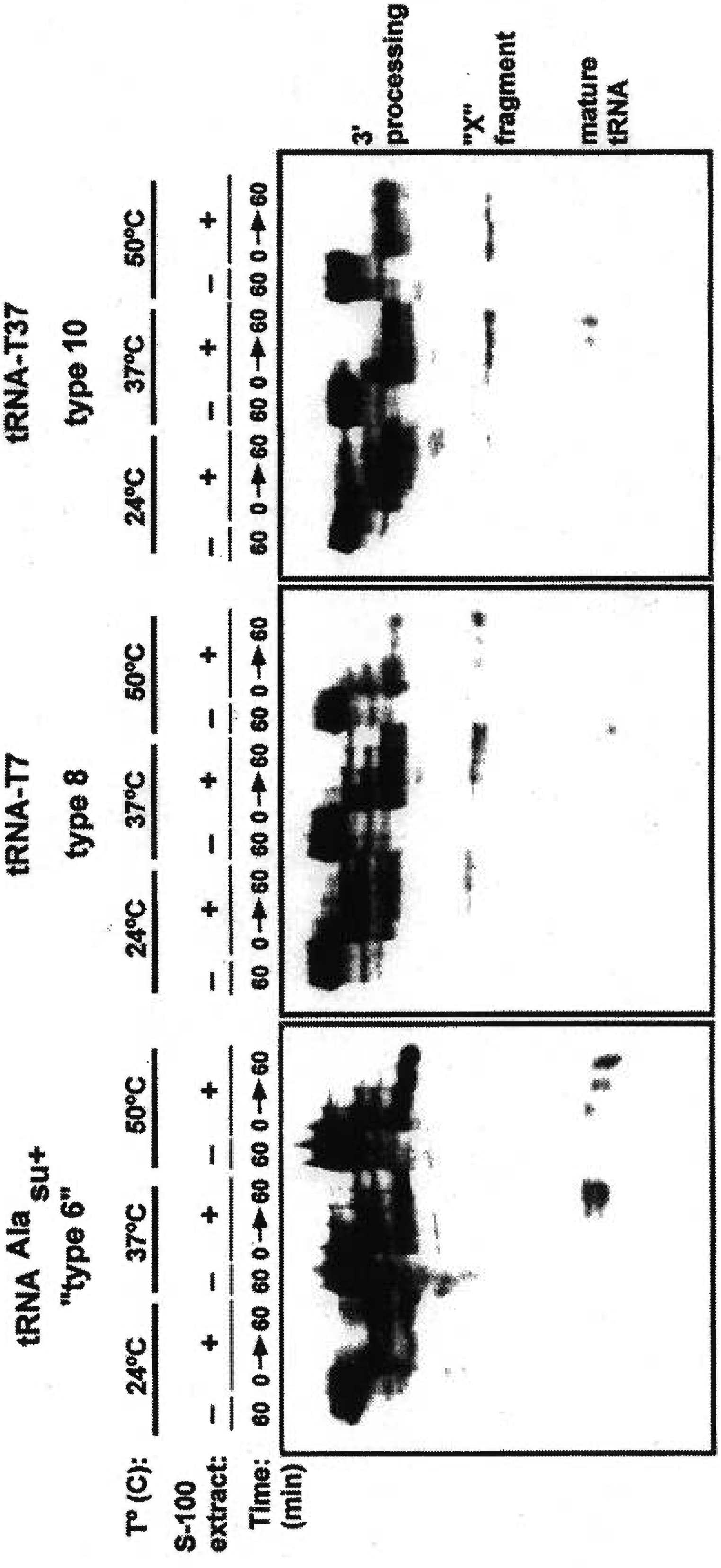

tRNA Ala su+
"type 6"
tRNA-T7
type 8
tRNA-T37
type 10
T° (C):
S-100 extract:
Time: (min)
24°C
37°C
50°C
− + − + − +
60 0→60 60 0→60 60 0→60
3' processing
"X" fragment
mature tRNA

**Figure 6.** *In vivo* identification of a fragment equivalent to the "X" fragment.

Left panel: Northern blot of 2 µg total RNA hybridized with a probe corresponding to the anticodon of the suppressor tRNAs (recognizing $tRNA^{Ala}_{su+}$, mt-like tRNA-T7 and mt-like tRNA-T37; schematized on top) as well as a probe for the 5S ribosomal RNA (rRNA; control). Right panel: Northern blot of 2 µg total RNA hybridized with a probe corresponding to the first 19 nucleotides of the *in vivo* mt-like tRNA primary transcripts (recognizing perfectly the mt-like tRNAs but partially $tRNA^{Ala}_{su+}$; schematized on top) as well as a probe for the 5S rRNA (control).

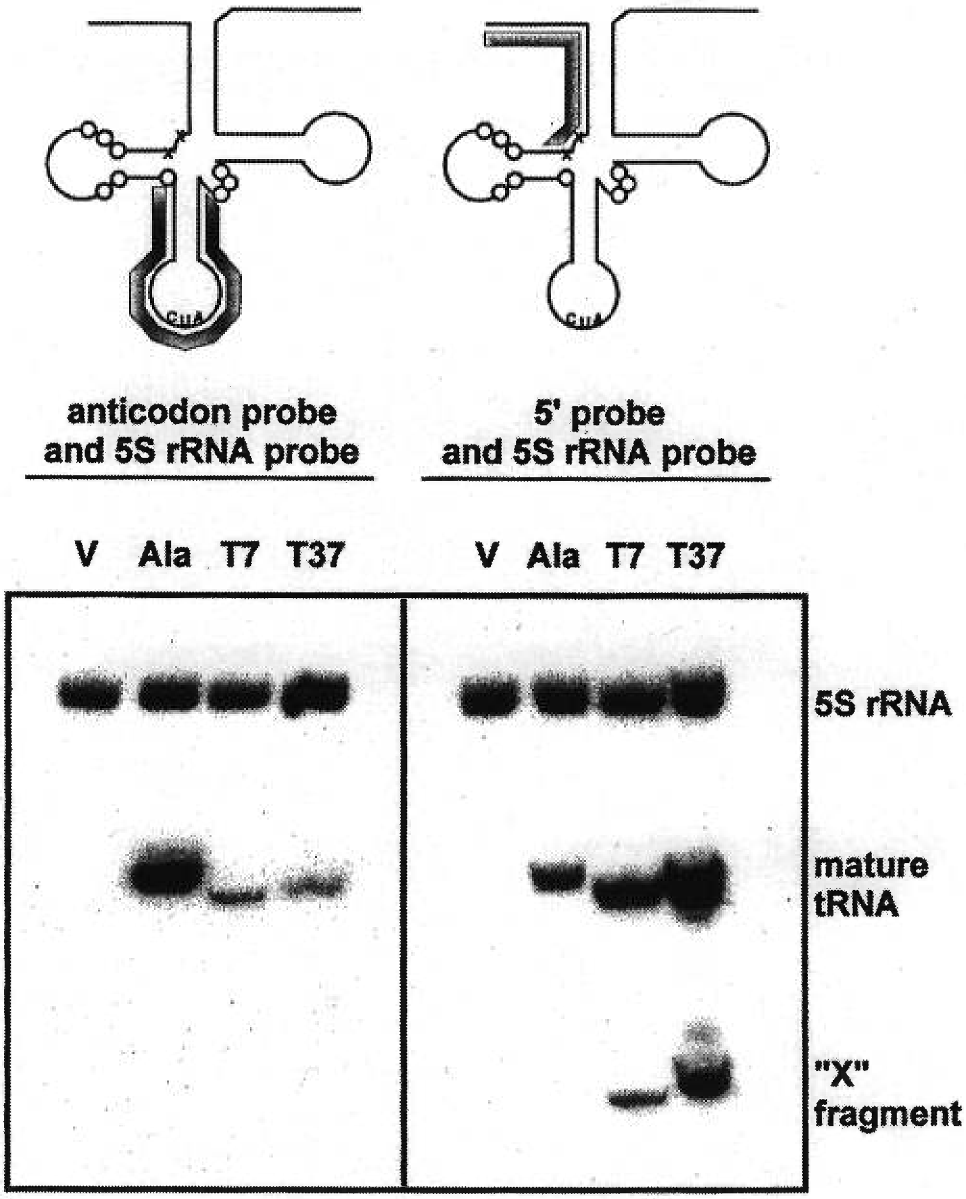
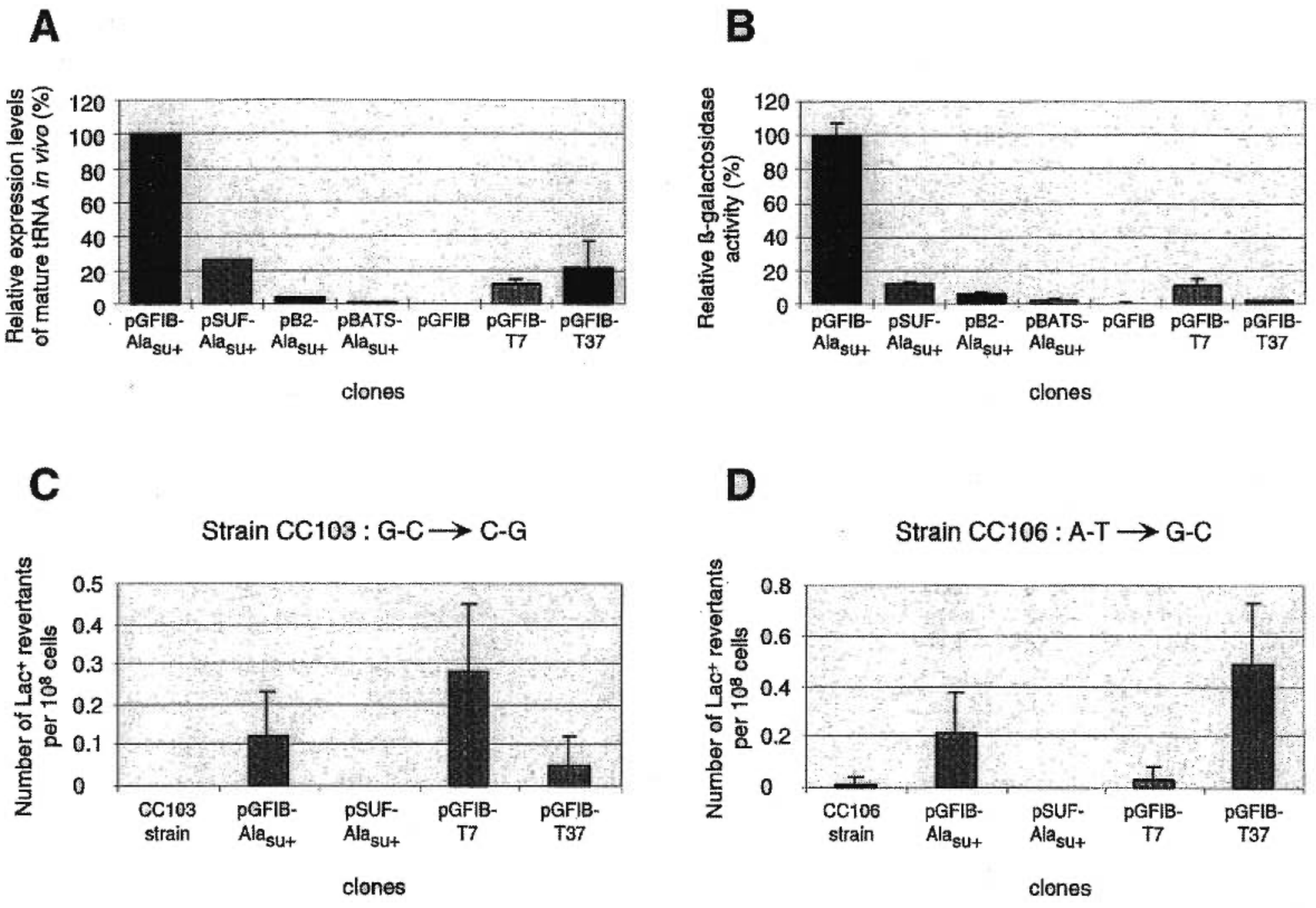

CUA
CUA
anticodon probe
and 5S rRNA probe
5' probe
and 5S rRNA probe
V
Ala
T7
T37
V
Ala
T7
T37
5S rRNA
mature
tRNA
"X"
fragment

**Figure 7.** Levels of expression of mature tRNA, suppression activity and mutator effect.

A) Histogram of the relative expression of the mature $tRNA^{Ala}_{su+}$ and the mt-like tRNAs expressed *in vivo* from different plasmids (values from ref. 7 and 14). B) Relative ß-galactosidase activity detected for different plasmids expressing $tRNA^{Ala}_{su+}$ and the mt-like tRNAs (experimental procedure in ref. 7). C) Frequency of reversion expressed by the number of $Lac^{+}$ revertants per $10^{8}$ cells obtained from an overnight culture in rich medium. For each clone, 11 separate cultures were used.


A
Relative expression levels of mature tRNA *in vivo* (%)
120
100
80
60
40
20
0
pGFIB-$Ala_{su+}$
pSUF-$Ala_{su+}$
pB2-$Ala_{su+}$
pBATS-$Ala_{su+}$
pGFIB
pGFIB-T7
pGFIB-T37
clones
B
Relative ß-galactosidase activity (%)
120
100
80
60
40
20
0
pGFIB-$Ala_{su+}$
pSUF-$Ala_{su+}$
pB2-$Ala_{su+}$
pBATS-$Ala_{su+}$
pGFIB
pGFIB-T7
pGFIB-T37
clones
C
Strain CC103 : G-C → C-G
Number of $Lac^+$ revertants per $10^8$ cells
0.5
0.4
0.3
0.2
0.1
0
CC103 strain
pGFIB-$Ala_{su+}$
pSUF-$Ala_{su+}$
pGFIB-T7
pGFIB-T37
clones
D
Strain CC106 : A-T → G-C
Number of $Lac^+$ revertants per $10^8$ cells
0.8
0.6
0.4
0.2
0
CC106 strain
pGFIB-$Ala_{su+}$
pSUF-$Ala_{su+}$
pGFIB-T7
pGFIB-T37
clones

# DISCUSSION

## 1. Distribution des motifs d'ARN dans GenBank (Chapitre 1 et 3)

Les recherches effectuées dans GenBank montrent une distribution très étendue des différents motifs structuraux d'ARN utilisés. Cette distribution s'étend à l'ensemble des organismes pour lesquels des séquences sont disponibles et à l'ensemble des structures génétiques pouvant être identifiées (région codante, promoteur, intron, ... ). Aucune influence par le biais de l'utilisation des codons (Ikemura, 1981; Jukes, 1990; Osawa *et al.*, 1990) ne semble affecter la distribution. La distribution de la majorité des motifs que nous avons utilisés semble donc aléatoire avec une fréquence semblable à une dérive sans pression évolutive. Quelques exceptions se sont présentées pour lesquelles une sélection positive ou négative apparaît à travers la dérive aléatoire : les motifs RBE, DNAzyme_8-17, boucle-UV, liant la valine et liant l'ATP (Chapitre 1).

