Université de Montréal

L'évolution de la mitochondrie, étudiée à travers son appareil de transcription et de traduction

par

Nicolas Cermakian

Département de biochimie

Faculté de Médecine

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de

Philosophiæ Doctor (Ph.D.)

en biochimie

Avril 1998

© Nicolas Cermakian, MCMXCVIII

W
4
U58
1998
V.031

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

L'évolution de la mitochondrie, étudiée à travers son appareil de transcription et de traduction

présentée par

Nicolas Cermakian

a été évaluée par un jury composé des personnes suivantes:

Guy Boileau	président-rapporteur
Robert Cedergren	directeur de recherche
Sylvie Mader	membre du jury
Brian Golding	examinateur externe
Guy Doucet	représentant du doyen

Thèse acceptée le: 03 AVR. 1998

SOMMAIRE

L'avènement des séquences moléculaires a profondément bouleversé tous les aspects de la biologie moléculaire, et en particulier l'évolution moléculaire. Le but des études relatées dans cette thèse est d'utiliser des données moléculaires pour étudier l'évolution des mitochondries, qui sont des compartiments cellulaires retrouvés chez la plupart des eucaryotes. La mitochondrie possède son propre génome et semble être d'origine endosymbiotique. L'évolution de cet organite est ici étudiée à travers les machineries de transcription et de traduction qui y sont retrouvées.

La synthèse d'ARN dans les mitochondries est réalisée par une enzyme très simple, apparentée à l'ARN polymérase (RNAP) de certains bactériophages. Pour découvrir quelle est la distribution de ce type d'enzyme chez les eucaryotes, une approche PCR a été mise au point et a permis d'isoler des séquences similaires à celles des RNAP de phages chez plusieurs eucaryotes très diversifiés (Chapitre 1, article 1). Ces résultats, appuyés par les analyses de séquences et constructions d'arbres phylogénétiques de l'article 2, permettent de conclure que ce type d'enzyme est très répandu chez les eucaryotes et a été recruté très tôt dans l'évolution, à peu près au moment de l'acquisition des mitochondries par leurs hôtes. Ces RNAP à une sous-unité ont ensuite remplacé très rapidement les enzymes plus complexes que les mitochondries avaient hérité de leurs ancêtres bactériens.

Une autre protéine très importante pour la transcription du génome mitochondrial des eucaryotes supérieurs est le facteur de transcription mitochondrial A (mtTFA). Une recherche dans les bases de données de séquences a permis de trouver un ADN complémentaire de la souris qui code pour une protéine très semblable à mtTFA, mais qui est nucléaire et spécifique aux testicules (Chapitre 2). Une série d'expériences de transcription inverse-PCR a ensuite montré que ces deux protéines sont produites à partir d'un même gène par sélection de promoteurs alternatifs et épissage différentiel subséquent. Cette découverte a des implications importantes non seulement pour l'évolution et la régulation de l'expression de mtTFA, mais aussi pour les mécanismes de communication entre le noyau et les mitochondries.

Enfin, les études du Chapitre 3 s'attardent sur les ARN de transfert mitochondriaux, qui présentent plusieurs caractéristiques atypiques. Pourquoi ces ARNt atypiques sont-ils absents des cytoplasmes mais n'ont pas été éliminés dans les mitochondries? Pour tenter de répondre à ces questions, un système *in vivo* a été mis au point dans la bactérie *Escherichia coli*. Les ARN de transfert atypiques, qui ne peuvent former une interaction tertiaire présumée importante (article 4) ou qui ont une tige de l'anticodon plus longue que la normale (article 5) peuvent fonctionner dans le cytoplasme de la bactérie, ce qui explique qu'ils puissent être utilisés dans les mitochondries, mais ont cependant une activité moindre qu'un contrôle ayant une structure normale, ce qui permet de comprendre pourquoi ils sont absents des cytoplasmes. Les mitochondries peuvent donc supporter une machinerie de traduction suboptimale.

Ces résultats sont discutés dans le cadre des publications récentes dans le domaine, et les implications plus générales sont également soulignées.

TABLE DES MATIÈRES

SOMMAIRE iii
TABLE DES MATIÈRES v
LISTE DES TABLEAUX x
LISTE DES FIGURES xii
LISTE DES SIGLES ET ABBRÉVIATIONS xiv
DÉDICACE xvi
REMERCIEMENTS xvii
INTRODUCTION 1
1. HISTOIRE GÉNÉRALE DE LA VIE ET CONTRIBUTION DES ÉTUDES MOLÉCULAIRES 1
1.1 Révolutions en évolution... 1
1.2 La phylogénie moléculaire 2
1.2.1 Les arbres phylogénétiques et les séquences 2
1.2.2 Deux exemples d'application de la phylogénie 6
1.3 Les horloges moléculaires 8
1.4 L'évaluation de la distribution d'un gène ou d'un processus 9
1.5 L'étude et la comparaison des génomes 10
1.6 Conclusion 10
2. LES ENDOSYMBIOSES DANS L'ÉVOLUTION 11
2.1 L'origine des chloroplastes et des mitochondries 12
2.2 L'origine du noyau et de la cellule eucaryote 14

2.3 Et les autres organites? 17
3. LA MITOCHONDRIE 18
3.1 Qu'est-ce que la mitochondrie? 18
3.2 Diversité des génomes mitochondriaux 19
3.3 Conclusion 21
4. LA TRANSCRIPTION DANS LES MITOCHONDRIES 22
4.1 Les ARN polymérases à plusieurs sous-unités 22
4.2 Les ARN polymérases à une sous-unité 24
4.3 Des ARN polymérases à une sous-unité dans les mitochondries 24
4.4 De la compagnie pour les ARN polymérases mitochondriales 25
4.4.1 Chez la levure 25
4.4.2 Chez les mammifères 28
5. LA TRADUCTION DANS LES MITOCHONDRIES 30
5.1 La machinerie de traduction et ses origines 30
5.2 Quelques particularités de la traduction dans les mitochondries 36
5.2.1 ARN ribosomiques minuscules ou en morceaux 36
5.2.2 Codes génétiques déviants et interactions codon-anticodon relâchées 37
5.2.3 ARN de transfert atypiques 38
5.2.4 Conclusion 39
6. SUJETS ET QUESTIONS ABORDÉS DANS CETTE THÈSE 40
6.1 L'évolution des ARN polymérases à une sous-unité (Chapitre 1) 40
6.2 La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris (Chapitre 2) 41
6.3 Particularités des ARN de transfert mitochondriaux (Chapitre 3) 41
CHAPITRE 1. L'évolution des ARN polymérases à une sous-unité 43
Article 1 44
Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage
Cermakian, N., Ikeda, T.M., Cedergren, R. et Gray, M.W.
Nucleic Acids Research **24**:648-654 (1996)

ABSTRACT 45
INTRODUCTION 45
MATERIALS AND METHODS 45
RESULTS 47
DISCUSSION 49
ACKNOWLEDGEMENTS 50
REFERENCES 51

Article 2 52

On the evolution of the single-subunit RNA polymerases
Cermakian, N., Ikeda, T.M., Miramontes, P., Lang, B.F., Gray, M.W. et Cedergren, R.
Journal of Molecular Evolution **45**:671-681 (1997)

ABSTRACT 54
INTRODUCTION 54
METHODS 56
RESULTS AND DISCUSSION 58
ACKNOWLEDGEMENTS 70
REFERENCES 70

CHAPITRE 2. La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris 87

Article 3 88

The relationship between the mitochondrial transcription factor A and the testis-specific HMG box protein in mouse
Cermakian, N. et Cedergren, R.

ABSTRACT 90
INTRODUCTION 90
MATERIAL AND METHODS 91
RESULTS 93
DISCUSSION 96
ACKNOWLEDGEMENTS 98
REFERENCES 98

CHAPITRE 3. Particularités des ARN de transfert mitochondriaux 115

Article 4 116

Nucleotide 47 in tRNAs: selection of the fittest is not always based on activity

Cermakian, N., McClain, W.H. et Cedergren, R.

ABSTRACT 118

INTRODUCTION 118

RESULTS 119

DISCUSSION 123

MATERIALS AND METHODS 130

ACKNOWLEDGEMENTS 132

REFERENCES 133

Article 5 144

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

Bourdeau, V., Steinberg, S.V., Ferbeyre, G., Émond, R., Cermakian, N. et Cedergren, R.

Proc. of the National Academy of Sciences of the USA **95**:1375-1380 (1998)

ABSTRACT 146

INTRODUCTION 146

MATERIALS AND METHODS 147

RESULTS 150

DISCUSSION 153

ACKNOWLEDGEMENTS 155

REFERENCES 155

DISCUSSION 169

1. LA "FRESQUE DE LA MITOCHONDRIE" 169

1.1 "Éveil de la Mitochondrie" 169

1.2 "Épanouissement et Déclin de la Mitochondrie" 171

1.3 "Angoisse de Vivre, ou plutôt de Respirer" 176

1.4 "Mort (ou Disparition) de la Mitochondrie" 178

2. Implications générales des travaux présentés dans cette thèse 179
2.1 L'origine des ARN polymérases 179
2.2 Épissage alternatif, gonades et communication mitochondrie-noyau 180
2.3 Petit ne signifie pas nécessairement simple, et simple, pas nécessairement vieux; le cas des ARN de transfert 181
3. Conclusion 183
BIBLIOGRAPHIE 185
ANNEXE xviii
Chapitre de livre xix
Modified nucleosides always were: an evolutionary hypothesis
Cermakian, N. et Cedergren, R.
In *Modification and editing of RNA : the alteration of RNA structure and function*, H. Grosjean et R. Benne (Ed.), Chapitre 29, American Society for Microbiology Press, Washington DC, Sous presse.
The evolutionary framework xxi
The phylogenetics of modified nucleosides xxii
The deep past xxiv
A model of the emergence of modified nucleosides xxvii
Conclusion xxxii
Acknowledgements xxxii
References xxxiii

LISTE DES TABLEAUX

INTRODUCTION

Tableau 1. ARN polymérase et facteurs de transcription connus chez les mitochondries de quelques organismes, au moment d'entreprendre les études des Chapitres 1 et 2 26

CHAPITRE 1. L'évolution des ARN polymérases à une sous-unité

Article 1

Table 1. Characteristics of PCR amplification products containing T3/T7 RNAP-like sequences 47

Table 2. Characteristics of deduced introns in T3/T7 RNAP-like sequences isolated by PCR amplification of eukaryotic DNA 49

Table 3. Characteristics of inferred intron junction sequences in T3/T7-like RNAP amplification products 49

Article 2

Table 1. RNA polymerase sequences used in this study 77

CHAPITRE 2. La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris

Article 3

CHAPITRE 3. Particularités des ARN de transfert mitochondriaux

Article 4

Article 5

Table 1. β-galactosidase activities and gene sequences of suppressor tRNAs 157

Table 2. Levels of expression and aminoacylation 159

DISCUSSION

BIBLIOGRAPHIE

ANNEXE

Chapitre de livre

Table 1. Modified nucleosides found at corresponding positions in tRNAs from eukaryotes, bacteria and archaea xxxvii

Table 2. Conserved modified nucleosides in rRNAs from eukaryotes, bacteria and archaea xxxviii

LISTE DES FIGURES

INTRODUCTION

Figure 1. Deux exemples d'arbres phylogénétiques 4

Figure 2. Deux hypothèses pour l'origine des eucaryotes 15

Figure 3. Le cycle d'élongation de la traduction 31

Figure 4. Structure secondaire et tertiaire des ARN de transfert 34

CHAPITRE 1. L'évolution des ARN polymérases à une sous-unité

Article 1

Figure 1. Design of PCR primers and strategy employed for amplification of putative mtRNAP sequences 46

Figure 2. Alignment of RNAP sequences 48

Figure 3. Positions of putative introns in *Acanthamoeba*, *Cryptomonas* and *Tetrahymena* PCR products 48

Article 2

Figure 1. Alignment of complete ssRNAP amino acid sequences 78

Figure 2. Maximum parsimony phylogenetic tree derived from the alignment of Fig. 1 81

Figure 3. Partial ssRNAP sequence obtained from the chytridiomycete fungus, *Rhizophydium* sp. #136 83

Figure 4. Maximum parsimony phylogenetic tree generated as in Fig. 2, but using an alignment of partial protein sequences, corresponding to the region shown in Fig. 3 85

CHAPITRE 2. La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris

Article 3

Figure 1. Position of gene-specific oligonucleotides used in this study 101

Figure 2. Phylogenetic tree of HMG box proteins 103

Figure 3. The amplification of m-tsHMG cDNA fragments 105

Figure 4. 5'-RACE with mouse embryo RNA 107

Figure 5. Amplification of the m-tsHMG intron 109

Figure 6. Amplification of splicing products and intermediates 111

Figure 7. Sequence of the 5' end of the mouse tsHMG/mtTFA gene; Model of the proposed alternative splicing of pre-mRNA giving rise to the mouse tsHMG and mtTFA isoforms 113

CHAPITRE 3. Particularités des ARN de transfert mitochondriaux

Article 4

Figure 1. Amber suppression assays 136

Figure 2. Growth assays with an *E. coli* strain lacking alanine $tRNA_{GGC}$ 138

Figure 3. tRNA aminoacylation assay 140

Figure 4. Long-term competition between cells with $U13_{am}$ or $C13_{am}$ 142

Article 5

Figure 1. The base-pairing pattern of mitochondrial tRNAs 160

Figure 2. Modifications to the $tRNA^{Ala}$ gene sequence 162

Figure 3. Properties of active tRNA clones 164

Figure 4. Secondary structure of isolated tRNAs 167

DISCUSSION

BIBLIOGRAPHIE

ANNEXE

Chapitre de livre

Figure 1. Nucleoside modifications likely to have been occuring in the cenancestor and in the RNA World xxxix

LISTE DES SIGLES ET ABRÉVIATIONS

ADN	Acide désoxyribonucléique
ADNmt	ADN mitochondrial
ARN	Acide ribonucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ARNt	ARN de transfert
ATP	Adénosine triphosphate
cp-...	Protéine chloroplastique
D-loop	*Displacement loop* (dans l'ADN mitochondrial d'animaux)
DNase I	Désoxyribonucléase I
EF-G	Facteur d'élongation (de la traduction) G
EF-Tu	Facteur d'élongation (de la traduction) Tu
HMG	*High Mobility Group*
h-mt....	Protéine mitochondriale humaine
HS	*Heavy strand* (brin lourd de l'ADNmt d'animaux)
HSP	*Heavy strand promoter* (promoteur du HS)
kb	kilobases
kDa	kilodaltons
LS	*Light strand* (brin léger de l'ADNmt d'animaux)
LSP	*Light strand promoter* (promoteur du LS)
msRNAP	ARN polymérase à plusieurs sous-unités

MTF1	*Mitochondrial transcription factor 1*
mtRNAP	ARN polymérase mitochondriale
mtTFA	Facteur de transcription mitochondrial A
mtTFB	Facteur de transcription mitochondrial B
NAD	Nicotinamide adénine dinucléotide
nt	nucléotide
RNAP	ARN polymérase
sc-mt....	Protéine mitochondriale de la levure *Saccharomyces cerevisiae*
ssRNAP	ARN polymérase à une sous-unité
TBP	*TATA-box binding protein*
tsHMG	*Testis-specific HMG protein*
xl-mt....	Protéine mitochondriale de l'amphibien *Xenopus laevis*

À mes parents.

À la vie,
la plus belle de toutes les richesses.
L'étudier, c'est lui rendre hommage...

REMERCIEMENTS

Premièrement, merci à Bob Cedergren, pour avoir "veillé sur moi" depuis plus de cinq ans. Bob est un exemple de dévouement, de courage et d'amour de la vie. Qu'est-ce que j'en ai appris, des choses, depuis cinq ans à ses côtés!

Merci également aux gens avec qui j'ai eu la chance de collaborer : Bill McClain et les gens de son laboratoire (Jay et Kay), Tatsuya Ikeda, Franz Lang, et par-dessus tout Mike Gray. Je me rappelle comme si c'était hier de ce samedi après-midi du début de 1995 où j'ai rencontré Mike par hasard dans un couloir du département; avec un peu de recul, je peux affirmer qu'il s'agit sans doute d'un point tournant dans mon doctorat!

Merci à tous mes confrères et consœurs de laboratoire pour les discussions, les rires, les collaborations, les encouragements, les enseignements... Un clin d'œil particulier à Benoit Cousineau, qui m'a beaucoup appris et avec qui j'ai tant échangé: solutions, idées, états d'âme... Je désire remercier Véronique Bourdeau pour avoir patiemment lu et commenté cette thèse, et pour la collaboration dans l'épopée des ARN de transfert. Et n'oublions pas Pedro Miramontes, qui nous a apporté le plus beau des cadeaux pour de jeunes biologistes: une nouvelle façon de voir la Nature!

Enfin, merci à ma famille et à mes amis, pour leur présence et leur appui. Mes "pèlerinages" réguliers à Trois-Rivières chez mes parents, ainsi que les discussions et sorties variées avec mon frère ont sans doute été aussi importants que les PCR et séquençages pour la réussite de mon doctorat!

INTRODUCTION

1. HISTOIRE GÉNÉRALE DE LA VIE ET CONTRIBUTION DES ÉTUDES MOLÉCULAIRES

1.1 RÉVOLUTIONS EN ÉVOLUTION...

Thomas Kuhn (1983) a décrit ce qu'il appelle "Révolutions scientifiques" comme des épisodes de l'histoire des sciences où le *paradigme* (modèle d'étude commun) d'un groupe de scientifique est remplacé par un autre, entraînant de nouvelles questions, de nouveaux outils et une nouvelle façon de voir le monde et les problèmes scientifiques. Dans ce cadre, une des plus grandes révolutions scientifiques à avoir eu lieu est sans doute celle qui a accompagné la publication en 1859 de *L'Origine des Espèces* de Charles Darwin (1962). Bien que l'idée d'évolution des espèces fût assez bien soutenue et acceptée même avant la parution de ce livre, ce dernier a tout de même causé une commotion au XIXè siècle puisqu'il impliquait une autre façon de voir le monde et la place de l'Homme dans celui-ci. L'être humain n'était plus le but de l'évolution, mais seulement une conséquence de la sélection naturelle. La théorie de la sélection naturelle a depuis fourni à des générations de biologistes un paradigme sur lequel baser leurs recherches.

Je vais tenter de montrer dans les prochaines pages que la biologie de l'évolution a été transformée dans le dernier quart de ce siècle par la disponibilité d'une nouvelle masse d'informations: les séquences d'acides nucléiques et de protéines. Je présenterai

comment peuvent être utilisées ces séquences en évolution moléculaire, et donner des exemples de ces méthodes d'analyse pour illustrer à quel point la disponibilité de ces séquences a révolutionné ce domaine en remettant en cause bien des concepts établis, et donc en établissant de nouveaux paradigmes.

1.2 La phylogénie moléculaire

1.2.1 Les arbres phylogénétiques et les séquences

C'est en 1866 que Ernst Haeckel invente le mot *phylogénie* alors qu'il publie un ensemble d'arbres représentant ce qui était connu des relations évolutives entre les êtres vivants (Hillis, 1997). Mais il faut attendre le début de ce siècle pour que voie le jour la *phylogénie moléculaire*, c'est-à-dire l'utilisation de données de biologie moléculaire pour établir des phylogénies (Li, 1997). Des données immunologiques, puis d'hybridation d'ADN et de digestion par des enzymes de restriction, permettaient de donner une nouvelle dimension à une discipline qui n'utilisait jusqu'alors que des critères anatomiques, physiologiques, morphologiques et paléontologiques.

L'arrivée des séquences de gènes et de protéines, en particulier à partir des années 1970, donna un nouveau souffle à la phylogénie moléculaire. Ces séquences présentent en effet de grands avantages pour construire des phylogénies: elles comportent une quantité énorme d'information (chaque acide aminé ou nucléotide pouvant être un caractère distinct), les caractères évoluent souvent de façon relativement indépendante, et ils sont aisément comparables (Ridley, 1993).

Pour qu'une séquence soit considérée adéquate pour ce genre d'analyses, elle doit remplir certaines conditions (Woese, 1987; Klenk et Zillig, 1994). Premièrement, un *marqueur évolutif* doit être présent chez toutes les espèces concernées. Deuxièmement, il doit être d'une longueur suffisante pour qu'on ait assez d'informations pour dresser la phylogénie. Troisièmement, il doit être assez conservé pour qu'on puisse aligner correctement les différentes séquences, mais pas trop (sinon,

on risque de manquer d'information); le degré de conservation doit être relié au niveau où doit se faire la phylogénie: une étude impliquant des espèces rapprochées requerra des séquences évoluant assez vite, tandis qu'un arbre incluant des organismes très éloignés nécessitera un marqueur qui a varié très peu au cours du temps. Après avoir dressé un alignement de ces séquences, différentes techniques de construction d'arbres peuvent être utilisées (Swofford et Olsen, 1990; Ridley, 1993). Certaines suivent le principe de la *parcimonie*, qui fait la supposition que le *vrai* arbre est celui qui est le plus *court*, c'est-à-dire celui qui nécessite le moins de changements ou transformations pour expliquer les différences entre tous les points de l'arbre (Stewart, 1993). D'autres méthodes se basent plutôt sur la distance entre les séquences, c'est-à-dire que les espèces partageant le plus grand nombre de caractères sont regroupées. Aucune méthode n'est parfaite, et il est bon de faire l'analyse avec quelques-unes pour vérifier que le résultat est le même (voir Chapitre 1; Iwabe et al., 1991; Klenk et Zillig, 1994).

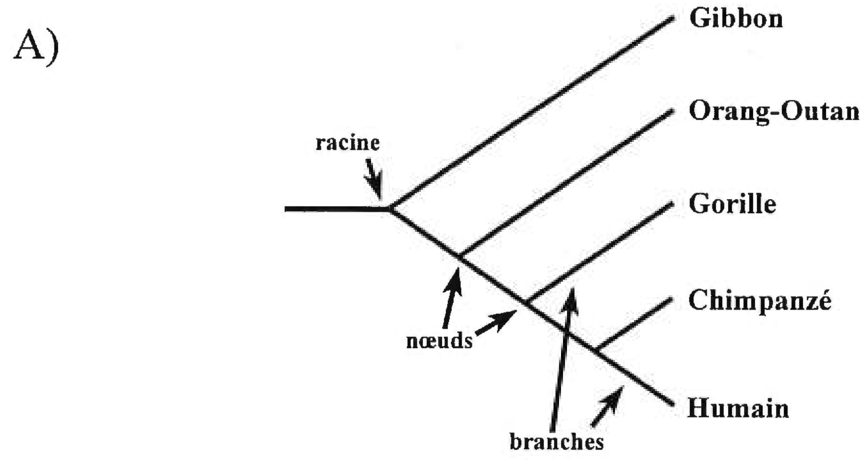
Deux exemples d'arbres phylogénétiques sont présentés à la Figure 1. On peut voir (Fig. 1A) que ces arbres comportent des *branches* et des *nœuds*, et que la pointe de chacune des branches représente un organisme ou une séquence. Un nœud correspond à l'ancêtre commun des organismes situés au bout des branches qui en émanent. Un arbre phylogénétique peut avoir une *racine* (par exemple à la Fig. 1A), c'est-à-dire un point correspondant à l'ancêtre commun de tous les organismes ou séquences de l'arbre. On peut situer une racine dans l'arbre en incluant dans l'analyse une séquence d'un organisme qu'on sait être hors du groupe étudié (par exemple la souris ou le mouton pour l'arbre de la Fig. 1A). Dans d'autres cas, en l'absence d'une telle séquence, on peut se retrouver avec un arbre sans racine, comme l'arbre de la Fig. 1B. Un autre concept qui sera important plus loin, est celui de *monophylie*. Un groupe est dit *monophylétique* si ses membres partagent un ancêtre commun ET qu'il contient tous les descendants de cet ancêtre: par exemple, les humains, les chimpanzés et les gorilles (Fig. 1A) forment un groupe monophylétique. Au contraire, dans notre exemple, les grands singes, si on en exclut les humains, ne sont pas monophylétiques; on dit plutôt qu'ils sont *paraphylétiques*, puisque leur ancêtre est partagé par les

Figure 1. Deux exemples d'arbres phylogénétiques.

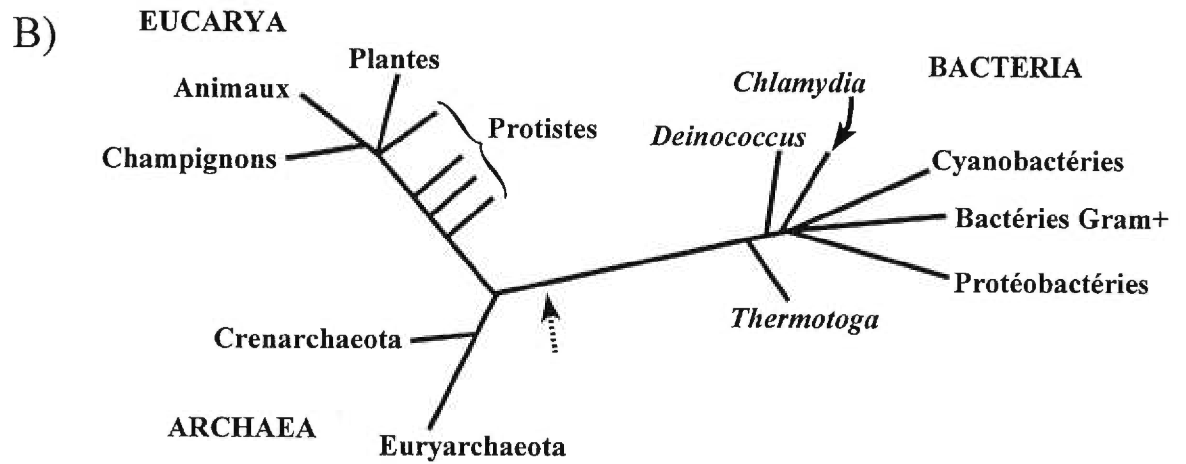
A) Arbre phylogénétique de l'humain et des grands singes. La racine de l'arbre, ainsi que des exemples de branches et de nœuds, sont mis en évidence. (Adapté de Li, 1997)

B) Arbre phylogénétique universel, tel qu'obtenu avec des séquences d'ARN ribosomique et certains marqueurs protéiques. Les trois domaines (eucarya, archaea et bacteria) sont indiqués, de même que quelques règnes de chaque domaine. La position présumée de la racine est illustrée par une flèche pointillée. Voir le texte pour plus de détails. (Adapté de Woese, 1987 et Woese et al., 1990)

A)



B)



humains. Par ailleurs, des organismes ou mécanismes ayant différentes origines sont dits *polyphylétiques*.

Différentes méthodes ont été mises au point pour évaluer la robustesse des arbres, le niveau de confiance qu'on peu avoir en eux (Swofford et Olsen, 1990). La plus utilisée est appelée *bootstrap*, et consiste en un rééchantillonnage statistique des données de l'alignement, permettant d'assigner à chaque regroupement de branches une certaine valeur donnant une idée de sa robustesse (plus sa valeur est proche de 100%, plus le regroupement est jugé fiable). Cependant, quelle que soit la robustesse de l'arbre, il est important de toujours distinguer entre un arbre de gènes (ou protéines), et l'arbre des organismes. Si le but est d'étudier l'évolution de ces organismes, il est important de faire des arbres avec différents marqueurs. D'un autre côté, on peut faire un arbre pour étudier l'histoire de la protéine ou du gène en question; c'est d'ailleurs ce qui a été fait dans les articles présentés dans les Chapitres 1 et 2.

1.2.2 Deux exemples d'application de la phylogénie

Mon premier exemple de bouleversements d'idées dus à la phylogénie moléculaire touche l'organisation générale des êtres vivants. Ceux-ci ont pendant longtemps été classés selon des caractéristiques morphologiques et physiologiques, soit en deux règnes (Margulis, 1996): les eucaryotes et les procaryotes; soit en cinq règnes (Margulis, 1988): les animaux, les plantes, les champignons, les protistes et les procaryotes. Cependant, une analyse phylogénétique de Woese et Fox (1977), utilisant des séquences d'ARN ribosomique (ARNr), a permis de suggérer une subdivision des êtres vivants plutôt en trois grands groupes: les eucaryotes, les eubactéries ("vraies" bactéries) et les archaebactéries (Fig. 1B). Par la suite, cette proposition a été confirmée par de nouvelles analyses incluant un bien plus grand nombre de séquences d'ARNr (Woese, 1987), et également des séquences d'autres gènes, notamment des gènes codant pour différentes protéines (Woese et al., 1990). On appelle maintenant ces trois grands groupes "Domaines", et ceux-ci ont été rebaptisés Eucarya, Archaea et Bacteria, pour souligner que les eubactéries et archaebactéries n'ont pas de parenté

directe (Woese et al., 1990) (Nous emploierons ici indifféremment les deux séries d'appellations.). De fait, Iwabe et al. (1989) ont montré de façon élégante, en utilisant des familles de gènes dupliqués, que les eucaryotes et archaebactéries sont probablement plus proches parents entre eux qu'avec les eubactéries (l'emplacement de la racine ainsi inférée est indiqué par une flèche dans la Fig. 1B); d'autres études ont appuyé ces résultats en utilisant des marqueurs différents (Brown et Doolittle, 1995) ou des ensembles de données plus grands (Baldauf et al., 1996).

Ainsi, comme Woese l'avait indiqué en 1977, la classification eucaryotes/procaryotes ne tient pas en termes de phylogénie. Par ailleurs, la classification en cinq règnes a encore moins de sens, puisqu'elle met un vaste groupe d'organisme (les procaryotes) sur le même pied que des groupements beaucoup plus semblables entre eux; de plus, les protistes ne forment pas un groupe monophylétique (voir Fig. 1B). La nouvelle division tripartite, bien que très répandue, ne fait toujours pas l'unanimité (Cavalier-Smith, 1993; Margulis, 1996), mais les arguments contre elle ne sont plus très forts. Selon Woese, "eucaryotes" et "procaryotes" sont des concepts valables, mais à un niveau cytologique plutôt que phylogénétique (Woese et al., 1990). Il vaut mieux d'abord réduire la complexe cellule eucaryote en unités phylogénétiques séparées (cytoplasme, organites) avant de comparer en fait des procaryotes entre eux (Woese et Fox, 1977). En fait, Margulis (1981) semble être d'accord sur ce point (« Several prokaryotes make a eukaryote »). De plus, les procaryotes ont été unis sur des bases négatives — l'absence de caractéristiques eucaryotes — qui ne sont pas informatives du point de vue de la phylogénie (Woese et al., 1990). La controverse se situerait donc peut-être au niveau du champ d'application des termes.

L'autre exemple que je souhaite présenter ici est de moins grande envergure, mais montre que la "révolution" a lieu à plusieurs niveaux. Dès l'enfance, on apprend à classer ensemble sous l'appellation "rongeurs" toutes ces petites bêtes aux grandes incisives: souris, cobayes, lapins, rats, etc. En fait, plusieurs caractères morphologiques semblent regrouper ces mammifères dans un groupe monophylétique. Cependant, deux articles récents suggèrent que les lapins et lièvres ne peuvent être groupés avec les

rongeurs (Graur et al., 1996) et que les cobayes ne sont pas des rongeurs (D'Erchia et al., 1996), en utilisant respectivement les séquences de 91 protéines et les génomes mitochondriaux. Leurs analyses sont assez convaincantes, mais la quantité de données analysées pas toujours très grande. Ainsi, plutôt que de nier cette analyse ou les données morphologiques disponibles, les recherches futures devront tenter de voir comment concilier les deux phylogénies: la traditionnelle (les données morphologiques pourraient être expliquées autrement) et la plus récente (elle devra être approfondie et confirmée)... Le cas des lapins et des cobayes n'est qu'un exemple du potentiel énorme des séquences moléculaires pour aider dans la résolution de la phylogénie des mammifères, un sujet complexe débattu depuis au moins deux siècles (Graur, 1993).

1.3 LES HORLOGES MOLÉCULAIRES

Les séquences de gènes évoluent au gré des mutations au cours du temps. Il a été proposé que cette vitesse d'évolution est à peu près constante dans certains cas, ce qui a mené au concept d'*horloge moléculaire* (Li, 1997). Si on connaît les dates de divergences de certaines lignées évolutives d'une part (par exemple en se basant sur des données paléontologiques) et la divergence de séquences de gènes ou de protéines de ces lignées d'autre part, il serait donc possible, en principe, de dater d'autres événements de divergence pour lesquelles on n'a que des données moléculaires.

Cette approche a été utilisée par Doolittle et al. (1996) pour proposer que les trois grands domaines de la vie (section 1.2.2 et Fig. 1B) ont divergé il y a environ deux milliards d'années, ce qui est bien loin de la valeur jusqu'alors admise, 3,0 ou 3,5 milliards d'années, découlant d'études biologiques (diversité de la vie) et géologiques (selon certains fossiles et processus géochimiques) (Golding, 1996). Un autre exemple est l'étude publiée en 1996 par Wray et al., repoussant à près d'un milliard d'année la divergence des différents groupes d'animaux, que les paléontologistes placent plutôt dans une courte fenêtre de temps il y a environ 600 millions d'années (Conway Morris, 1997), la fameuse explosion du Cambrien (Gould, 1991), à cause d'un manque de fossiles avant cette période. Bien que ces nouvelles visions méritent notre attention

(d'autant plus que les versions précédentes ne sont pas si fortement appuyées par des données), cette technique de l'horloge moléculaire comporte deux problèmes: 1) les vitesses d'évolution varient beaucoup selon les lignées, et même dans les lignées, alors que les auteurs des nouvelles études supposent que les vitesses sont constantes; 2) les dates utilisées pour calibrer le système doivent être bonnes (par exemple, Doolittle et al. utilisent la date conventionnelle de 600 millions d'années pour la divergence des animaux, mais leur évaluation finale aurait été poussée loin dans le passé s'ils avaient utilisé la valeur de Wray et al.). Les deux études utilisent cependant une grande quantité d'informations et de nombreux tests statistiques pour valider leur approche. D'ailleurs, de nouvelles études paléontologiques semblent maintenant aller dans le sens des études phylogénétiques, en repoussant au Précambrien l'origine d'au moins certains embranchements d'animaux (Fedonkin et Waggoner, 1997).

1.4 L'ÉVALUATION DE LA DISTRIBUTION D'UN GÈNE OU D'UN PROCESSUS

Une approche courante en biologie de l'évolution, et en particulier en évolution moléculaire maintenant que sont disponibles des quantités incroyables de séquences, est de déterminer la distribution d'un caractère, d'une protéine, d'un mécanisme donné parmi les êtres vivants. Cette approche va être centrale dans la discussion des articles du Chapitre 1, ainsi que dans l'essai présenté en annexe. Voici un exemple. Les promoteurs eucaryotes sont reconnus par des complexes contenant entre autres une protéine appelée TBP (pour *TATA-box binding protein*; Roeder, 1996). Cette protéine est présente chez tous les eucaryotes: on en conclut qu'elle a été recrutée par ceux-ci très tôt dans cette lignée évolutive. Les eubactéries reconnaissent différemment leurs promoteurs et n'ont pas de TBP. Cependant, une protéine très similaire à TBP a été trouvée chez des archaebactéries (March et al., 1994). On en conclut que la TBP ancestrale a été recrutée pour agir en transcription avant la divergence des archaebactéries et des eucaryotes. Il est à noter que les mécanismes de transcription sont en général plus semblables entre archaebactéries et eucaryotes qu'avec les eubactéries, ce qui est en accord avec les études phylogénétiques (voir section 1.2.2 et Fig. 1B).

1.5 L'ÉTUDE ET LA COMPARAISON DES GÉNOMES

Le premier génome d'un être vivant à être séquencé au complet fut celui de l'eubactérie *Haemophilus influenza* (Fleischmann et al., 1995), et depuis, onze autres génomes ont été séquencés (de sept eubactéries, trois archaebactéries et un eucaryote) (Pennisi, 1997). Et ce n'est pas fini: des dizaines d'autres séquences de génomes sont en voie d'être complétées (Pennisi, 1997). Il s'agit d'une énorme quantité de nouvelles informations brutes, qui seront d'une grande utilité dans l'étude des processus biologiques, mais également d'évolution. En effet, on entre dans l'ère où l'on comparera non seulement des gènes, mais aussi des génomes entre eux! Quel est l'ordre des gènes dans tel ou tel génome (voir la Discussion)? Quel est le contenu de ces génomes en termes de gènes de différentes familles fonctionnelles? De quoi pouvait avoir l'air le génome de l'ancêtre commun de tous les êtres vivants? Voici des questions parmi tant d'autres qui pourront être posées maintenant que des génomes entiers sont séquencés. Par exemple, l'analyse du génome de l'archaebactérie *Methanococcus jannaschii* (Bult et al., 1996) a permis de confirmer le caractère distinct des archaebactéries (voir sections 1.2.2 et 1.4).