Cette preuve que les séquences naturelles contiennent des motifs d'ARN évoluant à la dérive est importante dans le contexte où toutes les séquences théoriquement possibles ne se retrouvent pas dans un même organisme. Cependant, grâce à la propriété de dégénérescence de la structure des ARN, l'absence de certaines permutations ou de certaines séquences particulières n'empêche pas de retrouver des motifs structuraux d'ARN donnés.

Bien entendu, plusieurs repliements peuvent prévenir la formation des motifs potentiels que nous avons identifiés. Il est néanmoins important de concevoir que la seule présence du motif (ou de ses mutants ponctuels) implique la possibilité qu'il devienne important dans le cas où l'environnement change et lui permette de se former. Ces changements environnementaux incluent une mutation affectant le repliement normal ou simplement la présence du ligand du motif potentiel. Une fois le motif formé, même accidentellement, son utilité pourra se manifester et être conservée (ou perdue) selon la force de l'avantage (ou du désavantage).

## 2. Valeur des motifs potentiels trouvés (Chapitre 1, 2 et 3)

La possibilité que les motifs trouvés par nos recherches puissent se former nous a incités à faire ressortir dans les Chapitres 1, 2 et 3 plusieurs exemples particulièrement intéressants. Ces exemples démontrent à la fois la conservation ou fixation de certains motifs comme la ribozyme en tête-de-marteau dans les séquences d'ADN répété (Chapitre 1 et 3 ainsi que Epstein et Gall, 1987; Ferbeyre *et al.*, 1998) ou la dérive et l'implication fortuite d'un motif comme le motif de liaison de la protéine Tat dans l'ARN du virus de vaccinia (Chapitre 1) et comme celui associé au Sarcome de Kaposi (Chapitre 2). Une lettre nous est d'ailleurs parvenue à propos de ce dernier cas peu de temps après la publication de l'hypothèse d'une interaction entre la protéine Tat du HIV et le transcrit de la glycoprotéine H du virus du Sarcome de Kaposi (Chapitre 2). Le Dr Gallo partageait avec nous dans cette lettre (voir appendice I) sa curiosité et son intérêt face à notre hypothèse ainsi que des observations permettant de croire que l'interaction serait possible *in vivo*. Il est donc possible et vraisemblable d'établir, dans certains cas, des liens avec des observations antérieures plutôt incomprises.

Un point particulier des recherches que nous avons effectuées concerne les motifs d'ARN trouvés pouvant lier des protéines du virus du VIH ou du VIS (Virus d'Immunodéficience du Singe), que sont spécialement intéressants si on considère que ces virus sont récents. Trop peu de temps s'est donc écoulé pour que le génome des organismes pouvant être infectés puisse sélectionner contre des interactions accidentelles dommageables. Nos recherches ont identifié de nombreux ARNm possédant des sites potentiels de liaison par les protéines Rev et Tat et leur expression pourrait très bien être influencée par une liaison de ces protéines. Une telle influence a déjà été observée par Flores *et al.* (1993) avec une répression de 50% de l'expression de la superoxide dismutase des cellules HeLa en présence de la protéine Tat.

Il faudra du temps pour démontrer l'activité possible des différentes séquences trouvées pouvant former les motifs structuraux que nous avons cherchés. Une seule personne ou un seul groupe ne peut les tester tous. Cependant, comme mentionné précédemment, le seul potentiel impliqué par la présence de ces motifs est important. C'est pourquoi nous avons décidé de rendre accessible l'ensemble des résultats obtenus à travers une série de pages Internet. De cette façon, des chercheurs de tous les domaines pourront tester les motifs qui sont intéressants dans le contexte de leurs propres recherches. Une présentation de ces sites se trouve dans les Appendices II et III.

## 3. Importance de l'isolement d'ARNt atypiques actifs (Chapitre 4)

L'isolement des ARNt atypiques fonctionnels présenté au Chapitre 4 démontre encore une fois le niveau de dégénérescence que possèdent les ARN. Nous avons identifié de nouvelles structures primaires formant des structures secondaires non-canoniques semblables à celles retrouvées parmi les ARNt mitochondriaux. Ces nouvelles structures forment tout de même une structure tertiaire semblable à celle des ARNt normaux puisque les ARNt atypiques isolés sont actifs dans le système de suppression utilisé. Il existe donc plus de séquences que celles identifiées à ce jour qui peuvent remplir la fonction " d'adaptateur " dans la traduction bien qu'elles ne se retrouvent pas naturellement.

Le fait que ces ARNt atypiques soient actifs soulève de nouveau la question de leur absence dans le cytoplasme versus leur présence dans les mitochondries. Un manque d'activité ne peut plus être invoqué comme cause de cette absence. Toutefois, quelle(s) que soi(en)t la ou les raisons rendant les ARNt atypiques inappropriés pour la traduction dans le cytoplasme, les cellules semblent faire un réel effort pour les éviter. En effet, plusieurs travaux semblent indiquer que de nombreuses modifications des bases effectuées sur les ARNt pourraient servir à stabiliser leur structure fonctionnelle (Arnez, 1994; Derrick et Horowitz, 1993; Przykorska, 1995). En particulier, une forte corrélation a été établie entre le potentiel de certains ARNt à former des structures alternatives atypiques et la présence d'une modification : diméthylguanosine. Cette modification peut empêcher la formation des structures alternatives. Cette corrélation a été découverte suite à une recherche avec RNAMOT (Steinberg et Cedergren, 1995).

La question inverse à savoir pourquoi la présence d'ARNt atypiques est permise dans certaines mitochondries, soulève aussi de nombreuses questions. Malheureusement, seulement des hypothèses peuvent être émises pour y répondre. Il est possible que les mitochondries tolèrent mieux une synthèse protéique moins efficace en vitesse et/ou en précision. Les mitochondries possèdent un plus haut taux de mutations et d'évolution (Brown *et al.*, 1982) résultant peut-être du compromis entre un petit génome et le rendement optimal de toutes ses fonctions. Les ARNt atypiques résultent vraisemblablement de ce même compromis.

## 4. Significations possibles d'une activité de clivage des ARNt atypiques (Chapitre 5)

C'est en espérant en découvrir plus sur ce qui pourrait avoir influencé l'exclusion d'ARNt atypiques du système de traduction cytoplasmique que nous avons procédé aux expériences présentées au Chapitre 5. Deux résultats importants ont été obtenus, soit le clivage spécifique des ARNt atypiques par une activité cellulaire et l'induction *in vivo* de certaines mutations par les ARNt atypiques.

Tel que discuté dans le Chapitre 5, quelques enzymes ont déjà été identifiées pour leur capacité de cliver un ARNt à l'intérieur de sa séquence : i) une nucléase de l'anticodon ARNt$^{Lys}$-spécifique, ii) la colicine E5 coupant les ARNt–Tyr, –His, –Asp et –Asn comportant une modification queunine, et iii) la RNaseA, moins spécifique. Une protéine similaire à ces enzymes pourrait être impliquée dans l'activité que nous avons rapportée. Il est cependant intéressant de mentionner deux autres suspects contre lesquels les évidences sont faibles mais dont l'opportunité serait parfaite puisqu'ils interagissent avec les ARNt avant que ceux-ci ne deviennent actifs: le complexe de la RNase P et la glutaminyl-ARNt-synthétase. Tous deux entrent en contact avec des ARNt pour effectuer la maturation de l'extrémité 5' ou pour ajouter un acide aminé à l'extrémité 3', respectivement. Certains articles rapportent pour chacun d'eux une activité produisant (RNase P: Kikuchi et Sasaki, 1992; Kikuchi *et al.*, 1990) ou induisant (synthétase : Beresten *et al.*, 1992) un clivage à l'intérieur de l'ARNt mature. Dans chacun des cas, le clivage se produit dans un contexte spécifique mais l'implication de ces enzymes dans le cycle normal des ARNt typiques les positionnent avantageusement pour prévenir l'apparition d'ARNt atypiques nocifs dans une cellule.

La découverte du clivage amène aussi à se questionner sur l'envergure que pourrait prendre cette activité dans une cellule normale. Si cette activité n'a pour cible que des ARNt atypiques, elle aurait pu procurer un avantage évolutif à la cellule la contenant si un système de traduction avec des structures uniformes est effectivement plus efficace. Cependant, si cette activité peut agir sur des ARNt canoniques pour dégrader ceux qui auraient un mauvais repliement ou ceux qui seraient devenus trop “ vieux ”, alors il s'agit d'une découverte nous offrant des perspectives sur un des aspects du renouvellement des ARNt, qui est encore mal connue : la dégradation normale. À ce point, il ne s'agit que de spéculations mais il sera intéressant de voir ce que de futures recherches permettront de découvrir.

L'identification de niveaux plus élevés de mutations pour certaines transversions ou transitions indique que les ARNt atypiques pourraient bien être nocifs à long terme pour une cellule. Les niveaux observés sont bas à la fois à cause du faible niveau d'expression des ARNt mitochondriaux et du fait que l'acide aminé alanine est moins dommageable lorsqu'il est mal inséré dans une protéine que d'autres acides aminés avec des propriétés physiques ou chimiques plus particulières (i.e., la proline ou l'acide aspartique).

# 5. Implications générales des travaux de cette thèse

## 5.1 Observation générale

L'ensemble des Chapitres présentés dans cette thèse porte sur l'étude de la dégénérescence des différents niveaux de structure des ARN. À travers les recherches informatisées de motifs d'ARN dans GenBank, de nouvelles régions de séquences provenant de nombreux organismes différents ont été identifiées, démontrant le potentiel de formation de ces motifs là où ils n'étaient pas soupçonnés. Aussi, grâce à l'isolement et à la caractérisation de nouveaux ARNt atypiques fonctionnels chez *E. coli*, de nouvelles séquences ont été identifiées pour leur capacité à se replier et à agir telle que des ARNt typiques. Ces éléments prouvent la dégénérescence des structures des ARN dans les séquences naturelles entre la structure primaire et la structure secondaire ainsi que la dégénérescence entre les structures primaire, secondaire et tertiaire, à travers l'activité.

## 5.2 Avantages et désavantages de la dégénérescence des structures des ARN

### *i) Utilité de la dégénérescence*

Sans se laisser influencer par ce que nous connaissons, quelles propriétés devrait posséder une population de polymères pour avoir une chance de soutenir la vie? Si elle doit persister dans un monde changeant, dans un environnement hostile, la population d'un polymère donné devrait aussi être changeante. C'est donc dire qu'à l'intérieur de quelques cycles de réplication ou générations, une diversité doit s'établir qui permettra à au moins une partie de la population de survivre aux changements du milieu. La capacité de faire dériver la structure de base du polymère sans éliminer l'activité de la molécule est donc une vertu en or pour un monde naissant.

Comme souligné dans des travaux précédents (mentionnés à la section 4) ainsi que dans cette thèse par une approche différente, les ARN ont de façon intrinsèque cette possibilité de dériver dans l'ensemble des structures primaires possibles tout en conservant des structures secondaire et tertiaire similaires. Cette propriété permet à une population de molécules d'évoluer par dérive aléatoire de la séquence tout en continuant d'effectuer leur activité et ce jusqu'au moment où une utilité différente se voit sélectionnée par les conditions du milieu. Le Monde à ARN répond donc aux critères qui auraient pu être établis *a priori* pour décrire les caractéristiques nécessaires que doit avoir un polymère pour qu'il ait une chance de peupler le monde primitif. Cet aspect a probablement été un atout quand la nature a fait son choix entre les ADN, les ARN ou les protéines. Il est aussi fort probable que si un autre polymère a précédé le Monde à ARN, il possédait aussi cette propriété de dégénérescence ou de dériver dans l'ensemble des séquences théoriquement possibles.

### *ii) Inconvénient de la dégénérescence*

Au moment de l'émergence de la vie, la dégénérescence des structures des ARN devait avoir plus d'avantages que d'inconvénients. La possibilité de permettre la différenciation de la structure de base des polymères de la population sans risquer de détruire leur activité avait sûrement prédominance sur les risques que certains changements de séquence, obtenus par dérive, causent une diminution d'efficacité de l'activité. L'évolution étant telle qu'elle est et le temps n'y étant pas une commodité rare, une activité même faible devaient être préférable à une absence d'activité ou à une activité plus forte ne pouvant pas tolérer un changement d'environnement.

De nos jours cependant, la dégénérescence des structures des ARN ne semble pas avoir les avantages incontestables qu'elle a déjà eus. L'ARN a été remplacé par l'ADN pour le maintien de l'information génétique dans la majorité des organismes vivants. Il semble que la stabilité de l'ADN constitue une qualité plus importante dans le monde d'aujourd'hui que la dégénérescence des structures des ARN. Plus encore, les protéines ont délogé les ARN pour la majorité des activités catalytiques retrouvées dans les cellules. Leur versatilité et la présence de plus d'acides aminés différents semblent conférer aux protéines des propriétés plus avantageuses que celles que l'ARN possède. L'évolution au niveau des cellules et des organismes, contrairement à l'évolution moléculaire, semble demander une optimisation de toutes les fonctions possibles.

Les ARN sont malgré tout encore présents aujourd'hui. Ils sont responsables de quelques activités catalytiques comme l'activité de clivage des ribozymes en tête-de-marteau. Ils font partie de complexes qui remplissent des fonctions particulières comme l'ARNr dans le ribosome. Par contre, ils ne constituent pas la règle ou l'essentiel comme ils l'ont déjà fait. Aujourd'hui, la dégénérescence des structures des ARN est aussi un inconvénient. Les ARNt ont une structure secondaire uniforme dans les systèmes de traduction cytoplasmique ce qui implique une pression pour les garder ainsi. Il est possible que l'activité rapportée au Chapitre 5 contribue à prévenir l'apparition et la fixation dans la cellule d'ARNt atypiques. S'agit-il d'un vrai moyen de protection? Y en aurait-il d'autres à d'autres niveaux pour d'autres ARN?

Plusieurs systèmes complexes ont pour but le contrôle de l'expression appropriée des protéines (exemple avec la régulation de la dégradation des ARNm, revue chez les eucaryotes par Beelman et Parker, 1995; et chez les procaryotes par Regnier et Arraiano, 2000), la vérification de leur repliement (comme avec les chaperonnes, revue par Horwich *et al.*, 1999) et leur dégradation en temps opportun (exemple de la voie de dégradation par l'ubiquitine, revue par Kornitzer et Ciechanover, 2000). Les études dans ces domaines sont nombreuses. Par contre, des études équivalentes sur les ARN sont encore peu répandues, sauf pour la synthèse et la dégradation des ARNm. Il est pourtant possible que des systèmes importants contrôlent également les différents aspects de l'expression des ARN non-codants. À titre d'exemple, la synthèse et la maturation des ARNt (Deutscher, 1995; Martin, 1995; Mazzara et McClain, 1980) et des ARNr (Pace et Burgin, 1990; Srivastava et Schlessinger, 1990; van Nues *et al.*, 1995) sont relativement bien étudiées mais leur dégradation normale, un aspect du renouvellement de la population de ces ARN dans la cellule, est pour le moins obscure. Les travaux sur des mutants ou des structures atypiques pourraient faciliter de telles études comme semblent l'indiquer les résultats du Chapitre 5. Lorsque plus d'informations seront disponibles, il sera peut-être possible d'augmenter les parallèles entre les ARN et les protéines. Des parallèles existent déjà au niveau de la structure en motifs et du repliement structural (Brion et Westhof, 1997; Draper, 1996) mais il reste plus encore à découvrir.

## 6. Conclusion

Les résultats de cette thèse ont permis d'appuyer l'hypothèse de l'évolution à la dérive des motifs d'ARN ainsi que d'augmenter les preuves de la dégénérescence des structures des ARN. Ils ont aussi ouvert une porte sur des activités possibles de régulation de certains ARN fonctionnels.

Ce qui reste de nos jours du Monde à ARN n'est connu que partiellement. L'accroissement de nos connaissances sur les ARN, leurs fonctions et leurs propriétés en tant que molécule complète, aussi bien qu'en tant que motif faisant partie d'une plus grande molécule, promet encore de belles découvertes. À témoin sont les nombreux ARN récemment découverts ne codant pas pour des protéines mais fonctionnant directement comme ARN (revue par Eddy, 1999). Effectivement, à la liste des ARNt, des ARNr et des petits ARN nucléaires (snRNA) se sont ajoutés ces dernières années une multitude d'ARN à fonctions diverses. Par exemple :

- L'ARN de la particule de reconnaissance du signal participant à la translocation des protéines au réticulum endoplasmique (revue par Bovia et Strub, 1996).

- Les ARN guides pour l'édition des ARN (revue par Simpson, 1999).
- L'ARN du complexe de la télomérase (revue par O'Reilly *et al.*, 1999).
- L'ARN de transfert / messager (10Sa RNA ou tmRNA) permettant de compléter la traduction d'ARNm tronqués et ciblant les protéines résultantes pour la dégradation (revue par Keiler *et al.*, 1996; Williams, 1999).
- Les ARN antisens (revue par Delihas, 1995).
- Les ARN impliqués dans l'inactivation du chromosome X (revue par Panning et Jaenisch, 1998).

À cette liste non-exhaustive, doivent aussi s'ajouter les ARN non codant dont la fonction est encore inconnue comme l'ARN 6S de *E. coli*, le transcrit *hsr-omega* de *Drosophila* ou le transcrit *H19* humain (revue par Eddy, 1999).

Un groupe particulièrement en expansion doit aussi être mentionné : celui des petits ARN nucléolaires (snoRNA; Weinstein et Steitz, 1999). Leur fonction a récemment été découverte dans la modification des ARNr (Smith et Steitz, 1997) ainsi que dans celle des petits ARN nucléaires (Tycowski *et al.*, 1998). L'utilisation d'un programme considérant la structure secondaire minimale de ces ARN a déjà prouvé la supériorité de cette approche pour identifier de nouveaux membres (Lowe et Eddy, 1999). Nombre d'entre eux sont encore à découvrir. Attendons-nous à des surprises!

# RÉFÉRENCES

**– A –**

Arnez, J.G. et Steitz, T. A. (1994) Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry*, **33**, 7560-7567.

**– B –**

Balzer, M. et Wagner, R. (1998) Mutations in the leader region of ribosomal RNA operons cause structurally defective 30 S ribosomes as revealed by in vivo structural probing. *J. Mol. Biol.*, **276**, 547-557.

Banerjee, A.R., Jaeger, J.A. et Turner, D.H. (1993) Thermal unfolding of a group I ribozyme: the low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, **32**, 153-163.

Beelman, C.A. et Parker, R. (1995) Degradation of mRNA in eukariotes. *Cell*, **81**, 179-183.

Beresten, S., Jahn, M. et Soll, D. (1992) Aminoacyl-tRNA synthetase-induced cleavage of tRNA. *Nucleic Acids Res*, **20**, 1523-1530.

Besancon, W. et Wagner, R. (1999) Characterization of transient RNA-RNA interactions important for the facilitated structure formation of bacterial ribosomal 16S RNA. *Nucl. Acids Res.*, **27**, 4353-4362.

Beuning, P.J., Yang, F., Schimmel, P. et Musier-Forsyth, K. (1997) Specific atomic groups and RNA helix geometry in acceptor stem recognition by a tRNA synthetase. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 10150-10154.

Blancafort, P., Steinberg, S.V., Paquin, B., Klinck, R., Scott, J.K. et Cedergren, R. (1999) The recognition of a noncanonical RNA base pair by a zinc finger protein. *Chem. Biol.*, **6**, 585-597.

Bork, P. et Koonin, E.V. (1996) Protein sequence motifs. *Curr. Opin. Struct. Biol.*, **6**, 366-376.

Bovia, F. et Strub, K. (1996) The signal recognition particle and related small cytoplasmic ribonucleoprotein particles. *J. Cell. Sci.*, **109**, 2601-2608.

Bratty, J., Chartrand, P., Ferbeyre, G. et Cedergren, R. (1993) The hammerhead RNA domain, a model ribozyme. *Biochim. Biophys. Acta*, **1216**, 345-359.

Brion, P. et Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 113-137.

Brown, W.M., Prager, E.M., Wang, A. et Wilson, A.C. (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, **18**, 225-239.

Burkard, M.E., Turner, D.H. et Tinoco, I.J. (1999a) APPENDIX 2: Schematic Diagrams of secondary and tertiary structure elements. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 681-685.

Burkard, M.E., Turner, D.H. et Tinoco, I.J. (1999b) The interactions that shape RNA sturucture. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 233-264.

## – C –

Caput, D., Beutler, B., Hartog, K., Thayer, R., Brown-Shimer, S. et Cerami, A. (1986) Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 1670-1674.

Chang, K.Y., Varani, G., Bhattacharya, S., Choi, H. et McClain, W.H. (1999) Correlation of deformability at a tRNA recognition site and aminoacylation specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 11764-11769.

Coetzee, T., Herschlag, D. et Belfort, M. (1994) *Escherichia coli* proteins, including ribosomal protein S12, facilitate *in vitro* splicing of phage T4 introns by acting as RNA chaperones. *Genes Dev.*, **8**, 1575-1588.

Crick, F.H. (1968) The origin of the genetic code. *J. Mol. Biol.*, **38**, 367-379.

Crothers, D.M., Cole, P.E., Hilbers, C.W. et Shulman, R.G. (1974) The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J. Mol. Biol.*, **87**, 63-88.

## – D –

Dahlberg, J.E. (1980) tRNAs as primers for reverse transcriptases. Dans *TRANSFER RNA: Biological Aspects*. Soll, D., Abelson, J.N. et Schimmel, P.R. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 507-516.

Dandekar, T. et Hentze, M.W. (1995) Finding the hairpin in the haystack: searching for RNA motifs. *Trends Genet.*, **11**, 45-50.

Delihas, N. (1995) Regulation of gene expression by trans-encoded antisense RNAs. *Mol. Microbiol.*, **15**, 411-414.

Derrick, W.B. et Horowitz, J. (1993) Probing structural differences between native and in vitro transcribed *Escherichia coli* valine transfer RNA: evidence for stable base modification-dependent conformers. *Nucl. Acids Res.*, **21**, 4948-4953.

Deutscher, M.P. (1995) tRNA Processing Nucleases. Dans *tRNA: Structure, Biosynthesis and Function*. Soll, D. et Rajbhandary, U.L. (eds.), American Society for Microbiology, Washington, D.C., pp. 51-65.

Dirheimar, G., Keith, G., Dumas, P. et Westhof, E. (1995) Primary, secondary, and tertiary structures of tRNAs. Dans *tRNA: Structure, Biosynthesis, and Function*. Söll, D. et RajBhandary, Y. (eds.), American Society for Microbiology, Washington, DC, pp. 93-126.

Draper, D.E. (1996) Parallel worlds. *Nat. Struct. Biol.*, **3**, 397-400.

Dunstan, H.M., Young, L.S. et Sprague, K.U. (1994) TFIIIR is an isoleucine tRNA. *Mol. Cell. Biol.*, **14**, 3588-3595.

– E –

Eddy, S.R. (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev.*, **9**, 695-699.

Egli, M., Usman, N. et Rich, A. (1993) Conformational influence of the ribose 2'-hydroxyl group: crystal structures of DNA-RNA chimeric duplexes. *Biochemistry*, **32**, 3221-3237.

Ekland, E.H. et Bartel, D.P. (1996) RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature*, **382**, 373-376.

Ekland, E.H., Szostak, J.W. et Bartel, D.P. (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science*, **269**, 364-370.

Ellington, A.D. et Szostak, J.W. (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818-822.

Emerman, M., Vazeux, R. et Peden, K. (1989) The rev gene product of the human immunodeficiency virus affects envelope-specific RNA localisation. *Cell*, **57**, 1155-1165.

Eschenmoser, A. (1999) Chemical etiology of nucleic acid structure. *Science*, **284**, 2118-2124.

Epstein, L.M. et Gall, J.G. (1987) Self-cleaving transcripts of satellite DNA from the newt. *Cell*, **48**, 535-543.

**– F –**

Feig, A.L. et Uhlenbeck, O.C. (1999) The role of metal ions in RNA biochemistry. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 287-319.

Ferbeyre, G., Smith, J.M. et Cedergren, R. (1998) Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol. Cell. Biol.*, **18**, 3880-3888.

Flores, S.C., Marecki, J.C., Harper, K.P., Bose, S.K., Nelson, S.K. et McCord, J.M. (1993) Tat protein of human immunodeficiency virus type 1 represses expression of manganese superoxide dismutase in HeLa cells. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7632-7636.

Fontana, W., Konings, D.A., Stadler, P.F. et Schuster, P. (1993) Statistics of RNA secondary structures. *Biopolymeres*, **33**, 1389-1404.

**– G –**

Gait, M.J. et Karn, J. (1993) RNA recognition by the human immunodeficiency virus Tat and Rev proteins. *Trends Biochem. Sci.*, **18**, 255-259.

Gautheret, D., Konings, D. et Gutell, R.R. (1994) A major family of motifs involving G.A mismatches in ribosomal RNA. *J. Mol. Biol.*, **242**, 1-8.

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

Gilbert, W. (1986) The RNA world. *Nature*, **319**, 618.

Gilbert, W. et de Souza, S.J. (1999) Introns and the RNA world. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 221-231.

Giver, L., Bartel, D., Zapp, M., Pawul, A., Green, M. et Ellington, A.D. (1993a) Selective optimization of the Rev-binding element of HIV-1. *Nucl. Acid. Res.*, **21**, 5509-5516.

Giver, L., Bartel, D.P., Zapp, M.L., Green, M.R. et Ellington, A.D. (1993b) Selection and design of high-affinity RNA ligands for HIV-1 Rev. *Gene*, **137**, 19-24.

Gray, M.W. et Cedergren, R. (1993) The new age of RNA. *FASEB J.*, **7**, 4-6.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. et Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849-857.

## – H –

Hadzopoulou-Cladaras, M., Felber, B.K., Cladaras, C., Athanassopoulos, A., Tse, A. et Pavlakis, G.N. (1989) The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region. *J. Virol.*, **63**, 1265-1274.

Hammarskjold, M.L., Haimer, J., Hammarskjold, B., Sangwan, I., Albert, L. et Rekosh, D. (1989) Regulation of human immunodeficiency virus env expression by the rev gene product. *J. Virol.*, **63**, 1959-1966.

Hawkins, E.R., Chang, S.H. et Mattice, W.L. (1977) Kinetics of the renaturation of yeast tRNA3 leu. *Biopolymers*, **16**, 1557-1566.

Henkin, T.M. (1994) tRNA-directed transcription antitermination. *Mol. Microbiol.*, **13**, 381-387.

Hermann, T. et Patel, D.J. (1999) Stitching together RNA tertiary architectures. *J. Mol. Biol.*, **294**, 829-849.

Hilbers, C.W., Robillard, G.T., Shulman, R.G., Blake, R.D., Webb, P.K., Fresco, R. et Riesner, D. (1976) Thermal unfolding of yeast glycine transfer RNA. *Biochemistry*, **15**, 1874-1882.

Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R. et Zamir, A. (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462-1465.

Horwich, A.L., Weber-Ban, E.U. et Finley, D. (1999) Chaperone rings in protein folding and deradation. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 11033-11040.

Huynen, M.A. (1996) Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, **43**, 165-169.

Huynen, M.A., Konings, D.A. et Hogeweg, P. (1993) Multiple coding and the evolutionary properties of RNA secondary structure. *J. Theor. Biol.*, **165**, 251-267.

**– I –**

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**, 389-409.

## – J –

Joyce, G.F. (1999) APPENDIX 3: Reactions catalysed by RNA and DNA enzymes. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 687-690.

Joyce, G.F. et Orgel, L.E. (1999) Prospects for understanding the origin of the RNA world. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 49-77.

Jukes, T.H. (1990) Genetic code 1990 - Outlook. *Experiencia*, **46**, 1149-1157.

Keiler, K.C., Waller, P.R. et Sauer, R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, **271**, 990-993.