1.6 CONCLUSION

Ce qui ressort de ces quelques exemples, c'est que peu importe si les études en question nous convainquent ou non, il est indéniable que les séquences constituent une masse incroyable d'information et que leur analyse représente un puissant moyen pour éclaircir les mécanismes du monde vivant. À mesure que les nouvelles séquences afflueront et que les techniques seront améliorées, il sera possible de retirer le maximum d'informations de ces données brutes. Cependant, la phylogénie et les autres méthodes d'analyse de séquences n'ont pas réponse à tout et ne peuvent être une fin en soi. Leurs résultats doivent être interprétés en conjonction avec ce que nous disent la biologie moléculaire, l'écologie, la paléontologie ou la physiologie selon les cas. Par ailleurs, certaines questions en évolution moléculaire sont approchées de façon

différente. Par exemple, certains chercheurs qui se demandent de quoi avaient l'air les toutes premières formes de vie tentent de mettre en évidence de nouvelles activités catalytiques de l'ARN (Szostak, 1992; Szostak et Ellington, 1993), qui auraient pu être l'apanage d'organismes précoces où les rôles de génome et de catalyseur étaient joués par l'ARN (Darnell et Doolittle, 1986; Cedergren et Grosjean, 1987). D'autres approches que l'analyse de séquences sont également utilisées dans les chapitres 2 et 3 de cette thèse (expression de gènes, activité de molécules).

En fait, il ressort de la discussion des dernières pages que l'avènement de l'évolution moléculaire *n'est pas* une révolution scientifique dans le sens que Kuhn (1983) prête à ce phénomène, mais plutôt une avancée des techniques menant à une accumulation d'informations sans précédent. Cependant, ces nouveaux outils permettent de profonds changements de perspective, donc des révolutions scientifiques, dans divers champs d'études en évolution moléculaire et en biologie en général. Un peu comme un enfant qui grandit, ce qui lui permet un jour de regarder par la fenêtre et de jeter ainsi un regard nouveau sur le monde. C'est le regard vers l'extérieur qui est la révolution, mais celle-ci est permise par la croissance de l'enfant.

Des systèmes biologiques qui se prêtent bien aux analyses évolutives, vu leur importance dans la bonne marche des cellules, sont la transcription (synthèse d'ARN) et la traduction (synthèse protéique). Nous verrons que l'étude de différentes molécules impliquées dans ces mécanismes (ARN polymérases, facteurs de transcription, ARN de transfert) nous a permis de mieux comprendre l'évolution non seulement des mécanismes eux-mêmes, mais aussi des mitochondries et des êtres vivants en général. Auparavant, nous donnerons cependant un aperçu de ce qui est connu de l'évolution des mitochondries, organites dont l'histoire est pleine de rebondissements…

2. Les endosymbioses dans l'évolution

Une symbiose peut être définie comme une « association, pour des portions importantes de leur cycle de vie, d'individus qui sont membres d'espèces différentes ». Cette définition est de Lynn Margulis (1981) et est tirée d'un livre où, s'appuyant sur de multiples descriptions et exemples, elle souligne l'importance des symbioses dans l'évolution et les multiples innovations qu'elles permettent, en particulier chez les eucaryotes et dans la genèse des organites.

2.1 L'origine des chloroplastes et des mitochondries

Une des caractéristiques distinctives des cellules eucaryotes est la présence de compartiments (appelés *organites*) séparés par des membranes lipidiques, où sont circonscrites certaines activités de ces cellules. Les chloroplastes et les mitochondries sont deux organites entourés d'une double membrane et responsables de convertir l'énergie en des formes utilisables par la cellule (Alberts et al., 1994). Il est aujourd'hui généralement admis que des événements d'endosymbiose impliquant une cellule hôte eucaryote et des eubactéries sont à l'origine de la formation des mitochondries et des chloroplastes (Gray, 1992). Déjà dans les années 1890, Altmann faisait remarquer la ressemblance entre les mitochondries (qu'il appelait "bioblastes") et les bactéries (Lehninger, 1964). Depuis, plusieurs caractéristiques de ces organites ont souligné cette parenté avec les bactéries : double membrane lipidique, division semblable aux bactéries, génome généralement circulaire (probablement attaché à des repliements de la membrane interne), ribosomes très semblables aux ribosomes bactériens et sensibles aux antibiotiques antibactériens, synthèse protéique amorcée avec une N-formylméthionine (et non avec une méthionine comme dans le cytoplasme eucaryote) (Alberts et al., 1994). Finalement, l'origine eubactérienne des mitochondries (Cedergren et al., 1988) et des chloroplastes (Delwiche et al., 1995) a été fortement appuyée par des études phylogénétiques utilisant les séquences de différents gènes: ces études groupent invariablement les mitochondries avec les eubactéries α-pourpres, et les chloroplastes avec les cyanobactéries, un groupe de bactéries photosynthétiques.

Suite à cette endosymbiose, la majorité des gènes de ces organites auraient été soit perdus, soit transférés au noyau de la cellule hôte, les produits de ces gènes étant dorénavant produits dans le cytoplasme et importés dans les organites. Ceci explique les génomes très réduits des mitochondries et des chloroplastes actuels.

L'origine par endosymbiose est en fait bien plus évidente dans le cas des chloroplastes, ou des plastides en général, que dans le cas des mitochondries (Gray, 1992). Les chloroplastes utilisent une ARN polymérase de type bactérien (voir section 4.1). L'organisation de plusieurs opérons chloroplastiques reflète de façon évidente les unités de transcription bactériennes correspondantes. Les promoteurs chloroplastiques sont le plus souvent semblables au modèle bactérien, avec leurs boîtes -35 et -10. Un intron a été découvert à la même position dans un $ARNt^{Leu}$ de chloroplastes et de cyanobactéries. Enfin, les gènes chloroplastiques peuvent être transcrits par l'ARN polymérase de la bactérie *Escherichia coli* et les ARN messagers (ARNm) traduits par la machinerie de traduction de cette même bactérie (Alberts et al., 1994).

Le cas des mitochondries est plus problématique, et avant que les études phylogénétiques ne confirment leur origine par endosymbiose, cette hypothèse cohabitait avec d'autres suggérant une origine autogène des mitochondrie (comme la théorie de l'épisome), c'est-à-dire une origine de l'organite à partir de sources cellulaires (Munn, 1974; Gray, 1992). En effet, les mitochondries ont des génomes contenant beaucoup moins de gènes que les chloroplastes, et ces gènes ne sont pas transcrits par une ARN polymérase de type bactérien (voir section 4.3 et Chapitre 1).

Tandis qu'il est de plus en plus clair que les mitochondries sont monophylétiques (un ancêtre commun; une seule endosymbiose), la question n'est toujours pas réglée pour ce qui est des plastides (Gray, 1992; Delwiche, 1995). Bien que tous les plastides possèdent la chlorophylle *a*, leur pigment accessoire diffère (phicobiliprotéines, chlorophylle *b* ou chlorophylle *c*). Il a été proposé que des eubactéries différentes sont à l'origine de ces plastides avec pigments différents (polyphilie); il est cependant possible que la divergence des types de pigments se soit faite après l'endosymbiose

(monophylie) (Bryant, 1992; Gray, 1993). La résolution de ce problème nécessitera des analyses phylogénétiques plus poussées (Delwiche, 1995).

2.2 L'ORIGINE DU NOYAU ET DE LA CELLULE EUCARYOTE

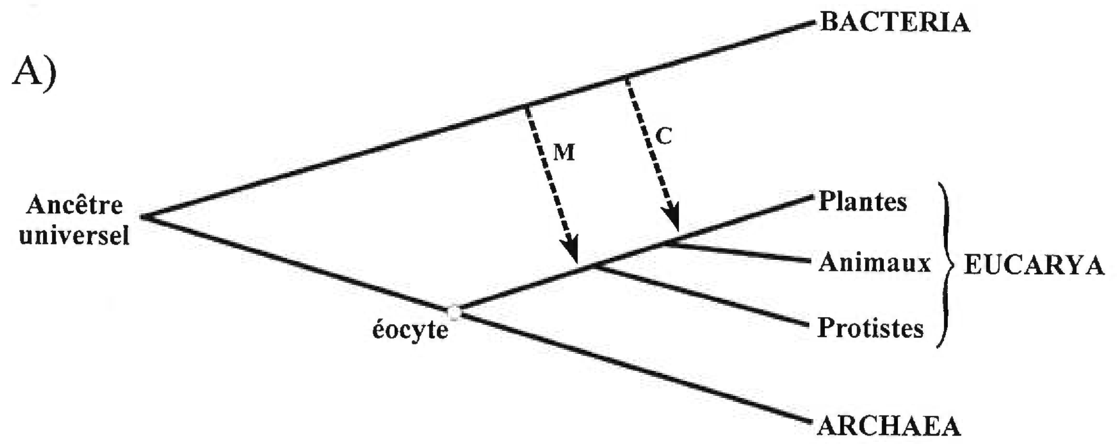
Une hypothèse très attrayante a été proposée par Zillig et al. (1991), puis développée et appuyée par Gupta et Golding (1993, 1996). Ils proposent que la cellule eucaryote est le résultat d'une symbiose entre une eubactérie gram-négative et une archaebactérie (Fig. 2B). Cette hypothèse va à l'encontre du modèle de Woese (voir section 1.2.2 et Fig. 2A), selon lequel les eucaryotes ont simplement dérivé d'une archaebactérie (éocyte). Cependant, on a vu que ce modèle avait été basé sur des arbres phylogénétiques faits à partir de séquences d'ARNr puis de quelques autres gènes seulement. L'analyse de Gupta et Golding (1996) est beaucoup plus approfondie puisqu'elle inclut des séquences de 24 protéines. Alors que certaines protéines mènent à un regroupement des eucaryotes et des archaebactéries (fidèlement à l'arbre de Woese), presque autant de séquences regroupent plutôt les eucaryotes et les eubactéries gram-négatives, laissant ensemble les archaebactéries et les eubactéries gram-positives. Les auteurs concluent que leur modèle de fusion (Fig. 2B) explique mieux ces résultats qu'un modèle tel que celui de la Figure 2A. Cette hypothèse implique donc que non seulement les endosymbioses sont importantes dans l'évolution des eucaryotes, mais l'origine elle-même de ce type cellulaire aurait été présidée par un tel événement. Mais les implications iraient plus loin: serait-ce que la "trichotomie" archaea/bactéries/eucaryotes n'est pas adéquate et qu'il faut revenir, même en phylogénie, à considérer les archaea et bactéries comme deux grandes branches de procaryotes et les eucaryotes comme un assemblage de procaryotes? On revient drôlement près, finalement, des thèses de Margulis (1981)... Mais cette hypothèse "chimérique" ne résout pas tous les problèmes (Doolittle, 1995; Pace, 1997) et est loin de faire l'unanimité (Roger et Brown, 1996). Des analyses supplémentaires seront donc nécessaires pour clarifier cette portion fondamentale de l'histoire de la vie.

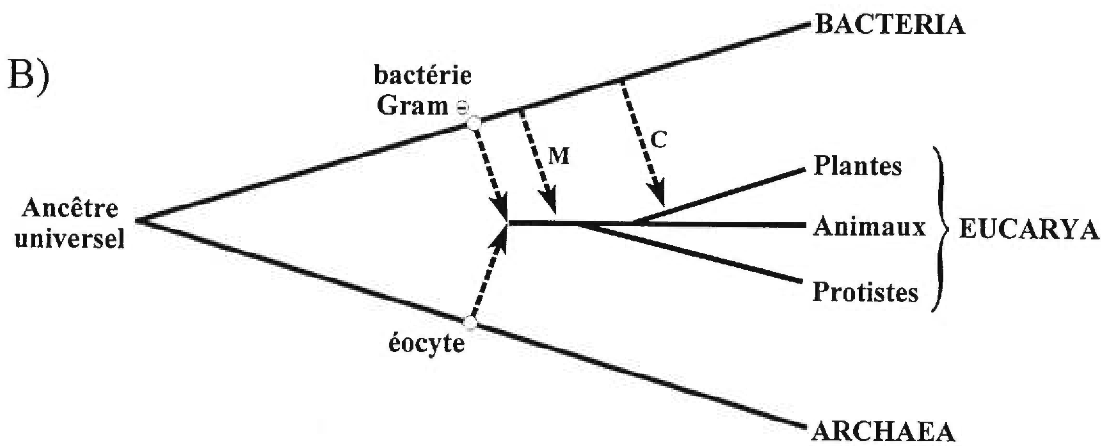
Figure 2. Deux hypothèses pour l'origine des eucaryotes.

A) Le modèle archaebactérien, selon lequel les eucaryotes ont simplement dérivé à partir d'une archaebactérie (éocyte). Cet arbre est équivalent à celui de la figure 1B.

B) Le modèle chimérique, selon lequel les eucaryotes ont pour origine une fusion entre une archaebactérie et une bactérie Gram-négative.

"M" et "C" représentent les événements d'endosymbiose ayant donné naissance aux mitochondries et aux chloroplastes, respectivement. (Figure adaptée de Gupta et Golding, 1996)


A)
BACTERIA
M
C
Ancêtre
universel
Plantes
Animaux
Protistes
EUCARYA
éocyte
ARCHAEA


B)
BACTERIA
bactérie
Gram ⊖
M
C
Ancêtre
universel
Plantes
Animaux
Protistes
EUCARYA
éocyte
ARCHAEA

2.3 ET LES AUTRES ORGANITES?

Une origine endosymbiotique a été proposée pour d'autres organites, mais aucun de ces cas n'est aussi convaincant que pour les mitochondries et les chloroplastes, en partie à cause du fait que ces deux organites sont les seules à avoir un génome, et donc à première vue à pouvoir se prêter à des analyses phylogénétiques.

Les peroxysomes, qui sont retrouvés chez tous les eucaryotes, ont pour rôle d'oxyder différents substrats, par exemple pour détoxiquer des substances (comme l'éthanol) ou pour dégrader des acides gras (β-oxydation). Certaines ressemblances avec les bactéries, dont le fait qu'ils se multiplient par croissance et division, ont fait des peroxysomes des candidats pour un scénario d'origine par endosymbiose. Les hydrogénosomes sont des organites retrouvés chez différents eucaryotes anaérobiques (Gray, 1992; Sogin, 1997). Leur rôle principal est la dégradation du pyruvate pour produire de l'ATP, de façon similaire à certaines bactéries anaérobiques. Ceci, en plus de leur double membrane et de leur mode de division, a mené à la suggestion que les hydrogénosomes ont pour ancêtre un endosymbiote eubactérien anaérobique. Cependant, chez les hydrogénosomes et les peroxysomes, l'absence de génome semble empêcher toute vérification de cette hypothèse; mais dans le cas des hydrogénosomes, quelques groupes ont contourné cet obstacle, et nous y reviendrons dans le cadre de la Discussion. Enfin, une origine endosymbiotique a également été proposée pour l'organite de motilité de certains eucaryotes, en se basant sur des similarités structurales avec des bactéries spirochètes (Margulis, 1981).

Les endosymbioses ont également eu lieu dans certains cas entre deux eucaryotes (Gray, 1992). Le cas le mieux connu est celui des algues cryptomonades. Chez *Cryptomonas* Φ, le plastide est enveloppé dans une troisième membrane, qui contient aussi une autre structure appelée "nucléomorphe", puisqu'il semble être le vestige du noyau de l'algue rouge qui aurait servi d'endosymbiote. Douglas et al. (1991) ont confirmé ces suppositions en séquençant les gènes d'ARNr du noyau et du nucléomorphe et en les plaçant dans un arbre phylogénétique.

Enfin, il est à noter que les endosymbioses ont été et sont toujours un événement courant chez les eucaryotes (Margulis, 1981). Les cas où elles ont abouti à un organite permanent ne sont que la pointe de l'iceberg, si on tient compte de toutes les endosymbioses "temporaires" qui ont eu lieu au cours de l'évolution. Il semble même que des gènes aient pu être transférés de ces endosymbiotes temporaires au noyau de l'hôte (Henze et al., 1995).

3. LA MITOCHONDRIE

3.1 QU'EST-CE QUE LA MITOCHONDRIE?

Dès les années 1850, Kölliker a décrit et étudié des granules dans le cytoplasme des cellules de muscle strié. D'autres ont ensuite découvert de telles structures dans différents types cellulaires, mais ce n'est qu'à partir des années 1890 que des techniques de coloration spécifique permirent une description détaillée de ces granules, entre autres dans l'ouvrage pionnier et visionnaire de Altmann, dont nous avons déjà parlé. Ces structures cytoplasmiques furent appelées *mitochondries* en 1898 par Benda, mais ce nom cohabitait alors avec une multitude d'autres appellations (blépharoblastes, chondriokontes, chondriomites, fila, Fädenkörner, et j'en passe!). Mais il fallut attendre les années 1910 pour qu'un rôle des mitochondries dans la respiration cellulaire fût proposé (Lehninger, 1964).

On sait maintenant que les mitochondries sont des organites à double membrane, dont le rôle est d'effectuer la *phosphorylation oxydative*, une série de réactions d'oxydation ayant pour but principal la synthèse d'ATP, forme d'énergie utilisable par la cellule. C'est dans l'intérieur de la mitochondrie (*matrice*) que se passe l'oxydation du pyruvate (produit de dégradation des glucides) et des acides gras en CO_2 (*cycle de Krebs* ou *de l'acide citrique*) et la réduction concomitante de transporteurs d'électrons, notamment le NAD. Les électrons sont transférés à de l'oxygène via une série de

complexes protéiques situés dans la membrane interne (la *chaîne respiratoire*), avec pour résultat l'expulsion de protons de la mitochondrie et l'établissement d'un gradient électrochimique. Enfin, ce gradient est utilisé pour la synthèse d'ATP, grâce à un complexe transmembranaire qui pompe les protons dans la mitochondrie et transforme l'ADP en ATP (Alberts et al., 1994).

3.2 DIVERSITÉ DES GÉNOMES MITOCHONDRIAUX

Plusieurs dizaines de polypeptides participent aux mécanismes très succinctement décrits ci-haut. Pourtant, le génome mitochondrial ne contient que quelques gènes codant pour des protéines. Nous avons vu à la section 2.1 que la mitochondrie a pour ancêtre un endosymbiote bactérien, dont presque tous les gènes ont été perdus, une grande quantité étant transférée au noyau. En plus de gènes d'ARNr et d'ARN de transfert (ARNt), le génome mitochondrial des vertébrés, une molécule circulaire de quelque 17 kb, ne code que pour 13 protéines, toutes essentielles pour les fonctions de la mitochondrie: sept sous-unités de la NADH déshydrogénase, l'apocytochrome b de l'ubiquinol cytochrome c réductase (*cob*), trois sous-unités de la cytochrome c oxydase (*cox1,2,3*) et deux sous-unités de l'ATP synthase (H^+-ATPase) (Gray, 1989). Les dizaines d'autres composantes de la machinerie de respiration mitochondriale sont synthétisées dans le cytoplasme (suite à la transcription de gènes nucléaires), avant d'être importées dans la mitochondrie. Puisque quelques gènes essentiels doivent tout de même être exprimés dans la mitochondrie, celle-ci a également besoin de tous les outils nécessaires pour la transcription, pour la maturation, l'épissage, l'édition des ARN, et pour la traduction. De plus, cette importation massive de protéine (et d'ARN) requiert une complexe machinerie de transport située dans les deux membranes, l'espace intermembranaire et la matrice. En tout, ce sont des centaines de protéines qui sont produites à partir de gènes nucléaires et dirigées vers la mitochondrie (Gray, 1992)!

Mais le petit ADN mitochondrial (ADNmt) des vertébrés est en fait très dérivé et pas du tout "typique". Alors ne soyons pas égocentriques et examinons donc d'autres

ADNmt. Lorsqu'on fait ainsi, on est vite frappé par la grande diversité des génomes mitochondriaux quant à leur taille, leur contenu en gènes, leur tempo et mode d'évolution et leurs régions régulatrices (Gray, 1989).

La taille des génomes mitochondriaux connus varie entre 6 kb (chez le protiste *Plasmodium falciparum*) et 2400 kb (chez la courge *Cucumis melo*)! Alors que l'ADNmt des animaux est petit (de 14 à 40 kb), avec des gènes organisés de façon très économique, les génomes mitochondriaux de plantes sont énormes, le plus petit comptant plus de 200 kb! Chez les champignons et les protistes, sauf exceptions, les génomes ont quelques dizaines de kb, et une taille assez variable d'une espèce à l'autre. Malgré ces divergences marquées pour la taille, le nombre de gènes dans les ADNmt ne varie pas énormément et se maintient d'ordinaire autour de quelques dizaines. Alors à quoi sont dues les différences de taille? Régions intergéniques non codantes, introns, séquences répétées et duplications de portions de génomes sont les principales sources de variation (Gray, 1989).

Si le nombre de gènes ne varie pas énormément, il en est autrement pour ce qui est de leur nature (Gray, 1989; Palmer, 1997b). Certains gènes reviennent souvent, mais seulement quatre ont été trouvés dans tous les ADNmt séquencés jusqu'ici: les deux grands ARNr, ainsi que *cox1* et *cob*. Ajoutez quelques protéines du métabolisme énergétique, assaisonnez (ou non) de protéines ribosomiques, une pincée ou un bouquet garni d'ARNt, et des cadres de lecture de fonction non identifiée; donc beaucoup de variété au menu… Ce qui est clair et constant, cependant, c'est que même les génomes mitochondriaux les plus riches en gènes n'arrivent pas à la cheville des génomes bactériens. Par exemple, *Mycoplasma genitalium* a un des plus petits génomes eubactériens connus, mais ce dernier contient néanmoins plus de 500 gènes (Fraser et al., 1995).

La vitesse et le mode d'évolution des ADNmt varie aussi beaucoup (Gray, 1989). On croit souvent que le taux de divergence de séquence de ces génomes est très rapide; c'est vrai chez les animaux, mais c'est loin de s'appliquer aux autres règnes. Encore

une fois, les plantes sont à l'opposé: le taux de substitution dans leur ADNmt est très faible. Cependant, les génomes mitochondriaux de plantes, et également ceux de champignons, subissent de nombreux et fréquents réarrangements, par recombinaison intramoléculaire impliquant des séquences répétées; leur introns (parfois même épissés en *trans* chez les plantes) contribuent également à la complexité de ces génomes. Les ADNmt des animaux, pour leur part, subissent peu de recombinaison et n'ont pas d'introns.

Enfin, les régions régulatrices, en particuliers les promoteurs mitochondriaux, ont des structures très différentes d'un organisme à l'autre, allant de petits promoteurs répartis dans tout le génome (levure, plantes), à de grandes régions avec un ou deux promoteurs pour chaque brin (animaux) (Tracy et Stern, 1995). Nous y reviendrons dans la prochaine section.

Les génomes mitochondriaux proprement dits ne constituent pas les seules molécules d'ADN des mitochondries. Chez les champignons et les plantes, on y retrouve des plasmides de différentes structures, linéaires ou circulaires, et dont le rôle exact n'a pas encore été défini (Meinhardt et al., 1990; Griffiths, 1995). En fait, très rares sont les cas où la présence de ces plasmides a été associée à un phénotype. Certains semblent impliqués dans des désordres dégénératifs comme la sénescence chez *Neurospora*. Plusieurs de ces plasmides contiennent des cadres de lecture codant généralement pour divers types de polymérases (Kempken et al., 1992; Griffiths, 1995). Une classe intéressante de tels plasmides mitochondriaux sont les plasmides linéaires avec protéines terminales, dont nous reparlerons plus loin.

3.3 CONCLUSION

Une bonne partie de cette thèse est consacrée à l'étude de l'évolution de ces organites pleins de surprises que sont les mitochondries, à travers leur machinerie de transcription et de traduction. Les deux sections suivantes donnent un aperçu de cette machinerie. Pourquoi étudier la transcription et la traduction à des fins évolutives?

Parce que ce sont des processus fondamentaux qui forment un véritable pivot dans les cellules, vu leur rôle dans l'expression de tous les gènes. Leur évolution est donc intimement liée à celle de la vie elle-même. On verra d'ailleurs dans la Discussion que les études menées au cours de mon doctorat, ainsi que les résultats obtenus, vont bien plus loin que la simple évolution des mitochondries.

4. LA TRANSCRIPTION DANS LES MITOCHONDRIES

4.1 LES ARN POLYMÉRASES À PLUSIEURS SOUS-UNITÉS

Avant de décrire les particularités de la transcription mitochondriale, il convient de jeter un coup d'œil sur ce qu'on retrouve dans d'autres systèmes. Les gènes de tous les êtres vivants sont transcrits par une ou des ARN polymérases à plusieurs sous-unités (msRNAP, pour *multi-subunit RNA polymerase*), dont certaines sous-unités sont très conservées entre les organismes. Chez *Escherichia coli*, l'ARN polymérase est constituée de deux grandes sous-unités appelées β' (la plus grande) et β, et de deux petites sous-unités α (Palenik, 1992). Les deux grandes sous-unités se partagent les rôles de liaison de l'ADN, liaison des nucléotides et formation des liens phosphodiesters (polymérisation). Les sous-unités α, quant à elles, se sont révélées être importantes pour la réponse à divers activateurs de la transcription, alors que pendant longtemps on avait cru que leur seul rôle était dans l'assemblage de l'enzyme (Russo et Silhavy, 1992). Ce *noyau* (*core*) de polymérase $\alpha_2\beta\beta'$ ne peut reconnaître par lui-même les promoteurs. Il doit pour cela s'associer à un facteur σ, pour former l'*holoenzyme* (Helmann et Chamberlin, 1988). Cet holoenzyme se lie aux éléments -35 et -10 (par rapport au site d'initiation) et initie la transcription. Le facteur σ est ensuite rapidement relâché, le noyau de l'enzyme entrant alors en mode d'élongation.

Les sous-unités β' et β ont des homologues chez tous les organismes, de même que dans les chloroplastes (Rowland et Glass, 1990). Elles sont parfois coupées en

deux, comme dans le cas de la β' de cyanobactéries et chloroplastes (Gray, 1992), ou l'équivalent de β' et β chez les archaebactéries (β' est scindée chez toutes les archaea; β seulement chez les archaea méthanogénes et halophiles; Zillig et al., 1993). Par ailleurs, certaines sous-unités ou facteurs de transcription eucaryotes ont révélé des régions de similarité avec α et σ (Jaehning, 1991; Archambault et Friesen, 1993). En fait, chez les eucaryotes, on n'a pas une mais trois msRNAP (I, II et III), ayant chacune leurs catégories spécifiques de gènes à transcrire. De plus, alors que chez les eubactéries cinq polypeptides sont suffisant pour amorcer la transcription, le même processus chez les eucaryotes est incroyablement plus compliqué, requérant, en plus de la polymérase, plusieurs facteurs de transcription de base (Roeder, 1996). Il est d'ailleurs apparu évident au cours des dernières années que la machinerie de transcription des archaebactéries est bien plus semblable à celle des eucaryotes qu'à celle des eubactéries: nous avons déjà parlé de TBP (section 1.4), mais la comparaison peut s'étendre à plusieurs sous-unités de msRNAP et facteurs d'initiation et d'élongation de la transcription (Klenk et Doolittle, 1994; Langer et al., 1995).

Les caractéristiques des sous-unités β' et β (distribution, conservation) en font *a priori* de très bons marqueurs phylogénétiques (voir section 1.2.1; Klenk et Zillig, 1994), et leur utilisation à ces fins a été très populaire au début de la décennie. Lorsque des arbres phylogénétiques sont construits avec la séquence de β' (Zillig et al., 1993) ou β (N. Cermakian, non publié), on obtient souvent une topologie étrange, où les trois RNAP eucaryotes ne sont pas regroupées (et donc apparemment polyphylétiques): II et III sont proches des archaebactéries et I des eubactéries. Cette topologie, plus d'autres observations, ont mené Zillig et ses collaborateurs à présenter leur hypothèse de l'origine des eucaryotes par fusion de deux procaryotes (Zillig et al., 1992; section 2.2). Cependant, Iwabe et al. (1991) ont publié une étude phylogénétique très complète avec les mêmes marqueurs, et ont ainsi mis en doute les résultats de Zillig: selon Iwabe et al., les trois polymérases eucaryotes sont bel et bien monophylétiques, et les résultats de l'autre groupe pourraient être un artefact de l'analyse par parcimonie, dû à une vitesse d'évolution très grande de la RNAP I. Il se pourrait donc que l'hypothèse de Zillig se soit basée en grande partie sur des prémisses erronées, mais fut plus tard

reprise et appuyée solidement par Gupta et Golding (1996). La Science emprunte parfois des chemins bien tortueux...

4.2 LES ARN POLYMÉRASES À UNE SOUS-UNITÉ

Plusieurs virus à ADN utilisent la RNAP cellulaire pour leur transcription (Watson et al., 1987). D'autres ont leur propre msRNAP, tel le virus de la vaccine (Bryoles et Moss, 1986). Mais un petit groupe de bactériophages, comprenant T7, T3 et SP6, sont plus originaux: ils transcrivent la majorité de leurs gènes avec une petite RNAP ne comportant qu'une sous-unité d'environ 100 kDa. Ces RNAP à une sous-unité (ssRNAP, pour *single-subunit RNAP*) reconnaissent un promoteur constitué d'une vingtaine de nucléotides, situé en amont des gènes tardifs du phage et chevauchant le site d'initiation (McGraw et al., 1985). Notons que le gène de RNAP de ces phages est transcrit par la RNAP bactérienne.

Ces ssRNAP n'ont pas de similarité de séquence avec les msRNAP. Il semble plutôt qu'elles soient apparentées à d'autres polymérases. Delarue et al. (1990) ont en effet montré que l'ARN polymérase du phage T3 partage trois petits motifs avec plusieurs ADN polymérases (DNAP) cellulaires et virales, ainsi qu'avec des transcriptases inverses et des RNAP dépendantes de l'ARN, et l'importance fonctionnelle de ces motifs a été démontrée par des expériences de mutagenèse (Bonner et al., 1992; Osumi-Davis et al., 1992). Cette similarité entre ssRNAP et DNAP a été appuyée lorsque la structure tridimensionnelle de la RNAP de T7 a été élucidée et comparée à celle du fragment Klenow de l'ADN polymérase I de *E. coli* : le repliement et la position de la plupart des éléments de structure secondaire de ces protéines sont presque identiques (Sousa et al., 1993; Sousa, 1996).

4.3 DES ARN POLYMÉRASES À UNE SOUS-UNITÉ DANS LES MITOCHONDRIES

Comme on l'a vu plus haut, les chloroplastes ont un système transcriptionnel très semblable à celui retrouvé chez les cyanobactéries, ce qui est en accord avec leur

origine endosymbiotique. Pour les mêmes raisons, on aurait donc pu s'attendre à trouver dans les mitochondries une RNAP de type bactérien. Or, cela ne semble pas être le cas. Le clonage et séquençage du gène nucléaire codant pour l'ARN polymérase mitochondriale (mtRNAP) de la levure *Saccharomyces cerevisiae* a causé une grande surprise: cette enzyme est homologue à celle des bactériophages T3 et T7 (Masters et al., 1987). La similarité s'étend à plusieurs grands blocs de séquence, où l'identité avec les séquences virales va de 49 à 77% (au total, les protéines ont 28% d'identité). La deuxième séquence de mtRNAP connue fut celle de la moisissure du pain *Neurospora crassa* (Chen et al., 1996). Puis, des séquences partielles de ssRNAP humaines, de nématode et de riz sont apparues dans les bases de données. Comme on le verra plus loin (Chapitre 1), ces découvertes, comme bien souvent en science, amènent dans leur sillage bien plus de questions que de réponses quant à l'évolution de ce type d'enzyme et les mécanismes de transcription dans les mitochondries. De plus, des gènes de ssRNAP ont également été trouvés sur une multitude de plasmides linéaires mitochondriaux de champignons et de plantes (section 3.2; Kempken et al., 1992). Bien que l'expression de quelques-uns de ces gènes plasmidiques ait été démontrée, le rôle de ces ssRNAP n'est pas connu.

4.4 De la compagnie pour les ARN polymérases mitochondriales

Une différence importante entre les ssRNAP de phages et les mtRNAP est que celles-ci ont besoin d'autres facteurs, eux aussi codés par des gènes nucléaires, pour mener à bien l'initiation de la transcription des gènes mitochondriaux (Tableau 1). C'est chez la levure et les mammifères que la transcription mitochondriale a été la mieux étudiée.

4.4.1 Chez la levure

Les deux brins de l'ADNmt de *S. cerevisiae* peuvent servir de matrice pour la transcription, et on retrouve des promoteurs pour la sc-mtRNAP ("sc" pour *S. cerevisiae*) à différents endroits de la molécule (Tracy et Stern, 1995). Les promoteurs présentent la séquence de 9 nucléotides 5'-ATATAAGTA(+1)-3'; cette séquence

Tableau 1. ARN polymérase et facteurs de transcription connus chez les mitochondries de quelques organismes, au moment d'entreprendre les études des Chapitres 1 et 2 [1]

Nom	*S. cerevisiae*	*N. crassa*	*H. sapiens*	*X. laevis*	Plantes
MtRNAP	ssRNAP [2]	ssRNAP	ssRNAP?	ssRNAP?	ssRNAP?
MtTFA (facteur à deux domaines HMG)	Oui (surtout un rôle structural)	?	Oui (surtout un rôle d'activateur)	Oui	?
MtTFB (facteur de spécificité)	Oui (légère ressemblance avec σ)	?	?	Oui	?

[1] D'autres facteurs d'initiation de la transcription pourraient être découverts dans le futur.

[2] Par "ssRNAP", on entend ici spécifiquement les enzymes homologues à la RNAP du bactériophage T7.

ressemble à l'élément -10 bactérien ou à la boîte TATA eucaryote, mais sa position par rapport au site d'initiation rappelle plutôt le cas du promoteur T7 (Jaehning, 1993; Tracy et Stern, 1995).

La sc-mtRNAP ne peut pas reconnaître ces promoteurs et y initier la transcription toute seule. Ces étapes nécessitent la participation d'un facteur appelé MTF1 (*mitochondrial transcription factor 1*) ou sc-mtTFB (S. cerevisiae *mitochondrial transcription factor B*). Il s'agit d'une protéine d'environ 40kDa, dont le gène a au départ été isolé génétiquement par Lisowsky et Michaelis (1988), comme suppresseur à haut nombre de copies d'un mutant de la sc-mtRNAP. Jang et Jaehning (1991) ont isolé une protéine conférant une spécificité d'initiation à la sc-mtRNAP; le microséquençage d'une partie de la protéine a permis de voir que cette protéine est celle codée par le gène cloné par Lisowsky et Michaelis (1988). Jang et Jaehning (1991) ont fait remarquer que sc-mtTFB présente une certaine similarité de séquence avec les facteurs de spécificité σ des eubactéries, mais cette similarité a été mise en doute par des expériences de mutagenèse (Shadel et Clayton, 1995) et surtout par l'analyse de la séquence de mtTFB d'autres levures (Carrodeguas et al., 1996). Par ailleurs, Xu et Clayton (1992) ont montré que sc-mtTFB était le seul facteur de spécificité de la sc-mtRNAP, c'est-à-dire que ces deux protéines sont suffisantes et nécessaires pour une initiation précise aux promoteurs mitochondriaux de levure. De plus, Mangus et al. (1994) ont démontré que la polymérase et le facteur ne lient efficacement l'ADN que lorsqu'ils sont ensemble; puis cet holoenzyme initie la transcription au promoteur et après quelques nucléotides, mtTFB est relâché. Ce mécanisme présente donc plusieurs points en commun avec le système bactérien (voir section 4.1).

Un autre facteur, auparavant appelé ABF2, HM ou p19/HM, mais maintenant plutôt sc-mtTFA (Xu et Clayton, 1992), permet d'atteindre un niveau de transcription plus élevé et est nécessaire au maintien de l'ADN mitochondrial (Parisi et al., 1993), mais ne sert pas de facteur de spécificité alternatif (Xu et Clayton, 1992). Il s'agit d'une petite protéine (19 kDa), contenant principalement deux boîtes HMG, domaines

compacts et basiques partagés par toute une famille de protéine liant l'ADN, avec des rôles variés (Grosschedl et al., 1994). Il semble que plusieurs protéines à domaines HMG courbent l'ADN; c'est le cas de sc-mtTFA, qui aurait entre autres rôles celui d'"empaquetage" de l'ADN mitochondrial, un peu comme les histones dans le noyau (d'où sa présence obligatoire pour le maintien de l'ADN mitochondrial).

4.4.2 Chez les mammifères

La transcription des gènes mitochondriaux des mammifères (ainsi que des autres animaux) se fait à partir d'une région régulatrice appelée *D-loop* (pour *displacement loop*), qui contient les promoteurs pour chacun des brins (LSP et HSP, pour *Light* et *Heavy Strand Promoter*) et l'origine de réplication du HS; l'origine pour le LS est ailleurs dans l'ADN mitochondrial (Clayton, 1991; Shadel et Clayton, 1993; Tracy et Stern, 1995). En fait, un point commun frappant entre ce système et celui de la mitochondrie de levure ou des bactériophages du type T7, est que dans tous ces cas, le transcrit naissant peut servir d'amorce pour la réplication (Masters et al., 1987).