## – K –

Kikuchi, Y. et Sasaki, N. (1992) Hyperprocessing of tRNA by the catalytic RNA of RNase P. Cleavage of a natural tRNA within the mature tRNA sequence and evidence for an altered conformation of the substrate tRNA. *J Biol Chem*, **267**, 11972-11976.

Kikuchi, Y., Sasaki, N. et Ando-Yamagami, Y. (1990) Cleavage of tRNA within the mature tRNA sequence by the catalytic RNA of RNase P: implication for the formation of the primer tRNA fragment for reverse transcription in copia retrovirus-like particles. *Proc Natl Acad Sci U S A*, **87**, 8105-8109.

Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kornitzer, D. et Ciechanover, A. (2000) Modes of regulation of ubiquitin-mediated protein degradation. *J. Cell. Physiol.*, **182**, 1-11.

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

**– L –**

Laferrière, A., Gautheret, D. et Cedergren, R. (1994) An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.*, **10**, 211-212.

Leibowitz, M.J. et Soffer, R.L. (1969) A soluble enzyme from Escherichia coli which catalyzes the transfer of leucine and phenylalanine from tRNA to acceptor proteins. *Biochem. Biophys. Res. Commun.*, **36**, 47-53.

Lewin, R. (1986) RNA catalysis gives fresh perspective on the origin of life. *Science*, **231**, 545-546.

Liiv, A., Tenson, T., Margus, T. et Remme, J. (1998) Multiple functions of the transcribed spacers in ribosomal RNA operons. *Biol. Chem.*, **379**, 783-793.

Littauer, U.Z. et Inouye, H. (1973) Regulation of tRNA. *Annu. Rev. Biochem.*, **42**, 429-470.

Lowe, T.M. et Eddy, S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168-1171.

– M –

Maizels, N. et Weiner, A.M. (1999) The genomic tag hypothesis: What molecular fossils tell us about the evolution of tRNA. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 79-111.

Malim, M.H., Hauber, J., Le, S.Y., Maizel, J.V. et Cullen, B.R. (1989) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature*, **338**, 254-257.

Martin, N.C. (1995) Organellar tRNAs: Biosynthesis and function. Dans *tRNA: Structure, Biosynthesis and Function*. Soll, D. et Rajbhandary, U.L. (eds.), American Society for Microbiology, Washington, D.C., pp. 127-140.

Mazzara, G.P. et McClain, W.H. (1980) tRNA synthesis. Dans *TRANSFER RNA: Biological Aspects*. Soll, D., Abelson, J.N. et Schimmel, P.R. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 3-27.

McClain, W.H., Chen, Y.-M., Foss, K. et Schneider, J. (1988) Association of transfer RNA acceptor identity with a helical irregularity. *Science*, **240**, 793-796.

Michel, F. et Westhof, E. (1996) Visualizing the logic behind RNA self-assembly. *Science*, **273**, 1676-1677.

Miramontes, P., Medrano, L., Cerpa, C., Cedergren, R., Ferbeyre, G. et Cocho, G. (1995) Structural and thermodynamic properties of DNA uncover different evolutionary histories. *J. Mol. Evol.*, **40**, 698-704.

Mojzsis, S.J., Krishnamurthy, R. et Arrhenius, G. (1999) Before RNA and after: geophysical and geochemical constraints on molecular evolution. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 1-47.

**– O –**

O'Reilly, M., Teichmann, S.A. et Rhodes, D. (1999) Telomerases. *Curr. Opin. Struct. Biol.*, **9**, 56-65.

Orgel, L.E. (1968) Evolution of the genetic apparatus. *J. Mol. Biol.*, **38**, 381-393.

Orgel, L.E. (1998) The origin of life--a review of facts and speculations. *Trends Biochem. Sci.*, **23**, 491-495.

Osawa, S., Muto, A., Ohama, T., Andachi, Y., Tanaka, R. et Yamao, F. (1990) Prokaryotic genetic code. *Experientia*, **46**, 1097-1106.

**– P –**

Pace, N.R. et Burgin, A.B. (1990) Processing and Evolution of the rRNAs. Dans *The Ribosome: Structure, Function & Evolution*. Hill, W.E., Dahlberg, A., Garrett, R.A., Moore, P.B., Schlessinger, D. et Warner, J.R. (eds.), American Society for Microbiology, Washington, D.C., pp. 417-425.

Pace, N.R. et Marsh, T.L. (1985) RNA catalysis and the origin of life. *Orig. Life Evol. Biosph.*, **16**, 97-116.

Panning, B. et Jaenisch, R. (1998) RNA and the epigenetic regulation of X chromosome inactivation. *Cell*, **93**, 305-308.

Perreault, J.P., Wu, T.F., Cousineau, B., Ogilvie, K.K. et Cedergren, R. (1990) Mixed deoxyribo- and ribo-oligonucleotides with catalytic activity. *Nature*, **344**, 565-567.

Peters, G.G., Harada, F., Dahlberg, J.E., Panet, A., Haseltine, W.A. et Baltimore, D. (1977) Low-molecular-weight RNAs of Moloney murine leukimia virus: Identification of the primer for RNA-directed DNA synthesis. *J. Virol.*, **21**, 1031-1041.

Provine, W.B. (1986) *Sewall Wright and evolutionary biology*. University of Chicago Press, Chicago.

Przykorska, A. (1995) Influence of modified nucleosides on tRNA structure as probed by two plant nucleases. *Biochimie*, **77**, 109-112.

**– Q –**

Quay, S.C. et Oxender, D.L. (1980) Role of tRNA-Leu in branched-chain amino acid transport. Dans *TRANSFER RNA: Biological Aspects.* Soll, D., Abelson, J.N. et Schimmel, P.R. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 481-491.

**– R –**

Regnier, P. et Arraiano, C.M. (2000) Degradation of mRNA in bacteria: emegence of ubiquitous features. *Bioessays*, **22**, 235-244.

**– S –**

Schimmel, P., Giege, R., Moras, D. et Yokoyama, S. (1993) An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 8763-8768.

Schimmel, P. et Ribas de Pouplana, L. (1995) Transfer RNA: from minihelix to genetic code. *Cell*, **1995**, 7.

Schon, A., Krupp, G., Gough, S., Berry-Lowe, S., Kannangara, C.G. et Soll, D. (1986) The RNA required in the first step of chlorophyll biosynthesis is a chloroplast glutamate tRNA. *Nature*, **322**, 281-284.

Schuster, P., Fontana, W., Stadler, P.F. et Hofacker, I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B Biol. Sci.*, **255**, 279-284.

Sharp, P.A. (1985) On the origin of RNA splicing and introns. *Cell*, **42**, 397-400.

Shaw, G. et Kamen, R. (1986) A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell*, **46**, 659-667.

Simpson, L. (1999) RNA editing - an evolutionary perspective. Dans *The RNA World*. Gesteland, R.F., Cech, T.R. et Atkins, J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 585-608.

Smith, C.M. et Steitz, J.A. (1997) Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell*, **89**, 669-672.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. et Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, **26**, 148-153.

Srivastava, A.K. et Schlessinger, D. (1990) rRNA processing in *Escherichia coli*. Dans *The Ribosome: Structure, Function & Evolution*. Hill, W.E., Dahlberg, A., Garrett, R.A., Moore, P.B., Schlessinger, D. et Warner, J.R. (eds.), American Society for Microbiology, Washington, D.C., pp. 426-434.

Steinberg, S. et Cedergren, R. (1994) Structural compensation in atypical mitochondrial tRNAs. *Nat. Struct. Biol.*, **1**, 507-510.

Steinberg, S. et Cedergren, R. (1995) A correlation between N2-dimethylguanosine presence and alternate tRNA conformers. *RNA*, **1**, 886-891.

Steinberg, S., Gautheret, D. et Cedergren, R. (1994) Fitting the structurally diverse animal mitochondrial tRNAs(Ser) to common three-dimensional constraints. *J. Mol. Biol.*, **236**, 982-989.

Steinberg, S., Leclerc, F. et Cedergren, R. (1997) Structural rules and conformational compensations in the tRNA L-form. *J. Mol. Biol.*, **266**, 269-282.

Strazewski, P., Biala, E., Gabriel, K. et McClain, W.H. (1999) The relationship of termodynamic stability at a G x U recognition to tRNA aminoacylation specificity. *RNA*, **5**, 1490-1494.

– T –

Tycowski, K.T., You, Z.H., Graham, P.J. et Steitz, J.A. (1998) Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell.*, **2**, 629-638.

– V –

Voet, D. et Voet, J.G. (1990) *Biochemistry*. John Wiley & sons, New York.

Waterman, M.S. (1978) Secondary structure of single-stranded nucleic acids. Studies on foundations and combinatorics. *Adv. Math. Suppl. Studies*, **1**, 167-212.

– W –

Waters, L.C., Mullin, B.C., Ho, T. et Yang, W.K. (1975) Ability of tryptophan tRNA to hybridize with 35S RNA of avian myeloblastosis virus and prime reverse transcription *in vitro*. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 2155-2159.

Watson, J.D. (2000) *A passion for DNA*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Weiner, A.M. et Maizels, N. (1987) tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: implications for the origin of protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 7383-7387.

Westhof, E., Dumas, P. et Moras, D. (1988) Hydration of transfer RNA molecules: a crystallographic study. *Biochimie*, **70**, 145-165.

White, H.B.d. (1976) Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.*, **7**, 101-104.

Williams, K.P (1999) The thRNA website. *Nucl. Acids Res.*, **27**, 165-166.

Woese, C. (1967) The evolution of the genetic code. *The Genetic Code.* Harper & Row, New York, pp. 179-195.

Woese, C.R., Winker, S. et Gutell, R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 8467-8471.

Wood, B. et Collard, M. (1999) The human genus. *Science*, **284**, 65-71.

Wu, M. et Tinico, I.Jr. (1998) RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11555-11560.

**– Y –**

Yokogawa, T., Watanabe, Y., Kumazawa, Y., Ueda, T., Hirao, I., Miura, K. et Watanabe, K. (1991) A novel cloverleaf structure found in mammalian mitochondrial tRNA(Ser) (UCN). *Nucl. Acids. Res.*, **19**, 6101-6105.

# APPENDICE I

Lettre de R. C. Gallo

INSTITUTE OF HUMAN VIROLOGY

OFFICE OF
THE DIRECTOR

February 17, 1998

Dr. Gerardo Ferbeyre
Dr. Veronique Bourdeau
Dr. Robert Cedergren
Department de Biochimie
Universite de Montreal
Montreal (Quebec) HC3 3J7
CANADA

Dear Drs. Ferbeyre, Bourdeau and Cedergren,

I read with interest your comments and data in TIBS regarding the Tat protein in HIV. For these ideas to be significant in vivo Tat must be available as an extra-cellular protein and be taken up by cells, since HHV-8 (KSHV) infected cells do not harbor HIV genes. I thought you would be interested to know that both have been shown, i.e., Tat is taken up by cells as originally shown by Frankel, et al and Tat is activity secreted to the extra-cellular space by HIV acutely infected T cells, as we first showed in papers which were (incidentally) relevant to Kaposi sarcoma.

On the other hand you should be aware that the Tat transgenic mouse results (Vogel, et al) which you referred to unfortunately have been widely irreproducible. I hope these comments will be useful to you as your published remarks have been to me.

Sincerely,

[redacted]

Robert C. Gallo, M.D.
Professor and Director

RCG:cici

# APPENDICE II

Site Internet des résultats du Chapitre 1

Les résultats de la recherche présentée dans le Chapitre 1 sont disponibles par Internet à l'adresse suivante :

http://www.centrcn.umontreal.ca/~bourdeav/Ribonomics

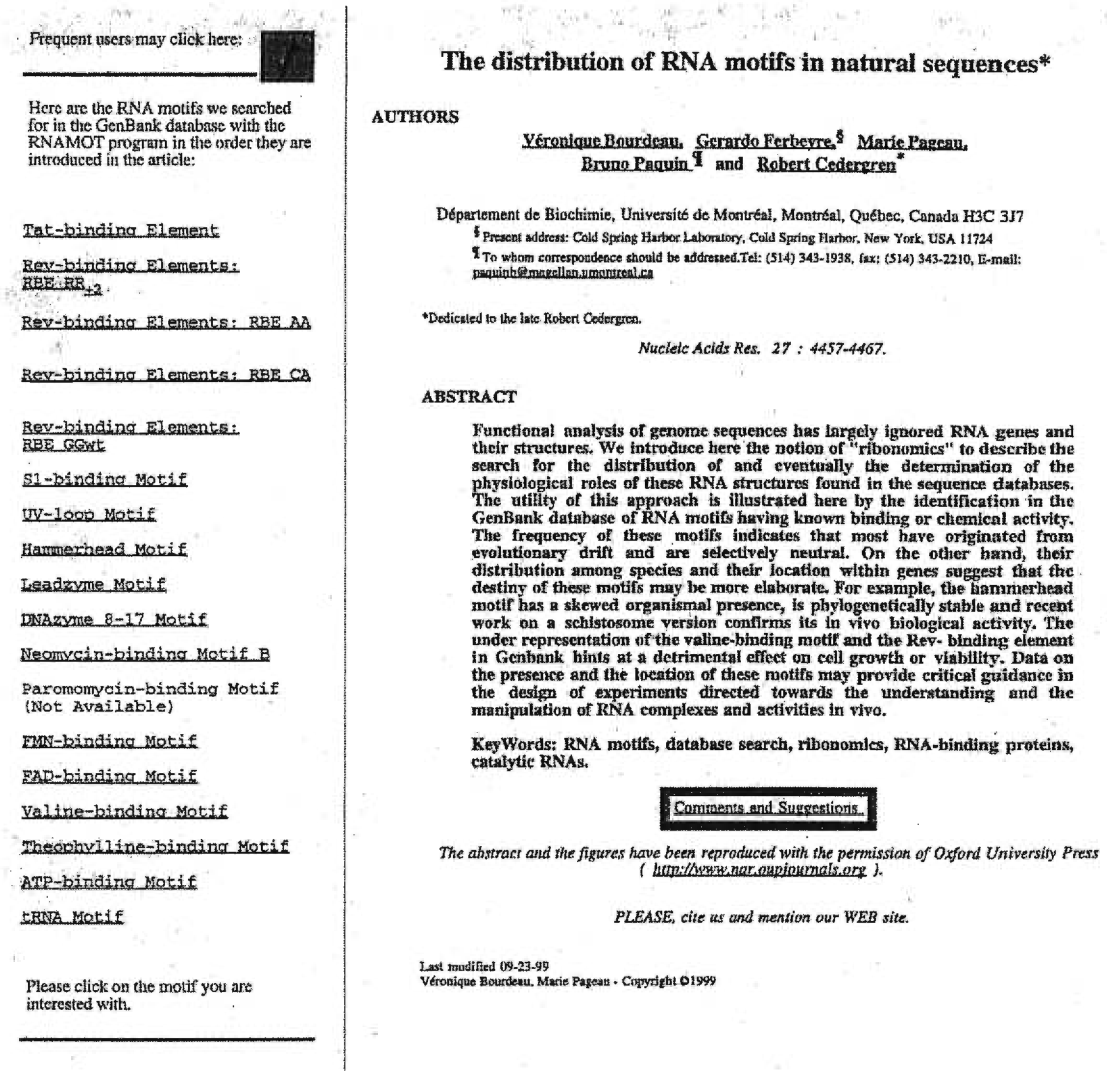
Cette appendice consiste en un bref survol de ce qui se retrouve sur le site Internet. Une impression de quelques unes des pages électroniques est disponible à la suite de l'explication du site.

*Explication du site*

La page d'accueil se compose de deux fenêtres. Celle de droite rappelle le titre du travail, les auteurs (avec liens au pages personnelles de chacun), le lieu de publication et un résumé. Dans la fenêtre de gauche, une liste des différents motifs est retrouvée avec un lien à une page contenant l'illustration du motif correspondant. À partir de la page de l'illustration du motif, l'utilisateur peut accéder à une page présentant le tableau des fichiers GenBank où les recherches ont été effectuées. Chaque nom de fichier réfère à la page des résultats correspondante. Cette page contient d'abord une petite légende dans une fenêtre particulière qui explique les sous-sections du motif dans le même format où les séquences trouvées sont présentées. Une seconde légende mentionne les symboles utilisés pour désigner la location du motif lorsque disponible. L'ensemble des séquences trouvées qui possèdent le potentiel de former le motif est ensuite énuméré.

Pour les utilisateurs fréquents, le lien à un tableau classant les motifs par type d'activité est disponible dès la page d'accueil. Le nom de ces motifs est relié à son illustration lorsque deux orientations sont possibles ou au tableau des fichiers GenBank (s'il n'y a qu'une orientation ou à travers la mention de l'orientation).

# Page d'Accueil



Frequent users may click here:

Here are the RNA motifs we searched for in the GenBank database with the RNAMOT program in the order they are introduced in the article:

Tat-binding Element

Rev-binding Elements: RBE RR$_{+2}$

Rev-binding Elements: RBE AA

Rev-binding Elements: RBE CA

Rev-binding Elements: RBE GGwt

S1-binding Motif

UV-loop Motif

Hammerhead Motif

Leadzyme Motif

DNAzyme 8-17 Motif

Neomycin-binding Motif B

Paromomycin-binding Motif (Not Available)

FMN-binding Motif

FAD-binding Motif

Valine-binding Motif

Theophylline-binding Motif

ATP-binding Motif

tRNA Motif

Please click on the motif you are interested with.

## The distribution of RNA motifs in natural sequences*

AUTHORS

Véronique Bourdeau, Gerardo Ferbeyre,[§] Marie Pageau,
Bruno Paquin [¶] and Robert Cedergren*

Département de Biochimie, Université de Montréal, Montréal, Québec, Canada H3C 3J7

[§] Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA 11724

[¶] To whom correspondence should be addressed.Tel: (514) 343-1938, fax: (514) 343-2210, E-mail: paquinb@magellan.umontreal.ca

*Dedicated to the late Robert Cedergren.

*Nucleic Acids Res. 27 : 4457-4467.*

ABSTRACT

Functional analysis of genome sequences has largely ignored RNA genes and their structures. We introduce here the notion of "ribonomics" to describe the search for the distribution of and eventually the determination of the physiological roles of these RNA structures found in the sequence databases. The utility of this approach is illustrated here by the identification in the GenBank database of RNA motifs having known binding or chemical activity. The frequency of these motifs indicates that most have originated from evolutionary drift and are selectively neutral. On the other hand, their distribution among species and their location within genes suggest that the destiny of these motifs may be more elaborate. For example, the hammerhead motif has a skewed organismal presence, is phylogenetically stable and recent work on a schistosome version confirms its in vivo biological activity. The under representation of the valine-binding motif and the Rev- binding element in Genbank hints at a detrimental effect on cell growth or viability. Data on the presence and the location of these motifs may provide critical guidance in the design of experiments directed towards the understanding and the manipulation of RNA complexes and activities in vivo.

KeyWords: RNA motifs, database search, ribonomics, RNA-binding proteins, catalytic RNAs.

Comments and Suggestions

*The abstract and the figures have been reproduced with the permission of Oxford University Press ( http://www.nar.oupjournals.org ).*

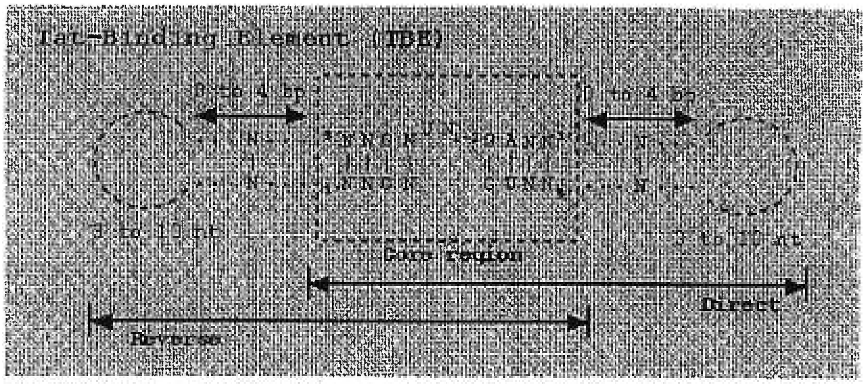
*PLEASE, cite us and mention our WEB site.*

Last modified 09-23-99
Véronique Bourdeau, Marie Pageau - Copyright ©1999

## Illustration des Motifs

### Exemple d'un motif à deux orientations : motif de liaison à la protéine Tat.

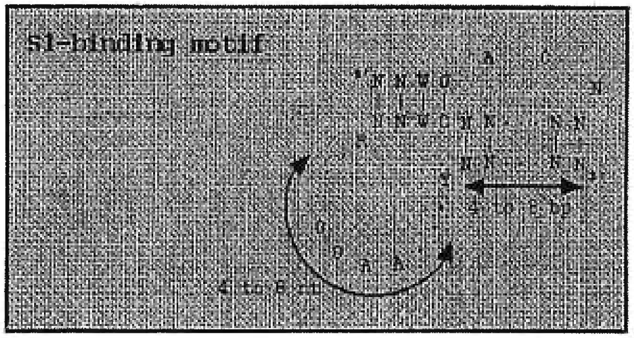
In the course of this work, we have searched the GenBank database using RNAMOT program with two orientations of a Tat-binding element as shown in the figure below.



*Please choose if you want to see the results of the direct motif or of the reverse one.*

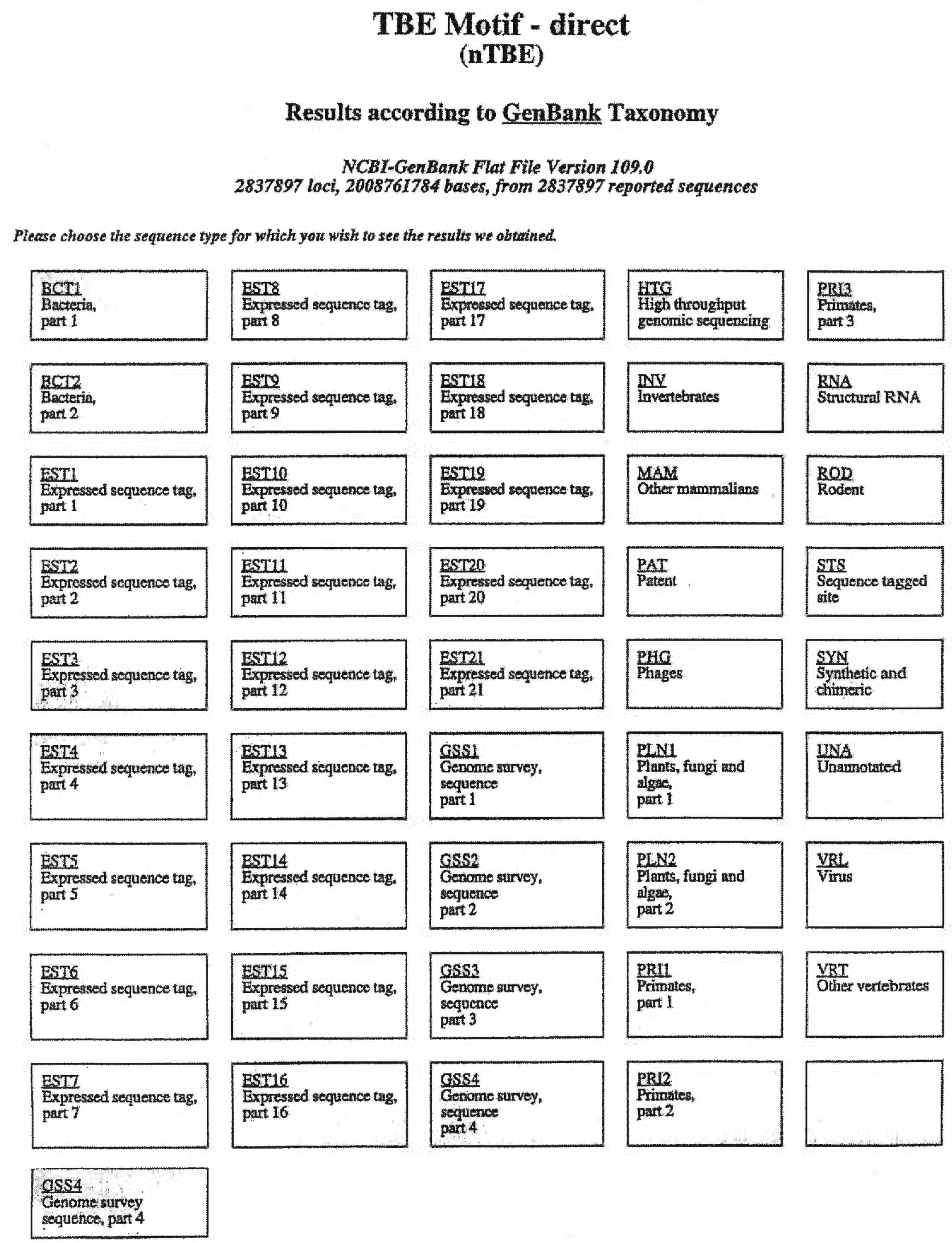
### Exemple d'un motif à une seule orientation : motif de liaison à la protéine S1.

In the course of this work, we have searched the GenBank database using RNAMOT program with a S1-binding motif as shown in the figure below.



## Tableau des Fichiers

Exemple avec le motif de liaison à la protéine Tat, directe orientation.



**TBE Motif - direct**
**(nTBE)**

**Results according to GenBank Taxonomy**

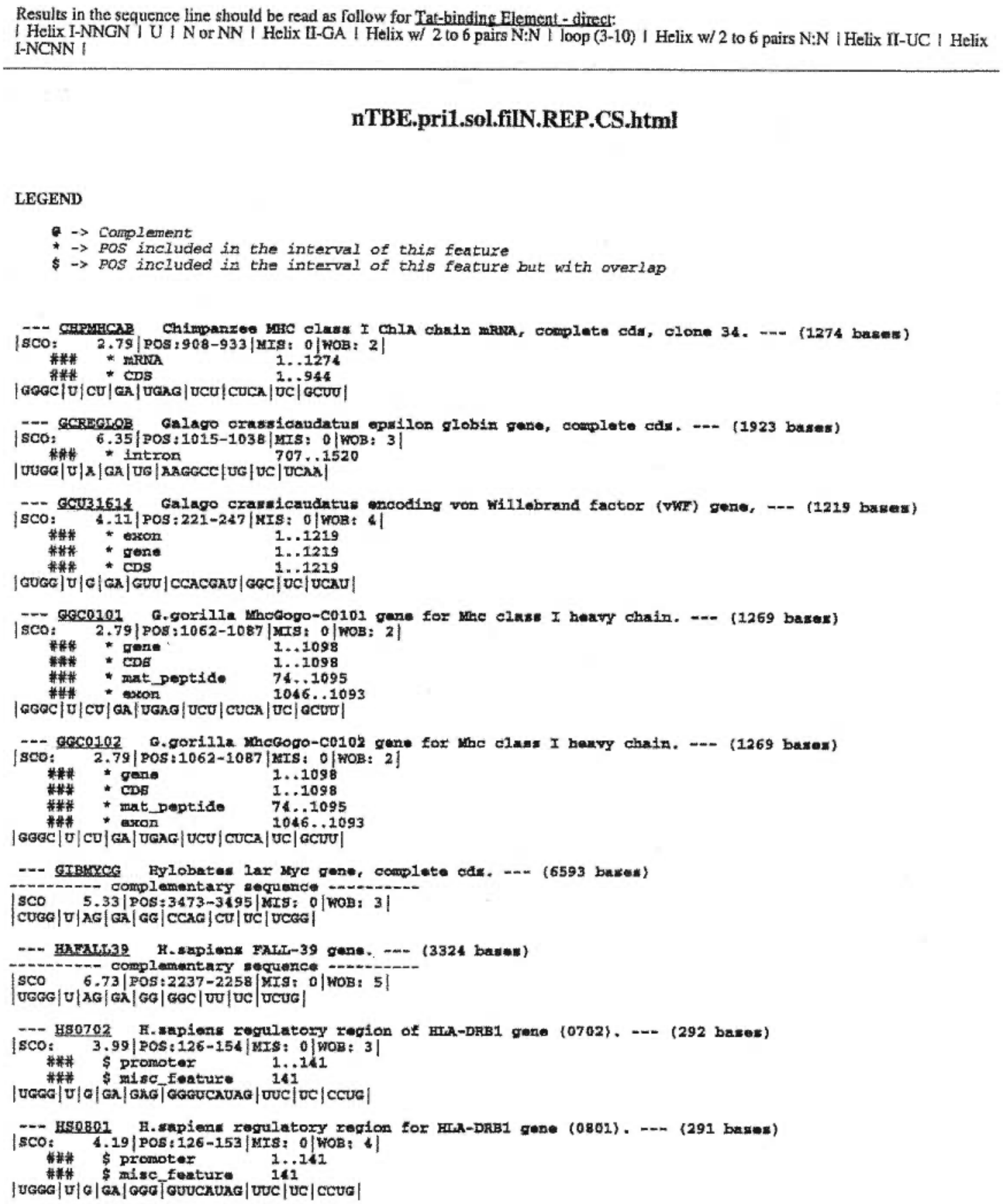
*NCBI-GenBank Flat File Version 109.0*
*2837897 loci, 2008761784 bases, from 2837897 reported sequences*