Une fraction protéique mitochondriale capable de transcrire à partir d'un promoteur mitochondrial humain a été séparée en une fraction ayant une activité transcriptionnelle non spécifique, et une autre conférant la spécificité à la première lorsque les deux sont réunies (Fisher et Clayton, 1985). Cette dernière fraction se lie au promoteur et y laisse une empreinte dans des tests de protection contre la digestion à la DNase I (Fisher et al., 1987). Les auteurs ont donc pu définir, par constructions de LSP et HSP mutants, les régions importantes de ces promoteurs pour l'initiation de la transcription et la liaison du facteur de spécificité. Ce dernier, nommé mtTF1, puis plus tard h-mtTFA (*human mitochondrial transcription factor A*), a été purifié (Fisher et Clayton, 1988) puis cloné et séquencé (Parisi et Clayton, 1991). Il s'agit d'une protéine de 25 kDa, qui se lie plus fortement au LSP qu'au HSP. Elle lie également l'ADN de façon non spécifique mais avec moins d'affinité, et cette propriété servirait pour remplir un rôle structurel, comme sc-mtTFA, mais à un niveau beaucoup moins important. En fait, h-mtTFA et sc-mtTFA semblent homologues, ou plutôt partagent

les caractéristiques d'avoir deux boîtes HMG et de courber l'ADN. D'ailleurs, h-mtTFA peut au moins en partie remplacer fonctionnellement sc-mtTFA lorsque exprimé dans une souche de levure avec le gène de sc-mtTFA muté (Parisi et al., 1993). Cependant, h-mtTFA est un puissant activateur de la transcription mitochondriale, ce qui semble être son rôle principal; à ce sujet, il est intéressant de noter que la protéine humaine possède deux régions que celle de levure n'a pas, entre les 2 boîtes HMG et à l'extrémité C-terminale, et que ces deux régions sont importantes pour sa fonction de transactivation (Dairaghi et al., 1995).

Mais h-mtTFA est-il vraiment le facteur de spécificité (rôle joué par mtTFB chez la levure)? Bien que les articles publiés sur ce facteur n'aient jamais été clairs à ce sujet, un examen plus approfondi de leurs résultats suggère que si ce facteur est nécessaire pour avoir des niveaux élevés de transcription, il ne l'est peut-être pas pour avoir un niveau de base. Il est à noter que la spécificité d'espèce pour la transcription, entre l'homme et la souris par exemple, n'est pas imputable à mtTFA, puisque les facteurs humains et murin peuvent être interchangés en conservant la capacité de transcrire à partir des promoteurs (Fisher et al., 1989). Cette spécificité d'espèce, et probablement aussi la spécificité vis-à-vis du promoteur de base, réside donc dans la fraction *polymérase*, dont la purification n'a pu être menée plus loin jusqu'à maintenant. Aucun homologue de sc-mtTFB n'a été découvert chez l'humain jusqu'ici; soit il n'y en a pas, soit ce facteur est dans la fraction *polymérase*.

Le cas d'autres animaux comme l'amphibien *Xenopus laevis* peut ici nous éclairer. *X. laevis* possède une polymérase de 140 kDa (taille voisine à celle de la sc-mtRNAP) (Bogenhagen et Insdorf, 1988), et deux facteurs appelés par analogie xl-mtTFA et xl-mtTFB ("xl" pour *X. leavis*; Antoshechkin et Bogenhagen, 1995). Xl-mtTFA est un activateur transcriptionnel, mais moins fort que h-mtTFA; d'ailleurs, la séquence de ce facteur amphibien est très semblable à celle de h-mtTFA (voir Chapitre 2). Xl-mtTFB est pour sa part le facteur de spécificité étant nécessaire pour l'initiation de la transcription; il n'a pas encore été séquencé, mais il s'agit d'une protéine de taille presque identique à sc-mtTFB (Bogenhagen, 1996). Ces résultats suggèrent que les

machineries de transcription mitochondriales animales sont très semblables à celle de la levure. Notons par ailleurs que les promoteurs de *X. laevis* et du poulet ont une structure et une complexité intermédiaires entre les promoteurs des mammifères et de la levure (Tracy et Stern, 1995).

5. LA TRADUCTION DANS LES MITOCHONDRIES

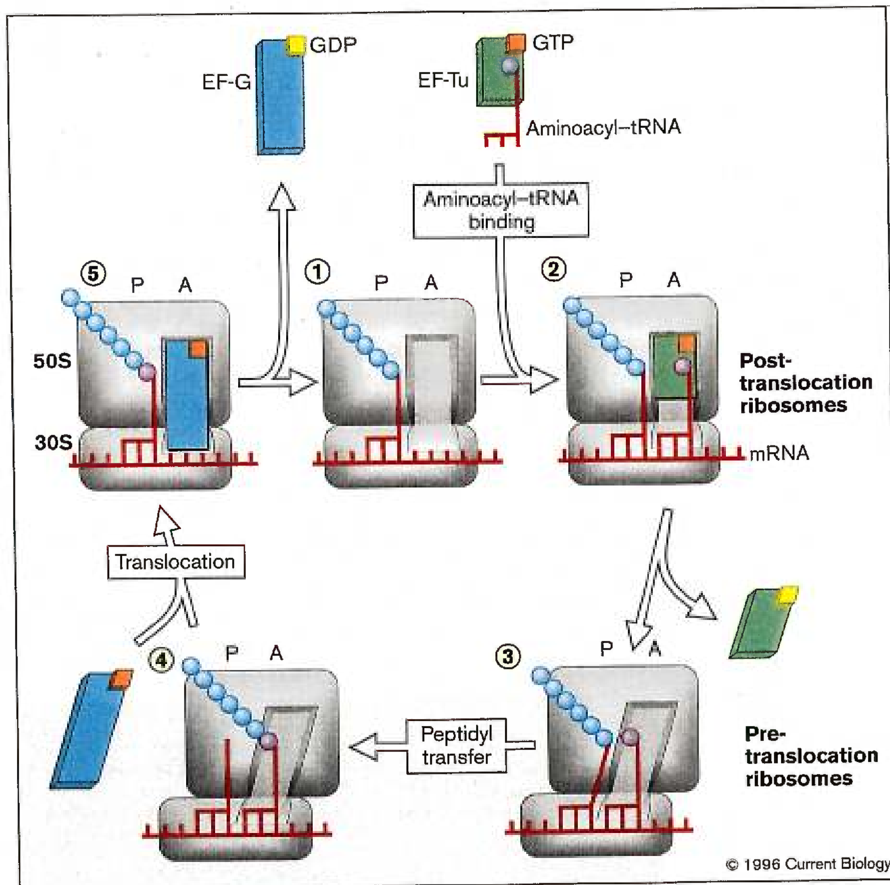
5.1 LA MACHINERIE DE TRADUCTION ET SES ORIGINES

La traduction est un processus par lequel un ARN messager (ARNm) est lu et décodé, en vue de produire un polypeptide dont la séquence d'acides aminés correspond à la séquence de nucléotides de l'ARNm (Alberts et al., 1994). La correspondance entre les deux "alphabets" se fait grâce aux ARNt, qui jouent le rôle d'adaptateurs entre des groupes de trois nucléotides (codons) dans l'ARNm et les acides aminés, selon une clef appelée *code génétique*. Ces ARNt sont *aminoacylés* ou *chargés* avec le bon acide aminé par des protéines appelées aminoacyl-ARNt synthétases. Les ARNt chargés sont ensuite pris en charge par un facteur d'élongation appelé EF-Tu chez les eubactéries (Figure 3). EF-Tu apporte l'ARNt aminoacylé au bon site sur le ribosome, qui est une énorme particule ribonucléoprotéique composée de deux grands ARN (plus un ou deux plus petits) et de dizaines de protéines. C'est sur le ribosome qu'a lieu la lecture de l'ARNm, le décodage de sa séquence et la synthèse protéique proprement dite. D'autres facteurs de traduction sont impliqués dans l'initiation, l'élongation et la terminaison de la traduction. Entre autres, notons EF-G, qui permet la translocation de l'ARNt d'un site à l'autre sur le ribosome (Figure 3).

Les ARNr et certains facteurs de traduction sont présents chez tous les organismes et sont très conservés, ce qui en fait de bons marqueurs moléculaires pour établir des phylogénies de tous les êtres vivants (voir section 1). Comme dans le cas de certaines sous-unités des ARN polymérases (section 4), on peut conclure que la plupart des composantes de la machinerie de traduction étaient présentes dans l'ancêtre commun

Figure 3. Le cycle d'élongation de la traduction.

Le facteur d'élongation Tu (EF-Tu)/GTP lie l'ARN de transfert (ARNt) aminoacylé et l'amène au site A vacant du ribosome (1,2). EF-Tu hydrolyse son GTP, le ribosome change de conformation et le complexe EF-Tu/GDP s'en dissocie (3). Puis a lieu la réaction de transfert du peptide de l'ARNt au site P à celui au site A (4), suivi de la translocation du nouveau peptidyl-ARNt vers le site P, assistée par un autre changement conformationnel du ribosome et l'arrivée du facteur d'élongation G (EF-G)/GTP au site A. Le cycle est refermé par l'hydrolyse du GTP de EF-G, qui libère alors le site A (1). (Figure tirée de Liljas, 1996)

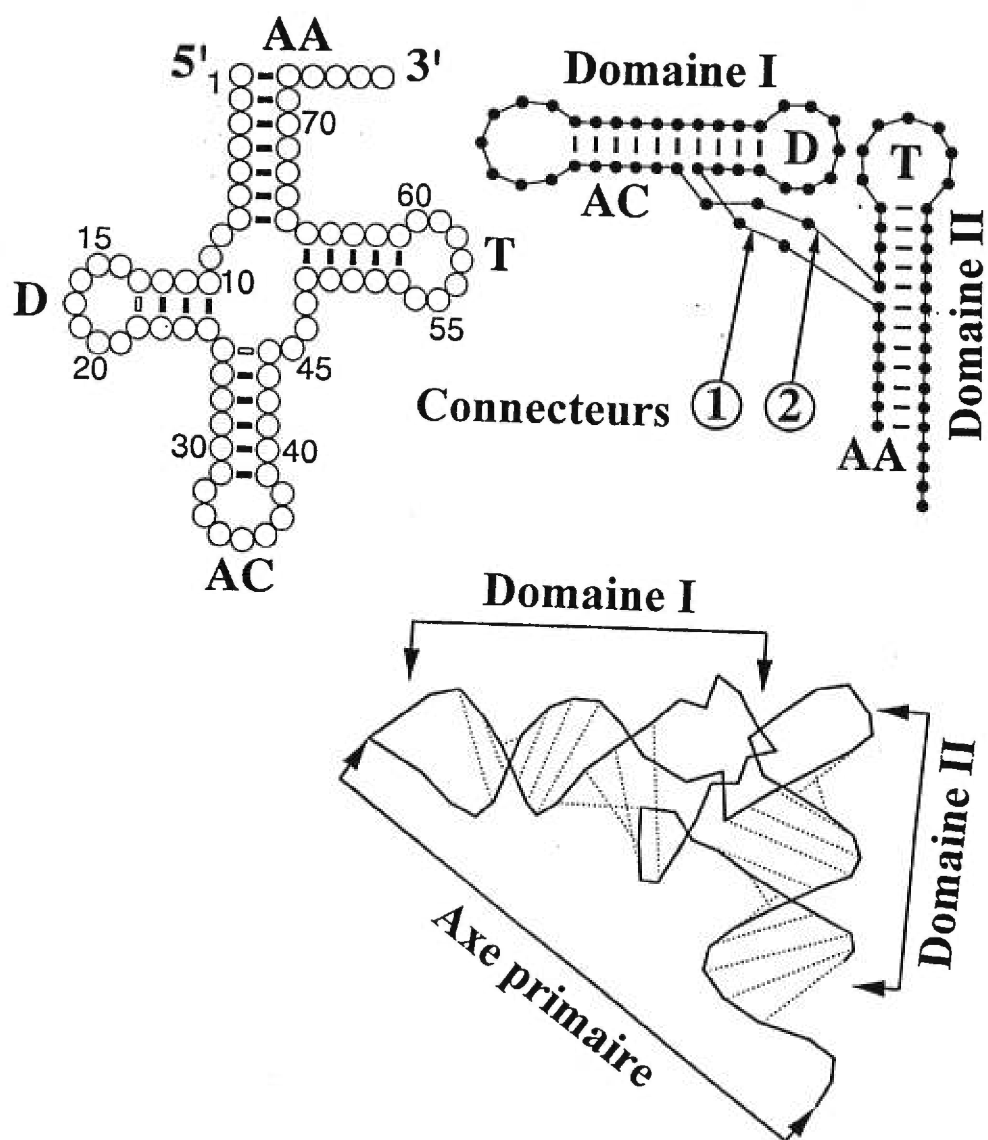

GDP
EF-G
GTP
EF-Tu
Aminoacyl–tRNA
Aminoacyl–tRNA binding
5
P
A
1
P
A
2
P
A
50S
30S
Post-translocation ribosomes
mRNA
Translocation
4
P
A
3
P
A
Peptidyl transfer
Pre-translocation ribosomes
© 1996 Current Biology

de tous les être vivants. Cousineau et al. (1997) ont récemment proposé un modèle pour l'origine de tous les facteurs de traduction via une série de duplications-fusions d'un même petit domaine ancestral. Plusieurs protéines ribosomiques pourraient même provenir du même domaine ancestral, ce qui unirait la plupart des protéines impliquées dans la traduction dans une même superfamille (Cousineau et al., 1997). Par ailleurs, les ARNr doivent remonter aux tout débuts de la synthèse protéique, au temps où les ribosomes n'étaient peut-être formés que d'ARN (Cedergren et Grosjean, 1987; Noller et al., 1992). Des expériences suggérant que l'activité peptidyl transférase pourrait être l'apanage de l'ARNr de la grande sous-unité du ribosome ont d'ailleurs été rapportées (Noller et al., 1992).

Les ARNt sont des molécules d'environ 76 nucléotides qui peuvent être repliées en une structure secondaire dite “en feuilles de trèfle” (Steinberg et Cedergren, 1994; Dirheimer et al., 1995; Figure 4). On reconnaît quatre parties dans cette structure: la tige acceptrice, qui reçoit l'acide aminé à son extrémité 3' (AA); la tige-boucle D; la tige-boucle de l'anticodon (AC), qui comme son nom l'indique comprend l'anticodon, complémentaire au codon correspondant à l'acide aminé chargé; et la tige-boucle T (ou TΨC). Comme on le voit à la Figure 4, grâce à de multiples interactions tertiaires, l'ARNt adopte en trois dimensions une structure en forme de L, dans laquelle la tige de l'anticodon et la tige D sont colinéaires, tout comme la tige acceptrice et la tige T. On a donc en fait deux domaines agencés à peu près à angle droit. Ces deux domaines sont séparés par deux connecteurs (voir Figure 4); le connecteur 2 est aussi désigné sous le nom de “boucle variable”, puisque sa longueur peut varier. Il a été proposé que les premiers ARNt étaient beaucoup plus simples, consistant en une seule minihélice correspondant au domaine II de la Figure 4 (Schimmel et Ribas de Pouplana, 1995). Il a en effet été montré qu'une telle minihélice peut être aminoacylée et être reconnue par EF-Tu *in vitro*. L'autre domaine se serait ajouté par la suite, d'abord en *trans*, puis en *cis* par fusion des deux gènes (Schimmel et Ribas de Pouplana, 1995). Il a par ailleurs été proposé que les ARNt sont tout d'abord apparus pour remplir un rôle dans la réplication, et ont par la suite été cooptés pour servir dans la synthèse protéique (Maizels et Weiner, 1994). Un autre point encore flou sur l'évolution de la synthèse

Figure 4. Structure secondaire et tertiaire des ARN de transfert.

En haut à gauche: structure secondaire dite “en feuilles de trèfle”.
En bas: structure tertiaire en L (obtenue par cristallographie).
En haut à droite: visualisation en deux dimensions de la structure tertiaire en L.
Les différentes régions, domaines, connecteurs sont indiqués.
(Figure adaptée de Steinberg et Cedergren, 1994)


AA
5'
3'
1
70
60
15
T
10
D
55
20
45
30
40
AC
Domaine I
D
T
AC
Domaine II
Connecteurs ① ②
AA
Domaine I
Domaine II
Axe primaire

protéique et des ARNt est l'origine du code génétique: comment les codons ont-ils été assignés à leurs acides aminés particuliers? Deux hypothèses s'affrontent (Osawa et al., 1992): l'hypothèse de *l'accident gelé* (*frozen accident*), selon laquelle l'établissement du code serait dû à des événements aléatoires, et que le code serait resté le même par la suite parce que tout changement serait létal (Crick, 1968); et l'hypothèse stéréochimique, qui promeut l'idée que des complémentarités structurelles entre acides aminés et codons ont été à la base du code génétique, et donc que celui-ci ne découle pas d'événements aléatoires (Woese, 1967; Cedergren et Miramontes, 1996). Il est en fait plutôt probable que rien n'est noir ou blanc dans cette histoire et que les deux hypothèses sont partiellement adéquates: certaines assignations de codons ont une origine stéréochimique, tandis que d'autres sont le fruit du hasard.

Quoique les mitochondries et les eubactéries aient certains points en commun au niveau de la traduction (initiation avec N-formylméthionine, sensibilité à certains antibiotiques, facteurs de traduction semblables), les particularités du système mitochondrial par rapport à l'ancêtre bactérien sont nombreuses (Gray, 1992). Nous allons survoler les plus frappantes.

5.2 QUELQUES PARTICULARITÉS DE LA TRADUCTION DANS LES MITOCHONDRIES

5.2.1 ARN ribosomiques minuscules ou en morceaux

La taille des ARNr mitochondriaux varie énormément d'un organisme à l'autre. Tandis que les ARNr des mitochondries de plantes et de champignons ont une taille semblable aux ARNr bactériens, ceux des animaux sont environ deux fois plus petits; le record appartient toutefois à certains protistes, par exemple *Trypanosoma brucei*, dont les ARNr mitochondriaux dépassent à peine le tiers de la taille de ceux des bactéries (Benne et Sloof, 1987)! Il faut cependant noter que ce qui est retenu dans ces minuscules ARNr correspond aux régions bien conservées parmi tous les ARNr, et que ces ARNr se replient de façon semblable à des ARNr de grandeur normale (Benne et Sloof, 1987). Une autre surprise à survenir au cours de l'étude des ARNr

mitochondriaux fut la découverte par Boer et Gray (1988) d'ARNr en morceaux dans les mitochondries de l'algue verte unicellulaire *Chlamydomonas reinhardtii.* La séquence de ces petits "modules" d'ARNr suggère qu'ils peuvent s'associer pour former une structure secondaire "normale" (Boer et Gray, 1988). Des ARNr en morceaux ont aussi été trouvés dans les cytoplasmes (Dover, 1988; Schnare et Gray, 1990), mais le cas des mitochondries est jusqu'ici unique, puisque les gènes codant pour les modules sont dispersés parmi d'autres gènes dans l'ADNmt (Boer et Gray, 1988). Une situation similaire est d'ailleurs retrouvée chez *P. falciparum* (Wilson et Williamson, 1997). D'autres particularités des mitoribosomes doivent être notées. Premièrement, alors que plusieurs protéines ribosomales ont des homologues bactériens, certaines sont uniques aux mitochondries et d'origine inconnue (Benne et Sloof, 1987; Gray, 1992). Par ailleurs, il ne semble pas y avoir d'interaction de type Shine-Dalgarno entre l'ARNr de la petite sous-unité et l'ARNm dans les mitochondries (Benne et Sloof, 1987; Palmer, 1997b).

5.2.2 Codes génétiques déviants et interactions codon-anticodon relâchées

Pendant plusieurs années, le code génétique a semblé être vraiment universel, ce qui supportait la proposition qu'il était "gelé" dans cet état. Mais ceci a été remis en cause par la découverte de codes génétiques alternatifs chez différents organismes (Barrell et al., 1979; Osawa et al., 1992; Watanabe et Osawa, 1995). Les variations dans le code sont très répandues dans les mitochondries (sauf chez les plantes). La modification la plus répandue est l'utilisation du codon UGA, normalement un codon d'arrêt, pour coder un tryptophane, mais plusieurs autres sont observées (Watanabe et Osawa, 1995). Il a été proposé que ces réorganisations du code dans les mitochondries sont un stratagème pour faciliter la réduction du nombre d'ARNt nécessaires (Kurland, 1992). Ce nombre est en effet très petit dans beaucoup de mitochondries, le record (et minimum possible) appartenant aux animaux: seulement 22 ARNt différents (Dirheimer et al., 1995)! L'autre "stratagème" permettant une si petite batterie d'ARNt est l'utilisation de règles de pairage codon-anticodon plus relâchées que les règles de

wobble qui prévalent ailleurs (Benne et Sloof, 1987; Kurland, 1992; Osawa et al., 1992).

5.2.3 ARN de transfert atypiques

Certains ARNt mitochondriaux d'animaux, de champignons et de protistes présentent une structure secondaire tout à fait atypique, avec une tige-boucle D ou T raccourcie ou inexistante (Dirheimer et al., 1995; Steinberg et al., 1997). Des membres de notre laboratoire ont réexaminé la structure de ces ARNt anormaux, pour se rendre compte que leur structure secondaire pouvait être réarrangée dans le but d'avoir un meilleur appariement (Steinberg et Cedergren, 1994): il suffit de supposer une tige de l'anticodon avec 5, 7, 8, 9 ou 10 paires de bases (alors que la normale est de 6). Cependant, afin de conserver une structure tertiaire normale, c'est-à-dire avec le même arrangement des deux domaines et un axe primaire de longueur adéquate (voir Figure 4), il faut inclure dans ces structures secondaires des compensations structurelles. Premièrement, la longueur de la tige D varie de façon à peu près inversement proportionnelle à celle de la tige de l'anticodon un peu comme une fermeture-éclair; deuxièmement, la longueur des connecteurs entre les deux domaines varie selon des règles simples (Steinberg et Cedergren, 1994; Steinberg et al., 1997). Ce qui est frappant, c'est que ce type de structure déviante n'est pas trouvé dans les ARNt cytosoliques. En fait, les ARNt non mitochondriaux ayant la possibilité de former de telles structures alternatives semblent posséder un ou des nucléotides modifiés incompatibles avec ces nouveaux appariements (Steinberg et Cedergren, 1995).

Un autre dilemme est le nucléotide 47, auquel aucun rôle n'a encore été assigné. L'examen attentif de base de données d'ARNt a permis de mettre en lumière une corrélation intéressante (Steinberg et Ioudovitch, 1996). Lorsque ce nucléotide (nt) de la boucle variable est absent, la paire de bases 13-22 dans la tige D, habituellement une paire Watson-Crick C-G, présente plutôt une interaction wobble U-G. Seulement quelques ARNt ont à la fois une paire C13-G22 et une absence de nt 47, et on sait

qu'au moins trois de ces ARNt sont inactifs ou ont seulement des rôles hors de la synthèse protéique. Cette corrélation entre le nt 47 et la paire 13-22 peut sembler étrange, mais elle s'explique facilement quand on réalise que ces deux régions de l'ARNt sont en contact étroit dans la structure tridimensionnelle (Figure 4). En fait, Steinberg et Ioudovitch (1996) ont tenté de modéliser un ARNt sans nt 47 et avec une paire C13-G22, et ont montré qu'il était impossible de conserver cette interaction tertiaire, puisque les deux régions de l'ARNt entraient en collision dans ce modèle (ce qui n'est pas le cas dans les situations satisfaisant la corrélation). Encore une fois, la mitochondrie se distingue, puisque seuls ses ARNt ne satisfont pas cette corrélation: la moitié des ARNt mitochondriaux analysés vont à l'encontre de celle-ci (Steinberg et Ioudovitch, 1996)! Cette interaction tertiaire n'est-elle pas importante? Mais alors qu'est-ce qui aurait permis l'établissement de la corrélation observée? Ou alors l'interaction tertiaire connecteur 2/tige D n'est-elle pas importante dans les mitochondries?

5.2.4 Conclusion

En conclusion, les particularités de l'appareil de traduction mitochondrial ont certains points en commun qui nous informent sur le mode d'évolution des mitochondries:

• Toutes ces modifications semblent concourir à un but commun, soit la réduction du génome mitochondrial: réduction de la taille des ARNr, réduction du nombre d'ARNt nécessaires, élimination de certaines portions d'ARNt. Même le cas du nucléotide 47 peut partiellement s'appliquer: seulement 35% des ARNt mitochondriaux examinés par Steinberg et Ioudovitch (1996) ont un tel nucléotide, contre 85% dans le cas des ARNt cytosoliques! Mais d'autres forces doivent être en jeu puisque certaines de ces modifications sont également trouvées, quoique à une faible fréquence, dans des mitochondries avec des génomes assez grands.

• Ces modifications ont eu lieu indépendamment dans différentes lignées, et semblent donc être des événements plutôt probables, ce qui indique que la force de réduction du génome est puissante.

• Ces modifications des ARN risquent d'entraîner un relâchement de la traduction, une dégradation de son efficacité et de son exactitude; un tel relâchement est peut-être toléré à cause du petit nombre de protéines à produire dans la mitochondrie, la plupart des gènes ayant été transférés au noyau.

6. SUJETS ET QUESTIONS ABORDÉS DANS CETTE THÈSE

Les mêmes questions fondamentales ont été au centre des différentes études que j'ai menées au cours des cinq dernières années: quelle fut l'histoire évolutive des machineries de transcription et de traduction? Quels mécanismes peuvent expliquer la situation actuelle de ces systèmes? Il est bien vite devenu apparent que la mitochondrie était un des systèmes qui offrait le plus de surprises, tant au niveau de la transcription que de la traduction. En guise de préambule aux chapitres qui vont suivre, je vais ici exposer les quelques sujets abordés et les questions qui y sont rattachées.

6.1 L'ÉVOLUTION DES ARN POLYMÉRASES À UNE SOUS-UNITÉ (CHAPITRE 1)

Nous avons vu que l'ARN polymérase mitochondriale des champignons et peut-être des autres eucaryotes supérieurs est en fait une ARN polymérase du type T7 (c'est-à-dire une ssRNAP). Or, l'ancêtre bactérien des mitochondries avait très probablement une ARN polymérase à plusieurs sous-unités. Quand et comment a eu lieu le remplacement? Nous nous pencherons sur l'évolution peu banale de ces ssRNAP en adoptant deux approches. Le premier article expose des études employant une approche PCR pour évaluer la distribution de séquences codant pour des ssRNAP chez les eucaryotes, en particuliers des protistes ayant divergé des autres eucaryotes très tôt dans l'évolution. Nous verrons que ce type de séquences est très répandu chez les eucaryotes. Le second article va un peu plus loin: l'analyse d'alignements et d'arbres phylogénétiques de ces séquences nous amènent à explorer l'origine de ces enzymes chez les eucaryotes, ainsi que l'origine du gène ancestral de ssRNAP. On verra que ces séquences ont été recrutées il y a très longtemps par les eucaryotes, et que le gène

ancestral a probablement pour origine un événement de duplication impliquant une ADN polymérase ou une transcriptase inverse virale.

6.2 La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris (Chapitre 2)

C'est dans ce chapitre qu'on se rendra compte qu'on pose parfois une question et qu'on découvre une réponse à une autre question non anticipée et de bien plus grande envergure. Les études relatées dans ce troisième article avaient en effet comme point de départ une recherche dans les bases de données se basant sur les questions: « Quelle est l'histoire évolutive des mtTFA? Peut-on trouver dans les bases de données des séquences de mtTFA intermédiaires entre celle de levure et celles des animaux supérieurs? ». Nous sommes finalement tombés sur une séquence de souris, qui code pour le mtTFA murin, mais également pour une protéine nucléaire spécifique aux spermatides dans les testicules, les deux protéines étant obtenues par sélection de promoteurs alternatifs et épissage différentiel subséquent. Ces recherches ont donc apporté non seulement des informations sur un événement intéressant de l'évolution de mtTFA, mais aussi et surtout sur un des mécanismes employés par certains types de cellules pour la communication entre les systèmes d'expression nucléaires et mitochondriaux.

6.3 Particularités des ARN de transfert mitochondriaux (Chapitre 3)

Nous avons vu que les ARNt mitochondriaux sont atypiques à plusieurs niveaux. À quel point les fonctions de ces ARNt sont-elles affectées par ces anomalies? Pourquoi ces ARNt atypiques sont-ils absents des cytoplasmes? Et à l'inverse, pourquoi n'ont-ils pas été éliminés dans les mitochondries? C'est à ces questions que nous avons décidé de nous attaquer, en mettant au point un système de suppression de codons d'arrêt *in vivo*, permettant d'analyser l'activité de ces ARNt atypiques. Nous avons d'un côté

étudié l'importance de l'interaction tertiaire entre les régions comprenant le nucléotide 47 et la paire de base 13-22 (premier article de ce chapitre): la différence d'activité *in vivo* est minime, et plusieurs générations sont nécessaires pour que l'effet s'en fasse ressentir. D'un autre côté, nous avons montré que des ARNt avec des tiges de l'anticodon plus longues que la normale pouvaient fonctionner *in vivo* (deuxième article), mais avec une activité inférieure à celle d'un contrôle avec une structure conventionnelle. Ces résultats peuvent expliquer que ces ARNt ne sont pas retrouvés ailleurs que dans les mitochondries et suggèrent que ces organites peuvent supporter un système de traduction moins efficace. Mais de façon inattendue, il appert que cette baisse d'exigences pour l'efficacité du système d'expression mitochondriale n'est pas très grande.

CHAPITRE 1

L'évolution des ARN polymérases à une sous-unité

Article 1

Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage

Article 2

On the evolution of the single-subunit RNA polymerases

ARTICLE 1

Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage

(Les séquences homologues aux ARN polymérases de la mitochondrie de levure et des bactériophages T3 et T7 sont très répandues parmi les eucaryotes)

Cermakian, N., Ikeda, T.M., Cedergren, R. et Gray, M.W. (1996) *Nucleic Acids Research* **24**:648-654.

(inclus avec la permission de Oxford University Press)

Les contributions de N.C. à cet article sont:

- La moitié des expériences (PCR, séquençage, traitement des séquences).
- Participation à la discussion et à la rédaction du manuscrit.

648–654 Nucleic Acids Research, 1996, Vol. 24, No. 4

© 1996 Oxford University Press

Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage

Nicolas Cermakian[1], Tatsuya M. Ikeda[2], Robert Cedergren[1] and Michael W. Gray[2,*]

Program in Evolutionary Biology, Canadian Institute for Advanced Research, [1]Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada and [2]Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Received October 30, 1995; Revised and Accepted January 5, 1996

GenBank accession nos U34283–U34286 and U34402–U34406

ABSTRACT

Although mitochondria and chloroplasts are considered to be descendants of eubacteria-like endosymbionts, the mitochondrial RNA polymerase of yeast is a nucleus-encoded, single-subunit enzyme homologous to bacteriophage T3 and T7 RNA polymerases, rather than a multi-component, eubacterial-type $\alpha_2\beta\beta'$ enzyme, as encoded in chloroplast DNA. To broaden our knowledge of the mitochondrial transcriptional apparatus, we have used a polymerase chain reaction (PCR) approach designed to amplify an internal portion of phage T3/T7-like RNA polymerase genes. Using this strategy, we have recovered sequences homologous to yeast mitochondrial and phage T3/T7 RNA polymerases from a phylogenetically broad range of multicellular and unicellular eukaryotes. These organisms display diverse patterns of mitochondrial genome organization and expression, and include species that separated from the main eukaryotic line early in the evolution of this lineage. In certain cases, we can deduce that PCR-amplified sequences, some of which contain small introns, are localized in nuclear DNA. We infer that the T3/T7-like RNA polymerase sequences reported here are likely derived from genes encoding the mitochondrial RNA polymerase in the organisms in which they occur, suggesting that a phage T3/T7-like RNA polymerase was recruited to act in transcription in the mitochondrion at an early stage in the evolution of this organelle.

INTRODUCTION

The evolutionary origin of the mitochondrial (mt) transcription system is puzzling. It is generally accepted that both mitochondria and chloroplasts arose from eubacteria-like endosymbionts, closely related to contemporary α-Proteobacteria and Cyanobacteria, respectively (1–3). Consistent with this view, chloroplast DNA is known to encode components of a eubacteria-like $\alpha_2\beta\beta'$ RNA polymerase (RNAP) (4). No such genes have been found in any of the mitochondrial genomes sequenced to date; instead, all of the genes for the mitochondrial transcriptional machinery appear to be encoded by nuclear DNA, with the protein products being imported into the organelle (5).

In yeast (*Saccharomyces cerevisiae*), the mitochondrial RNA polymerase (mtRNAP) is encoded by a nuclear gene, *RPO41* (6,7), and is a homolog of the single-polypeptide RNAPs of bacteriophages T3 and T7 (8). In view of the similar endosymbiotic origins of mitochondria and chloroplasts and the discovery of eubacteria-like RNAP genes in chloroplast DNA, this is a surprising finding. Moreover, this observation raises questions about the nature of the mitochondrial transcription system in other eukaryotes, and the evolutionary origin(s) of this system.

The identification of a second T3/T7-like mtRNAP sequence in *Neurospora crassa* (9), coupled with the recent appearance of homologous sequences in various expressed sequence tag (EST) databases (human, rice, *Caenorhabditis elegans*; Fig. 1), prompted us to devise a polymerase chain reaction (PCR) amplification strategy to search more widely within the eukaryotic lineage for phage T3/T7-like RNAP sequences. Information about the types and phylogenetic distribution of mtRNAPs, and their structural similarity to one another, is necessary to determine whether the mtRNAPs of different eukaryotes all arose from a single common ancestor and, if so, what the evolutionary source of this enzyme might have been. In particular, because the multicellular eukaryotes (animals, fungi, plants) represent relatively late radiations in the eukaryotic lineage (10), it is important to explore a phylogenetically broad range of earlier diverging unicellular eukaryotes (protists), in order to address the question of whether a T3/T7-like mtRNAP was acquired early or at a relatively late stage in the evolution of the mitochondrial transcription system. Comparative information about mtRNAPs and other transcriptional components in different eukaryotes is also essential for defining and understanding species-specific peculiarities in the biochemical mechanism of expression of mitochondrial genomes that can vary tremendously in size, base composition and organization.

MATERIALS AND METHODS

DNA and RNA preparations

Nuclear DNA from *Pycnococcus provasolii*, *Thraustochytrium aureum* and *Isochrysis* sp. Tahiti was isolated by B.F. Lang. The

* To whom correspondence should be addressed

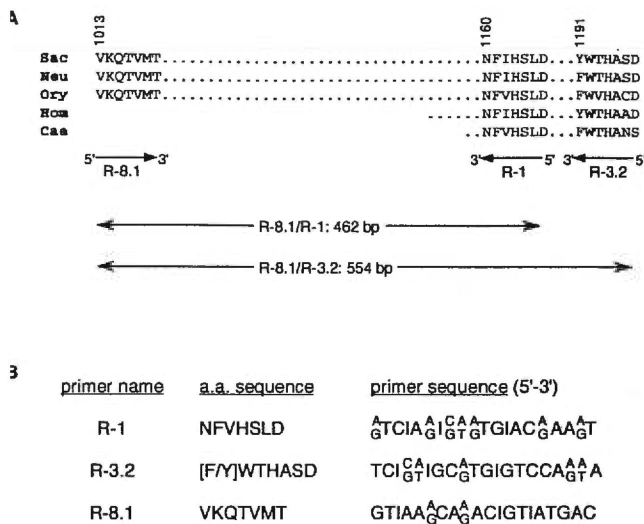


Figure 1. Design of PCR primers and strategy employed for amplification of putative mtRNAP sequences. (**A**) The amino acid sequences of established or putative mtRNAPs from five organisms were aligned, and oligonucleotide primers were designed against three highly conserved sequence motifs, as indicated. Abbreviations of organism names and GenBank accession numbers of the sequences used (square brackets) are: Sac, yeast (*Saccharomyces cerevisiae*) [M17539]; Neu, *Neurospora crassa* [L25087]; Ory, rice (*Oryza sativa*) [D24565, D23514]; Hom, human (*Homo sapiens*) [T97038, T93942, R31174, H03471, H03562]; Cae, *Caenorhabditis elegans* [D34229, D32914]. The three numbers above the aligned sequences (1013, 1160 and 1191) are the residue positions in the Sac sequence (8); the corresponding coordinates in the T7 RNAP sequence are 630, 781 and 814, respectively. Additional rice sequence beyond that in the database was generated as described in the text. (**B**) Designation and sequence of degenerate primers based on the alignment shown in (A) (I = inosine).

ɪrganisms were cultured axenically, with *Isochrysis* and *Pycnococ-*·*us* grown in sterilized seawater with additions (F/2 medium; for urther details, consult WWW site URL http://megasun.bch. ımontreal.ca/People/lang/FMGP/methods.html). Cells were broken nechanically by shaking with glass beads (11), DNA was olubilized in 0.5% SDS, and proteins were hydrolyzed in the ɪresence of 100 µg/ml proteinase K. After removal of detergent by alt precipitation (1 M NaCl, 1 h on ice), the isolated total cellular)NA was further purified by equilibrium centrifugation in CsCl lensity gradients (1.1 g/ml CsCl, 10 µg/ml Hoechst dye 33258 Serva); 40 000 r.p.m., 48 h). The main, lower band in the gradient, epresenting the nuclear DNA fraction in all three species, was used.