*Please choose the sequence type for which you wish to see the results we obtained.*

| | | | | |
|---|---|---|---|---|
| BCT1<br>Bacteria,<br>part 1 | EST8<br>Expressed sequence tag,<br>part 8 | EST17<br>Expressed sequence tag,<br>part 17 | HTG<br>High throughput<br>genomic sequencing | PRI3<br>Primates,<br>part 3 |
| BCT2<br>Bacteria,<br>part 2 | EST9<br>Expressed sequence tag,<br>part 9 | EST18<br>Expressed sequence tag,<br>part 18 | INV<br>Invertebrates | RNA<br>Structural RNA |
| EST1<br>Expressed sequence tag,<br>part 1 | EST10<br>Expressed sequence tag,<br>part 10 | EST19<br>Expressed sequence tag,<br>part 19 | MAM<br>Other mammalians | ROD<br>Rodent |
| EST2<br>Expressed sequence tag,<br>part 2 | EST11<br>Expressed sequence tag,<br>part 11 | EST20<br>Expressed sequence tag,<br>part 20 | PAT<br>Patent | STS<br>Sequence tagged<br>site |
| EST3<br>Expressed sequence tag,<br>part 3 | EST12<br>Expressed sequence tag,<br>part 12 | EST21<br>Expressed sequence tag,<br>part 21 | PHG<br>Phages | SYN<br>Synthetic and<br>chimeric |
| EST4<br>Expressed sequence tag,<br>part 4 | EST13<br>Expressed sequence tag,<br>part 13 | GSS1<br>Genome survey,<br>sequence<br>part 1 | PLN1<br>Plants, fungi and<br>algae,<br>part 1 | UNA<br>Unannotated |
| EST5<br>Expressed sequence tag,<br>part 5 | EST14<br>Expressed sequence tag,<br>part 14 | GSS2<br>Genome survey,<br>sequence<br>part 2 | PLN2<br>Plants, fungi and<br>algae,<br>part 2 | VRL<br>Virus |
| EST6<br>Expressed sequence tag,<br>part 6 | EST15<br>Expressed sequence tag,<br>part 15 | GSS3<br>Genome survey,<br>sequence<br>part 3 | PRI1<br>Primates,<br>part 1 | VRT<br>Other vertebrates |
| EST7<br>Expressed sequence tag,<br>part 7 | EST16<br>Expressed sequence tag,<br>part 16 | GSS4<br>Genome survey,<br>sequence<br>part 4 | PRI2<br>Primates,<br>part 2 | |
| GSS4<br>Genome survey<br>sequence, part 4 | | | | |

## Page des Résultats

Exemple des résultats du motif direct de liaison à la protéine Tat dans pri1 (premier fichier des primate).

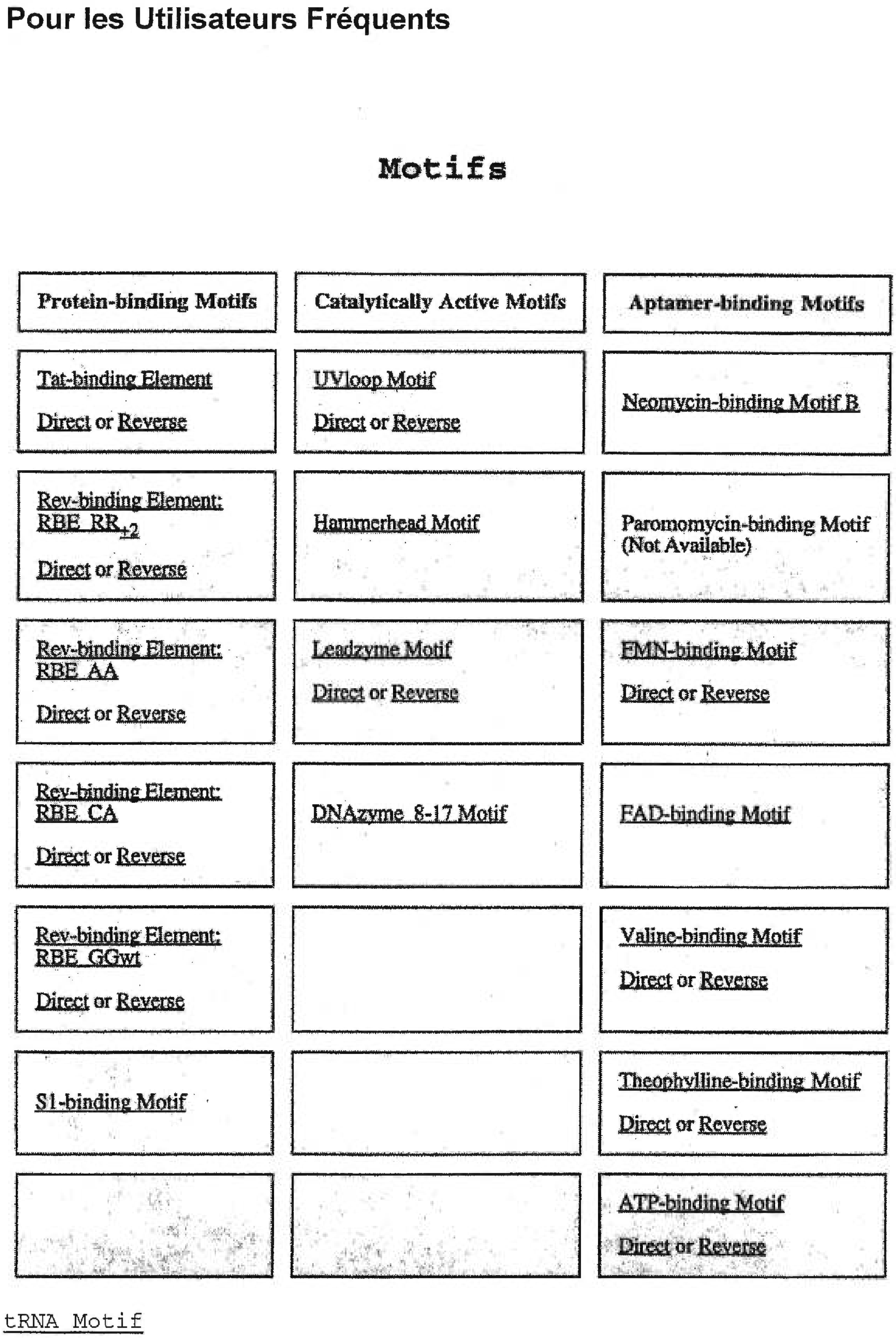


Results in the sequence line should be read as follow for Tat-binding Element - direct:
| Helix I-NNGN | U | N or NN | Helix II-GA | Helix w/ 2 to 6 pairs N:N | loop (3-10) | Helix w/ 2 to 6 pairs N:N | Helix II-UC | Helix I-NCNN |

**nTBE.pri1.sol.filN.REP.CS.html**

LEGEND

```
    @ -> Complement
    * -> POS included in the interval of this feature
    $ -> POS included in the interval of this feature but with overlap


 --- CHPMHCAB   Chimpanzee MHC class I ChlA chain mRNA, complete cds, clone 34. --- (1274 bases)
|SCO:     2.79|POS:908-933|MIS: 0|WOB: 2|
    ###    * mRNA               1..1274
    ###    * CDS                1..944
|GGGC|U|CU|GA|UGAG|UCU|CUCA|UC|GCUU|

 --- GCREGLOB   Galago crassicaudatus epsilon globin gene, complete cds. --- (1923 bases)
|SCO:     6.35|POS:1015-1038|MIS: 0|WOB: 3|
    ###    * intron             707..1520
|UUGG|U|A|GA|UG|AAGGCC|UG|UC|UCAA|

 --- GCU31614   Galago crassicaudatus encoding von Willebrand factor (vWF) gene, --- (1219 bases)
|SCO:     4.11|POS:221-247|MIS: 0|WOB: 4|
    ###    * exon               1..1219
    ###    * gene               1..1219
    ###    * CDS                1..1219
|GUGG|U|G|GA|GUU|CCACGAU|GGC|UC|UCAU|

 --- GGC0101   G.gorilla MhcGogo-C0101 gene for Mhc class I heavy chain. --- (1269 bases)
|SCO:     2.79|POS:1062-1087|MIS: 0|WOB: 2|
    ###    * gene               1..1098
    ###    * CDS                1..1098
    ###    * mat_peptide        74..1095
    ###    * exon               1046..1093
|GGGC|U|CU|GA|UGAG|UCU|CUCA|UC|GCUU|

 --- GGC0102   G.gorilla MhcGogo-C0102 gene for Mhc class I heavy chain. --- (1269 bases)
|SCO:     2.79|POS:1062-1087|MIS: 0|WOB: 2|
    ###    * gene               1..1098
    ###    * CDS                1..1098
    ###    * mat_peptide        74..1095
    ###    * exon               1046..1093
|GGGC|U|CU|GA|UGAG|UCU|CUCA|UC|GCUU|

 --- GIBMYCG   Hylobates lar Myc gene, complete cds. --- (6593 bases)
---------- complementary sequence ----------
|SCO    5.33|POS:3473-3495|MIS: 0|WOB: 3|
|CUGG|U|AG|GA|GG|CCAG|CU|UC|UCGG|

 --- HAFALL39   H.sapiens FALL-39 gene. --- (3324 bases)
---------- complementary sequence ----------
|SCO    6.73|POS:2237-2258|MIS: 0|WOB: 5|
|UGGG|U|AG|GA|GG|GGC|UU|UC|UCUG|

 --- HS0702   H.sapiens regulatory region of HLA-DRB1 gene (0702). --- (292 bases)
|SCO:     3.99|POS:126-154|MIS: 0|WOB: 3|
    ###    $ promoter           1..141
    ###    $ misc_feature       141
|UGGG|U|G|GA|GAG|GGGUCAUAG|UUC|UC|CCUG|

 --- HS0801   H.sapiens regulatory region for HLA-DRB1 gene (0801). --- (291 bases)
|SCO:     4.19|POS:126-153|MIS: 0|WOB: 4|
    ###    $ promoter           1..141
    ###    $ misc_feature       141
|UGGG|U|G|GA|GGG|GUUCAUAG|UUC|UC|CCUG|
```



## Pour les Utilisateurs Fréquents

### Motifs

| Protein-binding Motifs | Catalytically Active Motifs | Aptamer-binding Motifs |
|---|---|---|
| Tat-binding Element<br>Direct or Reverse | UVloop Motif<br>Direct or Reverse | Neomycin-binding Motif B |
| Rev-binding Element: RBE_RR$_{+2}$<br>Direct or Reverse | Hammerhead Motif | Paromomycin-binding Motif (Not Available) |
| Rev-binding Element: RBE_AA<br>Direct or Reverse | Leadzyme Motif<br>Direct or Reverse | FMN-binding Motif<br>Direct or Reverse |
| Rev-binding Element: RBE_CA<br>Direct or Reverse | DNAzyme_8-17 Motif | FAD-binding Motif |
| Rev-binding Element: RBE_GGwt<br>Direct or Reverse | | Valine-binding Motif<br>Direct or Reverse |
| S1-binding Motif | | Theophylline-binding Motif<br>Direct or Reverse |
| | | ATP-binding Motif<br>Direct or Reverse |

tRNA Motif

# APPENDICE III

## Site Internet des résultats du Chapitre 3

Les résultats de la recherche présentée dans le Chapitre 3 sont disponibles par Internet à l'adresse suivante :

http://www.centrcn.umontreal.ca/~bourdeav/HH

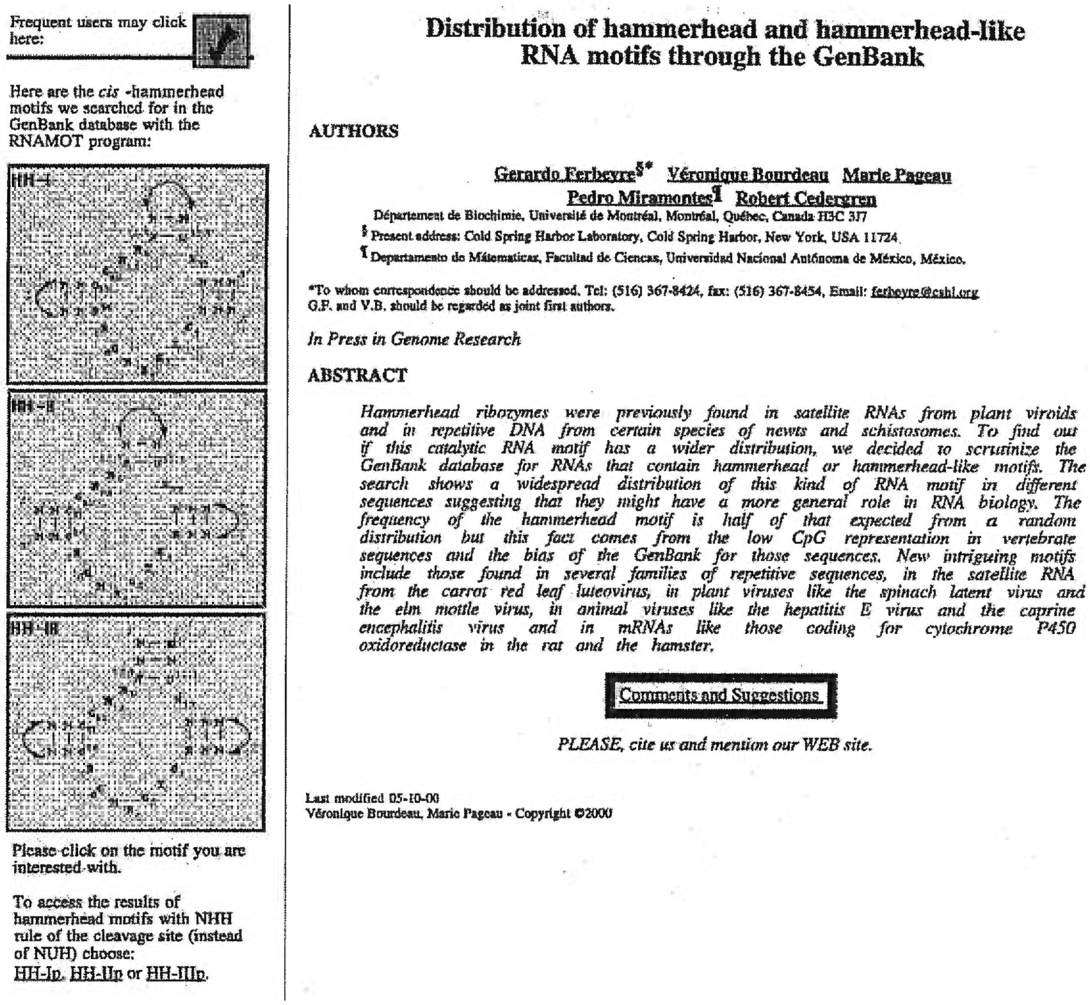
Cette appendice consiste en un bref survol de ce qui se retrouve sur le site Internet. Une impression de quelques unes des pages électroniques est disponible à la suite de l'explication du site.