Nuclear DNA from *Acanthamoeba castellanii* (strain Neff; ıxenic culture) was prepared by K.M. Lonergan from isolated ıuclei and further purified by two rounds of CsCl density-gradi-:nt centrifugation after RNase A treatment (12).

Cells from an axenic culture of *Cryptomonas* Φ were supplied ›y M.A. Ragan, and DNA and cellular RNA were isolated from hese by D.F. Spencer. Cells were lysed in 1% SDS containing M sodium perchlorate, following which the lysate was ex-racted with chloroform/isoamyl alcohol and nucleic acids ›recipitated with isopropanol. The pellet was redissolved and :xtracted several times with phenol–cresol, after which the ıucleic acids were twice precipitated from ethanol before further ›urification by CsCl density-gradient centrifugation in the ›resence of Hoechst dye 33258. The main-band (nuclear) DNA raction and RNA pellet were recovered.

DNA was prepared from isolated nuclei of an axenic culture of *Naegleria fowleri* (strain LEE) and further purified by CsCl density-gradient centrifugation (13). The *N.fowleri* DNA sample was obtained from R. N. Band via A. Roger and P. Keeling (Department of Biochemistry, Dalhousie University).

Cells of *Tetrahymena pyriformis* (axenic culture), from which total cellular DNA was prepared by phenol extraction followed by RNase treatment, were grown by J. Edqvist.

A wheat (*Triticum aestivum*) cDNA clone bank was graciously provided by B.G. Lane. The library was constructed with mRNA that had been isolated as described (14), using λgt11 as vector in a Y1090 host (B.G. Lane, personal communication).

Total cellular RNA from rice (*Oryza sativa* cv. Lacassine) was kindly provided by P. Gros. The RNA sample was prepared from leaves homogenized in the presence of guanidinium hydrochloride, followed by phenol extraction.

Oligonucleotide primers

Degenerate primers used in this study are listed in Figure 1. Primers specific for λgt11 DNA were used to pre-amplify cDNA inserts, as follows: forward, 5′-ACTCCTGGAGCCCGTCAG-TA-3′; reverse, 5′-CAGACCAACTGGTAATGGTA-3′. For amplification of rice cDNA, the following were used: primer 1, 5′-AACAACCGGAAGACCAACG-3′; primer 2, 5′-CGTCCA-TTTGACAGGGTGG-3′.

PCR amplification

With *Pycnococcus*, *Thraustochytrium* and *Isochrysis* DNAs, PCR amplification was performed in a total volume of 50 µl containing PCR reaction buffer (Pharmacia) and 100 ng DNA, 0.2 mM each dNTP, 1 µM each primer (R-3.2 and R-8.1) and 100 µg/ml BSA. Reactions were carried out in a Perkin Elmer GeneAmp PCR System 9600, according to the following protocol: denaturation at 95°C for 4 min; addition of 1 U *Taq* DNA polymerase at 75°C; 3 cycles of 95°C for 1 min, 40°C for 1 min, slow ramp to 72°C for 1 min, and hold at 72°C for 3 min; 30 cycles of 95°C for 1 min, 50°C for 1 min, slow ramp to 72°C for 1 min, hold at 72°C for 3 min; hold at 72°C for 4 min. PCR products were resolved by agarose gel electrophoresis and extracted using QIAquick columns (QIAGEN).

With *Acanthamoeba*, *Cryptomonas*, *Naegleria*, *Tetrahymena* and *Triticum* (wheat) DNAs, PCR was performed in a total volume of 50 µl containing reaction buffer [25 mM Tricine (pH 8.5), 16 mM $(NH_4)_2SO_4$, 2 mM $MgCl_2$], 100 ng DNA, 0.1 mM each dNTP, 2 µM each primer (R-3.2 or R-1 and R-8.1), 1 U *Taq* DNA polymerase (Gibco-BRL) and 0.01 U *Pfu* DNA polymerase (Stratagene). Reactions were performed in a Perkin Elmer GeneAmp PCR System 2400, according to the following protocol: denaturation at 94°C for 3 min; 35 cycles of 94°C for 30 s, 50°C for 1 min, 55°C for 20 s, 60°C for 10 s and 72°C for 2–3 min, followed by an extension at 72°C for 10–15 min. PCR products were isolated after electrophoresis in low-melting-point agarose gels, as above. All PCR fragments were cloned into pT7Blue (Novagen) and sequenced using Sequenase version 2.0 (USB).

From a wheat cDNA library, inserts were pre-amplified in a total volume of 50 µl containing reaction buffer [25 mM Tricine (pH 8.5), 16 mM $(NH_4)_2SO_4$, 2 mM $MgCl_2$], 100 ng of DNA, 0.1 mM each dNTP, 0.2 µM each primer specific for λgt11 DNA (forward and reverse), 1 U *Taq* DNA polymerase and 0.01 U *Pfu* DNA polymerase. The reaction was performed according to the

following protocol: denaturation at 94°C for 3 min; 25 cycles of 94°C for 30 s, 55°C for 30 s and 72°C for 2 min, followed by an extension at 72°C for 5 min. A 100-fold dilution of amplified cDNAs was used for the PCR reaction following the same method as used for *Acanthamoeba*, *Cryptomonas*, *Naegleria* and *Tetrahymena* genomic DNAs.

Additional rice sequence was obtained by rapid amplification of cDNA ends (RACE) with construction of specific primers based on a rice EST (accession no. D23514). Rice total cellular RNA (5 μg) was used as template for reverse transcription by MuMLV (NEB) in the presence of 1 mM dNTP (Pharmacia), 15 U RNAguard (Pharmacia) and 50 ng of primer 1 in 20 μl containing 50 mM Tris–HCl (pH 8.3), 8 mM $MgCl_2$, 10 mM dithiothreitol for 1 h at 37°C, 30 min at 42°C and 15 min at 52°C. The resulting cDNAs were separated from primer using a Centricon-30 concentrator (Amicon) and extended at the 3′-end by terminal deoxynucleotidyl transferase (Gibco-BRL) in the presence of dATP. A 200-fold dilution of polyadenylated cDNAs was taken for two nested PCR runs, using primer 1 in the first run and primer 2 in the second. The conditions of the two runs and the sequences of the non-specific primers used were as described (15).

RESULTS

Identification of sequence homologs of T3/T7 RNA polymerase genes in eukaryotic DNA

From available protein sequences, we designed a set of degenerate oligonucleotide primers targeted to DNA sequences encoding regions highly conserved among known or putative mtRNAPs (all homologous to T3 and T7 RNAPs). The latter include rice, human and *C.elegans* expressed sequence tags (ESTs) that have recently appeared in public domain databases (Fig. 1). Primer combinations were tested in PCR amplification experiments with total cellular DNA preparations from a phylogenetically broad range of eukaryotes. The characteristics of positive PCR amplification products are summarized in Table 1. In a number of cases (including all red algae tested and several early diverging, amitochondriate eukaryotes), negative results were obtained with cellular DNA samples: i.e. either no discrete amplification products were obtained, or ones that were generated proved to be unrelated to T3/T7 RNAPs. Aside from the possibility that the gene in question may actually be absent in these negative cases, failure to generate a positive PCR product could result from any number of factors, including sub-optimal concentration of the target sequence in selected DNA preparations (high genomic complexity), spurious presence of competing sequences that sequester PCR primers, presence of introns that interfere with PCR amplification, and sequence divergence at primer binding sites.

An alignment of T3/T7-like RNAP sequences amplified from eight eukaryotes (seven of them protists) is shown in Figure 2. Also included are known bacteriophage (T7, K11, SP6) and fungal mitochondrial (yeast, *Neurospora crassa*) RNAP sequences, as well as human, rice and *C.elegans* EST homologs. The alignment excludes several blocks (the sizes of which are indicated in Fig. 2) that display pronounced sequence and length variation. Within the remaining alignable blocks (I–III), a high degree of amino acid sequence identity is evident, with a number of positions universally conserved or almost so among the eukaryotic sequences, and to a lesser extent between the eukaryotic and phage ones. Despite the phylogenetic breadth of the source organisms, the extent and degree of positional identity leaves little doubt that all of these sequences are related by descent from a common ancestral sequence (i.e. are homologous).

The amplified region (encompassing T7 positions 637–813) comprises part of the palm and fingers domain of the recently determined, hand-shaped T7 RNAP crystal structure (16). The most highly conserved portions of the alignment correspond mainly to regions facing the template-binding cleft of the enzyme. All of the sequences shown in Figure 2 contain the invariant and catalytically essential aspartate residue equivalent to D812 in T7 RNAP (17), as well as the adjacent and catalytically significant H811 (17). It should be noted that these two functionally important residues, as well as the equivalents of T7 RNAP residues Y639 and G440, are also found in DNA polymerases (18,19); however, overall sequence similarity clearly identifies the non-phage sequences in Figure 2 as RNAP, not DNAP, homologs.

Table 1. Characteristics of PCR amplification products containing T3/T7 RNAP-like sequences[a]

Organism	Abbreviation	Phylum	Type	Product (bp)	Cristal type	Introns
Triticum aestivum	Tri	Angiospermatophyta	land plant	500	flattened	?[b]
Pycnococcus provasolii	Pyc	Chlorophyta	green alga (prasinophyte)	692	flattened	0
Acanthamoeba castellanii	Aca	Rhizopoda	amoeboid protozoon	820	tubular	3
Isochrysis sp. Tahiti	Iso	Haptophyta	haptomonad alga	524	tubular	0
Thraustochytrium aureum	Thr	Heterokonta	slime net	581	tubular	0
Cryptomonas Φ	Cry	Cryptista	cryptomonad alga	824	flattened	6
Tetrahymena pyriformis	Tet	Ciliophora	ciliate protozoon	728	tubular	3
Naegleria fowleri	Nae	Percolozoa	amoeboid protozoon	411	discoidal	0

[a]All of these PCR products were obtained using the primer combination R-8.1/R-3.2 (see Fig. 1) except for the *N.fowleri* one, which amplified only with the R-8.1/R-1 combination. Negative results were obtained with *Dennstaedtia punctilobula* (hay-scented fern); *Chlamydomonas reinhardtii* (green alga); *Palmaria palmaria*, *Chondrus crispus*, *Mastocarpus stellatus* and *Gracilaria sordida* (red algae); *Cyanophora paradoxa* (glaucocystophyte); *Goniomonas truncata* Stein (heterotrophic nanoflagellate); *Crypthecodinium cohnii* (dinoflagellate); *Crithidia fasciculata* (kinetoplastid flagellate); *Euglena gracilis* (euglenoid flagellate); *Reclinomonas americana* and *Jakoba libera* (jakobid flagellates); and two amitochondriate eukaryotes (*Giardia lamblia*, *Trichomonas vaginalis*). Negative results were also obtained with a soybean cDNA library (Clontech).

[b]Sequence determined from cDNA clone.

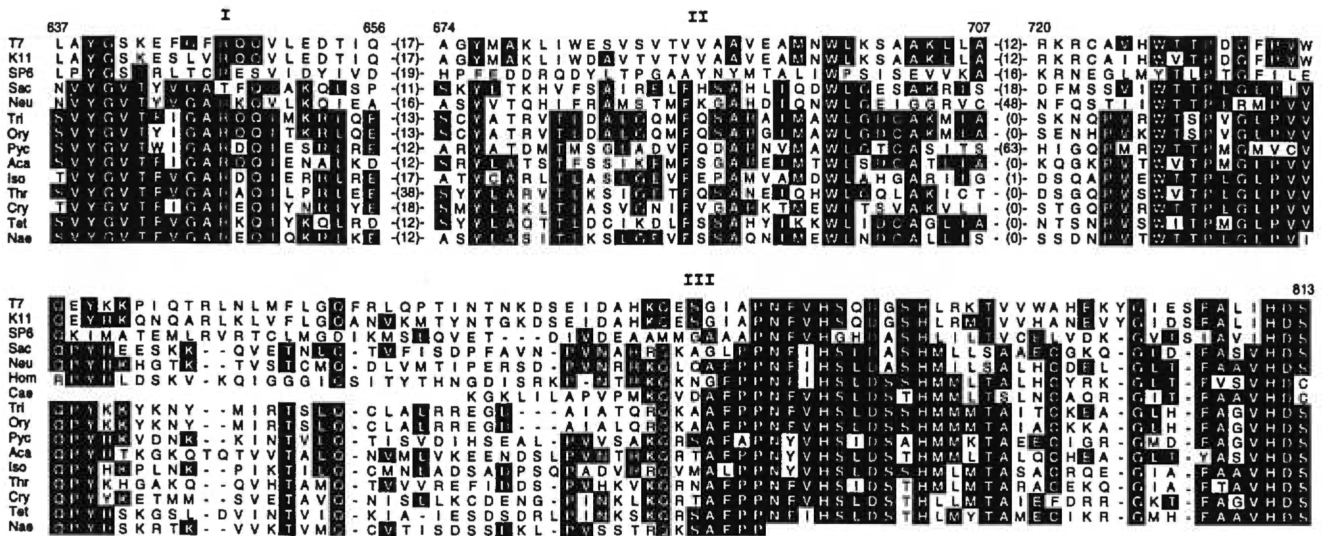


Figure 2. Alignment of RNAP sequences (sequences and abbreviations of organism names are listed in Figure 1 and Table 1). Sequences determined in the present study have been deposited in GenBank (accession numbers U34283–86 and U34402–06). Accession numbers for sequences taken from the database are listed in Figure 1, with additional ones as follows: T7 [M38308], K11 [X53238], SP6 [Y00105]. Note that EST databases contain several human and *C.elegans* sequences homologous to the yeast mtRNAP (see Fig. 1), only one of which is included here in each case (Hom, T97038; Cae, D34229). The alignment was constructed using the Multalin program (41) and refined manually; residue numbering refers to the bacteriophage T7 RNAP sequence. Only blocks of residues that can be unambiguously aligned (I, II and III, corresponding to positions 637–656, 674–707 and 720–813 in the T7 mtRNAP) are shown; numbers in parentheses indicate the sizes of less well conserved regions that vary considerably in length and sequence. Amino acids that are positionally identical in at least five of the sequences are highlighted (white on black). The complete alignment is available from the authors upon request (contact M.W.G.; Email address: mgray@ac.dal.ca).

At several positions (e.g. 638, 641, 646, 689, 738 and 810), the yeast mtRNAP and its PCR-amplified homologs are distinguished as a group from the phage RNAP sequences. For example, at T7 residue 641, all of the non-phage sequences have Val rather than the Ser found in the three phage sequences. Interestingly, an S641A mutant of T7 RNAP displays substantial DNA polymerase activity (20). Although they lack Ser at position 641, all of the non-phage sequences in Figure 2 have another hydroxy amino acid (Thr) at the immediately adjacent position (642).

Intron sequences in PCR amplification products

In several cases, PCR amplification generated positive products that were substantially larger than expected (compare Fig. 1 and Table 1). In these instances (*Acanthamoeba castellanii*, *Cryptomonas* Φ and *Tetrahymena pyriformis*), we infer the presence of small intron sequences in the PCR products (Fig. 3 and Table 2). Exon/intron boundaries were assigned based on consideration of optimal amino acid sequence alignment and maintenance of open reading frames, as well as the assumption that introns begin with GT and end in AG. The deduced intron sequences display distinctive base compositions relative to their flanking exons, being particularly C+T-rich (61–69%) in *Acanthamoeba* and A+T-rich (81–85%) in *Tetrahymena* (Table 2); the latter feature is characteristic of spliceosomal-type nuclear introns in this organism (21). Intron junction sequences also correspond closely to known consensus sequences at intron splice sites in *Acanthamoeba* and *Tetrahymena* nuclear genes (Table 3). A well-defined consensus sequence is also evident at *Cryptomonas* intron junctions; in this case, however, no published nuclear intron sequences are available for comparison.

In the case of *Cryptomonas*, RT–PCR using total cellular RNA gave a product having a size expected for amplification from a spliced message; sequencing of this product confirmed the absence of putative intron sequences and verified the inferred splice junctions (data not shown). For the *Acanthamoeba* and *Tetrahymena* PCR products, the actual existence of the deduced introns and the validity of the assigned exon–intron boundaries remain to be confirmed by cDNA analysis. Although the positions of several of the introns are quite close to one another (e.g. the final intron in each case), none of the assigned insertion sites is precisely shared among the three organisms.

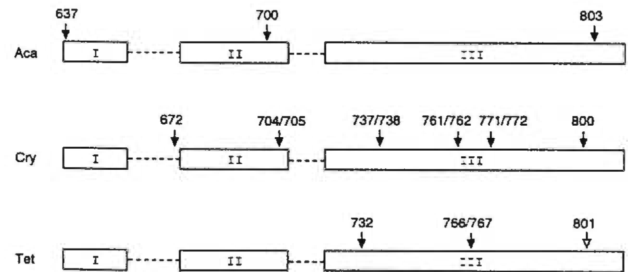


Figure 3. Positions of putative introns (arrows) in *Acanthamoeba* (Aca), *Cryptomonas* (Cry) and *Tetrahymena* (Tet) PCR products, deduced from considerations of amino acid sequence alignment, as described in the text. Intron positions are numbered with reference to the T7 RNAP sequence, with insertion sites falling either within (single numbers) or between (double numbers) the indicated codons. Introns that disrupt the RNAP reading frame are denoted by solid arrows, whereas one that is continuous with upstream and downstream exons and contains no termination codons is indicated by an open arrow. Rectangles labelled I, II and III refer to regions of the sequence specifying highly conserved blocks of amino acid residues; dashed lines refer to more variable regions (see Fig. 2).

Table 2. Characteristics of deduced introns in T3/T7 RNAP-like sequences isolated by PCR amplification of eukaryotic DNA

Organism	Intron position[a]	Size (bp)	G+C (%)	C+T (%)
A.castellanii	637	104	51.0	69.2
	700	97	47.4	67.0
	803	97	51.4	60.7
			(61.4)[b]	(49.3)[c]
Cryptomonas Φ	672	52	50.0	61.5
	704/705	47	53.2	63.8
	737/738	57	60.7	42.9
	761/762	49	51.0	69.4
	771/772	57	53.2	57.4
	800	52	65.4	55.8
			(52.2)[b]	(49.9)[c]
T.pyriformis	732	67	14.9	42.3
	766/767	62	19.4	54.8
	801	93	16.1	63.4
			(35.6)[b]	(51.9)[c]

[a]Numbering refers to the homologous codon(s) in the T7 RNAP sequence.
[b]G+C content (%) of flanking exons.
[c]C+T content (%) of flanking exons.

DISCUSSION

The T3/T7-like sequences reported here, which are also homologous at the amino acid level to the yeast mtRNAP sequence, represent a phylogenetically broad sampling within the eukaryotic lineage. In several cases (e.g. *Cryptomonas*, wheat), we have direct evidence that the PCR-amplified gene is expressed. In the remaining instances, additional work will be required to verify that the gene we have identified is expressed, and that it encodes a functional mitochondrial RNAP. Whether or not all of these sequences ultimately prove to be functional, their widespread occurrence throughout the eukaryotic lineage raises intriguing questions about their evolutionary origin, and their present or former function.

The species from which positive PCR products were obtained include members of a number of major protist phyla, representing all three mitochondrial cristal types (discoidal, tubular, flattened) (Table 1). The organisms branch widely within a eukaryotic phylogenetic tree [see e.g. Cavalier-Smith (22)]. Of particular interest is our finding of a mtRNAP homolog in the protozoon, *Naegleria fowleri*. *Naegleria* is thought to be one of the earliest diverging genera among mitochondria-containing eukaryotes (22). It is also striking that sequences homologous to yeast mtRNAP could be isolated from organisms displaying very diverse patterns of mitochondrial gene organization and expression, such variability being a hallmark of mitochondrial genomes (23). Allowing that the PCR-derived T3/T7-like sequences described here likely encode a portion of the mtRNAP in these organisms, this would suggest that transcriptional mechanisms

Table 3. Characteristics of inferred intron junction sequences in T3/T7-like RNAP amplification products[1]

Acanthamoeba castellanii														
consensus[2]				G100	T100	A89	C82	G75	T45 C39			C73	A100	G100
intron (637)				G	T	G	C	G	G			C	A	G
intron (700)				G	T	A	C	G	T			C	A	G
intron (803)				G	T	A	T	G	C			C	A	G
Tetrahymena pyriformis														
consensus[3]	A52 T33	A52 T33	G56	G100	T100	A93	A74			A67 T32	A64	T67	A100	G100
intron (732)	G	G	G	G	T	A	A			A	A	T	A	G
intron (766/767)	A	G	A	G	T	A	A			A	A	T	A	G
intron (801)	A	C	G	G	T	A	A			A	T	T	A	G
***Cryptomonas* Φ**														
consensus[4]		A/C	G	G	T						G/C	C	A	G
intron (672)		C	G	G	T	G	C			C	C	C	A	G
intron (704/705)		A	G	G	T	C	C			G	G	C	A	G
intron (737/738)		A	G	G	T	G	A			T	C	C	A	G
intron (761/762)		A	G	G	T	T	G			T	G	C	A	G
intron (771/772)		A	G	G	T	T	G			G	G	C	A	G
intron (800)		C	G	G	T	C	T			T	G	C	A	G

[1]Numbers to the right of each nucleotide indicate frequency of occurrence (%) at that position. Universal GT and AG dinucleotides at the beginning and end of each intron are underlined.
[2]Based on 44 *A.castellanii* nuclear introns (40).
[3]In *Tetrahymena thermophila* (21).
[4]Based on the six sequences shown.

ıd the transcriptional machinery itself (24) may have more ·atures in common than the diversity of mitochondrial transcrip- ɔnal patterns might suggest.

Phylogenetic trees (not shown) constructed using the aligned nino acid sequences in Figure 2 did not display robust branching ıtterns, a consequence of the limited information content in the ·quence data currently available. Not surprisingly, the wheat and ɔe protein sequences are highly similar (142 identical residues ver 153 positions), and associate strongly in the trees. The same true of the *Neurospora* and yeast protein sequences, although ·anch lengths are much longer in this case, reflecting a greater ·gree of sequence divergence. Other than these two affiliations, e only other notable feature of this analysis is that, as a group, e putative mtRNAP sequences are more similar to one another an to the phage sequences, consistent with the idea that the tRNAPs diverged from a more recent common ancestor than ey shared with the phage RNAPs. The fact that none of the nplified sequences branches with the yeast-*Neurospora* clade akes fungal contamination of the DNA preparations used for CR analysis an unlikely possibility.

Because PCR products were generated using total cellular NA or partially purified nuclear DNA fractions as target, the ·nomic localization of each of the amplified sequences remains be definitively established. In the case of *A.castellanii* and *pyriformis*, no T3/T7-like RNAP sequences are present in the ɔmpletely sequenced mitochondrial genomes of these organ- ms (25; Burger,G., Zhu,Y., Littlejohn,T., Greenwood,S.J., ·hnare,M.N. and Gray,M.W., in preparation), whereas the ·esence of small, splicesosomal-like intron sequences in the ·spective PCR products supports a nuclear location for the tRNAP homologs. Moreover, in translating the *Tetrahymena* ɪon sequences for the alignment, it was necessary to use the ıodified genetic code (UAA and UAG decoded as Gln) that is ıown to be employed for nucleus-encoded mRNAs in this ·ganism (26).

Introns in the *Cryptomonas* Φ PCR product also suggest that ıese RNAP sequences are localized in the nuclear genome. owever, a complicating factor in this case is the additional ·esence of a nucleomorph genome that is the evolutionary ·mnant of the nuclear genome of an endosymbiotic alga !7–29). In Southern hybridization experiments with total DNA ·solved by pulse-field gel electrophoresis, the *Cryptomonas* CR product specifically hybridized with nuclear DNA (data not ıown). In the case of wheat, the mtRNAP-homologous sequence ɔuld not be amplified from purified mtDNA; moreover, no ·quences homologous to T3/T7 RNAPs have been found in ·veral completely sequenced chloroplast DNAs, including those om rice (a monocotyledon closely related to wheat) (30) and ıbacco (a dicotyledon) (31). These observations indicate that the ·heat and rice homologs are also nuclear genes.

Fungal and plant mitochondria commonly contain linear lasmids encoding single-subunit RNAPs (32). These plasmid- ıcoded RNAP sequences form a distinct group that is only istantly related to the clade of phage T3 and T7 RNAPs and to ıe nucleus-encoded yeast mtRNAP (32), as well as to the CR-derived RNAP sequences listed in Figure 2. In pairwise ɔmparisons, the non-phage RNAP sequences listed in Figure 2 ·e clearly more closely related to one another than to either their hage- or plasmid-encoded RNAP homologs (data not shown). Ioreover, as a consequence of sequence divergence at the inding sites against which PCR primers were constructed (Fig. 1), we would not expect that the equivalent region in mitochondrial plasmid-encoded RNAP genes would be amplified under the conditions employed in the current study (particularly with the R-8.1/R-3.2 primer combination used to recover all but the *N.fowleri* sequence; Table 1). For these reasons, we are confident that that sequences shown in Figure 2 do not originate from linear plasmids of the type characterized in fungal and plant mitochondria, the evolutionary origin of which is also obscure.

In photosynthetic organisms, there is evidence for the existence of a second chloroplast RNAP activity, not encoded in chloroplast DNA (33–35). In spinach chloroplasts, RNAP activity has been associated with a 110 kDa polypeptide that has some phage RNAP-like properties (36). Thus, it is possible that there are separate nuclear genes encoding distinct mitochondrial and chloroplast enzymes, each homologous to T3/T7 RNAPs, and that in certain cases, the gene for the chloroplast enzyme might be amplified preferentially over the gene for the mitochondrial enzyme under our conditions. In this regard, it may be significant that among the amplified sequences shown in Figure 2, the most divergent is the one from the prasinophyte *Pycnococcus provasolii*, a primitive green alga. On the other hand, in no case was more than one T3/T7 RNAP-homologous PCR product obtained from any of the DNA samples analyzed. In addition, a Southern hybridization experiment not only verified that the *Pycnococcus* amplification product comes from *Pycnococcus* DNA, but also indicated that the corresponding gene is single copy (data not shown). If a chloroplast homolog does exist, it is formally possible that both it and its mitochondrial counterpart are encoded by the same nuclear gene, with the protein products being targeted to the respective organelles. Precedent exists for targeting of the protein products of a single nuclear gene to different subcellular compartments (37). Additional experimentation will be required to establish the genomic location of each of the T3/T7-like RNAP genes identified in this study and the subcellular location and function of their encoded protein products.

The observations reported here support the thesis that a T3/T7-like RNAP was recruited to act as a mtRNAP at an early stage in the evolution of the mitochondrion. So far, however, we have few clues as to the evolutionary origin of the gene encoding this enzyme. In addition to negative results with amitochondriate eukaryotes, we were unable to amplify homologous sequences from DNAs of eubacteria or archaebacteria. Moreover, a search of the recently published complete genome sequences from *Haemophilus influenzae* (38) and *Mycoplasma genitalium* (39) failed to reveal any T3/T7-like RNAP sequences. Thus, it is not yet clear whether the gene for a T3/T7-like enzyme was acquired from a eubacteria-like symbiont or was provided by the ancestor of the nucleus-containing eukaryotic host cell. Other questions that warrant further investigation are whether mtRNAP genes of the phage T3/T7 type are present in any mitochondrial genomes not yet characterized, and whether such an enzyme is used as the mtRNAP in all eukaryotes. The results summarized here provide a useful entry to the study of mtRNAP evolution and function in a broad range of eukaryotes, particularly protists.

ACKNOWLEDGEMENTS

We are indebted to R. N. Band (Department of Zoology, Michigan State University), E. Denovan-Wright (Department of Biology, Dalhousie University), P. Gros (Department of Biochemistry, McGill University), B.F. Lang (Département de Biochimie,

Université de Montréal), and K.M. Lonergan and D.F. Spencer (Department of Biochemistry, Dalhousie University) for generous gifts of DNA and/or RNA; to M.A. Ragan (Institute for Marine Biosciences, National Research Council, Halifax) and J. Edqvist (Department of Biochemistry, Dalhousie University) for gifts of *Cryptomonas* Φ and *T. pyriformis* cells, respectively; to B.G. Lane (Department of Biochemistry, University of Toronto) for provision of a wheat cDNA library; and to G.I. McFadden (School of Botany, University of Melbourne) for a blot of *Cryptomonas* Φ chromosomes separated by pulse field gel electrophoresis. This work was supported by grants from MRC Canada (to M.W.G.; MT-4124) and from NSERC Canada (to R.C.), and by an NSERC 1967 Science and Engineering Scholarship (to N.C.). Salary and interactions support from the Canadian Institute for Advanced Research are gratefully acknowledged by M.W.G. and R.C., who are Fellows in the Program in Evolutionary Biology.

REFERENCES

1 Gray,M.W. (1989) *Trends Genet.* **5**, 294–299.
2 Gray,M.W. (1992) *Int. Rev. Cytol.* **141**, 233–357.
3 Gray,M.W. (1993) *Curr. Opin. Genet. Dev.* **3**, 884–890.
4 Bogorad,L. (1991) In Bogorad,L. and Vasil,I.K. (eds), *The Molecular Biology of Plastids*. Academic Press, Inc., San Diego, pp. 93–124.
5 Shadel,G.S. and Clayton,D.A. (1993) *J. Biol. Chem.* **268**, 16083–16086.
6 Greenleaf,A.L., Kelly,J.L. and Lehman,I.R. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 3391–3394.
7 Kelly,J.L., Greenleaf,A.L. and Lehman,I.R. (1986) *J. Biol. Chem.* **261**, 10348–10351.
8 Masters,B.S., Stohl,L.L. and Clayton,D.A. (1987) *Cell* **51**, 89–99.
9 Chen,B., Kubelik,A.R., Mohr,S. and Breitenberger,C.A. (1996) *J. Biol. Chem*, in press. L25087.
10 Patterson,D.J. and Sogin,M.L. (1992) In Hartman,H. and Matsuno,K. (eds), *The Origin and Evolution of the Cell*. World Scientific, Singapore, pp. 13–46.
11 Lang,B., Burger,G., Doxiadis,I., Thomas,D.Y., Bandlow,W. and Kaudewitz,F. (1977) *Anal. Biochem.* **77**, 110–121.
12 Lonergan,K.M. and Gray,M.W. (1993) *Science* **259**, 812–816.
13 Hu,W.-N., Kopachik,W. and Band,R.N. (1992) *Infect. Immunity* **60**, 2418–2424.
14 Haffner,M.H., Chin,M.B. and Lane,B.G. (1978) *Can. J. Biochem.* **56**, 729–733.
15 Frohman,M.A. (1994) *PCR Methods Applic.* **4**, S40-S58.
16 Sousa,R., Chung,Y.J., Rose,J.P. and Wang,B.-C. (1993) *Nature* **364**, 593–599.
17 Osumi-Davis,P.A., de Aguilera,M.C., Woody,R.W. and Woody,A-Y.M. (1992) *J. Mol. Biol.* **226**, 37–45.
18 Delarue,M., Poch,O., Tordo,N., Moras,D. and Argos,P. (1990) *Protein Engng.* **3**, 461–467.
19 Gross,L., Chen,W.-J. and McAllister,W.T. (1992) *J. Mol. Biol.* **228**, 488–505.
20 Kostyuk,D.A., Dragan,S.M., Lyakhov,D.L., Rechinsky,V.O., Tunitskaya,V.L., Chernov,B.K. and Kochetkov,S.N. (1995) *FEBS Lett.* **369**, 165–168.
21 Csank,C., Taylor,F.M. and Martindale,D.W. (1990) *Nucleic Acids Res.* **18**, 5133–5141.
22 Cavalier-Smith,T. (1993) *Microbiol. Rev.* **57**, 953–994.
23 Gray,M.W. (1989) *Annu. Rev. Cell Biol.* **5**, 25–50.
24 Tracy,R.L. and Stern,D.B. (1995) *Curr. Genet.* **28**, 205–216.
25 Burger,G., Plante,I., Lonergan,K.M. and Gray, M.W. (1995) *J. Mol. Biol.* **245**, 522–537.
26 Osawa,S., Jukes,T.H., Watanabe,K. and Muto,A. (1992) *Microbiol. Rev.* **56**, 229–264.
27 Douglas,S.E., Murphy,C.A., Spencer,D.F. and Gray,M.W. (1991) *Nature* **350**, 148–151.
28 McFadden,G.I., Gilson,P.R. and Douglas,S.E. (1994) *J. Cell Sci.* **107**, 649–657.
29 McFadden,G. and Gilson,P. (1995) *Trends Ecol. Evol.* **10**, 12–17.
30 Hiratsuka,J., Shimada,H., Whittier,R., Ishibashi,T., Sakamoto,M., Mori,M., Kondo,C., Honji,Y., Sun,C.-R., Meng,B.-Y., Li,Y.-Q., Kanno,A., Nishizawa,Y., Hirai, A., Shinozaki,K. and Sugiura,M. (1989) *Mol. Gen. Genet.* **217**, 185–194.
31 Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K. *et al.* (1986) *EMBO J.* **5**, 2043–2049.
32 Kempken,F., Hermanns,J. and Osiewacz,H.D. (1992) *J. Mol. Evol.* **35**, 502–513.
33 Morden,C.W., Wolfe,K.H., dePamphilis,C.W. and Palmer,J.D. (1991) *EMBO J.* **10**, 3281–3288
34 Falk,J., Schmidt,A. and Krupinska,K. (1993) *J. Plant Physiol.* **141**, 176–181.
35 Hess,W.R., Prombona,A., Fieder,B., Subramanian,A.R. and Börner,T. (1993) *EMBO J.* **12**, 563–571.
36 Lerbs-Mache,S. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5509–5513.
37 Martin,N.C. and Hopper,A.K. (1994) *Biochimie* **76**, 1161–1167.
38 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science* **269**, 496–512.
39 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J. Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) *Science* **270**, 397–403.
40 Wong,J.M., Liu,F. and Bateman,E. (1992) *Nucleic Acids Res.* **20**, 4817–4824.
41 Corpet,F. (1988) *Nucleic Acids Res.* **16**, 10881–10890.

ARTICLE 2

On the evolution of the single-subunit RNA polymerases

(De l'évolution des ARN polymérases à une sous-unité)

Cermakian, N., Ikeda, T.M., Miramontes, P., Lang, B.F., Gray, M.W. et Cedergren, R. (1997) *Journal of Molecular Evolution* **45**:671-681.

(inclus avec la permission de Springer-Verlag New-York, Inc.)

Les contributions de N.C. à cet article sont:

- Isolation et séquençage du fragment de gène de *Rhizophydium*.
- Alignements et analyses phylogénétiques.
- Rédaction du manuscrit.

On the Evolution of the Single-Subunit RNA Polymerases

Nicolas Cermakian[1†], Tatsuya M. Ikeda[2†], Pedro Miramontes[3], B. Franz Lang[1], Michael W. Gray[2] and Robert Cedergren[1*]

Program in Evolutionary Biology, Canadian Institute for Advanced Research

[1] Département de biochimie, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada

[2] Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

[3] Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico

*To whom correspondence should be addressed.
Fax: (514) 343-2210.
E-mail: ceder@medcn.umontreal.ca

Street address: Département de biochimie, Université de Montréal, 2900 boul. Édouard-Montpetit, Pavillon principal, local D-334, Montréal QC H3T 1J4, Canada

†These authors contributed equally to this work.

August 7, 1997

ABSTRACT

Many eukaryotic nuclear genomes as well as mitochondrial plasmids contain genes displaying evident sequence similarity to those encoding the single-subunit RNA polymerase (ssRNAP) of bacteriophage T7 and its relatives. We have collected and aligned these ssRNAP sequences and have constructed unrooted phylogenetic trees that demonstrate the separation of ssRNAPs into three well-defined and non-overlapping clusters (phage-encoded, nucleus-encoded and plasmid-encoded). Our analyses indicate that these three subfamilies of T7-like RNAPs shared a common ancestor; however, the order in which the groups diverged cannot be inferred from available data. On the basis of structural similarities and mutational data, we suggest that the ancestral ssRNAP gene may have arisen via duplication and divergence of a DNA polymerase or reverse transcriptase gene. Considering the current phylogenetic distribution of ssRNAP sequences, we further suggest that the origin of the ancestral ssRNAP gene closely paralleled in time the introduction of mitochondria into eukaryotic cells through a eubacterial endosymbiosis.

KEY WORDS: RNA polymerase - Evolution - Mitochondria - Bacteriophage - Linear plasmid - Phylogenetic tree

ABBREVIATIONS: RNAP, RNA polymerase; ssRNAP, single-subunit RNA polymerase; mtRNAP, mitochondrial RNA polymerase; DNAP, DNA polymerase; RT, reverse transcriptase.

INTRODUCTION

Considering the pivotal role played by the synthesis of RNA (transcription) in the expression of genetic information, knowledge of the origin and evolution of this process could provide crucial insights into the origin of cellular life itself.

Transcription is generally carried out by complex, multi-component RNA polymerases (RNAPs) consisting of two highly conserved subunits and a number of smaller polypeptides (Palenik 1992). However, in bacteriophages related to the coliphage T7, most genes are transcribed by a single-subunit RNA polymerase (ssRNAP; McAllister 1993), encoded by the phage genome. This enzyme not only initiates transcription, but also functions in elongation and termination during RNA synthesis. The origin of the ssRNAP is an evolutionary enigma, as is its relationship to the multi-subunit RNAPs, with which it shares no obvious sequence similarity. In fact, ssRNAPs possess short sequence motifs (Delarue et al. 1990) and the 'hand-like' three-dimensional structure (Sousa et al. 1993) of the Klenow fragment of *Escherichia coli* DNA polymerase (DNAP) I (Sousa 1996).

The mystery surrounding the origin of the ssRNAP gene deepened with the discovery that the mitochondrial RNAP (mtRNAP) from *Saccharomyces cerevisiae* displays evident sequence similarity to the phage T7 RNAP (Masters et al. 1987). To test the premise that other mitochondria-containing eukaryotes might harbor ssRNAP-like genes, we designed PCR primers to scan eukaryotic genomes for T7-like ssRNAP genes. We were able to show that such sequences are widespread within the eukaryotic lineage (Cermakian et al. 1996). Moreover, the mitochondria of certain plants and fungi contain linear plasmids that bear RNAP genes strongly resembling the one encoding the T7 ssRNAP (Kempken et al. 1992; Griffiths 1995). These plasmid-like DNAs are characterized by double-stranded structure, terminal inverted repeats and open reading frames specifying both a DNAP and a ssRNAP (Meinhardt et al. 1990; Griffiths 1995). The two polymerases are encoded either on a single plasmid, as in various fungi (Kempken et al. 1992), or on distinct plasmids, as in the case of the S-1 and S-2 plasmids of maize (Kuzmin et al. 1988) and, presumably, the pLm9 and pLm10 plasmids of *Leptosphaeria maculans* (Lim and Howlett 1994). It has been proposed that mitochondrial linear plasmids are evolutionarily related to bacteriophages (Meinhardt et al. 1990; Kempken et al. 1992). Both their protein-terminated double-stranded DNA structure (Meinhardt et al. 1990) and B-type DNAPs (Braithwaite and Ito 1993) are features shared with several bacteriophages, including

Φ29 and PRD1 (Salas 1988; Blanco and Salas 1996). Although phages Φ29 and PRD1 diverge greatly from each other (Ackermann and DuBow 1987), phylogenies based on DNAP sequences show a close relationship between them and linear plasmid DNAPs (Kempken et al. 1992; Rohe et al. 1992). In both of these phages, the terminal protein gene is located just upstream of the DNAP gene (Yoshikawa and Ito 1982; Grahn et al. 1994), whereas in the case of linear plasmids, it has been suggested that this role is assumed by the amino-terminal domain of the plasmid-encoded DNAP, consistent with the view that linear plasmids may be remnants of an ancestral bacteriophage.

The existence of two RNAP families (single- and multi-subunit) raises fundamental questions about the origin of transcription. In particular, we would like to know whether the two families share a common history, and which (if either) was responsible for transcription at the early stages in the evolution of this process. A further issue relevant to the ssRNAPs is whether the apparent similarity of these gene sequences from diverse genetic sources (eukaryotic nuclei, mitochondrial plasmids and bacteriophage genomes) belies a true evolutionary relationship. Here we address these questions through phylogenetic analysis of trees constructed from an alignment of the ssRNAP sequences, and discuss these and other data on ssRNAPs with a view to considering how this critical protein family may have originated and subsequently evolved, and how the known phylogenetic distribution of ssRNAP genes can be rationalized.

METHODS

Rhizophydium sp. #136, an un-named representative of the Chytridiales (chytridiomycete fungi), was isolated from garden soil by Dr. J. Longcore, University of Maine. Sporangia were grown on PmTG (Barr 1986) nutrient agar. Cells were broken with glass beads (Lang et al. 1977) and DNA was isolated as described in Cermakian et al. (1996). PCR experiments and subsequent cloning and sequencing

procedures were also performed as described in Cermakian et al. (1996), using the PCR primers R-8.1 and R-3.2.

The single-letter code for amino acids is used throughout this discussion and residues are numbered with reference to their position in the bacteriophage T7 RNAP sequence. Alignments of the ssRNAP sequences were performed using the Multalin program (Corpet 1988). Because the ssRNAP database consists of a diverse collection of sequences, we first performed separate alignments of the nucleus-encoded, plasmid-encoded and phage-encoded enzymes, before all sequences were assembled into a globally aligned collection. Partial sequences were added manually to the collection based on the alignment of Fig. 1 and that of Cermakian et al. (1996). Phylogenetic trees were constructed from this alignment using parsimony (PAUP; Swofford 1993) and distance methods (SEQBOOT, PROTDIST, FITCH, NEIGHBOR and CONSENSE of the PHYLIP package; Felsenstein 1993). For all trees, 100 bootstrap resamplings of the data were performed. All alignments and trees are available on request from the authors.

A Monte Carlo analysis of the sequence data was made to verify the null hypothesis of randomness in the distance scores (Manly, 1991). One hundred sets of 16 random sequences of the same amino acid content and length as the *S. cerevisiae* ssRNAP sequence shown in Fig. 1 were generated. These data sets were then used to generate 100 distance matrices using PROTDIST of PHYLIP package (with "categories distance model" included in this program). The statistics chosen to test the null hypothesis were d_{max}, the maximum of distances between elements in each set of 16 sequences, and the radius (r) of the distance matrix (d_{max}-d_{min}). The first parameter measures the worst possible case in a distance matrix and the second the degree of inhomogeneity among the sequences in a single distance matrix.

Results and Discussion

Alignments and Phylogenies of ssRNAP Sequences

The sequences used in this study (Table 1) include those employed previously (Kempken et al. 1992), as well as the nucleus-encoded mitochondrial ssRNAP sequences from *Neurospora crassa* (Chen et al. 1996), *Chenopodium album* (Weihe et al., 1997) and *Arabidopsis thaliana* (acc. no. Y09432), the linear plasmid sequences from *Gelasinospora* sp. (Griffiths 1995) and *Claviceps purpurea* strain T5 (Oeser et al. 1993), and the bacteriophage sequences from *Klebsiella* phage K11 (Dietz et al. 1990) and *Salmonella* phage SP6 (Kotani et al. 1987). To these we have added all partial sequences previously published by our group (Cermakian et al. 1996) as well as the new *Rhizophydium* sequence reported here.

The most conserved blocks in the alignment (Fig. 1) are similar to the regions delineated by Masters et al. (1987) and Chan et al. (1991), with the exception that we have eliminated the N-terminal region from the alignment because it is not well conserved in our more extensive data set. We also use an additional block (F) from the carboxy-terminal region of the protein sequence.

One of the most striking aspects of the alignment is the considerable length variation in the N-terminal region of the sequence, which suggests that this region could be involved in species-specific processes (Ikeda and Richardson 1987; Ikeda et al. 1993). This length variation accounts for most of the overall length variation of ssRNAP sequences (Fig. 1). On the other hand, the higher degree of conservation in the C-terminal region together with experimental data support the idea that this portion of the protein is involved in general RNAP functions such as template binding and nucleotide polymerization (Sousa 1996). Another obvious feature of this alignment is the existence of three distinct classes of ssRNAPs. This might have been anticipated from the fact that we had aligned each group separately prior to generating the global alignment of Fig. 1; however, intragroup similarities are markedly stronger than would

be expected from such an alignment bias. For example, the penultimate amino acid (882 in T7 RNAP) is F in all phage and nuclear sequences, but not in the plasmid-borne polymerases. Only bacteriophage sequences have L at position 637, and only the nuclear sequences have P at position 738.

Phylogenetic trees were derived from the alignment of Fig. 1 and the different phylogenetic treeing techniques gave almost identical results. One such phylogeny is shown in Fig. 2. As indicated by low bootstrap values, the branching order within the plasmid-borne RNAP gene representatives is poorly defined; this may reflect looser structural and functional constraints on this class. This observation is underscored as well by the presence of several unusual amino acids at a number of positions that are otherwise highly conserved among ssRNAP sequences. The overall topology is robust with the three treeing methods: the phage-encoded, nucleus-encoded and plasmid-encoded sequences form three well-separated phylogenetic clusters, confirming what had been observed in the course of tree building and in analyzing the alignment. Even in the case where there is a nucleus-encoded and a plasmid-encoded sequence, i.e. *N. crassa*, the two sequences branch separately, each being akin to sequences encoded by the same type of genome.

To examine the effect of adding a number of branches to the nuclear clade, phylogenetic trees were then generated from an alignment that included the eight partial RNAP sequences obtained by PCR amplification (Cermakian et al. 1996) as well as a new sequence from the chytridiomycete fungus *Rhizophydium* sp. #136 (Fig. 3). The use of PCR to generate these gene sequences limits this particular analysis to the amplified region. One of the trees derived from this abbreviated alignment is presented in Fig 4. Although the SP6 sequence is not grouped with those from the other phages, the monophyly of phage sequences is evident when a longer sequence alignment is used in the phylogenetic analysis (see Fig. 2). The plasmid-encoded sequences are monophyletic, as are the nucleus-encoded ones; for these branches, the bootstrap values, although satisfactory, are not as high as in Fig. 2, most likely because the sequences are shorter. Finally, this topology indicates that the plasmid-encoded

and nucleus-encoded ssRNAPs (both of which are found in eukaryotic cells) do not share a recent common ancestor, although rapid sequence divergence among the plasmid-borne genes makes it impossible to rule out other, less parsimonious conclusions.

RELATIONSHIPS AMONG SSRNAPS

The clear separation of ssRNAP sequences into three clusters raises the question of whether they actually form an homologous family, or instead represent three distinct origins with subsequent convergent evolution. Close examination of the sequences and the alignment in Fig. 1 highlights several important aspects of the ssRNAP sequence database. In concert with tertiary structure considerations, these sequence features strongly support the conclusion that the three ssRNAP classes are derived from a common ancestral sequence.

The prototype of this family, the bacteriophage T7 RNAP, has been extensively studied to identify amino acid residues important for catalysis. Comparison of the three-dimensional structure of the T7 ssRNAP (Sousa et al. 1993) with those of other RNAPs (Sousa 1996) and examination of polymerase alignments (Delarue et al. 1990) have directed structure/function studies to several highly conserved amino acids. Among these, D537 and D812 are involved in the binding of two divalent metal ions at the active site and are, therefore, essential for catalysis (Osumi-Davis et al. 1992; Woody et al. 1996). Residues K631 and H811 are implicated in phosphodiester bond formation, and the latter residue appears to be involved in nucleotide binding as well (Osumi-Davis et al. 1992). In light of these data, it is significant that these signature residues are found in all sequences compiled in Fig. 1. Only in the RNAP encoded by the kalilo plasmid of *N. intermedia* has D812 been substituted (by E). Although this is a conservative change, the fact that RNAP activity is decreased 3400-fold in a D812E mutant of T7 RNAP (Bonner et al. 1992) suggests either that the kalilo-encoded enzyme is not expressed or that it has low RNAP activity. In this case, expression of the ssRNAP gene has been demonstrated (Vickery and Griffiths 1993).

Another mutation, Y639P, is located at a highly conserved site in these RNAPs (Delarue et al. 1990) and disturbs the dNTP/rNTP discrimination of the enzyme (Sousa and Padilla 1995), whereas a S641A mutation directs T7 RNAP to use dNTPs instead of rNTPs (Kostyuk et al. 1995). The double mutant Y639F/S641A can use both sets of nucleoside triphosphates (Kostyuk et al. 1995). Residue Y639 is present in all ssRNAPs, whereas S641 occurs exclusively in the phage enzymes; however, another hydroxy amino acid (T) is found at the adjacent position in all of the available mtRNAP sequences (Fig. 1; Cermakian et al. 1996).

The penultimate amino acid, F882, is important for rNTP binding (Patra et al. 1992; Sousa et al. 1993) and is present in phage and nucleus-encoded polymerases (Fig. 1); however, it is replaced by another large hydrophobic amino acid (I or L) in all plasmid sequences. Other positions that are characteristic of plasmid-encoded genes again raise the possiblity that these gene products might have little or no activity, although compensatory changes in other regions of the molecule could be operating in these cases. Moreover, at this stage, editing of mRNA to generate a required amino acid residue cannot be ruled out.

If it was the case that only those amino acids important for activity are conserved among the ssRNAP sequences, we might question the homology or monophyly of these sequences, suggesting as an alternative convergent evolution to a common structure and functionality. However, as can be seen in Fig. 1, sequence similarity extends well beyond these few catalytic motifs, and relatively large blocks of conserved sequence can be delineated in the alignment. Moreover, these blocks appear in all the sequences in the same linear order.

In order to support the homology of the sequences in our alignment, we performed a Monte Carlo analysis on these data. We calculated the basic statistics of the Monte Carlo randomization for each of the two parameters described in Methods. Our results reject the null hypothesis of randomness. The d_{max} value of the distance

matrix for ssRNAP sequences (7.7) is 10 standard deviations (S_D) from the mean determined on the random sequences (108.8, with a S_D of 10.2) and none of the individual maxima is smaller than the d_{max} of real sequences. This d_{max} is even significantly smaller than the mean of d_{min} in random sequence matrices (14.9, with a S_D of 1.0). On the other hand, the radius of the matrix of ssRNAP sequences (7.6) is less than that of any of the randomized sequences matrices and significantly smaller than their mean (93.8, with a S_D of 10.3). These significant results show that similarity in the ssRNAP alignment is well above the noise level.

Even though these results support common ancestry of the ssRNAP sequences, the possibility remains, however, that they originated through a type of modular evolution (Li, 1997). This possibility could be discounted if parts of the alignment of Fig. 1 showed topological congruency upon phylogenetic analysis. We thus divided the sequence alignment into four parts of approximately equal length (blocks A+B, C+D, E+F, G+H+I+J+K+L) and constructed phylogenetic trees from the four subsets, using parsimony and distance methods. All subsets yielded the same overall topology as in Fig. 2 and 4, with three well defined clusters. The only deviation from congruency is the SP6 sequence, which did not group with the T7/T3/K11 clade in the C+D and E+F phylogenies (with distance method, but not parsimony), similar to its behavior in Fig. 4.

Thus, there is little doubt that these three classes of ssRNAP gene originated from a common ancestral sequence. However, because the phylogenetic trees are unrooted, the order of this divergence cannot be inferred from the available data. On another hand, even if blocks in our alignment share common ancestry, modular evolution still can be invoked and is even likely in some cases. For instance, fungal nuclear sequences (from *S. cerevisiae* and *N. crassa*) are much longer than the others, mainly due to a N-terminal extension and an insertion between blocks K and L. In fact, the three-dimensional structure of the T7 RNAP suggests that these two regions are outside of the polymerase core (Sousa et al. 1993; Sousa 1996).

Origin of the Ancestral ssRNAP Gene

A major issue in the evolution of the ssRNAPs is what type of cell or genetic system could have given rise to their genes, and how this origin might explain their present-day distribution pattern. Because the phylogenetic and structural data support a unique origin for ssRNAPs, it is useful to consider possible evolutionary scenarios.

In addition to the several sequence motifs discussed above (Delarue et al. 1990), the T7 RNAP possesses an intriguingly high structural similarity to the Klenow fragment of DNAP I (Sousa 1996; Sousa et al. 1993). Many secondary structure elements are in the same order in the sequence and can readily be aligned at the level of tertiary structure. Sequence motifs found in ssRNAPs and DNAPs do not exist in multi-subunit RNAPs (Delarue et al. 1990). Moreover, as previously noted, single amino-acid changes are able to lower the ability of the ssRNAP of T7 phage to discriminate between rNTPs and dNTPs (Sousa and Padilla 1995), and can even change the substrate specificity from rNTPs to dNTPs (Kostyuk et al. 1995). More recently, Gao et al. (1997) have shown that mutation of a single residue in a reverse transcriptase (RT) results in a variant enzyme now capable of acting as an RNAP. These observations strongly suggest that ssRNAPs and DNAP I-like enzymes are homologous, i.e., that they arose by divergent evolution, in spite of the fact that homology is not clearly evident at the level of primary sequence similarity, which is limited to the few instances of sequence conservation within the catalytic pocket.

What can be said about the timing of the origin of the ancestral ssRNAP gene? To date, no member of this family has been found in any of the available complete prokaryotic genomic sequences (four eubacterial and one archaeal species; Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Kaneko et al. 1996; Blattner et al. 1997), nor has such a sequence emerged from the extensive database of partial eubacterial and archaeal genome sequences. Moreover, it is not clear what purpose a ssRNAP would have served even in the last common ancestor (cenancestor) of the three domains of life, considering that the transcription system in this entity

almost certainly utilized a multi-component RNAP (Klenk and Doolittle 1994). Among many prokaryotic viruses studied to date, only four (phages T3, T7, K11 and SP6) are known to encode a ssRNAP; moreover, the host organisms for these phages (*E. coli*, *Klebsiella*, *Salmonella*) are all relatively closely related members of the same subgroup of γ-Proteobacteria. Otherwise, ssRNAP sequences are limited in occurrence to the eukaryotic nucleus (where they encode the mtRNAP) and to plasmid-like DNAs in the mitochondria of certain plants and fungi.

Taken at face value, the available data do not support the view that the ssRNAP gene originated early in evolutionary history, although the data do not rigorously exclude this view, either. A relatively late origin is more consistent with the known, phylogenetically limited distribution of ssRNAP sequences; conversely, an early origin better explains the extent of sequence divergence among the three classes of ssRNAP gene and between these and DNAP genes. An early-origin scenario (prior to the separation of the three domains of life, or even early in eubacterial evolution) would demand numerous independent losses of an ssRNAP gene from prokaryotic genomes (assuming vertical descent of the nucleus-encoded gene from a homolog in a cenancestral genome) and/or the sequestration of this gene in a quasi-independent infectious entity, such as a bacterial virus, which as far as we know displays only a very narrow host range. At present, there is little we can say with confidence about the timing of the appearance of the ancestral ssRNAP gene and whether this origin occurred in a cellular or virus-like genome.

Possible Scenarios for the Origin of the Mitochondrial RNA Polymerase

Given the expectation that the proto-mitochondrial, endosymbiont genome should have encoded a multi-component, $\alpha_2\beta\beta'$-type enzyme, as the chloroplast genome still does (Bogorad 1991; Reith 1995), it is rather surprising that a nucleus-encoded ssRNAP serves as the mitochondrial transcription enzyme. In this regard, it is notable that the mitochondrial transcription factor required by yeast mitochondria for accurate

initiation of transcription has some sequence similarities with bacterial σ factors (Jang and Jaehning 1991). However, of greater significance is the fact that four eubacteria-like *rpo* genes have recently been discovered (Lang et al. 1997) in the mtDNA of *Reclinomonas americana* (Flavin and Nerad 1993), an early diverging mitochondria-containing relative of certain amitochondriate eukaryotes (retortamonads) (O'Kelly 1993; Brugerolle and Mignot 1990). This finding in particular lends credence to the view that the mitochondrial transcription system initially employed a eubacteria-like, multi-subunit RNAP, but that this mtDNA-encoded system was subsequently lost upon recruitment of a nucleus-encoded ssRNAP to act as the mtRNAP. This replacement must have occurred early in the evolution of the eukaryotic lineage because a T7-like RNAP gene has been identified in *Naegleria fowleri* (Cermakian et al. 1996), one of the earliest diverging of mitochondria-containing eukaryotes (Cavalier-Smith 1993). A very recent study provides evidence for the existence of a phage-like ssRNAP in chloroplasts of *Arabidopsis thaliana*, whose chloroplast genome still encodes a bacterial-like RNAP (Hedtke et al. 1997). Such data suggest that partial replacement of the latter type of RNAP by a ssRNAP has also taken place in chloroplasts.

In considering the evolutionary source of the nucleus-encoded ssRNAP gene, several scenarios can be entertained.

(1) *An ssRNAP gene was already present in the nuclear genome of the eukaryotic cell that served as host for the mitochondrial endosymbiosis.* The origin of the eukaryotic genome is currently controversial (Doolittle 1996), with scenarios ranging from direct descent from a common ancestor with the Archaea (Keeling and Doolittle 1995) to a fusion of two prokaryotic cells, one a gram-negative eubacterium and the other an archaeon (Gupta and Golding 1996). In either case, as noted above, there is currently no evidence of an ssRNAP homolog in either eubacterial or archaeal genomes, and our attempts to amplify such a gene by PCR from the DNA of early diverging eukaryotes (including amitochondriate ones) have so far been unsuccessful. Although there are many technical reasons why PCR amplification might fail in particular instances, the available data do suggest that a ssRNAP gene was introduced

into the eukaryotic nuclear genome at some early stage in the evolution of this lineage, subsequent to the acquisition of mitochondria.

(2) *An ssRNAP gene was present in the DNA of the eubacterial endosymbiont that contributed the mitochondrial genome, and was subsequently transferred to the nucleus.* There is abundant evidence that mitochondria originated from the rickettsial subdivision of the α-Proteobacteria (see Gray and Spencer 1996), and mitochondrion-to-nucleus gene transfer is a well-established feature of eukaryotic cell evolution (Gray 1992). Transfer during a cryptic endosymbiosis, such as that suggested (Henze et al. 1995) for the cytosolic glyceraldehyde-3-phosphate dehydrogenase of some protists, is also a possibility. Relatively little is known about the rickettsial group of α-Proteobacteria and their genomes, but the arguments outlined in (1), above, can also be advanced here against the idea that an ssRNAP gene was selectively present in the genome of the proto-mitochondrial endosymbiont. In previous work (Cermakian et al. 1996), we were unable to amplify an ssRNAP sequence from several α-proteobacterial DNAs, including those of *Rickettsia*, *Ehrlichia* and *Agrobacterium*. More significant perhaps is that in sequenced mtDNAs, including 11 complete protist mtDNA sequences recently determined by the Organelle Genome Megasequencing Program (see http://megasun.bch.umontreal.ca/ogmp/projects.html), there is no evidence of remnants of a proto-mitochondrial ssRNAP gene that might have served as the source of the nucleus-encoded mtRNAP gene. Although ssRNAP sequences have been identified in some fungal mtDNAs (Robison et al. 1996, 1997), these are clearly related to plasmid-encoded, not nucleus-encoded, ssRNAP sequences. Their presence in mtDNA is best rationalized as the result of incorporation of mitochondrial plasmid-like DNA sequences into mtDNA via a recombination event.

(3) *An ssRNAP gene was contributed by a bacterial virus that accompanied the proto-mitochondrial endosymbiont.* It has previously been suggested that a T7-like phage may have brought the ssRNAP gene into the eukaryotic host cell, perhaps by virtue of being present within the α-proteobacterial endosymbiont that contributed the mitochondrial genome (Gray 1989; Schinkel and Tabak 1989). One could imagine

that a phage genome might have served as a convenient vector for transfer of the ssRNAP gene to the nucleus. Arguing against this scenario is the fact that whereas T7-like phages are known to infect a select group of Proteobacteria (enterobacteria or other γ-Proteobacteria such as *Pseudomonas putida* and *Caulobacter crescentus*; Hausmann 1988), they are not known to be present in rickettsial-type α-Proteobacteria. A relationship with phage T7 has been suggested for the cyanobacterial virus LPP-1 on morphological and biochemical grounds (Sherman and Haselkorn 1970), raising the possibility that the range of T7-like phage hosts is in fact broader than currently supposed. However, at the moment there are no compelling molecular data showing that an ssRNAP is encoded by the genome of LPP-1 or any other member of the Podoviridae (the morphological group that includes phage T7), outside of T7 and its close relatives.

An additional consideration is that T7 and its relatives are virulent phages, whose multiplication leads to destruction of the eubacterial host cell. This makes it unlikely that a eubacterial endosymbiont harboring such a phage could have persisted for the length of time required to effect its evolutionary conversion to a mitochondrion. In this regard, it would seem that a temperate phage accompanying the proto-mitochondrial endosymbiont would be a much more likely source of a nucleus-encoded mtRNAP gene than a virulent phage, and it is entirely possible that such an entity exists within the α-Proteobacteria. Finally, although the nucleus-encoded and phage-encoded ssRNAPs clearly derive from a common ancestor, our data provide no support for the view that these two groups are *specifically* related to the exclusion of the plasmid-encoded sequences.

(4) *An ssRNAP gene was contributed by a plasmid-like agent that accompanied the proto-mitochondrial endosymbiont.* If mitochondrial linear plasmids are indeed of bacteriophage ancestry (see Introduction), it is possible that such an entity, rather than a phage *per se*, was present in the mitochondrial endosymbiont, and that a plasmid ssRNAP gene was the direct precursor of the nuclear gene. The plasmid-borne ssRNAP gene could have been transferred to the nucleus along with many other

mitochondrial genes. Known mitochondrial plasmids are stably maintained within the organelle, with no obvious effect on mitochondrial function, and such long-term maintenance would presumably have favored plasmid-to-nucleus gene transfer during the early stages of the mitochondrial endosymbiosis. However, as in the case of phage-encoded ssRNAP sequences, the phylogenetic evidence does not support a *specific* common ancestry of plasmid-encoded and nucleus-encoded ssRNAP genes. Moreover, mitochondrial plasmids that contain an ssRNAP gene have so far been identified only in certain angiosperms and fungi. This limited phylogenetic distribution, coupled with evidence of horizontal transfer in some cases, makes it difficult to argue that mitochondrial linear plasmids represent a transitional form in the acquisition of an ssRNAP gene by the nuclear genome, although it is entirely possible that a different plasmid-like DNA element, not directly related to contemporary mitochondrial plasmids, brought an ssRNAP gene into eukaryotes.

SUMMARY AND FUTURE PROSPECTS

From the existing data, we can draw a number of inferences about the origin and evolutionary relationships of known ssRNAP sequences. We suggest that the most likely origin of the ssRNAP gene was via duplication and divergence of a DNAP or reverse transcriptase gene. Phylogenetic and structural analysis supports the existence of three distinct classes of ssRNAP (phage-encoded, plasmid-encoded, and nucleus-encoded) sharing a common ancestor; however, the order of divergence of these clades, and therefore their specific relationships to one another, cannot be ascertained at present. The nuclear ssRNAP gene (encoding the mitochondrial RNAP in most eukaryotes) is clearly of monophyletic origin, and appears to have arisen early in the evolution of the eukaryotic cell, coincident with or shortly after the mitochondrial endosymbiosis. It is possible to propose scenarios in which the source of the mtRNAP gene was either a T7-like phage or a plasmid-like DNA, but the data are not strongly supportive of either of these possibilities. It is equally possible that either the phage-encoded or the plasmid-encoded ssRNAP gene, or both, are derived in evolution from the nuclear gene: current data are simply not able to distinguish the directionality of

possible gene transfers. It does seem unlikely that an ssRNAP gene was already present in the host nuclear genome prior to the mitochondrial endosymbiosis, or that it was introduced into the eukaryotic cell via the genome of a eubacteria-like endosymbiont.

Current data on phylogenetic distribution are also not strongly supportive of the idea that the ssRNAP gene is the remnant of a primitive transcription system that pre-dated and was the evolutionary precursor of the multi-component RNAPs used in the transcription of eubacterial, archaeal and eukaryotic nuclear genomes. Rather, it would appear that this gene emerged relatively late in evolution, around the time of the origin of mitochondria. This raises the possibility that the origin of the nucleus-encoded mtRNAP and the origin of the ssRNAP gene *per se* may be temporally related: i.e., that the postulated duplication and divergence of a DNAP or RT gene may have occurred either in the proto-mitochondrial genome (with subsequent transfer to the nucleus) or in the eukaryotic nuclear genome, perhaps after transfer of an initially mtDNA-encoded DNAP or RT gene to the nucleus. Subsequent horizontal transfer events might have led to incorporation of a copy of the nuclear ssRNAP gene into a select group of bacteriophage genomes (those of T7 and its relatives) and/or into linear mitochondrial plasmid-like DNAs, with subsequent sequence divergence (which would have had to have been both rapid and extensive) of the three classes of ssRNAP. In this regard, it is worth emphasizing that genes encoding single-subunit DNAPs are much more widely distributed among both bacteriophage and plasmid DNAs than are ssRNAP genes. In addition, DNAP genes of this type have now been found in several protist mitochondrial genomes (see http://megasun.bch.umontreal.ca/ogmp/projects.html), and reverse transcriptases are commonly encoded by mitochondrial group II introns and some fungal mitochondrial plasmids (Xiong and Eickbush 1990). Further exploration of a possible evolutionary link between RNAP and DNAP and/or RT genes would seem worthwhile, as would further study of those α-Proteobacterial species (especially minimally derived, free-living ones) that tree closest to mitochondria.

ACKNOWLEDGEMENTS

We thank Alice B. Rae (Université de Montréal) for critical reading of this manuscript, and Hans-Wolfgang Ackermann (Université Laval, Québec) and David F. Spencer (Dalhousie University) for helpful discussions. This work was supported by grants from NSERC Canada (to R.C.) and MRC Canada (MT-4124, to M.W.G, and MT-14028, to B.F.L.) and by an NSERC 1967 Science and Engineering Scholarship and a FCAR Scholarship (to N.C.). R.C., M.W.G. and B.F.L., who are fellows in the Program in Evolutionary Biology of the Canadian Institute for Advanced Research, gratefully acknowledge salary and interactions support from CIAR.

REFERENCES

Ackermann H-W, DuBow MS (1987) Viruses of prokaryotes, Volume 2, Natural groups of bacteriophages. CRC Press, Boca Raton, pp 72-81, 174-178

Barr DJS (1986) *Allochytridium expandens* rediscovered: morphology, physiology and zoospore ultrastructure. Mycologia 78:439-448

Blanco L, Salas M (1996) Relating structure to function in Φ29 DNA polymerase. J Biol Chem 271:8509-12

Blattner FR, Plunkett G III, Mayhew GF, Perna NT, Glasner FD (1997) National Center for Biotechnology Information; locus ECOLI; accession number U00096

Bogorad L (1991) Replication and transcription of plastid DNA. In: Bogorad L, Vasil IK (eds) The molecular biology of plastids, Acadaemic Press Inc., SanDiego, CA, pp 93-124

Bonner G, Patra D, Lafer EM, Sousa R (1992) Mutations in T7 RNA polymerase that support the proposal for a common polymerase active site structure. EMBO J 11:3767-3775

Braithwaite DK, Ito J (1993) Compilation, alignment and phylogenetic relationships of DNA polymerases. Nucl Acids Res 21:787-802

Brugerolle G, Mignot J-P (1990) Phylum Zoomastigina, Class Retortamonadida. In: Margulis L, Corliss JO, Melkonian M, Chapman DJ (eds) Handbook of Protoctista. Jones and Bartlett, Boston, MA, pp 259-265

Bult CJ et al (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. Science 273:1058-1073

Cavalier-Smith T (1993) Kingdom Protozoa and its 18 phyla. Microbiol Rev 57:953-994

Cermakian N, Ikeda TM, Cedergren R, Gray MW (1996) Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucl Acids Res 24:648-654

Chan BS, Court DA, Vierula PJ, Bertrand H (1991) The *kalilo* linear senescence-inducing plasmid of *Neurospora* is an invertron and encodes DNA and RNA polymerases. Curr Genet 20:225-237

Chen B, Kubelik AR, Mohr S, Breitenberger CA (1996) Cloning and characterization of the *Neurospora crassa cyt-5* gene: a nuclear-coded mitochondrial RNA polymerase with a polyglutamine repeat. J Biol Chem 271:6537-6544

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucl Acids Res 16:10881-10890

Court DA, Bertrand H (1992) Genetic organization and structural features of *maranhar*, a senescence-inducing linear mitochondrial plasmid of *Neurospora crassa*. Curr Genet 22:385-397

Delarue M, Poch O, Tordo N, Moras D, Argos P (1990) An attempt to unify the structure of polymerases. Protein Eng 3:461-467

Dietz A Weisser HJ, Kossel H, Hausmann R (1990) The gene for *Klebsiella* bacteriophage K11 RNA polymerase: sequence and comparison with the homologous genes of phages T7, T3 and SP6. Mol Gen Genet 221:283-286

Doolittle WF (1996) Some aspects of the biology of cells and their possible evolutionary significance. In: Roberts DM, Sharp P, Alderson G, Collins M (eds)

Evolution of microbial life, Cambridge University Press, Cambridge, UK, pp 1-21

Felsenstein J (1993) PHYLIP: Phylogeny Inference Package, Version 3.5c, University of Washington, Seattle, WA

Flavin M, Nerad TA (1993) *Reclinomonas americana* N. G., N. Sp., a new freshwater heterotrophic flagellate. J Euk Microbiol 40:172-179

Fleischmann RD et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. Science 269:496-512

Fraser CM et al (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403

Gao G, Orlova M, Georgiadis MM, Hendrickson WA, Goff SP (1997) Conferring RNA polymerase activity to a DNA polymerase: a single residue in reverse transcription controls substrate selection. Proc Natl Acad Sci USA 94:407-411

Grahn AM, Bamford JKH, O'Neill MC, Bamford DH (1994) Functional organization of the bacteriophage PRD1 genome. J Bacteriol 176:3062-3068

Gray MW (1989) The evolutionary origins of organelles. Trends Genet 5:294-299

Gray MW (1992) The endosymbiont hypothesis revisited. Int Rev Cytol 141:233-357

Gray MW and Spencer DF (1996) Organellar evolution. In: Roberts DM, Sharp P, Alderson G, Collins M (eds) Evolution of microbial life, Cambridge University Press, Cambridge, UK, pp 109-126

Griffiths AJF (1995) Natural plasmids of filamentous fungi. Microbiol Rev 59:673-685

Gupta RS, Golding GB (1996) The origin of the eukaryotic cell. Trends Biochem Sci 21:166-171

Hausmann R (1988) The T7 group. In: Calendar R (ed) The bacteriophages. Plenum Press, New York London, Vol 1, pp 259-289

Hedtke B, Börner T, Weihe A (1997) Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*. Science 277:809-811

Henze K, Badr A, Wettern M, Cerff R, Martin W (1995) A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc Natl Acad Sci USA 92:9122-9126

Hermanns J, Osiewacz HD (1992) The linear mitochondrial plasmid pAL2-1 of a long-lived *Podospora anserina* mutant is an invertron encoding a DNA and RNA polymerase. Curr Genet 22:491-500

Ikeda RA, Richardson CC (1987) Enzymatic properties of a proteolytically nicked RNA polymerase of bacteriophage T7. J Biol Chem 262:3790-3799

Ikeda RA, Chang LL, Warshamana GS (1993) Selection and characterization of a mutant T7 RNA polymerase that recognizes an expanded range of T7 promoter-like sequences. Biochemistry 32:9115-9124

Jang SH, Jaehning JA (1991) The yeast mitochondrial RNA polymerase specificity factor, MTF1, is similar to bacterial σ factors. J Biol Chem 266:22671-22677

Kaneko T et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109-136

Keeling PJ and Doolittle WF (1995) Archaea: narrowing the gap between prokaryotes and eukaryotes. Proc Natl Acad Sci USA 92:5761-5764

Kempken F, Hermanns J, Osiewacz HD (1992) Evolution of linear plasmids. J Mol Evol 35:502-513

Klenk H-P, Doolittle WF (1994) Archaea and eukaryotes versus bacteria? Curr Biol 4:920-922

Kostyuk DA, Dragan SM, Lyakhov DL, Rechinsky VO, Tunitskaya VL, Chernov BK, Kochetkov SN (1995) Mutants of T7 RNA polymerase that are able to synthesize both RNA and DNA. FEBS Lett 369:165-168

Kotani H, Ishizaki Y, Hiraoka N, Obayashi A (1987) Nucleotide sequence and expression of the cloned gene of bacteriophage SP6 RNA polymerase. Nucl Acids Res 15:2653-2664

Kuzmin EV, Levchenko IV, Zaitseva GN (1988) S2 plasmid from cms-S-maize mitochondria potentially encodes a specific RNA polymerase. Nucl Acids Res 16:4177

Lang BF, Burger G, Doxiadis I, Thomas DY, Bandlow W, Kaudewitz F (1977) A simple method for the large-scale preparation of mitochondria from microorganisms. Anal Biochem 77:110-121

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-497

Levings CS III, Sederoff RR (1983) Nucleotide sequence of the S-2 mitochondrial DNA from the S cytoplasm of maize. Proc Natl Acad Sci USA 80:4055-4059

Li, W-H (1997) Molecular evolution. Sinauer Associates Inc, Sunderland MA

Lim L, Howlett BJ (1994) Linear plasmids, pLm9 and pLm10, can be isolated from the phytopathogenic ascomycete *Leptosphaeria maculans* by pulsed-field gel electrophoresis. Curr Genet 26:276-280

Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall, London New York Tokyo Melbourne Madras