*Explication du site*

La page d'accueil se compose de deux fenêtres. Celle de droite rappelle le titre du travail, les auteurs (avec liens au pages personnelles de chacun), le lieu de publication et un résumé. Dans la fenêtre de gauche, un schéma des trois motifs de base est retrouvé avec un lien à une page contenant l'illustration détaillée du motif correspondant. Sur cette illustration ou dans le texte en dessous, l'utilisateur peut choisir le type de résultat qui l'intéresse : le motif initial, un mutant ponctuel ou un mutant des paires de bases. Dans le cas des mutants d'une paire de bases, un tableau est présenté donnant le choix de la paire de bases désirée avant d'accéder au tableau des fichiers GenBank. Chaque nom de fichier de ce dernier tableau réfère à la page des résultats correspondante. Cette page contient d'abord une petite légende dans une fenêtre particulière qui explique les sous-sections du motif dans le même format où les séquences sont présentées. L'ensemble des séquences trouvées qui possèdent le potentiel de former le motif est ensuite énuméré.

Pour les utilisateurs fréquents, le lien à un tableau nommant tous les motifs recherchés est disponible dès la page d'accueil. Le nom de ces motifs est relié à son tableau des fichiers du GenBank.

# Page d'Accueil



Frequent users may click here: ✓

Here are the *cis* -hammerhead motifs we searched for in the GenBank database with the RNAMOT program:

Please click on the motif you are interested with.

To access the results of hammerhead motifs with NHH rule of the cleavage site (instead of NUH) choose:
HH-Ip, HH-IIp or HH-IIIp.

## Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank

### AUTHORS

Gerardo Ferbeyre[§]* Véronique Bourdeau Marie Pageau
Pedro Miramontes[¶] Robert Cedergren
Département de Biochimie, Université de Montréal, Montréal, Québec, Canada H3C 3J7
[§] Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA 11724.
[¶] Departamento de Mátematicas, Facultad de Ciencas, Universidad Nacional Autónoma de México, México.

*To whom correspondence should be addressed. Tel: (516) 367-8424, fax: (516) 367-8454, Email: ferbeyre@cshl.org
G.F. and V.B. should be regarded as joint first authors.

*In Press in Genome Research*

### ABSTRACT

*Hammerhead ribozymes were previously found in satellite RNAs from plant viroids and in repetitive DNA from certain species of newts and schistosomes. To find out if this catalytic RNA motif has a wider distribution, we decided to scrutinize the GenBank database for RNAs that contain hammerhead or hammerhead-like motifs. The search shows a widespread distribution of this kind of RNA motif in different sequences suggesting that they might have a more general role in RNA biology. The frequency of the hammerhead motif is half of that expected from a random distribution but this fact comes from the low CpG representation in vertebrate sequences and the bias of the GenBank for those sequences. New intriguing motifs include those found in several families of repetitive sequences, in the satellite RNA from the carrot red leaf luteovirus, in plant viruses like the spinach latent virus and the elm mottle virus, in animal viruses like the hepatitis E virus and the caprine encephalitis virus and in mRNAs like those coding for cytochrome P450 oxidoreductase in the rat and the hamster.*

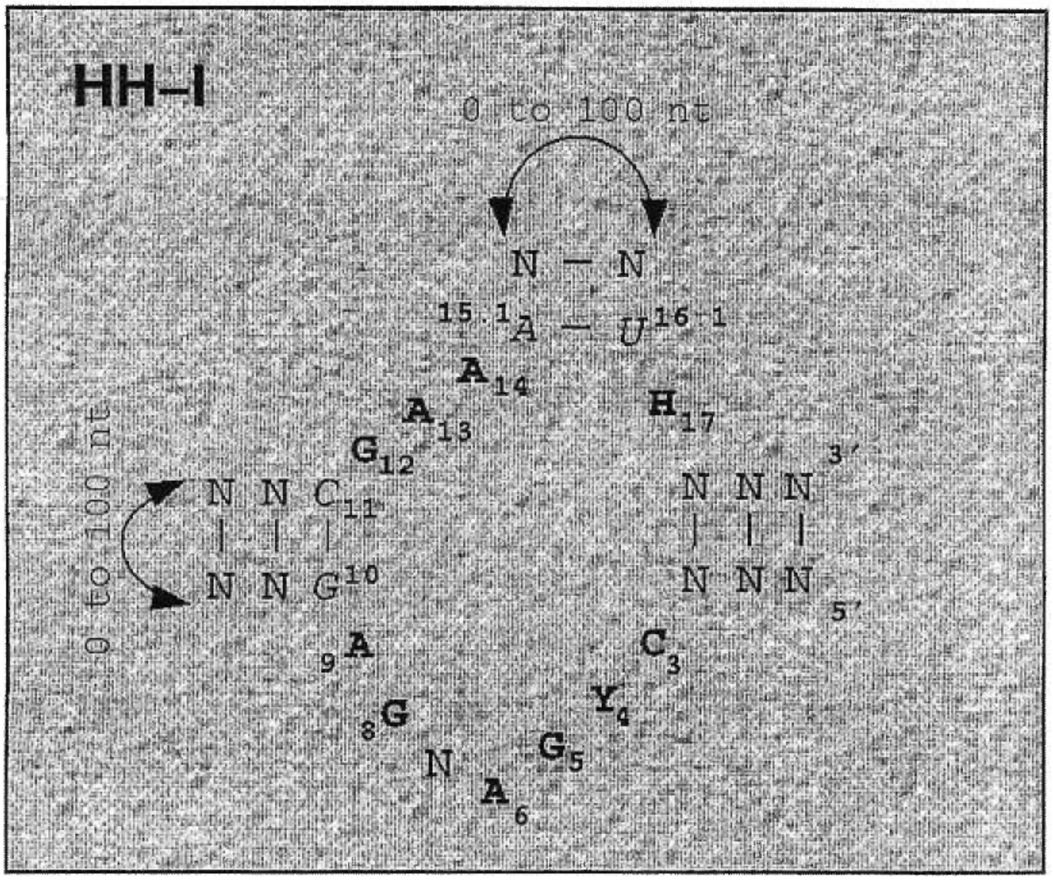
Comments and Suggestions

*PLEASE, cite us and mention our WEB site.*

## Illustration Détaillée d'un Motif de Base

Exemple du motif HH-I.

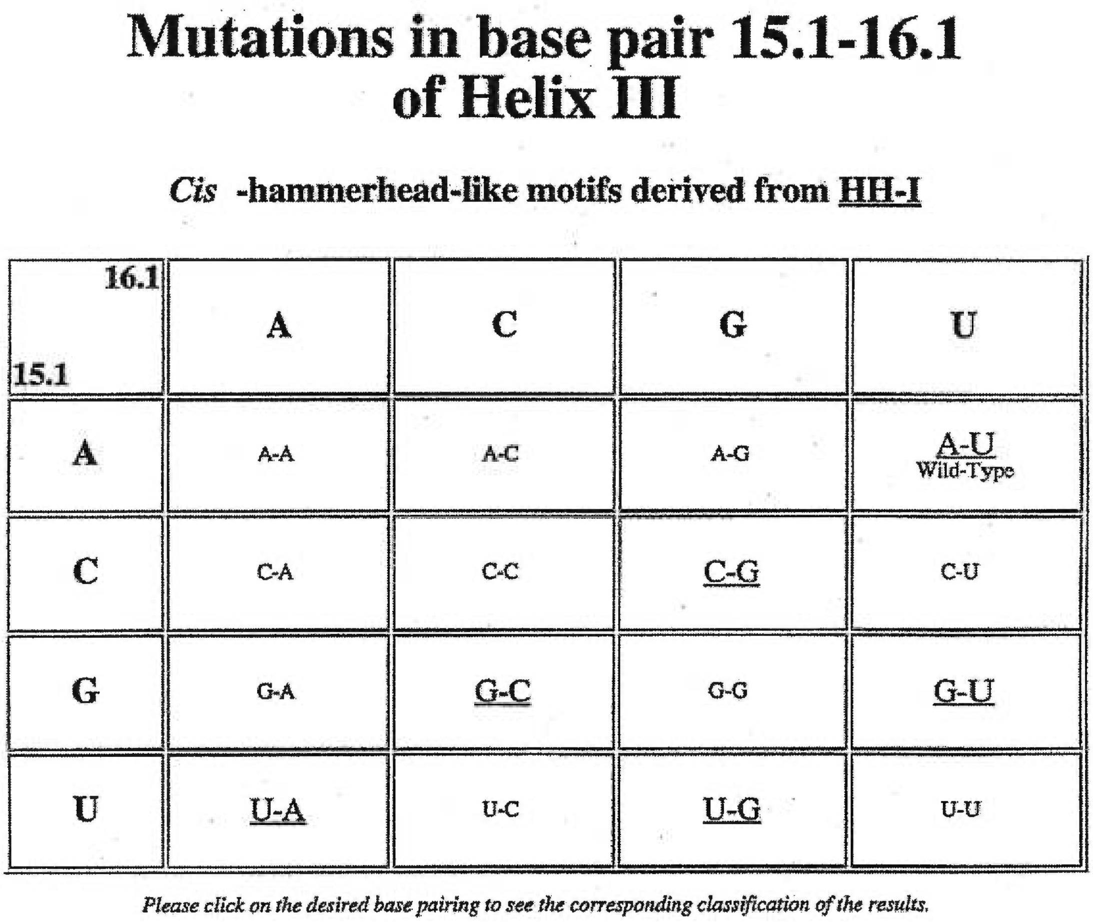
In the course of this work, we have searched the GenBank database using RNAMOT with the "wild-type" *cis* -hammerhead motif shown below. We also search for *cis* -hammerhead-like motifs which are all the motifs resulting from the replacement of one of the positions shown in maroon by any nucleotide: "N", or by a different base pairing.



*Please click on* HH-I *to access the classification of the results for the "wild-type" cis -hammerhead motif or on positions* $C_3$, $Y_4$, $G_5$, $A_6$, $G_8$, $A_9$, $G_{10}$-$C_{11}$, $G_{12}$, $A_{13}$, $A_{14}$, $A_{15.1}$-$U_{16.1}$ *and* $H_{17}$ *to access the classification of the results for the specific cis -hammerhead-like motif you wish.*

## Choix de la Paire de Bases

Exemple pour le motif HH-I avec mutations en position 15.1-16.1.



# Mutations in base pair 15.1-16.1 of Helix III

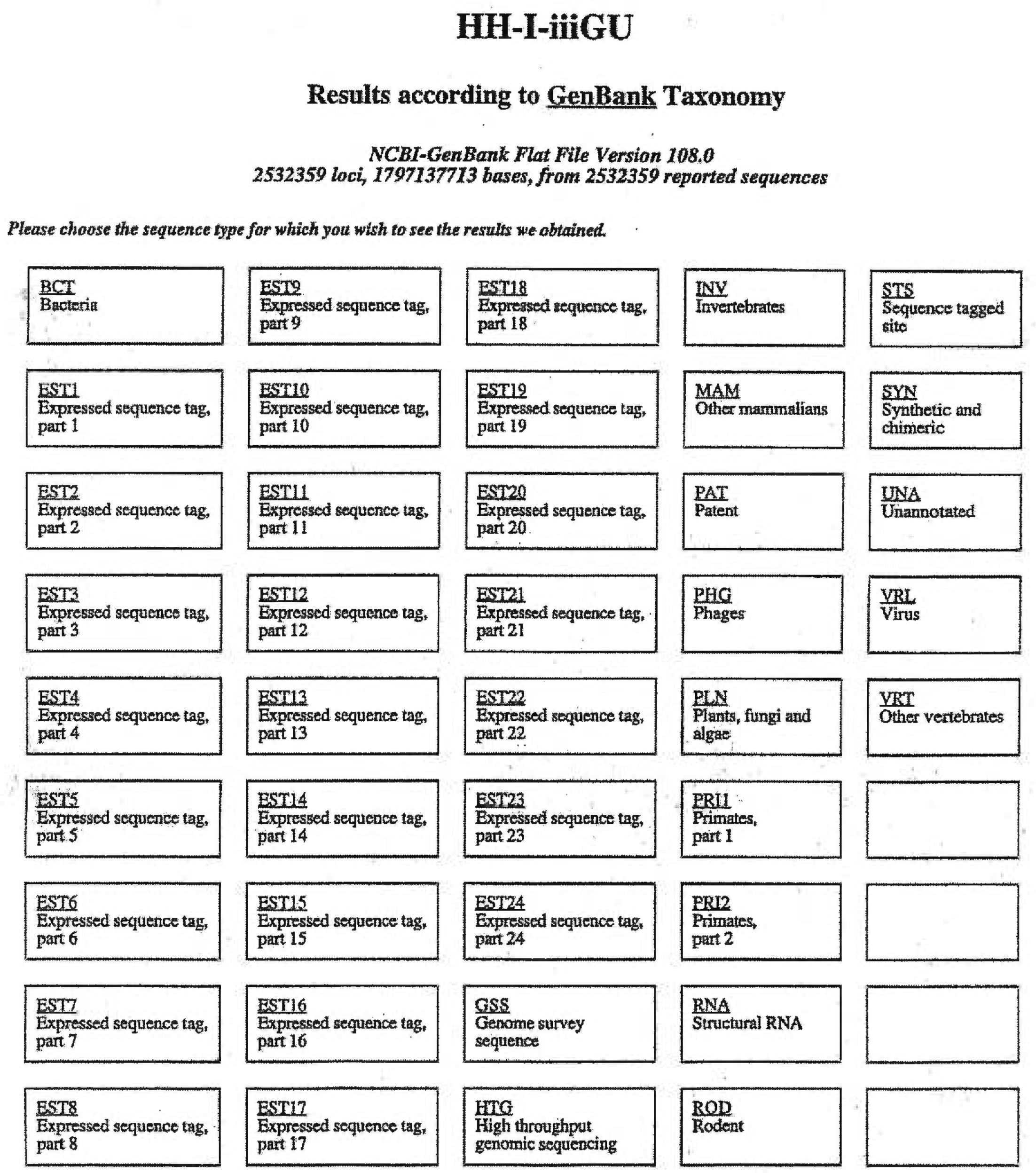
### *Cis* -hammerhead-like motifs derived from HH-I

| 16.1<br>15.1 | A | C | G | U |
|---|---|---|---|---|
| A | A-A | A-C | A-G | A-U<br>Wild-Type |
| C | C-A | C-C | C-G | C-U |
| G | G-A | G-C | G-G | G-U |
| U | U-A | U-C | U-G | U-U |

*Please click on the desired base pairing to see the corresponding classification of the results.*

## Tableau des Fichiers

Exemple avec le choix du motif HH-I possédant la paire de bases G-U en position 15.1-16.1.