Masters BS, Stohl LL, Clayton DA (1987) Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. Cell 51:89-99

McAllister WT (1993) Structure and function of the bacteriophage T7 RNA polymerase (or, the virtues of simplicity). Cell Mol Biol Res 39:385-391

McGraw NJ, Bailey JN, Cleaves GR, Dembinski DR, Gocke CR, Joliffe LK, MacWright RS, McAllister WT (1985) Sequence and analysis of the gene for bacteriophage T3 RNA polymerase. Nucl Acids Res 13:6753-6766

Meinhardt F, Kempken F, Kämper J, Esser K (1990) Linear plasmids among eukaryotes: fundamentals and application. Curr Genet 17:89-95

Moffatt BA, Dunn JJ, Studier FW (1984) Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase. J Mol Biol 173:265-269

Oeser B, Tudzynski P (1989) The linear mitochondrial plasmid pClK1 of the phytopathogenic fungus *Claviceps purpurea* may code for a DNA polymerase and an RNA polymerase. Mol Gen Genet 217:132-140

Oeser B, Gessner-Ulrich K, Deing P, Tudzynski P (1993) pClK1 and pClT5 - two linear mitochondrial plasmids from unrelated *Claviceps purpurea* strains: a comparison. Plasmid 30:274-280

O'Kelly CJ (1993) The jakobid flagellates: structural features of *Jakoba*, *Reclinomonas* and *Histiona* and implications for the early diversification of eukaryotes. *J Euk Microbiol* 40:627-636

Osumi-Davis PA, de Aguilera MC, Woody RW, Woody A-YM (1992) Asp537, Asp812 are essential and Lys631, His811 are catalytically significant in bacteriophage T7 RNA polymerase activity. J Mol Biol 226:37-45

Palenik B (1992) Polymerase evolution and organism evolution. Curr Opin Genet Dev 2:931-936

Patra D, Lafer EM, Sousa R (1992) Isolation and characterization of mutant bacteriophage T7 RNA polymerases. J Mol Biol 224:307-318

Rechinsky VO, Kostyuk DA, Lyakhov DL, Chernov BK, Kochetkov SN (1993) Random mutagenesis of the gene for bacteriophage T7 RNA polymerase. Mol Gen Genet 238:455-458

Reith M (1995) Molecular biology of rhodophyte and chromophyte plastids. Annu Rev Plant Physiol Plant Mol Biol 46:549-575

Robison MM, Royer JC, Horgen PA (1991) Homology between mitochondrial DNA of *Agaricus bisporus* and an internal portion of a linear mitochondrial plasmid of *Agaricus bitorquis*. Curr Genet 19:495-502

Robison MM, Horgen PA (1996) Plasmid RNA-polymerase-like mitochondrial sequences in *Agaricus bitorquis*. Curr Genet 29:370-376

Robison MM, Kerrigan RW, Horgen PA (1997) Distribution of plasmids and a plasmid-like mitochondrial sequence in the genus *Agaricus*. Mycologia 89:43-47

Rohe M, Schründer J, Tudzynski P, Meinhardt F (1992) Phylogenetic relationships of linear, protein-primed replicating genomes. Curr Genet 21:173-176

Salas M (1988) Phages with protein attached to the DNA ends. In: Calendar R (ed) The bacteriophages. Plenum Press, New York London, Vol 1, pp 169-191

Schinkel AH, Tabak HF (1989) Mitochondrial RNA polymerase: dual role in transcription and replication. Trends Genet 5:149-154

Sherman LA, Haselkorn R (1970) LPP-1 infection of the blue-green alga *Plectonema boryanum*. J Virol 6:841-846

Sousa R (1996) Structural and mechanistic relationships between nucleic acid polymerases. Trends Biochem Sci 21:186-190

Sousa R, Padilla R (1995) A mutant T7 RNA polymerase as a DNA polymerase. EMBO J 14:4609-4621

Sousa R, Chung YJ, Rose JP, Wang B-C (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. Nature 364:593-599

Swofford DL (1993) PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1, Illinois Nat. Hist. Surv., Champaign, IL

Vickery DB, Griffiths AJ (1993) Transcription of the kalilo linear senescence plasmid from *Neurospora intermedia*. Plasmid 29:180-192

Weihe A, Hedtke B, Börner T (1997) Cloning and characterization of a cDNA encoding a bacteriophage-type RNA polymerase from the higher plant *Chenopodium album*. Nucleic Acids Res 25 :2319-2325

Woody A-YM, Eaton SS, Osumi-Davis PA, Woody RW (1996) Asp537 and Asp812 in bacteriophage T7 RNA polymerase as metal ion-binding sites studied by EPR, flow-dialysis, and transcription. Biochemistry 35:144-152

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362

Yoshikawa H, Ito J (1982) Nucleotide sequence of the major early region of bacteriophage Φ29. Gene 17:323-335

Table 1. RNA polymerase sequences used in this study

Organism name	Gene name	Gene source	Reference	Acc. No.
bacteriophage T7	gene 1	phage genome	Moffat et al. 1984	M38308
bacteriophage T3	gene 1	phage genome	McGraw et al. 1985	X02981
bacteriophage K11	gene 1	phage genome	Dietz et al. 1990	X53238
bacteriophage SP6	gene 1	phage genome	Kotani et al. 1987	Y00105
Saccharomyces cerevisiae	*RPO41*	nuclear genome	Masters et al. 1987	M17539
Neurospora crassa	*cyt-5*	nuclear genome	Chen et al. 1996	L25087
Chenopodium album	-	nuclear genome	Weihe et al. 1997	Y08067
Arabidopsis thaliana	-	nuclear genome	unpublished	Y09432
Triticum aestivum	-	nuclear genome	Cermakian et al. 1996	U34402
Oryza sativa	-	nuclear genome	Cermakian et al. 1996	U34283
Pycnococcus provasolii	-	nuclear genome	Cermakian et al. 1996	U34286
Acanthamoeba castellanii	-	nuclear genome	Cermakian et al. 1996	U34405
Isochrysis sp. Tahiti	-	nuclear genome	Cermakian et al. 1996	U34284
Thraustochytrium aureum	-	nuclear genome	Cermakian et al. 1996	U34285
Cryptomonas Φ	-	nuclear genome	Cermakian et al. 1996	U34404
Tetrahymena pyriformis	-	nuclear genome	Cermakian et al. 1996	U34406
Naegleria fowleri	-	nuclear genome	Cermakian et al. 1996	U34403
Rhizophydium sp.	-	nuclear genome	this study	AF000226
Neurospora intermedia	ORF-1	plasmid kalilo	Chan et al. 1991	X52106
Neurospora crassa	ORF-1	plasmid maranhar	Court and Bertrand 1992	X55361
Gelasinospora sp.	ORF-1	plasmid kalilo	Griffiths 1995	L40494
Agaricus bitorquis	ORF2	plasmid pEM	Robison et al. 1991	X63075
Zea mays	ORF1	plasmid S-2	Levings and Sederoff 1983 Kuzmin et al. 1988	J01426
Claviceps purpurea	ORF2	plasmid pClK1	Oeser and Tudzynski 1989	X15648
Claviceps purpurea	ORF2	plasmid pClT5	Oeser et al. 1993	X68490
Podospora anserina	ORF2	plasmid pAL2-1	Hermanns & Osiewacz 1988	X60707

Figure 1. Alignment of complete ssRNAP amino acid sequences. Blocks of conserved sequence are delineated, with the number of amino acids in the more variable regions between each block given in parentheses. Sequences from linear plasmids are indicated by a (p) beside the name of the organism. The number of amino acids upstream of block A in sequences from *A. bitorquis*, *Z. mays* and *C. purpurea* plasmids is tentative. In all cases, block L represents the 17 last amino acids of the protein, except in *Z. mays* RNAP, where there is an additional residue at the end (Y). Numbers at the top of each block refer to positions in the phage T7 RNAP sequence. Residues discussed in the text are set in bold. Gaps are denoted by dashes (-). Residues identical and similar in more than 80% of the sequences are indicated under the alignment by a '*' and a '^', respectively. The total number of amino acids in each sequence is given after block L.

```
                                308                 A                  340    414                 B                  450
T7                       (307) YEDVY-MPEVYKAINIAQNTAWKINKKVLAVANV (73) IWFPYNMDWRGRVYA-VSM-FNPQGNDMTKGLL-TLAKGK (19)
T3                       (308) YEDVY-MPEVYKAVNLAQNTAWKINKKVLAVVNE (73) IWFPYNMDWRGRVYA-VPM-FNPQGNDMTKGLL-TLAKGK (19)
K11                      (330) YADVH-MPEVYKAVNLAQNTPWKVNKKVLAVVNE (73) IWFPYNMDWRGRVYA-VSM-FNPQGNDMTKGSL-TLAKGK (19)
SP6                      (280) LTQKQ-MPKVYKAINALQNTQWQINKDVLAVIEE (90) IYFVYAMDSRSRVYVQSST-LSPQSNDLGKALL-RFTEGR (20)
S. cerevisiae            (713) SDNGD-IDRVYDGLNVLGRTPWTVNRKVFDVVSQ (71) LYFPHNLDFRGRAYP-LSPHFNHLGNDMSRGLL-IFWHGK (19)
N. crassa                (668) IARGD-MDQVFKGLDVLGKTGWKINSPVFKVMLD (71) FYFPHNVDFRGRAYP-IPPYLNHMGADHVRGLM-LFAKGK (19)
A. thaliana              (444) TPKEQ-LEPVYEALDTLGNTKWKINKKVLSLVDR (70) FYYPHNVDFRGRAYP-IHPYLNHLGSDLCRGIL-EFCEGK (21)
C. album                 (456) ARDNN-LCTVFEALDTLGNTKWRVNKKVLSIIDR (70) FYYPHNLDFRGRAYP-MHPHLNHLGSDLCRGIL-EFAEGR (21)
N. intermedia (p)        (279) RHKIAITDKLYNVINKLNAFKFKINGDLLSYLQN (34) FYLNVNIDWRGRIYT-QSFYLDYQGSELSLALINLFEGKK (22)
N. crassa (p)            (345) KIPYTINLKLLDFITRDEFINRDKKDNVIIYKHI (47) FYMTVFIDWRGRFYT-SSCALNIQGGELARSLL-LFKEGQ (19)
Gelasinospora (p)        (279) SHKIAISDKLYSVINKLNAFKFKINADLLSYLQN (34) FYLNVNIDWRGRIYT-QSFYLGYQGSELSLALINLFEGQK (22)
A. bitorquis (p)         (289) RDHTEASEMLYSSINYLNSIKFKINKAMLNFILA (55) FYLPVYVDFRGRVYP-LSNYISYQGGDLARSLI-LFADTK (29)
Z. mays (p)              (480) HIDRERSQPVLSAVKKLQRQPYRINKLVYDFIQK (75) IYFPIFLDFRGRNYR-HGPF-HFHERDLVRSLI-IFDESD (33)
C. purpurea K1 (p)       (352) SRIQE-KNIIYESINKMMATAFQINKKLLKYLLE (60) LYFPLKLDHRGRLYP-ITAYFHYQSSELAKALL-LFAIPD (24)
C. purpurea T5 (p)       (360) STILE-NNVIYESINNMMSTSFKINSVLLRYLVE (60) LYFPLRLDHRGRLYP-VTAYFHYQSCELAKALL-LFAVPD (24)
P. anserina (p)          (380) TQILH-FNQIFDLVNNLSSIGYKINTDVLDFIEL (58) FFIPVNADFRGRVYC-TPEYLNYQSTDLAKSLL-LFSKPG (22)
                                           ^^  ^         ^ ^*  ^^  ^        ^     ^*^*** *            ^^ ^ *^   *
```

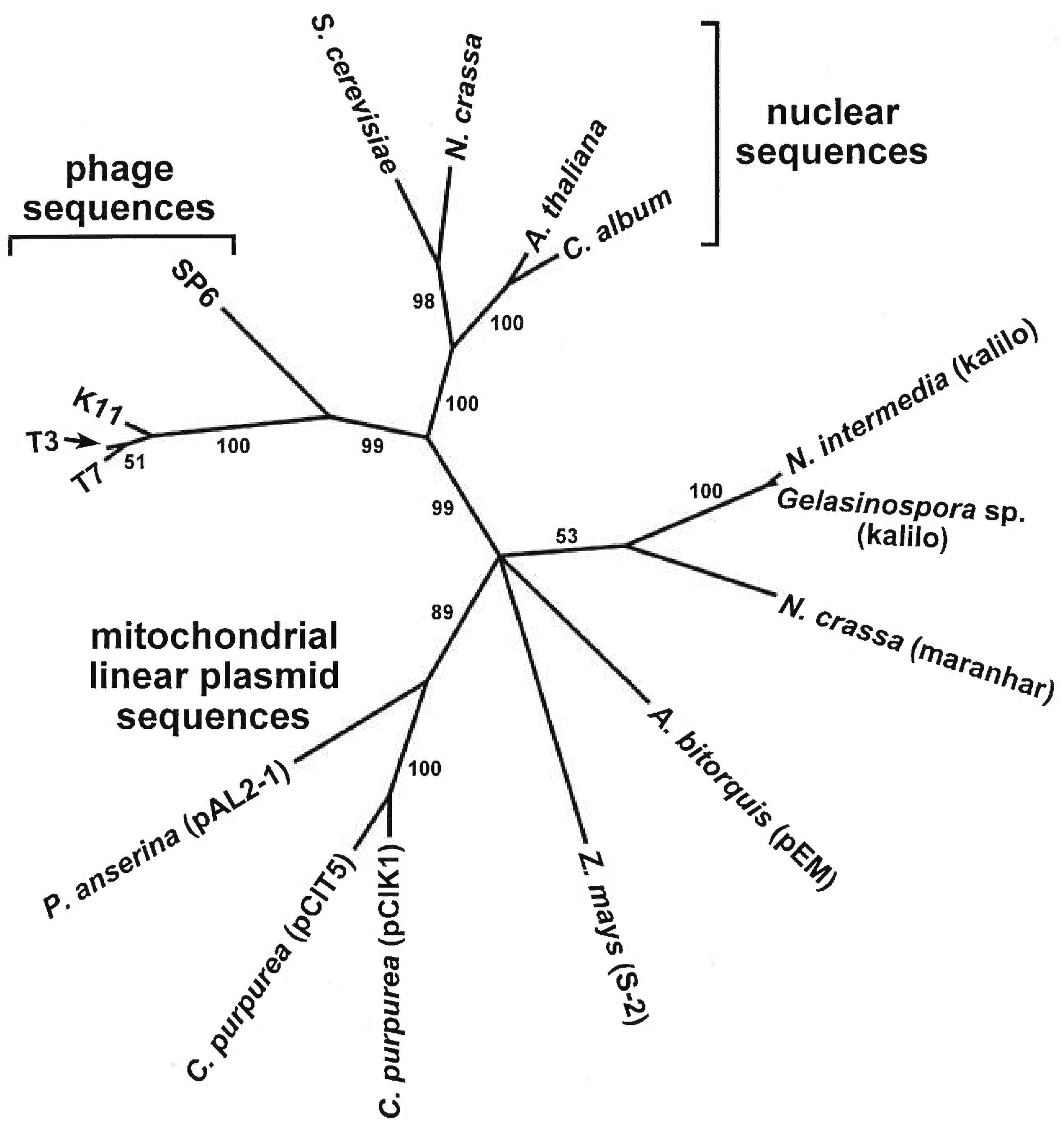
```
                          470          C              499   508                    D
T7                        VDKVPFPERI-KFIEENHENI-MACAKSPLEN (8) PFCFLAFCFE----YA---GVQHHGL-S--YNCSLPLAFDGSCSGIQHFS
T3                        VDKVPFPERI-AFIEKHVDDI-LACAKDPINN (8) PFCFLAFCFE----YA---GVTHHGL-S--YNCSLPLAFDGSCSGIQHFS
K11                       VDKVPFPERI-KFIEENEGNI-LASAADPLNN (8) PFCFLAFCFE----YA---GVKHHGL-N--YNCSLPLAFDGSCSGIQHFS
SP6                       WDKKTFDVRVSNVLDEEFQDMCRDIAADPLTF (8) PYEFLAWCFE----YAQYLDLVDEGR-ADEFRTHLPVHQDGSCSGIQHYS
S. cerevisiae             FDKLPLKDRV-AFTESHLQDI-KDSAEKSLTG (9) PWQALATCFE---LNE---VMKMDNPEE--FISHQPVHQDGTCNGLQHYA
N. crassa                 FDKASLQERQ-DFADENIENI-RDSVNNPLNG (9) PWQCLATCFE---LAA---ALELEDPTK--YVSHLPIHQDGTCNGLQHYA
A. thaliana               VDKLAYEDRI-AFTESHLEDI-FDSSDRPLEG (9) PFQCLAACIN---LSE---ALRSPFPEA--AISHIPIHQDGSCNGLQHYA
C. album                  VDKLSYEGRV-AFTENRLDDV-FDSADRPLEG (9) PFQCLAACIN---LTD---ALRSSSPED--YLSHIPIHQDGSCNGLQHYA
N. intermedia (p)         FSKKSFQDRF-NWVVENLDNI-IAMDKEFILK (3) PTLFAAFCLTMRKLKE---NPDYPVFN--------PIFLDATCSGVQHFA
N. crassa (p)             LDKRSKEERL-DWVEQNLHKI-IDIDNYEIWR (4) PLLFLACALELKGYNE---DPNFISHL--------PILMDATCNGLQHLS
Gelasinospora (p)         FSKKSFQDRF-NWVVENLDNI-IAMDKDFILK (3) PTLFAAFCLTMRKLKE---NPDHPVFN--------PIFLDATCSGVQHFA
A. bitorquis (p)          LTKVDGIIK--EYLESNEISN-TKYIEDN-ID (4) PFQFISI-MYAKLLSI---SNPKANISN-------PILFDASCSGIQHIA
Z. mays (p)               REALEFIYN--K-IEDMQSKP-TFFEKDIFVD (6) PFQYISSCISLKTYAD---TKDLSVLRY------TPVFQDASASAYQIMS
C. purpurea K1 (p)        LDKKSYDKRL-QWVDANWENI-INYRNNVLLE (4) KFMFLAYCLEIERLNQ---FLNDDSIEN--FKTYLPLQLDGTCNGFQHIV
C. purpurea T5 (p)        LDKKSYDKRL-QWVNENWANI-IDYKNNILLE (4) KFLFLSYCIEMVRLNQ---FFNDDSIQS--FQTYLPLQLDGTCNGFQHIA
P. anserina (p)           LDKLSANDRI-KWVDKNLNNI-KNYRNGKLIK (4) KFLFLAFCIEYNRYLN---CLDNHDVSW--FDTYLPIQMDATCNGFQHLS
                           ^*      *    ^ ^       ^         ^       ^  ^* *                          *^  *^^* * **
```

```
                                       D                574   619                 E                 656      676  F
T7                       AMLRDEVGGRAVNLLP------SETVQDIYGIV (44) QWLAHGVTRSVTKRSVMTLAYGSKEFGFRQQVLEDTIQ (19) YMAKLIWES
T3                       AMLRDEVGGRAVNLLP------SETVQDIYGIV (44) QWLAYGVTRSVTKRSVMTLAYGSKEFGFRQQVLDDTIQ (19) YMAKLIWDA
K11                      AMLRDSIGGRAVNLLP------SDTVQDIYKIV (44) QWLQYGVTRKVTKRSVMTLAYGSKESLVRQQVLEDTIQ (19) YMAKLIWDA
SP6                      AMLRDEVGAKAVNLKP------SDAPQDIYGAV (37) AWDSIGITRSLTKKPVMTLPYGSTRLTCRESVIDYIVD (14) HPFEDDRQD
S. cerevisiae            ALGGDVEGATQVNLVP------SDKPQDVYAHV (19) KILKDKITRKVVKQTVMTNVYGVT-YVGATFQIAKQLS (19) YLTKHVFSA
N. crassa                ALGGDTWGAQQVNLVP------GDRPADVYSAV (19) KAMHGKITRKVVKQTVMTNVYGVT-YVGARKQVLKQIE (19) YVTQHIFRA
A. thaliana              ALGRDKLGADAVNLVT------GEKPADVYTEI (25) KLMLDQVDRKLVKQTVMTSVYGVT-YSGARDQIKKRLK (16) YAAKITLKA
C. album                 ALGRDKLGAEAVNLVA------GGKPADVYSGI (25) KLLLNQVDRKLVKQTVMTSVYGVT-YIGVRDQIKRRLK (16) YAAKTTMTA
N. intermedia (p)        AMLLDLELGKYVNLINS-----GESVNDFYSQL (19) KFSDISLNRSLLKKVIMTKSYNVTTYGITEQLKSKLEK (18) VPTKNGDFV
N. crassa (p)            AMVNDFVLAEKVNLLKST---ENDNPRDLYSEV (18) NLERINVERCLVKRGLMTITYGATERGIYDQIVSKFFQ (15) DSDIAPKDV
Gelasinospora (p)        AMLLDLELGKYVNLINS-----GESVNDFYSQL (19) KFSDISLNRSLLKKVIMTKSYNVTTYGITEQLKSKLEK (18) VPTKNGKFV
A. bitorquis (p)         ALTLEKELASNVNLYTDSSNPKEDYPQDFYTYA (14) ELRDIKLNRKIIKRSVMTIPYNISMAGIGEHLMEHFTV (13) SATISSKDV
Z. mays (p)              YFLLDIDYGIHTNLLKKT-NTDGRYIRDIYEFM (32) AKIVSIFDRNVVKKMFMPMMYGKTDYTLKKDVEDL---  (0) LKGKSDSEG
C. purpurea K1 (p)       LLSNEVKLYSQLNLDG---KSKNQDPSDFYSFM (25) RLINLGLSRANIKQAIMTKPYNAKDFTVAKYIIETLLI (18) YKVNETMTN
C. purpurea T5 (p)       LLSNEVRLYEKLNLDG---KPKSQDPSDFYTFM (25) RLINLGLSRANIKHAIMTKPYNAKDFTVTEYIIDTMLI (18) YKVNDTMTN
P. anserina (p)          LLSLDSNLSKELNLSE---STWDDVPKDFYTFL (27) RLINMKIIREIIKKGIMTIPYNVSNFSLINYIREGFEL  (7) YIYKNEPSI
                          ^   *    ^  ^**                * *  ^             ^ * ^ *  ^**  *
```

```
                                   F                708   723     G    736   737   H   748   766        I
T7                       --VSVTVVAAVEAMNWLKSAAKLLAA (14) CAVHWVTPDGFPVW (0) QEYKKPIQTRLN (17) DSEIDAHKQESGIAPNFVHSQ
T3                       --VSVTVVAAVEAMNWLKSAAKLLAA (14) CAVHWTTPDGFPVW (0) QEYRKPLQKRLD (17) DSGIDAHKQESGIAPNFVHSQ
K11                      --VTVTVVAAVEAMNWLKSAAKLLAA (14) CAIHWVTPDGFPVW (0) QEYRKQNQARLK (17) DSEIDAHKQESGIAPNFVHSQ
SP6                      YLTPGAAYNYMTALIWPSISEVVKAP (18) EGLMYTLPTGFILE (0) QKIMATEMLRVR (14) TDIVDEAAMMGAAAPNFVHGH
S. cerevisiae            --IRELFHSAHLIQDWLGESAKRISK (20) SSVIWTTPLGLPIV (0) QPYREESKKQVE (13) VNPVNARRQKAGLPPNFIHSL
N. crassa                --MSTMFKGAHDIQNWLGEIGGRVCR (50) STIIWTTPLRMPVV (0) QPYRKHGTKTVS (13) SDPVNRRKQLQAFPPNFIHSL
A. thaliana              --LEEMFEAARAIKSWFGDCAKIIAS  (2) NAVCWTTPLGLPVV (0) QPYRKPGRHLVK (11) TDKVMARRQMTAFAPNFIHSL
C. album                 --LEEMFEAARSIMGWLGDCAKIIAS  (2) QPARWSTPLGLPVV (0) QPYRNHGRRLVK (11) TDKVMAKRQRTAFPPNFVHSL
N. intermedia (p)        --V-LDTFEVETLASIINDNIFNQFP (20) IPLSWSTPDGLELT (0) QRYNLSKVKKLT (15) VNEKDSRREVQAIIPNIIHSL
N. crassa (p)            --V-FTQKNILLLSKIIYNSLFKIHP (20) LPVNWVTPYGLVIQ (0) QKYNKFTKYNET (16) TTSISKQKQIQAFIPNFVHSM
Gelasinospora (p)        --V-LDTFEVETLASIINDNIFNQFP (20) IPISCSTPDGLELT (0) QRYNLSKVQKLT (15) VSDKDSRREVPAIIPNIIHSL
A. bitorquis (p)         --Y-LDYSKYGQLCKIIYFVLTKELP (20) IPITWVTPSGLKIK (1) TNIKFKPQ-KVK (15) TDSLDVLSTKRSFMPNFIHSL
Z. mays (p)              --I-NLISKHISTYWKVNFGKMKDLM (20) TPY-WVTLQTYKRR (1) RVKMKIQYETTK (16) LNDNDIRKSSTSTFANFIHQK
C. purpurea K1 (p)       --V-ACFRDIYLLAKCINEILFLNYP (20) LPITWSLPTGLEVH (0) QQYMRSKLKKTK (15) KIKMDKVKQQIALMPNLIHSL
C. purpurea T5 (p)       --V-ACYKDIYLLSRCINEILFLNYP (20) LPIVWNLPSGLQVH (0) QKYMESLTTKTR (15) KENINKSKQSIALSPNLIHSL
P. anserina (p)          --R-LKSLDFTVLGKGLRKVLHEKFH (20) IPIIWTLPTGLVVR (0) QSYMIRDEVRIK (15) KTQFNNSKQITAFMPNLVHSL
                            ^              ^             ^ *  * *  ^      * *                          ^   ^  ** ^**
```

	I 794		804 J 818		819 K 837		867 L 883	Total
T7	DGSHLRKT	(9)	IESFALI**HD**SFGTIP	(0)	ADAANLFKAVRETMVDTYE	(29)	KGNLNLRDILESDFA**F**A	883
T3	DGSHLRMT	(9)	IESFALI**HD**SFGTIP	(0)	ADAGKLFKAVRETMVITYE	(29)	KGNLNLQDILKSDFA**F**A	884
K11	DGSHLRMT	(9)	IDSFALI**HD**SSGTIP	(0)	ADAGNLFKAVRETMVKTYE	(29)	KGDLNLRDILESDFA**F**A	906
SP6	DASHLILT	(8)	VTSIAVI**HD**SFGTHA	(0)	DNTLTLRVALKGQMVAMYI	(26)	QGEFDLNEIMDSEYV**F**A	874
S. cerevisiae	DASHMLLS	(7)	GLDFASV**HD**SYWTHA	(0)	SDIDTMNVVLREQFIKLHE	(120)	KGDFDVTVLRNSQYF**F**S	1351
N. crassa	DASHMILS	(7)	GLTFAAV**HD**SFWTHA	(0)	SDIDSMNAVLRDAFIRIHS	(201)	KGDFDVRSLKDSTYF**F**S	1422
A. thaliana	DGSHMMMT	(7)	GLSFAGV**HD**SFWTHA	(0)	CDVDVMNTILREKFVELYE	(25)	RGDFDLRKVLESTYF**F**N	976
C. album	DGSHMMMT	(7)	GMNFAGV**HD**SYWTHA	(0)	CDVDKMNQILREKFVELYE	(25)	RGDFDLREVLESPYF**F**N	988
N. intermedia (p)	DASHLTMI	(4)	DSYILPI**HE**CFGTHP	(0)	NDMYKLAEQVRECFILLYS	(51)	GELNVEDIRDMGKYM**I**S	811
N. crassa (p)	DGSNIVLL	(9)	KINFASI**HD**CFATHA	(0)	NDTAWLSWYVKQSFIRIYS	(56)	KDNKIYKEILHSEYF**I**N	896
Gelasinospora (p)	DAAHLPML	(4)	DSYILPI**HD**CFGTHP	(0)	NDMFKLAEQVRECFILLYS	(51)	GELNVEDIRDMGKYM**I**S	811
A. bitorquis (p)	DA-SNVHL	(8)	NLPVYTV**HD**CFASTA	(4)	KLEK-LV---KNAFINIYF	(80)	NINEFVKGILNSKYF**I**G	874
Z. mays (p)	DAFTAIQL	(11)	SIPIYAV**HD**NFITMP	(15)	RMGHPLIIINKFLFDHILI	(125)	GTQADSLDKGEDDYC**I**H	1159
C. purpurea K1 (p)	DAASLIML	(11)	VVNFYSV**HD**CYGVTA	(0)	KYIDLLISQLRAVYIELYS	(55)	HIRKAYEELAKANMF**I**K	936
C. purpurea T5 (p)	DAASLILL	(11)	PVNFYGV**HD**CYGVTA	(0)	KFIDLLISHLRAVYIELYS	(56)	LVHKAYEELSKANMF**I**K	945
P. anserina (p)	DAASLTLL	(10)	VKNIYTI**HD**CFAVPA	(0)	NKMECLISLLKLTYIKLYS	(53)	PSDFDFNVLKKSSYI**L**N	948
	*^ ^ ^		^** ^ ^		^ ^^ ^ ^*		^ ^	

Figure 2. Maximum parsimony phylogenetic tree derived from the alignment of Fig. 1. The bootstrap values (100 resamplings) are indicated on the branches. Branches with bootstrap values lower than 50 were collapsed. The three groups of ssRNAP sequences discussed in the text are indicated.

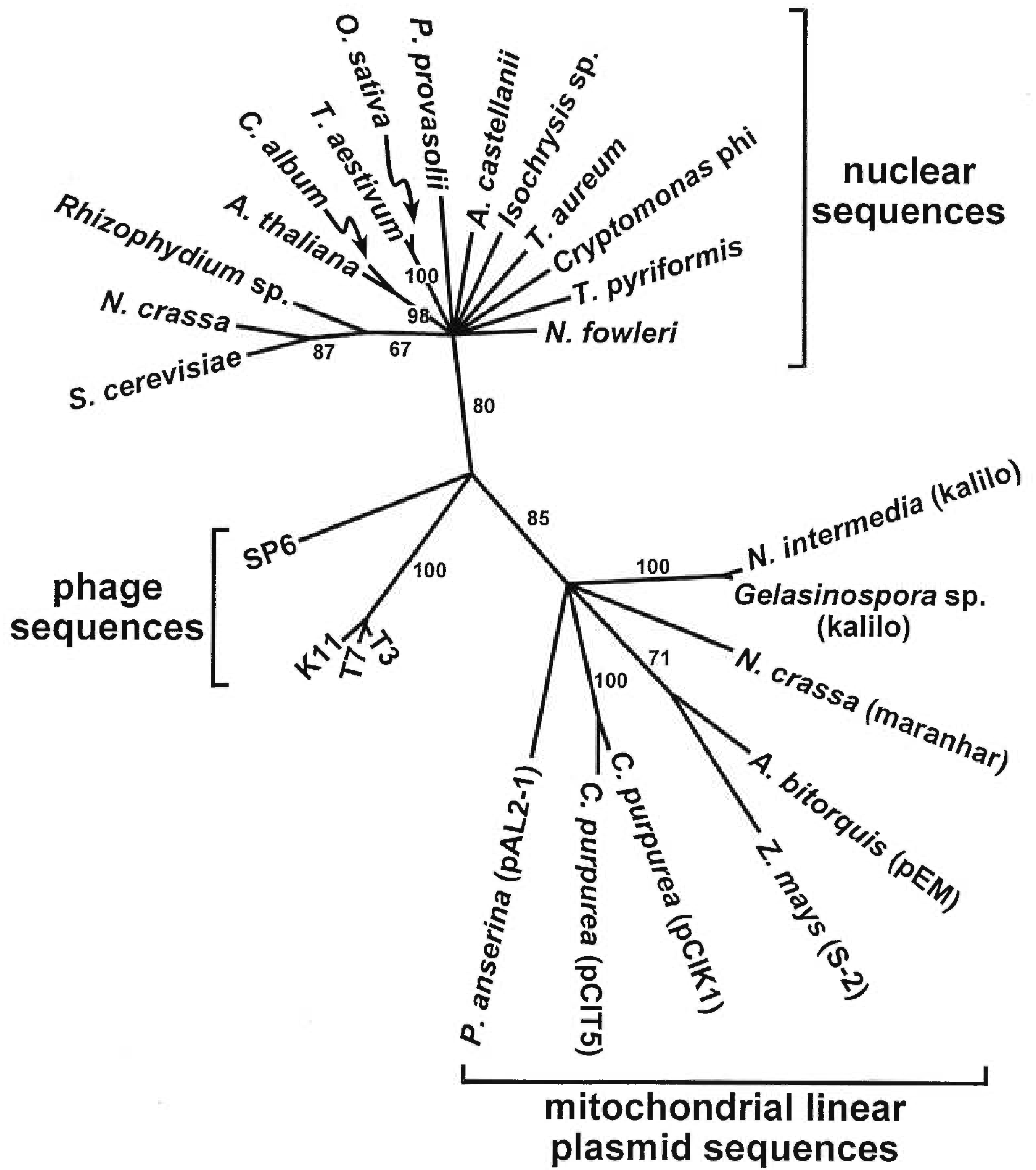


phage sequences
nuclear sequences
mitochondrial linear plasmid sequences
S. cerevisiae
N. crassa
A. thaliana
C. album
SP6
K11
T3
T7
51
100
99
98
100
100
99
N. intermedia (kalilo)
Gelasinospora sp. (kalilo)
100
53
N. crassa (maranhar)
89
A. bitorquis (pEM)
Z. mays (S-2)
P. anserina (pAL2-1)
100
C. purpurea (pCIT5)
C. purpurea (pClK1)

Figure 3. Partial ssRNAP sequence obtained from the chytridiomycete fungus, *Rhizophydium* sp. #136.The protein sequence was deduced from the sequence of a PCR-amplified DNA product obtained using *Rhizophydium* DNA, as described in the text. The sequence is aligned with that of the corresponding region in *Saccharomyces cerevisiae* mtRNAP and bacteriophage T7 RNAP, and identical and similar amino acids are highlighted white on black and gray respectively. Numbers under the alignment refer to positions in the phage T7 RNAP sequence. Three putative small introns were inferred, based on conservation of reading frame, optimization of sequence alignment and identification of GU-AG intron boundaries. The positions of these introns are denoted by arrows, with the accompanying numbers indicating intron size in bp. The *Rhizophydium* sequence has been deposited in the GenBank database of NCBI (acc. no. AF000226).

```
                 41
                 ↓
Rhizophydium sp. NTYGVTEVGARDQIFNR-LK---EARAKC--------ALYITRKVFESMSDMFEGARAIQN
S. cerevisiae    NVYGVTYVGATFQIAKQ-LSPIFDDRKESL-----DFSKYLTKHVFSAIRELFHSAHLIQD
T7               LAYGSKEFGFRQQVLEDTIQPAIDSGKGPMFTQPNQAAGYMAKLIWESVSVTVVAAVEAMN
                 637                                                        697

                 118                                      22
                  ↓                                        ↓
Rhizophydium sp. WLSKGAN-----FRMQR-----RNSKAPPKMTSMIWTSPLGFPIVQPYRKLGFDHVK-TFM
S. cerevisiae    WLGESAKRISKSIRLDVDEKSFKNGNKPDFMSSVIWTTPLGLPIVQPYREESKKQVE-TNL
T7               WLKSAAK------LLAAEVKDKKTGEILRKRCAVHWVTPDGFPVWQEYKKPIQTRLNLMFL
                 698                                                        752

Rhizophydium sp. QTFSI---IDDKKPSPVNSMKQASAFPPNFVHSLDASHMMLTAM-ACLAQNVT-FAAVHDS
S. cerevisiae    QTVFI---SDPFAVNPVNARRQKAGLPPNFTHSLDASHMLLSAA-ECGKQGLD-FASVHDS
T7               GQFRLQPTINTNKDSEIDAHKQESGIAPNFVHSQDGSHLRKTVVWAHEKYGIESFALIHDS
                 753                                                        813
```

Figure 4. Maximum parsimony phylogenetic tree generated as in Fig. 2, but using an alignment of partial protein sequences, corresponding to the region shown in Fig.3. Branches with bootstrap values lower than 50 were collapsed. The *N. fowleri* sequence used in the alignment is shorter than the other sequences (31 fewer amino acids; see Cermakian et al. 1996).


nuclear sequences
O. sativa
P. provasolii
A. castellanii
Isochrysis sp.
T. aureum
Cryptomonas phi
T. pyriformis
N. fowleri
C. album
T. aestivum
A. thaliana
Rhizophydium sp.
N. crassa
S. cerevisiae
100
98
87
67
80
85
phage sequences
SP6
100
K11
T7
T3
N. intermedia (kalilo)
100
Gelasinospora sp. (kalilo)
71
N. crassa (maranhar)
100
A. bitorquis (pEM)
Z. mays (S-2)
C. purpurea (pClK1)
C. purpurea (pClT5)
P. anserina (pAL2-1)
mitochondrial linear plasmid sequences

CHAPITRE 2

La relation entre un facteur de transcription mitochondrial et une protéine nucléaire spécifique aux testicules chez la souris

Article 3

The relationship between the mitochondrial transcription factor A and the testis-specific HMG box protein in mouse

ARTICLE 3

The relationship between the mitochondrial transcription factor A and the testis-specific HMG box protein in mouse

(La relation entre le facteur de transcription mitochondrial A et la protéine à boîtes HMG spécifique aux testicules chez la souris)

Cermakian, N. et Cedergren, R.

Les contributions de N.C. à cet article sont:

- Toutes les expériences.
- Rédaction du manuscrit.

The relationship between the mitochondrial transcription factor A and the testis-specific HMG box protein in mouse

Nicolas Cermakian and Robert Cedergren*

Département de biochimie, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal (Québec), Canada, H3C 3J7

* Corresponding author: Tel.: (514) 343-6320; Fax: (514) 343-2177
E-mail: ceder@bch.umontreal.ca

July 19, 1996

ABSTRACT

A protein identified as a mouse testis-specific HMG box protein (m-tsHMG) is here shown to be very similar to the human mitochondrial transcription factor A (h-mtTFA). Using mouse DNA and primers designed from animal mtTFA sequences, a DNA fragment was amplified which corresponded to the mtTFA gene and contained an intron upstream of the first HMG box as in the h-mtTFA gene. Two additional exons are located upstream of this intron: the first encoding part of the m-tsHMG sequence and the second encoding a putative mitochondrial targeting sequence. Our data suggest that production of the m-tsHMG and the m-mtTFA is due to alternative splicing of the same pre-mRNA.

KEY WORDS: Mitochondria, testis, alternative splicing, HMG box protein, evolution.

ABBREVIATIONS: HMG, High Mobility Group; h-, xl- and sc-mtTFA, mitochondrial transcription factor A from human, *Xenopus laevis* and *Saccharomyces cerevisiae*, respectively; tsHMG, testis-specific HMG protein; PCR, Polymerase Chain Reaction; RT, reverse transcription; bp, base pair; 5'-RACE, rapid amplification of cDNA ends; P_{mt}, promoter for the mitochondrial isoform of the protein; P_t, promoter for the testis-specific isoform.

INTRODUCTION

The HMG box, a basic, α-helical structural domain of proteins, was first recognized in non histone high-mobility group proteins (HMG proteins) from chromatin [1,2]. Since this discovery, many proteins have been shown to contain one or more of these domains. Members of this family of proteins have been found in different subcellular localizations (nucleus, mitochondria), and fulfill a diverse range of cellular functions such as transcriptional activation, DNA packaging, etc. [1,2]. A common functional

theme is the capacity of these proteins to bind, bend and wrap DNA [1,3]. The mitochondrial transcription factor A from human and *Xenopus laevis* (h- and xl-mtTFA), which activate transcription from the light- and heavy-strand promoters of mitochondrial DNA, are examples of HMG-box proteins [4-6]. These mitochondrial factors contain two HMG boxes, separated by a hinge region, and a C-terminal tail, all of which are required for DNA binding and transcriptional activation [7]. Being encoded in the nucleus, they contain a signal sequence to target them to the mitochondria.

In 1993, Boissonneault and Lau [8] isolated a mouse cDNA encoding an HMG box protein during a search for testis-specific transcription factors and DNA-binding proteins. Intriguingly, their testis-specific HMG protein (tsHMG) contained two HMG boxes, a central hinge and a C-terminal tail, as in h- and xl-mtTFA. The structural and sequence similarities among these proteins prompted us to examine their relationship. We present here evidence that both mtTFA and tsHMG proteins in mouse are encoded by the same gene and that they are the product of alternative splicing involving the first two exons of the gene.

MATERIAL AND METHODS

Total RNA from 11-day old mouse embryos and mouse tail DNA were gifts of Dr. Mark Featherstone. Embryo polyA+ RNA was purified using Promega's PolyATract mRNA isolation system III. Mouse testis polyA+ RNA was purchased from Clontech.

Reverse transcription was done in 20 μL with 250 ng of polyA+ RNA in the presence of 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 0.5 mM of each dNTP, 15U RNAguard (Pharmacia), 10 mM DTT, 200U Superscript II reverse transcriptase (Gibco BRL) and 50 ng of a primer. Incubations were at 42°C for 35 min, then at 50°C for 15 min. The solution was heated at 70°C for 15 min and then treated with 2U of *E. coli* RNase H (Pharmacia) at 37°C for 20 min. Finally the

volume was adjusted to 100 μL. In the case of 5'-RACE, the solution was concentrated using a Centricon-30 concentrator (Amicon). The resulting cDNA solution was treated with 15U terminal deoxynucleotidyl transferase (Gibco-BRL) in a reaction volume of 25 μL.

PCR experiments were done in 100 μL containing 1 % of the RT reaction, 200 μM of each dNTP, 1 μM of each phosphorylated primer, 1X Taq or Vent DNA polymerase commercial buffer, 100 μg/mL BSA and 2.5U Taq polymerase (Pharmacia) plus 0.02U Vent polymerase (NEB). The Q_T primer (40 nM) supplemented the above reaction solution for the first PCR amplification in 5'-RACE experiments. In the case of DNA substrates, 2U of Vent polymerase only was used. After 5 min at 95°C, enzymes were added at 75°C, followed by 30 cycles of: 95°C, for 30 sec; 50 or 55°C, for 30 sec; 72°C, for 30-60 sec. Finally, the solution was heated at 72°C for 4 min. In the case of the first PCR amplification of the 5'-RACE experiments, a cycle of 2 min at 50°C and 40 min at 72°C was implemented prior to the 30 regular cycles. PCR experiments with DNA used 100 ng of DNA. When consecutive PCR amplifications were performed 1 μl of the first amplification was added to the solution of the second and no other template was used.

Gene-specific primer sequences (Fig. 1) used for the above experiments were:
TFA5' : 5'(T/C)TN(T/C)TNGGNAA(A/G)CCNAA(A/G)(A/C)GNCC3';
(derived from the amino acid sequence: ^{151}LLGKPKRP^{158}; numbers refer to positions in h-mtTFA precursor)
TFA3' : 5'CAT(C/T)TG(C/T)TC(C/T)TCCCAN(C/G)(A/T)(C/T)TTCAT(C/T)TC3';
(derived from the amino acid sequence : ^{214}EMSKWEEQM^{222})
tshmg1: $^{5'}$AGCCGAATCATCCTTTGCCTCCT$^{3'}$;
tshmg2: $^{5'}$GATACATAAATGTTATATGCTGAACG$^{3'}$;
tshmg5: $^{5'}$ATGGCGCTGTTCCGGGGAATGTGG$^{3'}$;
tshmg6: $^{5'}$CTTTGGATAGCTACCCATGCTGGA$^{3'}$;
tshmg8: $^{5'}$GGTATGAACTCATAGGTTTCTTTGG$^{3'}$;

tshmg9: $^{5'}$CTATGGACTGCCACACTACCAGC$^{3'}$.

The non-specific primers used in the 5'-RACE experiments were [9]:
Q_T: $^{5'}$CCAGTGAGCAGAGTGACGAGGACTCGAGCTCAAGC(T_{17})$^{3'}$
Q_0: $^{5'}$CCAGTGAGCAGAGTGACG$^{3'}$
Q_I: $^{5'}$GAGGACTCGAGCTCAAGC$^{3'}$

Other standard techniques were performed as in [10]. PCR products were cloned into pBluescript SK+ (Stratagene) and both strands were sequenced by the ddNTP method using Sequenase 2.0 (USB). Sequence alignments were performed using Multalin [11]. These alignments were refined manually and used to build phylogenetic trees with the parsimony method in PAUP [12] and the SEQBOOT, PROTDIST, FITCH and CONSENSE tools in PHYLIP [13]. The sequence presented in this work is available from GenBank under accession number U63712.

RESULTS

HIGH SIMILARITY OF H-MTTFA AND M-TSHMG

Searching the DNA sequence database for sequences similar to the mitochondrial transcription factor A (mtTFA), we discovered that the m-tsHMG sequence was very similar to that of h-mtTFA. Mouse tsHMG had 63% and 48% positional identity with the human and amphibian mtTFA, respectively. These sequences were then aligned with the sequences of other HMG-box proteins such as xl-mtTFA and the yeast mitochondrial DNA-binding protein ABF2 or sc-mtTFA [14,15], and the alignment served to build a phylogenetic tree using the parsimony method in PAUP [12] (Fig. 2). H-mtTFA and m-tsHMG were similar enough to be directly linked in this tree with a bootstrap value of 100, indicating that they were closer to each other than to any other protein used in the analysis. The divergence of the amphibian protein (xl-mtTFA)

from the mammalian mtTFAs must, therefore, have taken place prior to the divergence of the testis-specific HMG box protein.

Identity of m-tsHMG and m-mtTFA

Two degenerate oligonucleotides (TFA5' and TFA3') were designed from the most conserved regions of the h-, xl-mtTFA and m-tsHMG sequences (Fig. 1), and an RT-PCR experiment with these primers and mouse embryo RNA was performed to isolate the mouse mtTFA cDNA. Fig. 3 shows that a single band was amplified having a length, 216 bp, consistent with the other transcription factors. This fragment was cloned, and seven independent isolates were sequenced. The sequences were invariably identical to that of m-tsHMG.

Next, 5'-RACE experiments using the m-tsHMG-specific oligonucleotides tshmg1 and tshmg2 were initiated on mouse embryo RNA to pursue the search for RNAs which might encode the m-mtTFA and m-tsHMG proteins. Amplification products (Fig. 4A) were cloned and sequenced, and again, only clones having the identical sequence to that of the m-tsHMG gene were observed; however, these sequences contained a 5'-terminal extension encoding a putative signal peptide. The alignment of this peptide sequence with the corresponding sequence of the h-mtTFA signal sequence [5] demonstrates their high similarity (Fig. 4B). Since this N-terminal extension could form an amphipathic α-helix [16], this sequence is likely the signal peptide which directs the protein to the mitochondria, thus supporting the notion that we had indeed amplified a part of m-mtTFA mRNA.

Introns in the putative m-mtTFA gene

The h-mtTFA gene harbors introns, one of which is located upstream of the first HMG box, but just downstream of the mitochondrial signal peptide sequence [17]. We thus performed PCR amplification on mouse DNA, using primers tshmg5 and tshmg6 located on either side of a putative mouse intron. Sequencing of the PCR products

shown in Fig. 5 demonstrated that an intron in the m-tsHMG gene is present at the same position as intron 1 in the h-mtTFA gene.

Since the presence of this intron suggested the possibility that m-tsHMG and m-mtTFA mRNAs could be derived from the same pre-mRNA by alternative splicing, a 5'-RACE experiment was performed in parallel on mouse embryo and testis RNAs using primer tshmg8 for the RT and the first nested PCR, and tshmg6 or tshmg9 for the second amplification (Fig. 6). As expected, embryo RNA gave only products bearing the putative signal sequence (Fig.6, lane 4 and 5). In contrast, amplification using primer tshmg6 with testis RNA gave a shorter product (Fig. 6, lane 6), the sequence of which revealed that it was identical to the product amplified from embryo RNA, except for the region upstream of the intron 1 site. This novel fragment (exon 1') extends 45 bp upstream of the intron 1 site and contains a reading frame encoding an N-terminal methionine and 5 amino acids in phase with the frame of exon 2 (see Fig. 7). Since the four independent clones bearing exon 1' had identical 5'-termini, they are possibly the true 5' terminus of the mRNA. A signal sequence-bearing fragment was not obtained in this experiment with testis RNA; however, when using RT primer tshmg1 and PCR primers tshmg2 and tshmg6, the putative m-mtTFA sequence from testis could be amplified, albeit at lower levels than the testis-specific mRNA.

Using primer tshmg9 and testis RNA, an additional 300-bp long fragment was found (Fig. 6, lane 7). Sequencing showed this fragment to be an amplification product spanning the putative 5' terminus of exon 1' to the beginning of exon 1. Comparison with the amino acid sequence of m-mtTFA shows the 165-bp long region between the two exons to be an intron whose ends respect the GT-AG rule as well as the consensus delineated by Tominaga et al. [17] for the other h-mtTFA introns (See Fig. 7).

DISCUSSION

The results presented here strongly suggest that m-mtTFA and m-tsHMG are encoded by a single gene, and that the mature mRNA encoding these proteins is obtained *via* alternative splicing events. Fig. 7 presents the sequence of the 5' region of this complex gene and a model consistent with our data for the origin of m-mtTFA and m-tsHMG proteins. Transcription in the testis during spermiogenesis could start from either the putative promoter P_t or the promoter P_{mt}. In the former case, splicing would remove intron 1', exon 1 and intron 1, thereby joining exons 1' and 2 and giving rise to the tsHMG mRNA. Transcription from P_{mt} would lead to the excision of intron 1 in order to join exons 1 and 2 producing the mtTFA mRNA. Transcription in other tissues would normally start from P_{mt} and only mtTFA would be produced. Selection of different 5' exons by differential use of promoters as suggested here have been shown for other genes [18], including genes encoding isoforms of a protein directed to different organelles [19] and tissue-specific proteins [19,20]. Our model involving differential selection between two promoters and alternative splicing does not, however, explain how the promoters P_{mt} and P_t are chosen. Whereas the mitochondrial transcription factor is an essential protein whose expression from P_{mt} should be constitutive, the P_t promoter must be regulated. Whether a testis-specific factor activates P_t transcription at the appropriate stage in spermatocyte or spermatid development is not known at this time.

One possible inconsistency in this model is that no TATA box can be found in the region upstream of exon 1 (Fig. 7A). However, many promoters lack a TATA box and in these cases promoter recognition and transcription initiation depends on a second element overlapping the transcription start site, the initiator (consensus : PyPyA+1NA/TPyPy)[21,22]. Moreover, binding sites for transcription factor Sp1 have been identified in many TATA-less promoters [21], and Tominaga et al. [23] have found such sites in the region upstream of the h-mtTFA gene. In our case, potential initiator sequences are present.

Although the testis-specific protein was previously found in the testis supporting the hypothesis that it might regulate the expression of testis-specific genes [8], we find that fragments identical to m-tsHMG cDNA can be amplified from RNA of mouse embryos as well. These experimental results are not necessarily incompatible: Northern blots and primer extension, used to show that the expression of m-tsHMG was restricted to the testis of postpuberal mice [8], are less sensitive than the RT-PCR used in this study.

Why a protein, otherwise implicated in general mitochondrial transcription, might be recruited for a specific nuclear role might be addressed by the observation that the m-tsHMG is expressed in late spermiogenesis [8]. This developmental phase witnesses many important organisational changes in the cytoplasm, not the least of which affect the structure, number and organization of mitochondria [24,25]. On the other hand, Tominaga et al. [17] have reported the existence of another widely distributed splicing isoform of h-mtTFA lacking exon 5. This isoform is inactive in transcriptional activation and might regulate the action of h-mtTFA [7]; however, the regulation of m-mtTFA is an unlikely role for tsHMG, since the splicing event that we envisage adds an exon and gives rise to a correctly expressed protein without the mitochondrial signal sequence.

Finally, that the two proteins m-mtTFA and m-tsHMG directed to different subcellular localizations arise from a single gene, suggests that either 1) the ancestral protein was mitochondrial and recruited to act in the nucleus during the genesis of a specific tissue; or 2) the ancestral protein was a stage- and tissue-specific nuclear protein, which was subsequently recruited to function as a mitochondrial factor. At this point, it is useful to consider the structure of the yeast mtTFA, since strong structural similarity among all mtTFA proteins would be most suggestive of a single common origin in support of scenario 1. However the lack of a hinge and a C-terminal region in the yeast protein and its implication in packaging of the yeast mitochondrial genome [14,15], raises serious questions pertaining to its membership in

the mtTFA family. Also, phylogenetic analyses (Fig. 1, and [2,26]) do not cluster the sequences of sc-mtTFA with h-mtTFA. The tree of Fig. 1, for example, associates sc-mtTFA with a yeast cisplatin-DNA adduct binding protein called Ixr1. Moreover, Laudet *et al.* [26] proposed a genesis for h-mtTFA *via* the association of two diverged boxes, which differs greatly from the sc-mtTFA origin which could be due to an internal duplication of one HMG box. Therefore at this point in time the origin of the animal mtTFA and tsHMG proteins remains obscure.

While this manuscript was in preparation, Larsson *et al.* published the results of different experiments which led to conclusions similar to those reported here about the alternative splicing of the m-mtTFA/m-tsHMG pre-mRNA [27].

ACKNOWLEDGEMENTS

We are grateful to Dr. Mark Featherstone of the McGill Cancer Center, Montreal for his generous gift of mouse RNA and DNA. We thank Benoit Cousineau for helpful discussions. This work was supported by a grant from NSERC Canada (to R.C.), and by a NSERC 1967 Science and Engineering Scholarship (to N.C.). Salary and interaction support from the Canadian Institute for Advanced Research are gratefully acknowledged by R.C., who is a Fellow of the Program in Evolutionary Biology.

REFERENCES

[1] Grosschedl, R., Giese, K. and Pagel, J. (1994) Trends Genet. 10, 94-100.

[2] Baxevanis, A.D. and Landsman, D. (1995) Nucleic Acid Res. 23, 1604-1613.

[3] Fisher, R.P., Lisowsky, T., Parisi, M.A. and Clayton, D.A. (1992) J. Biol. Chem. 267, 3358-3367.

[4] Shadel, G.S. and Clayton, D.A. (1993) J. Biol. Chem. 268, 16083-16086.

[5] Parisi, M.A. and Clayton, D.A. (1991) Science 252, 965-969.

[6] Antoshechkin, I. and Bogenhagen, D.F. (1995) Mol. Cell. Biol. 15, 7032-7042.

[7] Dairaghi, D.J., Shadel, G.S. and Clayton, D.A. (1995) J. Mol. Biol. 249, 11-28.

[8] Boissonneault, G. and Lau, Y.-F.C. (1993) Mol. Cell. Biol. 13, 4323-4330.

[9] Frohman, M.A. (1994) PCR Methods Applic. 4, S40-S58.

[10] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning : A Laboratory Manual, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[11] Corpet, F. (1988) Nucleic Acid Res. 16, 10881-10890.

[12] Swofford, D.L. (1993) PAUP : Phylogenetic Analysis Using Parsimony, Version 3.1, Illinois Nat. Hist. Surv., Champaign, Illinois.

[13] Felsenstein, J. (1993) PHYLIP : Phylogeny Inference Package, Version 3.5c, University of Washington, Seattle, Washington.

[14] Diffley, J.F.X. and Stillman, B. (1991) Proc. Natl. Acad. Sci. USA 88, 7864-7868.

[15] Parisi, M.A., Xu, B. and Clayton, D.A. (1993) Mol. Cell. Biol. 13, 1951-1961.

[16] Schatz, G. and Dobberstein, B. (1996) Science 271, 1519-1526.

[17] Tominaga, K., Hayashi, J.-I., Kagawa, Y. and Ohta, S. (1993) Biochem. Biophys. Res. Commun. 194, 544-551.

[18] Breitbart, R.E., Andreadis, A. and Nadal-Ginard, B. (1987) Ann. Rev. Biochem. 56, 467-495.

[19] Borgese, N., D'Arrigo, A., De Silvestris, M. and Pietrini, G. (1993) FEBS Lett. 325, 70-75.

[20] Ball, H.J., Shine, J. And Herzog, H. (1995) J. Biol. Chem. 270, 27272-27276.

[21] Hernandez, N. (1993) Genes Dev. 7, 1291-1308.

[22] Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. and Smale, S.T. (1994) Mol. Cell. Biol. 14, 116-127.

[23] Tominaga, K., Akiyama, S., Kagawa, Y. and Ohta, S. (1992) Biochim. Biophys. Acta 1131, 217-219.

[24] Grootegoed, J.A. and Den Boer, P.J. (1990) in: Cellular and molecular events in spermiogenesis (Hamilton, D.W. and Waites, G.M.H., eds), pp. 193-216, Cambridge University Press, Cambridge, New York, Melbourne.

[25] Munn, E.A. (1974) The Structure of Mitochondria, Academic Press, London, New York.

[26] Laudet, V., Stehelin, D. and Clevers, H. (1993) Nucleic Acid Res. 21, 2493-2501.

[27] Larsson, N.-G., Garman, J.D., Oldfors, A., Barsh, G.S. and Clayton, D.A. (1996) Nature Genet. 13, 296-302.

Figure 1. Position of gene-specific oligonucleotides used in this study. The coding sequences of the mature mtTFA mRNA are indicated by bold lines. Major structural features are represented by rectangles. S.p. is the signal peptide, HMG are the HMG boxes and int1 is the site of intron 1.

int1
s.p.
HMG1
HMG2
tshmg9
tshmg5
tshmg6
tshmg8
TFA5'
tshmg2
tshmg1
TFA3'

Figure 2. Phylogenetic tree of HMG box proteins. The tree topology was derived from an alignment of the HMG box regions of theses proteins, using the program PAUP (parsimony; 100 bootstrap replicas, heuristic search). The number on tree branches indicate bootstrap values. Similar results were obtained when distance-based methods were used to build the tree (PHYLIP program; data not shown). Ixr1 is a yeast cisplatin-DNA adduct binding protein; m-HMG1 and 2 are the mouse HMG1 and 2 proteins and HMG-T is the trout HMG protein. The alignments are available from the authors upon request.

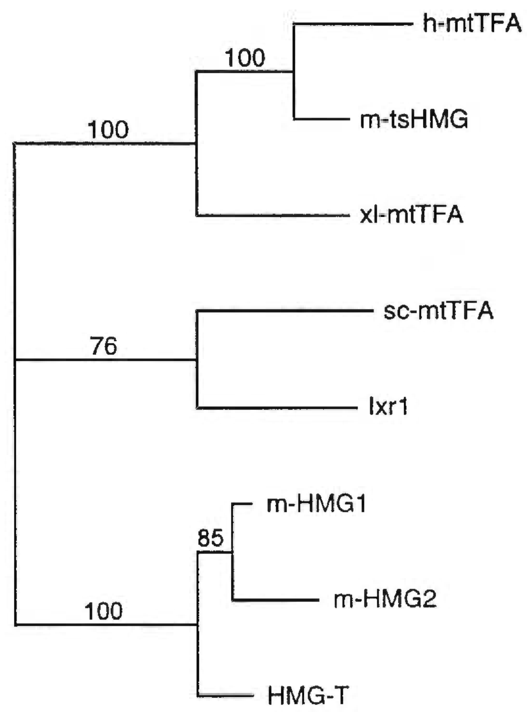

h-mtTFA
100
m-tsHMG
100
xl-mtTFA
sc-mtTFA
76
Ixr1
m-HMG1
85
m-HMG2
100
HMG-T

Figure 3. The amplification of m-tsHMG cDNA fragments. Two consecutive PCR amplifications with the same pair of primers TFA5' and TFA3' (T_{anneal}=50°C) were done after an RT experiment on mouse embryo RNA with the poly-T primer Q_T, analyzed on a 1.5% agarose gel. Lane 1, molecular weight marker (in bp); lane 2, control PCR amplification with no added cDNA; and lane 3, PCR amplification with cDNA.

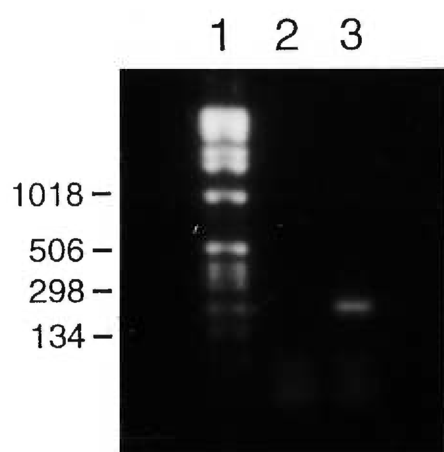
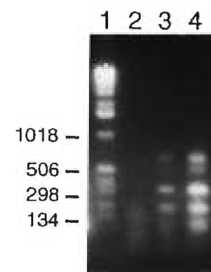

1
2
3
1018
506
298
134

Figure 4. 5'-RACE with mouse embryo RNA.

A) The result of an RT experiment with mouse embryo RNA using the primer tshmg1, followed by two consecutive PCR amplifications using tshmg1 and Q_o in the first amplification and tshmg2 and Q_I in the second (T_{anneal}=55°C), and analyzed on a 1.5% agarose gel. Lane 1, molecular weight marker (in bp); lane 2, control PCR without cDNA; lane 3 and 4, PCR amplification with cDNA; less template was used in the second PCR of lane 3.

B) Alignment of the putative signal peptide with the corresponding sequence in h-mtTFA.

A



B

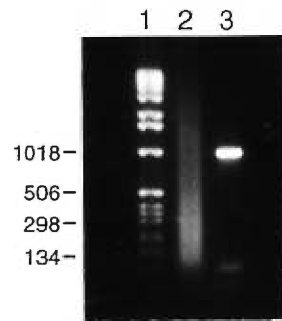
```
              1            14
m-tsHMG       MALFRGMWSVLKAL
              **::* ** ** **
h-mtTFA       MAFLRSMWGVLSAL

              15           28
              GRTGVEMCAGCGGR
              **:*:*:* *** *
              GRSGAELCTGCGSR

              29           42
              IPSSISLVCIPKCF
              : * :*:* :*: *
              LRSPFSFVYLPRWF
```

Figure 5. Amplification of the m-tsHMG intron.

A) PCR amplification using primers tshmg5 and 6 with mouse DNA analyzed on a 1.5% agarose gel. Lane 1, molecular weight marker (in bp); lane 2, control PCR without added DNA; lane 3, PCR with DNA.

B) Schematic comparison of the intron site in m-tsHMG and h-mtTFA. The intron sequence is in small letters.

A



B

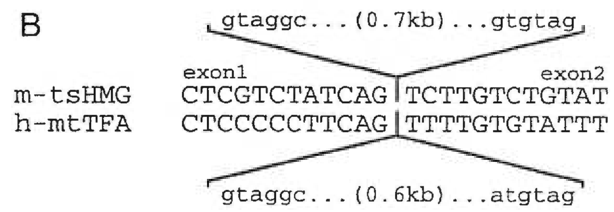
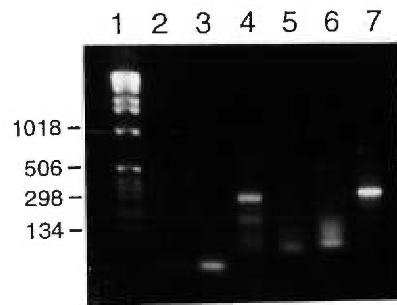


Figure 6. Amplification of splicing products and intermediates. Following the RT experiment with RNA from mouse embryos in lanes 4 and 5, or testis in lanes 6 and 7 both using the tshmg8 primer, two consecutive PCR amplifications were done. The first with primers tshmg8 and Q_o, and the second with primer Q_I and either primer tshmg6 (lanes 2, 4, 6) or primer tshmg9 (lanes 3, 5, 7) (T_{anneal}=55°C) analyzed on a 2.0% agarose gel. Lane 1, molecular weight marker (in bp); lanes 2 and 3, control PCR without cDNA. The bottom shows a schematic drawing of the content of the PCR fragments.


1 2 3 4 5 6 7
1018
506
298
134

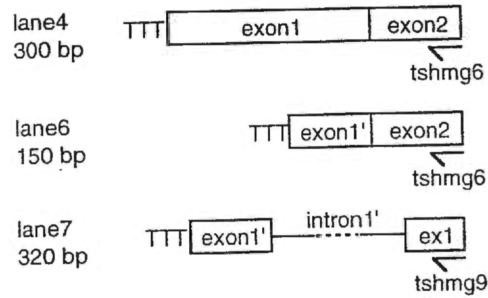
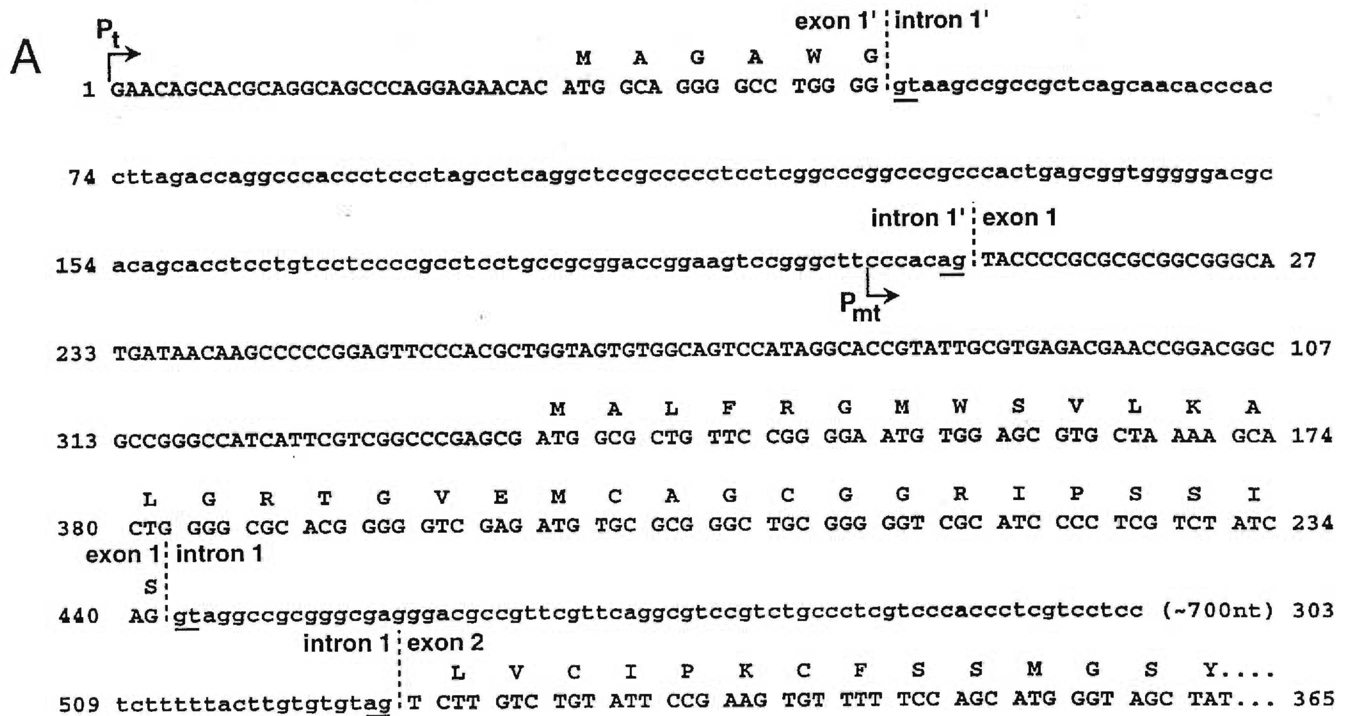
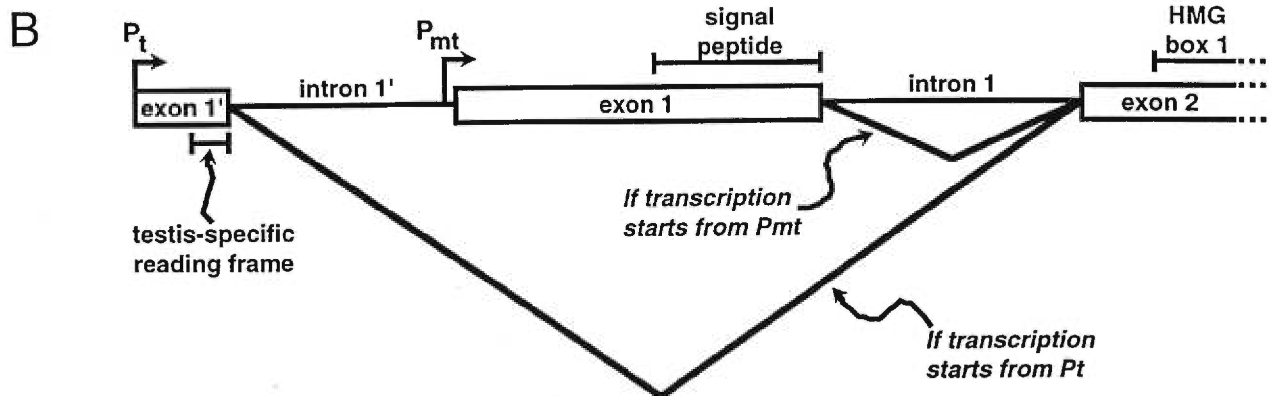

lane4
300 bp
TTT
exon1
exon2
tshmg6
lane6
150 bp
TTT
exon1'
exon2
tshmg6
lane7
320 bp
TTT
exon1'
intron1'
ex1
tshmg9

Figure 7. A) Sequence of the 5' end of the mouse tsHMG/mtTFA gene. Capital letters indicate exon sequences and small letters, intron sequences. The GT-AG intron junctions are underlined. Putative transcription start sites are indicated by P_t and P_{mt}. The nucleotide sequence is numbered in the left and right side margin from the P_t and P_{mt} promoters, respectively. Amino acid sequences are written above the nucleotide sequence in the one-letter code.

B) Model of the proposed alternative splicing of pre-mRNA giving rise to the mouse tsHMG and mtTFA isoforms.


A
Pt
exon 1' intron 1'
M A G A W G
1 GAACAGCACGCAGGCAGCCCAGGAGAACAC ATG GCA GGG GCC TGG GG gtaagccgccgctcagcaacacccac
74 cttagaccaggcccaccctccctagcctcaggctccgccccctcctcggcccggcccgcccactgagcggtgggggacgc
intron 1' exon 1
154 acagcacctcctgtcctccccgcctcctgccgcggaccggaagtccgggcttcccacag TACCCCGCGCGCGGCGGGCA 27
Pmt
233 TGATAACAAGCCCCCGGAGTTCCCACGCTGGTAGTGTGGCAGTCCATAGGCACCGTATTGCGTGAGACGAACCGGACGGC 107
M A L F R G M W S V L K A
313 GCCGGGCCATCATTCGTCGGCCCGAGCG ATG GCG CTG TTC CGG GGA ATG TGG AGC GTG CTA AAA GCA 174
L G R T G V E M C A G C G G R I P S S I
380 CTG GGG CGC ACG GGG GTC GAG ATG TGC GCG GGC TGC GGG GGT CGC ATC CCC TCG TCT ATC 234
exon 1 intron 1
S
440 AG gtaggccgcgggcgagggacgccgttcgttcaggcgtccgtctgccctcgtcccaccctcgtcctcc (~700nt) 303
intron 1 exon 2
L V C I P K C F S S M G S Y....
509 tctttttacttgtgtgtag T CTT GTC TGT ATT CCG AAG TGT TTT TCC AGC ATG GGT AGC TAT... 365


B
Pt
Pmt
signal peptide
HMG box 1
exon 1'
intron 1'
exon 1
intron 1
exon 2
testis-specific reading frame
If transcription starts from Pmt
If transcription starts from Pt

CHAPITRE 3

Particularités des ARN de transfert mitochondriaux

Article 4

Nucleotide 47 of tRNAs: selection of the fittest is not always based on activity

Article 5

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

ARTICLE 4

Nucleotide 47 of tRNAs: selection of the fittest is not always based on activity

(Le nucléotide 47 des ARNt: la sélection du mieux adapté ne se base pas toujours sur l'activité)

Cermakian, N., McClain, W.H. et Cedergren, R. (1998)

À soumettre à *RNA*.

Les contributions de N.C. à cet article sont:

- Toutes les expériences.
- Rédaction du manuscrit.

Nucleotide 47 in tRNAs: selection of the fittest is not always based on activity

NICOLAS CERMAKIAN, WILLIAM H. MCCLAIN* AND ROBERT CEDERGREN§

Département de biochimie, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal (Québec) H3C 3J7, Canada

* Department of bacteriology, University of Wisconsin, Madison, WI 53706, USA

§ Corresponding author. Tel.: (514) 343-6320. Fax: (514) 343-2210. E-mail: ceder@bch.umontreal.ca.

December 15, 1997

ABSTRACT

Analysis of tRNA sequences suggests a correlation between the absence of a nucleotide at position 47 (nt 47) in the extra loop and the presence of a G22:U13 pair in the D-stem. We have evaluated the significance of this by determining the *in vivo* activity of tRNAs with a G22:C13 or a G22:U13 base pair in the background of a tRNA with and without nt 47. First, the tRNA constructs without nt 47 were shown to have an activity in amber suppression in the cases of both base pairs at position 13-22. These suppressor activities were similar to the level of suppression by the wild-type tRNA containing nt 47. In addition, equivalent constructs with an anticodon GGC permitted the growth of an *Escherichia coli* strain deleted in alanine $tRNA_{GGC}$ genes as does the wild type $tRNA^{Ala}$. However, in competitive growth experiments between *E. coli* strains harboring tRNA constructs lacking nt 47, but containing either a G22:C13 or a G22:U13, a long-term difference could be observed between the clones. After 16 days, strain harboring the G22:U13 tRNA could be seen to taking over the culture. Clearly, activity is only one criterion in judging whether tRNA mutations become fixed in populations.