# HH-I-iiiGU

### Results according to GenBank Taxonomy

*NCBI-GenBank Flat File Version 108.0*
*2532359 loci, 1797137713 bases, from 2532359 reported sequences*

*Please choose the sequence type for which you wish to see the results we obtained.*

| | | | | |
|---|---|---|---|---|
| BCT<br>Bacteria | EST9<br>Expressed sequence tag, part 9 | EST18<br>Expressed sequence tag, part 18 | INV<br>Invertebrates | STS<br>Sequence tagged site |
| EST1<br>Expressed sequence tag, part 1 | EST10<br>Expressed sequence tag, part 10 | EST19<br>Expressed sequence tag, part 19 | MAM<br>Other mammalians | SYN<br>Synthetic and chimeric |
| EST2<br>Expressed sequence tag, part 2 | EST11<br>Expressed sequence tag, part 11 | EST20<br>Expressed sequence tag, part 20 | PAT<br>Patent | UNA<br>Unannotated |
| EST3<br>Expressed sequence tag, part 3 | EST12<br>Expressed sequence tag, part 12 | EST21<br>Expressed sequence tag, part 21 | PHG<br>Phages | VRL<br>Virus |
| EST4<br>Expressed sequence tag, part 4 | EST13<br>Expressed sequence tag, part 13 | EST22<br>Expressed sequence tag, part 22 | PLN<br>Plants, fungi and algae | VRT<br>Other vertebrates |
| EST5<br>Expressed sequence tag, part 5 | EST14<br>Expressed sequence tag, part 14 | EST23<br>Expressed sequence tag, part 23 | PRI1<br>Primates, part 1 | |
| EST6<br>Expressed sequence tag, part 6 | EST15<br>Expressed sequence tag, part 15 | EST24<br>Expressed sequence tag, part 24 | PRI2<br>Primates, part 2 | |
| EST7<br>Expressed sequence tag, part 7 | EST16<br>Expressed sequence tag, part 16 | GSS<br>Genome survey sequence | RNA<br>Structural RNA | |
| EST8<br>Expressed sequence tag, part 8 | EST17<br>Expressed sequence tag, part 17 | HTG<br>High throughput genomic sequencing | ROD<br>Rodent | |

## Page des Résultats

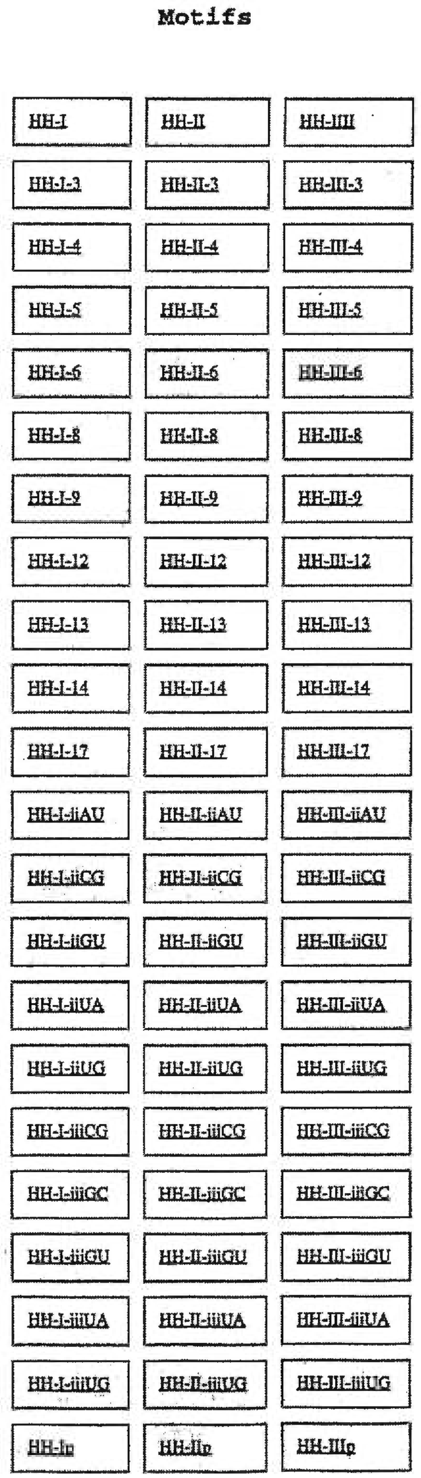
Exemple du motif HH-I possédant la paire de bases G-U en position 15.1-16.1 retrouvé dans le fichier vrt : autres vertébrés.

```
Results in the sequence line should be read as follow for HH-I motif:
| Helix I-NNN | CYGANGA | Helix II-GNN | loop (0-100) | Helix II-NNC | GAA | Helix III-AN | loop (0-100) | Helix III-NU | H | Helix I-NNN |
```

### HH-I-iiiGU.vrt.sol.filN.html

```
--- AB010101  Oryzias latipes gene for tyrosinase precursor, complete cds. --- (9791 bases)
|SCO:   2.68|POS:4630-4756|MIS: 0|WOB: 1|
|CAA|CUGAUGA|GAU|CUUCGUUUUGAUCUUCACAAUAAACUAAAUGUUAUUUUACGCAAAAAUG|AUC|GAA|GA|.. 50 nuc ..|UU|U|UUG|

--- CHKATE  Chicken erythroid anion transporter mRNA, complete cds. --- (3407 bases)
---------- complementary sequence ----------
|SCO:   3.19|POS:1540-1724|MIS: 0|WOB: 3|
|AGA|CUGAAGA|GCA|.. 99 nuc ..|UGC|GAA|GU|.. 57 nuc ..|GU|A|UUU|

--- CHKEATP  Chicken erythrocyte anion transport protein (band3) mRNA, complete --- (3007 bases)
---------- complementary sequence ----------
|SCO:   3.19|POS:1535-1719|MIS: 0|WOB: 3|
|AGA|CUGAAGA|GCA|.. 99 nuc ..|UGC|GAA|GU|.. 57 nuc ..|GU|A|UUU|

--- CHKIGMM  chicken ig heavy chain secreted mu constant region mrna. --- (1164 bases)
|SCO:   1.02|POS:946-1070|MIS: 0|WOB: 2|
|GGG|CCGAAGA|GUG|GGGCGCCGGCAACGUCUACACGUGCCUGGUGGGC|CAC|GAA|GC|.. 63 nuc ..|GU|C|UCC|

--- CHKVITAA  Chicken vitronectin receptor alpha subunit mRNA, complete cds. --- (3495 bases)
---------- complementary sequence ----------
|SCO:   2.44|POS:3131-3289|MIS: 0|WOB: 1|
|CCA|CUGAUGA|GAC|UUAAAUUCCAGUGGAUCAUCAGGAGCAAAAUCUCUGUUGCCUGUGGA|GUC|GAA|GA|.. 84 nuc ..|UU|U|UGG|

--- GGIGMUCH  Chicken mRNA for mu immunoglobulin heavy chain C region. --- (1398 bases)
|SCO:   1.02|POS:1181-1305|MIS: 0|WOB: 2|
|GGG|CCGAAGA|GUG|GGGCGCCGGCAACGUCUACACGUGCCUGGUGGGC|CAC|GAA|GC|.. 63 nuc ..|GU|C|UCC|

--- ONHPSOL  Chum salmon mRNA for somatolactin, complete cds. --- (2318 bases)
|SCO:   2.25|POS:1699-1817|MIS: 0|WOB: 2|
|UAA|CUGAAGA|GCU|.. 81 nuc ..|AGC|GAA|GU|UAAAGGAAGU|AU|C|UUG|

--- ONHSGSL3  Chum salmon gene for somatolactin, exon 4 and exon 5. --- (2671 bases)
|SCO:   2.25|POS:1810-1928|MIS: 0|WOB: 2|
|UAA|CUGAAGA|GCU|.. 81 nuc ..|AGC|GAA|GU|UAAAGGAAGU|AU|C|UUG|

--- QULTNC  Quail troponin C mRNA, 5' end. --- (480 bases)
---------- complementary sequence ----------
|SCO:   2.05|POS:176-303|MIS: 0|WOB: 1|
|AUC|CUGAAGA|GAU|.. 69 nuc ..|AUC|GAA|GU|CCACGGUGCCGCUGCCAUCCUCAUCCACCUC|AU|C|GAU|

--- RANTYRS  Frog mRNA for tyrosinase, complete cds. --- (3511 bases)
|SCO:   1.07|POS:899-1004|MIS: 0|WOB: 2|
|GGC|CUGAAGA|GUA|UAACAGCCUGAGAAUUAUAUGUAAUGGUACAAAUGAAGGUCCCCUAC|UGC|GAA|GC|CCUGGGCGUCACGACAGGAAC

--- SANKCC1  Squalus acanthias bumetanide-sensitive Na-K-Cl cotransport protein --- (5260 bases)
|SCO:   1.12|POS:1153-1229|MIS: 0|WOB: 2|
|CAG|CUGAAGA|GGC|UGU|GUC|GAA|GC|AUGUUGCUGACAAUAAAGGUGUUGUAAAGUUUGGCUGGAUUAAAGGU|GU|U|CUG|

--- U72484  Fugu rubripes RNA helicase (RNA-H) gene, partial cds; calcium --- (61901 bases)
|SCO:   1.94|POS:3203-3356|MIS: 0|WOB: 5|
|AGU|CUGAGGA|GUU|UUAAGGCCAGUGCAUGAGACAUCCUACGAUAGAGAAACAUUC|GGC|GAA|GC|.. 84 nuc ..|GU|U|GUU|
```

## Pour les Utilisateurs Fréquents



**Motifs**

| | | |
|---|---|---|
| HH-I | HH-II | HH-IIII |
| HH-I-3 | HH-II-3 | HH-III-3 |
| HH-I-4 | HH-II-4 | HH-III-4 |
| HH-I-5 | HH-II-5 | HH-III-5 |
| HH-I-6 | HH-II-6 | HH-III-6 |
| HH-I-8 | HH-II-8 | HH-III-8 |
| HH-I-9 | HH-II-9 | HH-III-9 |
| HH-I-12 | HH-II-12 | HH-III-12 |
| HH-I-13 | HH-II-13 | HH-III-13 |
| HH-I-14 | HH-II-14 | HH-III-14 |
| HH-I-17 | HH-II-17 | HH-III-17 |
| HH-I-iiAU | HH-II-iiAU | HH-III-iiAU |
| HH-I-iiCG | HH-II-iiCG | HH-III-iiCG |
| HH-I-iiGU | HH-II-iiGU | HH-III-iiGU |
| HH-I-iiUA | HH-II-iiUA | HH-III-iiUA |
| HH-I-iiUG | HH-II-iiUG | HH-III-iiUG |
| HH-I-iiiCG | HH-II-iiiCG | HH-III-iiiCG |
| HH-I-iiiGC | HH-II-iiiGC | HH-III-iiiGC |
| HH-I-iiiGU | HH-II-iiiGU | HH-III-iiiGU |
| HH-I-iiiUA | HH-II-iiiUA | HH-III-iiiUA |
| HH-I-iiiUG | HH-II-iiiUG | HH-III-iiiUG |
| HH-Ip | HH-IIp | HH-IIIp |

# Remerciements

Je voudrais d'abord exprimer ma profonde gratitude à Robert Cedergren pour m'avoir donné l'opportunité de travailler dans son laboratoire et pour m'avoir fait découvrir le monde de la recherche dans un environnement quasi international. Sa passion et son courage ont été un exemple inoubliable. Je lui dois beaucoup à plus d'un niveau.

Je remercie du fond du cœur tous les anciens membres du laboratoire Cedergren pour des interactions scientifiques et sociales enrichissantes. En particulier, je remercie ceux qui ont collaboré de près avec moi : Nicolas Cermakian, Rémi Émond, Gerardo Ferbeyre, Fabrice Leclerc, Marie Pageau, Bruno Paquin, Alice Rae et Sergey Steinberg. Je suis reconnaissante également à Pedro Miramontes de l'UNAM (Universidad Nacional Autónoma de México) et à Nicholas Delihas, Theodore Smirlis et Dawn Ashburn du State University of New York (SUNY) à Stony Brook. J'exprime un merci particulier au Dr Delihas pour un accueil chaleureux dans son laboratoire.

Je remercie les membres du Département de Biochimie et de l'Université de Montréal pour l'entraide et l'atmosphère amicale. Je mentionne un merci spécial à Mme Léa Brakier-Gingras pour avoir acceptée d'être ma codirectrice. Je suis reconnaissante aussi au Conseil National de Recherches du Canada pour le financement qu'il m'a fourni.

J'exprime enfin des remerciements tout particuliers à mon mari, ma famille et mes ami(e)s pour leur présence et leurs encouragements. Je dis merci aussi à Gerardo, Galia, Éric et Géraldine pour leur aide à la correction de cette thèse.