KEYWORDS: tRNA; structure; evolution; amber suppression; competitive growth.

ABBREVIATIONS: m^7G, 7-methylguanosine; mtDNA, mitochondrial DNA; nt 47, nucleotide 47; rRNA, ribosomal RNA; tRNA, transfer RNA.

INTRODUCTION

Massive sequence and structural databases are now readily available, so the structural and functional analysis of these data is being prioritized more and more. In the case of tRNAs, where the art of sequencing has been practiced for over 30 years, much of the correspondence between structure and function remains obscure, in spite of the

availability of the three-dimensional structure (Dirheimer et al., 1995). For some time we have been interested in gathering from the tRNA sequence database information regarding tRNA structure and function. Previously, correlations between D-stem and extra loop have been proposed (Dirheimer et al., 1995). Recently, Steinberg and Ioudovitch (1996) observed that the vast majority of cytoplasmic tRNAs which lack nucleotide 47 (nt 47) contained a G13:U22 in the D-stem. Furthermore, all exceptions to this rule were $tRNA^{Gly}$, none of which have been proven to be involved in ribosomal translation. Some in fact are known not to promote protein synthesis. The same analysis revealed that mitochondrial (mt) tRNAs do not follow this simple rule, since almost half of the tRNAs examined contain both the G:C base pair at 13-22 and nt 47. These authors then showed by molecular modeling that the absence of nt 47 affects the conformation of the region such that a normal Watson-Crick base pair at 13-22 can not be accommodated without a disruption of the tertiary interaction between the D stem and the variable loop (Steinberg and Ioudovitch, 1996).

We show here that tRNAs lacking nt 47 are equally active in suppression and maintaining cell growth whether the 13-22 base pair is G:C or G:U. However, in experiments where cells harboring tRNAs without nt 47 and with either of the two base pair are mixed in culture, the G13:U22 variant outcompetes the alternate after many rounds of cell division.

RESULTS

AMBER SUPPRESSION ASSAYS

The amber suppression test in *Escherichia coli* was used to evaluate the *in vivo* activity of different tRNA constructs derived from $tRNA^{Ala}{}_{UGC}$ (Gabriel et al., 1996). The wild type $tRNA^{Ala}$ suppressor (the "WT_{am}" sequence) contained only a modification of the anticodon loop to CUA in order to recognize the amber codon UAG. Two variants of the $tRNA^{Ala}$ sequence both lacking nt 47 were also constructed: one had a C13 (the

"$C13_{am}$" sequence), as in the wild type sequence, and the other a U13 (the "$U13_{am}$" sequence). These latter two correspond to tRNAs lacking nt 47 and containing either a Watson-Crick base pair G22:C13 or a wobble base pair G22:U13. Figure 1A presents the sequence of these molecules. The three potential suppressors, WT_{am}, $C13_{am}$ and $U13_{am}$ were amplified by PCR and cloned in the vector pGFIB as described in Material and Methods. The three suppressors as well as the plasmid alone were transformed into the *E. coli* strain XAC-1, which contains two reporter genes, a *lacI-lacZ* fusion and the *ArgE* gene, both of which contain an amber stop codon near the 5' end of their reading frame (Normanly et al., 1986).

After transformation of the bacteria with the vectors containing the tRNA constructs, the level of suppression of each transformant was determined. According to the predictions derived from the sequence analysis (Steinberg and Ioudovitch, 1996), we expected that the $C13_{am}$ construct might have less suppressor activity than the $U13_{am}$ tRNA. However, Figure 1B shows that both the $C13_{am}$ and the $U13_{am}$ tRNAs suppress the amber mutation in *lacI-lacZ* gene almost as efficiently as the WT_{am} tRNA, as indicated by assays of β-galactosidase activities. In fact, $U13_{am}$ tRNA was reproducibly less active in suppression than $C13_{am}$ or WT_{am}. These tests were repeated using the medium-copy plasmid pMPM-A5Ω (Mayer, 1995) yielding the activity order $WT_{am} \approx C13_{am} > U13_{am}$ as above (data not shown). We also tested the ability of these clones to confer arginine prototrophy to the XAC-1 strain: all three clones performed equally well in suppressing the amber codon in the *ArgE* gene (Figure 1C).

GROWTH ASSAY IN AN *E. COLI* ALANINE TRNA KNOCK-OUT STRAIN

In order to confirm the preceding results, we adopted a strategy to assess tRNA function in normal protein biosynthesis. The strain K45s1Δ*relA* (Gabriel et al., 1996) deleted for the two $tRNA^{Ala}_{GGC}$ genes requires a vector-bound $tRNA^{Ala}$ gene with the anticodon GGC for normal growth. We therefore designed tRNA genes composed of the wild-type alanine tRNA (the sequence of $tRNA^{Ala}_{UGC}$) except for the anticodon

loop which was from $tRNA^{Ala}_{GGC}$ and two variants having this same sequence, but lacking nt 47 and having either a C13 or a U13 (see Figure 2A). These constructions were made as the ones for suppressors. Growth of cells with these plasmids was compared in various culture media and temperatures, all conditions giving the same results, i.e. all three tRNA constructions permit cell growth to approximately the same level and rate (Figure 2B).

Further characterization of this system involved the determination of the level of *in vivo* aminoacylation of the tRNA constructs using acid gel electrophoresis. As shown in Figure 3, the amount of aminoacylation in the case of each tRNA construct is similar enough to be indistinguishable; however both tRNAs lacking nt 47 exhibit distinctly lower steady state levels in the cells. These lower levels may be due to lower expression or faster degradation, but since no electrophoresis band was seen at high molecular weight, tRNA processing is not a likely explanation for these results. Another potentially significant observation is that the $C13_{GGC}$ construct has a distinctly faster gel migration than the other tRNAs. A similar, increased migration is also seen in the case of the aminoacylated $C13_{GGC}$ species (lane 6), which would indicate that a defect in 3' processing is an unlikely explanation for the migration. Moreover, primer extension experiments of this tRNA confirmed that 5' processing was not faulty in either this or the other tRNA constructs (data not shown). The difference in gel mobility of the $C13_{GGC}$ species suggested the presence of charge or conformational difference between this tRNA and the others. Inspection of the wild type sequence demonstrated the presence of a 7-methylguanosine (m^7G) residue at position 46 (Steinberg et al., 1993). This, and the fact that m^7G is a frequent modification at position 46 of tRNAs, suggest that a defect in the modification at this position might explain the aberrant migration of C13. However, experiments with the aim of cleaving tRNAs with sodium borohydride/aniline at m^7G (Zueva et al., 1985) and subsequent primer extension to evaluate the extent of methylation of G46, did not yield a clear difference between our various constructs (data not shown). We also assessed tRNA mobility on gel (as in Figure 3) after expression in the GM18 strain which is mutated in *trmB*, a gene encoding a tRNA methyltransferase specific for m^7G (Marinus et al.,

1975), but the same mobility was observed as in K45s1Δ*relA* (data not shown), indicating either that the product of *trmB* is not responsible for $tRNA^{Ala}$ methylation in *E. coli*, or that a defect in this process is not the cause of the aberrant migration of C13.

LONG TERM GROWTH

Sequence correlations such as that relating the lack of nt 47 to the identity of a base pair in the D-stem are produced by natural selection acting upon a population of cells with different variants of tRNAs. The selection process, however, may not be based solely on the activity of a molecule, but rather the result of a combination of criteria, including such properties as long term stability, and or stability under different growth conditions. Nevertheless, none of the experiments performed up until now have properly evaluated the long-term effect of the presence in the cellular environment of a tRNA lacking nt 47 and containing a C13. In order to reveal any long term effects, we fashioned a competition experiment by serial dilution of a bacterial culture composed of bacteria harboring either the $C13_{am}$ or the $U13_{am}$ tRNAs, a strategy similar to that used to see the effect of non lethal mutations in yeast (Li et al., 1996). In this way we hoped to ascertain any long-term advantage of the mutant tRNAs.

XAC-1 cells which had been transformed with either $C13_{am}$ or $U13_{am}$ were grown overnight, and were diluted and mixed together in minimal medium lacking arginine (selection for suppression in the *ArgE* gene). To be sure that the outcome of the competition was not simply the consequence of a small difference in the number of living cells inoculated in this first step that might favor clone $U13_{am}$, the volume of the $C13_{am}$ culture was 1.5 times that of the $U13_{am}$ culture inoculate. The culture mixture was incubated overnight at 30°C and then diluted in fresh selective medium, and the incubation was continued at 30°C. Such dilutions were performed every day for three weeks. Plasmid DNA was then extracted from cells, digested with PvuII, Southern blotted and hybridized with a probe complementary to the anticodon stem-loop or the acceptor and T stems. The restriction enzyme PvuII cuts the pGFIB on both sides of

the multicloning site, as well as in the D-stem of the $C13_{am}$ variant, but not in the $U13_{am}$ insert. This fact provided a simple way to evaluate the ratio of the two variants in a culture by simple comparison of the gel bands highlighted by hybridization (Figure 4A).

After 21 days, and several hundred generations, 17 of the 24 competition experiments performed in parallel showed that the proportion of the $U13_{am}$ variant had attained the level of 70% or greater. Four other cultures contained between 50 and 70% of the clone U13, whereas only 3 cultures had more $C13_{am}$ than $U13_{am}$ (data not shown). Figure 4B presents the combined results for these 24 independent competitions. The proportion of the clone U13 in the culture rises gradually and finally becomes significantly above the initial ratio ($p \ll 0.001$ after 16 days, according to "t" test). We conclude that this competition experiment could be evidence of an advantage conferred to cells by the suppressor with a U13 over the one with a C13, in the absence of nt 47.

DISCUSSION

Steinberg and Ioudovitch (1996) put forward the hypothesis that the absence of tertiary interaction between nts 22 and 46 renders the tRNA inactive, and that this explains the almost complete absence from cytosols of tRNAs with a Watson-Crick base pair 13-22 and without nt 47. The results presented here imply that this hypothesis has to be reevaluated. In fact, we were able to show that a suppressor lacking a nt 47 but retaining the G22:C13 base pair is as active as a control with nt 47 in suppressing stop codons in *lacZ* and *argE* genes, within the limits of sensitivity of these tests (Figure 1). Since it could be argued that suppressor assays could not faithfully mimic tRNA function in elongation and that most of the tRNAs compiled by Steinberg and Ioudovitch (1996) are elongating tRNAs, we performed analogous experiments using tRNAs with a GGC anticodon and a strain with knocked-out $tRNA^{Ala}_{GGC}$ genes. These

experiments confirmed suppression tests: tRNAs with or without nt 47, and with G22:C13 or G22:U13 supplement the knockout strain equally well (Figure 2). Therefore, the reason that tRNAs with WC 22-13 and no nt 47 have been counterselected during evolution must be more subtle and must have needed many generations to take effect. This hypothesis is confirmed by the *in vivo* competition experiment, in which such a selection could be observed after many generations (Figure 4). While the previous experiments were done with cultures after a few hours of growth, i.e. after only a minimal number of generations, this latter experiment let the cells grow for hundreds of generations. Moreover, as pointed out by Li et al. (1996), such a protocol including a variety of stages of growth is a good representation of a natural environment.

Various factors could contribute to the selection of one tRNA structure over the other, and this puzzle could be hard to unravel given the tenuous nature of this loss. We showed that the level of aminoacylation does not seem to be affected, whereas misaminoacylation is unlikely to be the cause since the correlation between nts 13 and 47 is valid for all species of tRNAs and that different tRNAs have very different determinants for aminoacylation (Martinis and Schimmel, 1995; McClain, 1995; Pallanck et al., 1995). Post-transcriptional processing of the tRNA was called into question following the observation of aberrant migration on gel of our tRNA with WC 22-13 and no nt 47; however, this hypothesis has been discarded by a primer extension experiment, leaving abnormal base modification or stability of the tRNA as the only obvious explanations for this aberrant migration, and possible explanations for the elimination of such tRNAs from cytoplasms. One modified nucleotide that could be of some importance here is the m^7G often found at position 46 and which is present in *E. coli* alanine tRNAs. While the importance of m^7G46 is underscored by its widespread occurrence in many evolutionary lineages (Steinberg et al., 1993), the methylation of G46 does not seem to be essential for protein synthesis. Indeed, an *E. coli* strain lacking the methylase catalyzing the formation of m^7G exhibits no abnormal phenotype (Marinus et al., 1975). However, our results suggest that neither of our variant tRNAs present a sharp defect in base G46 methylation. Whatever the defect is,

it must be very general, since this correlation is commonly found in various tRNA species in all domains of life (Steinberg and Ioudovitch, 1996; N. Cermakian, unpublished observations). An intriguing hypothesis is that tRNAs lacking some tertiary interactions are less stable and thus prone to be degraded more rapidly. This effect could be very slight and thus require many generations to show up. In any case, our results show that activity per se is not the only factor involved in the selection or rejection of a tRNA feature in the course of evolution. Other factors that can be more tenuous and dependent on the cellular environment, such as overall tightness of tertiary structure and resistance to nuclease, are taken into account.

The reason why mitochondrial tRNA sequences do not follow the correlation observed for cytosolic tRNAs (Steinberg and Ioudovitch, 1996) might have remained mysterious if tRNAs with WC 22-13 and no nt 47 (and thus no tertiary interaction 22-46) were inactive. However, the exceptions to the rule in mitochondrial tRNAs can be explained much more easily in light of our results. We show that the defect of these abnormal tRNAs is far from being substantial. Therefore, only a slight lowering of the threshold for required efficiency or accuracy in tRNA function would be necessary to allow retention of these abnormal structures. This model implies that tRNA sequence variation occurs both in cytoplasm and in mitochondria, but that it is countered in the former but not in the latter. This model also explains other abnormalities of mitochondrial tRNAs, like absence of the D or the T stem (Dirheimer et al., 1995; Steinberg et al., 1997), as well as shortened or elongated anticodon stem (Steinberg et al., 1997). In the latter case, it has been shown that such mitochondria-like tRNAs can have substantial suppression activity in *E. coli* (Bourdeau et al., 1998).

This relaxation of constraints in mitochondria could be due to the lack of some enzymatic activities. For example, the effect on stability of structural defects in the tRNA molecule could be observable only in a cellular compartment with degradation enzymes for these molecules that could sense these defects. It could be that mitochondria are lacking such enzymes.

A more careful inspection of the data of Steinberg and Ioudovitch (1996) can provide clues as to how mitochondria came to have tRNAs lacking the correlation between the nts 47 and 13. First, these data show that the absence of the nt 47 might in itself be to a certain extent undesirable: only 15% of cytosolic tRNAs lack this nucleotide. Our results seem to point in this direction, in view of repeatedly observed differences between clones with and without this U47: lower suppression activity of suppressors without nt 47, lower expression or higher degradation of the corresponding tRNAs. On another hand, mitochondrial tRNAs lack nt 47 four times more often, which may occur simply because this deletion contributes to the reduction of the size of the mitochondrial genome. At least two mechanisms have been proposed for this shrinkage of mitochondrial DNA (mtDNA). Mitochondria may be seen as a population of microorganisms dividing and competing against each other in the cytoplasm (Kurland, 1992). Therefore, any factor allowing faster replication of mitochondrial genome confers an advantage to the mitochondrion and can be selected (Kurland, 1992). Another mechanism, called Muller's ratchet, involves fixation of mutations due to statistical fluctuations in small populations, including the accumulation of deletions. Various observations do indeed point to the occurrence of Muller's ratchet in mitochondrial genome evolution (Andersson and Kurland, 1995; Lynch, 1996).

The reduction in size of the mitochondrial genome is reflected in various oddities in the structure of mitochondrial RNAs. First, highly shrunk ribosomal RNAs (rRNAs) have been found in mitochondria of different lineages, sometimes with a length one third that of bacterial rRNAs (Benne and Sloof, 1986). Second, many tRNAs lacking entire domains (D- or T-domain) have been found in mitochondria (Dirheimer et al., 1995; Steinberg et al., 1997). Third, non universal genetic code and relaxed rules of anticodon-codon interaction allowed the reduction of the number of tRNAs that are necessary in mitochondria, especially in animals, where there is the minimal number of 22 tRNAs (Kurland, 1992; Osawa et al., 1992). All these modifications of the translational machinery have allowed an appreciable reduction of the mitochondrial genome. In fact, if these same mechanisms apply in the loss of nt 47, we can infer that

the reductive force is strong enough to act on single nucleotides. In view that these atypical features are mainly found in animals, it is interesting to note that almost all of the tRNAs without nt 47 in the data of Steinberg and Ioudovitch are from animals. In fact, more than 90% of the animal mitochondrial tRNAs lack this nucleotide (N. Cermakian, unpublished observations).

We saw above that almost two thirds of mitochondrial tRNAs in general have no nt 47, whereas this is the case in only 15% of cytosolic tRNAs. If one looks further into the data of Steinberg and Ioudovitch (1996), it appears that almost half of mitochondrial tRNAs do not follow the correlation between nt 47 and base pair 22-13 (and much more in animal mitochondria), whereas only 2% of cytosolic tRNAs are refractory to it. We notice in these observations that there are proportionally 50% more tRNAs lacking nt 47 in mitochondria than in cytosol; on another hand, there are 47% more tRNAs going against the correlation in mitochondria than in the cytosol. This suggests that in mitochondria, there is almost never a compensation in the nature of nt 13 for the loss of nt 47. This supports our assertion that mitochondria have reduced requirements regarding tertiary interactions between the variable loop and the D stem of tRNAs, and that this is reflected in a lower threshold of acceptance of the combination G22:C13/no nt 47. Our results strongly suggest that this threshold does not have to be very different from the one in the cytosol: the relaxation of constraints can be almost undetectable. In fact, Cedergren et al. (1981) have stated that base triples in tRNAs might not be particularly crucial in determining tRNA conformation.

Lang et al. (1997) have sequenced and analyzed the mitochondrial genome of the early-diverging protist *Reclinomonas americana*. They showed that this genome has been better able to resist the force of reduction of the mitochondrial genome and thus able to retain many more bacterial features than any other mitochondrial genome known to date. We inspected *R. americana* mitochondrial tRNA sequences, following the same procedure as Steinberg and Ioudovitch (1996), and found that only 11% of *R. americana* mitochondrial tRNAs lack nt 47, which is very close to the value for cytosolic tRNAs (15%), but far below the ratio in other, more derived mitochondria

(65%). One of the two tRNAs lacking nt 47 has a uridine in position 13, which satisfies the Steinberg and Ioudovitch correlation, and therefore allows the tertiary interaction between the D stem and the variable loop. The other one, however, does not satisfy the rule and has a WC pair G22:C13 instead. It is striking that this atypical tRNA is a glycine tRNA, as are all nine cytoplasmic exceptions observed (Steinberg and Ioudovitch, 1996). This raises the question of whether this $tRNA^{Gly}{}_{UCC}$ is active but not enough to participate in mitochondrial translation or whether the translation is relaxed enough in *R. americana* mitochondria to accommodate such deviant tRNA structures. A hint may be found in the fact that there is only one other $tRNA^{Gly}$ encoded by *R. americana* mitochondrial genome, and its anticodon is GCC. This anticodon would hardly be able to read all codons of box GGX (the four codons are used in these mitochondria; N. Cermakian, unpublished observations). Instead, it is more reasonable to propose that this $tRNA^{Gly}{}_{GCC}$ decodes glycine codons ending with a pyrimidine, while the deviant $tRNA^{Gly}{}_{UCC}$ decodes codons ending with a purine, as often occurs with split codon boxes in many mitochondria (Osawa et al., 1992). The implication is that this latter tRNA is indeed used by the *R. americana* mitochondrial translation apparatus. The possibility also exists, however, that a normal $tRNA^{Gly}{}_{UCC}$ is encoded in the nucleus and then imported into the *R. americana* mitochondria. In summary, these observations in *R. americana* indicate that a clear distinction should be made between the force directing the shrinkage of mitochondrial DNA, which might have not yet begun to act on nt 47 in this organism, and the relaxation of constraints for tRNA structure and translation in general in its mitochondria, which might already be relaxed enough to allow the functioning of a tRNA lacking the triple base pair 13-22-46. The data set is too small to draw firm conclusions.

However, the force of reduction of the mitochondrial genome can not by itself completely explain the high rate of sequence and structure variation in tRNAs of more derived mitochondria. First, atypical tRNAs are found in larger mitochondrial genomes, namely in mitochondria of plant, fungi and protist. Second, in animals, the rate of nucleotide substitution in tRNAs is 5 to 25 times greater in mitochondria than in the nucleus (Lynch, 1996), and Kumazawa and Nishida (1993) listed various

possible factors for this rate difference: (1) higher mutation rate of mtDNA, possibly due to insufficiency of damage-repairing mechanisms; (2) absence of promoter elements within tRNA genes, in contrast with nuclear RNA polymerase III promoters; (3) absence of some processing sequence signals; (4) simpler and hence less constrained interaction with aminoacyl-tRNA synthetases and with translation factors. To these can be added the fact that tRNAs fill other roles in cytoplasms but apparently do not in mitochondria (Söll, 1993). How does the mitochondrial protein synthesis machinery tolerate this high tRNA gene variation? It has been proposed that the expression of the content of genomes with reduced coding capacity could accommodate a suboptimally designed translational machinery (Hasegawa et al., 1984), including atypical or incomplete tRNAs. Moreover, since all replicational, transcriptional and translational proteins of animal mitochondria are encoded in the nucleus, mutation in tRNAs are less prone to lead to an error cascade in these organelles (Hasegawa et al., 1984). We have seen that such an explanation, which is in fact an indirect consequence of genome shrinkage, can be used to explain our results and those of Bourdeau et al. (1998), and that only a slight difference needs to exist between the cytosolic and mitochondrial systems.

An alternative mechanism is possible, that of structural compensatory changes in the tRNA molecule. Drastic structural compensations have been proposed to fit atypical mitochondrial tRNAs lacking a D stem into the L form recognized by the translation apparatus (Steinberg and Cedergren, 1994; Steinberg et al., 1997). Moreover, the absence or low occurrence of some tertiary structure correlations in mitochondrial tRNAs led other groups to suggest a replacement or a compensation of these by other structural features in the molecule (Kumazawa and Nishida, 1993; De Giorgi et al., 1996). Steinberg and Ioudovitch (1996) based their structural study on the modeling of a yeast $tRNA^{Phe}$ whose U47 had been removed. The impossibility of modeling this tRNA while preserving the tertiary interaction between the D stem and the variable loop could be due to the absence of compensating features elsewhere in this molecule, whereas such a compensation would occur in the *E. coli* $tRNA^{Ala}$, which served as the basis for the design of our tRNAs and suppressors. Such a compensatory

mechanism might thus conciliate the theoretical studies of Steinberg and Ioudovitch (1996) with *in vivo* results presented here.

MATERIAL AND METHODS

CONSTRUCTIONS OF PLASMIDS CARRYING SUPPRESSOR AND TRNA GENES AND OTHER MOLECULAR BIOLOGY TECHNIQUES

All suppressor and tRNA genes were amplified by PCR from DNA oligonucleotide templates corresponding to the sequences in Figures 1A and 2A plus EcoRI and PstI sites. A mixture containing 100 ng of this template, 100 pmol of each primer (designed from positions 1-12 and 63-76), 200 µM of dNTPs and 2U of Vent DNA polymerase (NEB), was heated 4 min at 95°C, followed by 30 cycles of: 30 s at 95°C, 30 s at 72°C; and then 4 min at 72°C. The product was cloned between EcoRI and PstI sites in the pGFIB1 plasmid (Masson and Miller, 1986). The sequence of the clones was confirmed by sequencing, using the dideoxy termination method (Pharmacia). All standard protocols were performed as in Sambrook et al. (1989). Freshly transformed cells were used for all the experiments.

The evaluation of the aminoacylation of the tRNAs in vivo was performed by running total RNA extracted under acid conditions through a 6.5% polyacrylamide gel, transferring it to a nylon membrane (Amersham) and hybridizing with an oligonucleotide probe complementary to the anticodon stem-loop, as described in Bourdeau et al. (1998). The strain GM18 was obtained from the *E. coli* Genetic Stock Center (New Haven, CT; CGSC #5384).

Primer extension experiments were performed as in Sambrook et al. (1989), using a probe complementary to positions 76 to 54 in the $tRNA^{Ala}{}_{UGC}$. Fragmentation of RNA with sodium borohydride and aniline was performed as in Zueva et al. (1985),

with modifications. In brief, 6-8 μg of total RNA (in 1 or 2 μL) was treated with 1 μL Tris (pH 9) 0.5 M for 30 min at 37°C to uncharge the tRNAs. 9.5 μL of Tris (pH 8) 1 M and 8 μL of $NaBH_4$ 1.25M were then added, and the mixture incubated for 5 min at 20°C in the dark. After ethanol precipitation, the pellet was resuspended in aniline:acetic acid:water 1:3:7 and incubated 10 min at 60°C. The RNA was precipitated, and either put on gel or subjected to primer extension.

SUPPRESSION AND tRNA COMPLEMENTATION ASSAYS

The strain used for suppression assays is XAC-1 (F' $lacI_{373}lacZ_{m118\ am}\ proB^{+}$ / F^{-} Δ(*lac-proB*)$_{XIII}$ *nalA rif* $argE_{am}$ *ara*) (Normanly et al., 1996). β-galactosidase activity was determined using the method of Miller (1972), after an overnight growth at 37°C in LB medium of followed by a 1:50 dilution in minimal medium A containing 0.4% glucose, 1 mM $MgSO_4$, 20 μg/mL vitamin B1, 50 μg/mL ampicillin and 50 μg/mL arginine, and further growth at 30°C or 37°C. Suppression of the stop codon in the *ArgE* gene was assessed by growing the cells with suppressor clones in a minimal medium A of the same composition but lacking arginine, following a 200-fold dilution of an overnight culture in LB medium at 37°C, and by monitoring OD_{600}.

The *Ala2* knockout strain used is K45s1Δ*ArelA* (*glyV55* Δ(*tonB trpAB17*) Δ(*ArgF lac*)::*TN10-5*, *relA1*, Δ*rrnA*::*camR*, Δ*ala2*::*kanR* / F' *trpA*(*UGA15*)) (Gabriel et al., 1996). The growth properties of these cells expressing wild-type and mutant $tRNA^{Ala}_{GGC}$ were evaluated as in Gabriel et al. (1996), by spreading from left to right 2 μL of an early-log phase (OD_{600} = 0.3-0.5) culture on split-agar plates with LB medium on the right side, and minimal medium A (containing 0.2% glucose or glycerol, 1 mM $MgSO_4$, 5 μg/mL vitamin B1, 40 μg/mL L-arginine, 40 μg/mL L-tryptophane, 50 μg/mL ampicillin and 0.012% bromothymol blue) on the left side. These plates were then incubated at 30°C, 33°C or 37°C.

IN VIVO COMPETITION EXPERIMENTS

Twenty-four competitions were carried out in parallel in sterile 24-well plates (Falcon 3047), using XAC-1 cells transformed with the clones $U13_{am}$ or $C13_{am}$. Twenty-four overnight LB cultures of each clone were diluted and mixed in 500 µL minimal medium A containing 0.4% glucose, 1 mM $MgSO_4$, 20 µg/mL vitamin B1, 50 µg/mL ampicillin, and without arginine. Each preculture in LB was inoculated with a different colony. $U13_{am}$ was diluted 0.8:1000 and $C13_{am}$ 1.2:1000. The number of viable cells in precultures was counted after plating dilutions on LB, to confirm that more $C13_{am}$ than $U13_{am}$ had been inoculated. The cultures were incubated with agitation at 30°C and an aliquot was diluted 1:500 in fresh medium every day. After 1, 6, 11, 16 and 21 days, cells were pelleted and plasmid DNA isolated and then digested with PvuII. This restriction enzyme cuts at CAGCTG, which is found on each side of the multicloning site in pGFIB, but also in *E. coli* $tRNA^{Ala}$ genes with a C13 (from positions 13 to 18), but not in the clone U13, where the corresponding sequence is TAGCTG. The digestions were run on 1.5% agarose gels, Southern blotted to Hybond-N nylon membrane (Amersham) and hybridized with an oligonucleotide probe downstream of the PvuII site in the tRNA gene, yielding an approximately 197-bp band for C13 and a 333-bp band for U13. Films were scanned and the relative amount of each clone in each competition for all days mentioned was evaluated using the ImagePC program (Scion Corporation, release Beta 1).

ACKNOWLEDGEMENTS

We thank Alice Rae for critical reading of this manuscript, and other members of Cedergren's laboratory for helpful and lively discussions. We are grateful to William McClain, Kay Gabriel and Jay Schneider (University of Wisconsin, Madison) for precious technical help, discussion and gift of bacterial strains. This work has been supported by a NSERC grant to RC, a NIH grant (GM42123) to WHM and by a FCAR

scholarship (Québec) to NC. RC is the Richard Ivey Fellow of the Canadian Institute for Advanced Research.

References

Andersson SG, Kurland CG. 1995. Genomic evolution drives the evolution of the translation system. *Biochem Cell Biol 73*:775-787.

Benne R, Sloof P. 1986. Evolution of the mitochondrial protein synthetic machinery. *Biosystems 21*:51-68.

Bourdeau V, Steinberg SV, Ferbeyre G, Emond R, Cermakian N, Cedergren R. 1998. Amber suppression in *Escherichia coli* by unusual mitochondria-like transfer RNAs. *Proc Natl Acad Sci USA* In press.

Cedergren RJ, Sankoff D, LaRue B, Grosjean H. 1981. The evolving tRNA molecule. *CRC Crit Rev Biochem 11*:35-104.

Dirheimer G, Keith G, Dumas, P, Westhof E. 1995. Primary, secondary, and tertiary structure of tRNAs. In: Söll and RajBhandary, ed. *tRNA: Structure, Biosynthesis and Function*, American Society for Microbiology, Washington DC, pp 93-126.

Gabriel K, Schneider J, McClain WH. 1996. Functional evidence for indirect recognition of G.U in tRNA(Ala) by alanyl-tRNA synthetase. *Science 271*:195-197.

Hasegawa M, Yano T, Miyata T. 1984. Evolutionary implications of error amplification in the self-replicating and protein-synthesizing machinery. *J Mol Evol 20*:77-85.

Kumazawa Y, Nishida M. 1993. Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. *J Mol Evol 37*:380-398.

Kurland CG. 1992. Evolution of mitochondrial genomes and the genetic code. *Bioessays 14*:709-714.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature 387*:493-497.

Li B, Vilardell J, Warner JR. 1996. An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. *Proc Natl Acad Sci U S A 93*:1596-1600.

Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol 13*:209-220.

Marinus MG, Morris NR, Soll D, Kwong TC. 1975. Isolation and partial characterization of three Escherichia coli mutants with altered transfer ribonucleic acid methylases. *J Bacteriol 122*:257-265.

Martinis SA, Schimmel P. 1995. Small RNA oligonucleotide substrates for specific aminoacylations. In: Söll and RajBhandary, ed. *tRNA: Structure, Biosynthesis and Function*, American Society for Microbiology, Washington DC, pp 349-370.

Masson JM, Miller JH. 1986. Expression of synthetic suppressor tRNA genes under the control of a synthetic promoter. *Gene 47*:179-183.

Mayer MP. 1995. A new set of useful cloning and expression vectors derived from pBlueScript. *Gene 163*:41-46.

McClain WH. 1995. The tRNA identity problem: past, present and future. In: Söll and RajBhandary, ed. *tRNA: Structure, Biosynthesis and Function*, American Society for Microbiology, Washington DC, pp 335-347.

Normanly J, Masson JM, Kleina LG, Abelson J, Miller JH. 1986. Construction of two Escherichia coli amber suppressor genes: $tRNA^{Phe}_{CUA}$ and $tRNA^{Cys}_{CUA}$. *Proc Natl Acad Sci U S A 83*:6548-6552.

Osawa S, Jukes TH, Watanabe K, Muto A. 1992. Recent evidence for evolution of the genetic code. *Microbiol Rev 56*:229-264.

Pallanck L, Pak M, Schulman LH. 1995. tRNA discrimination in aminoacylation. In: Söll and RajBhandary, ed. *tRNA: Structure, Biosynthesis and Function*, American Society for Microbiology, Washington DC, pp 371-394.

Sambrook, J., Fritsh, E.F. & Maniatis, T. (1989) *Molecular cloning, a laboraotry manual, 2nd edition.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Söll D. 1993. Transfer RNA: an RNA for all seasons. In: Gesteland RF, Atkins JF, eds. *The RNA World*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 157-184.

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

Steinberg S, Ioudovitch A. 1996. A role for the bulged nucleotide 47 in the facilitation of tertiary interactions in the tRNA structure. *RNA 2*: 84-87.

Steinberg S, Misch A, Sprinzl M. 1993. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res 21*:3011-3015.

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

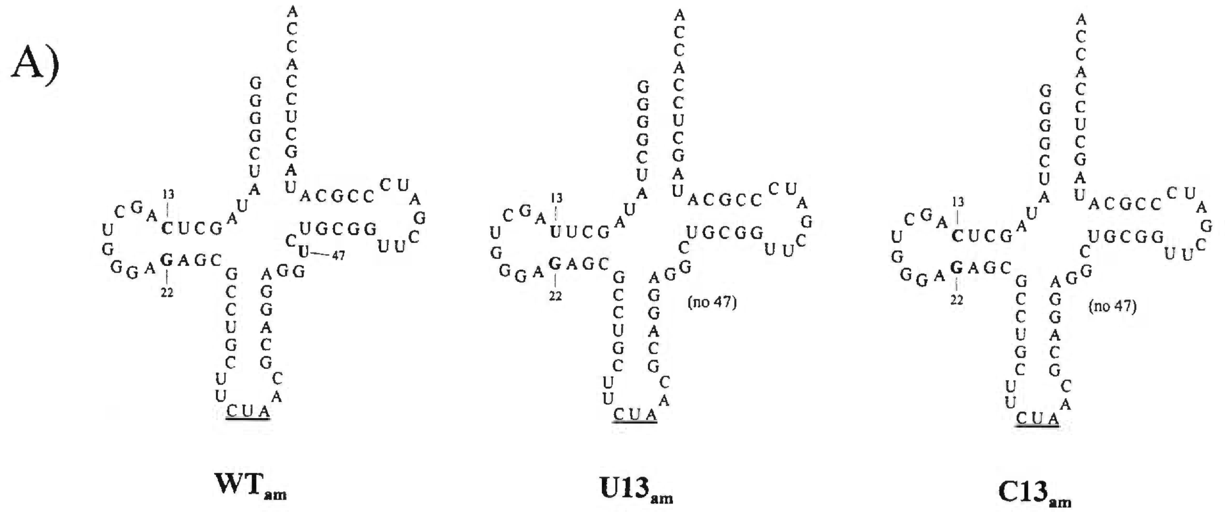
Zueva VS, Mankin AS, Bogdanov AA, Baratova LA. 1985. Specific fragmentation of tRNA and rRNA at a 7-methylguanine residue in the presence of methylated carrier RNA. *Eur J Biochem 146*:679-687.

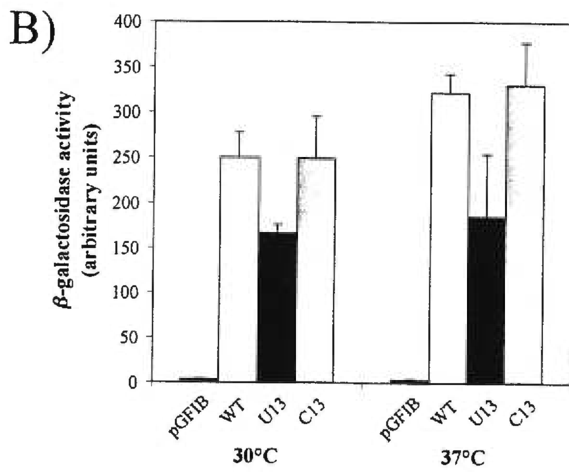
Figure 1. Amber suppression assays

A) The sequence of the three suppressor tRNAs used in this study, with nucleotides (nts) 13, 22 and 47 in bold and the amber anticodon underlined. The WT_{am} suppressor has the same sequence as *Escherichia coli* $tRNA^{Ala}{}_{UGC}$ except for the anticodon, whereas $U13_{am}$ and $C13_{am}$ lack nt 47 and have a U or a C in position 13, respectively. Oligonucleotides corresponding to these sequences were amplified by PCR and cloned in pGFIB.

B) β-galactosidase assays using *E. coli* XAC-1 strain carrying a suppressor of (A) or the plasmid pGFIB. The results shown are the mean of four independent experiments. Similar results were obtained at 25°C and 42°C.

C) Growth in minimal medium A without arginine, of cells carrying a suppressor of (A) or the plasmid pGFIB. Cells were diluted 200-fold from an overnight LB culture and grown at 37°C. Similar results were obtained at 30°C.


A)
WT$_{am}$
U13$_{am}$
C13$_{am}$
(no 47)
(no 47)


B)
β-galactosidase activity (arbitrary units)
pGFIB
WT
U13
C13
30°C
37°C

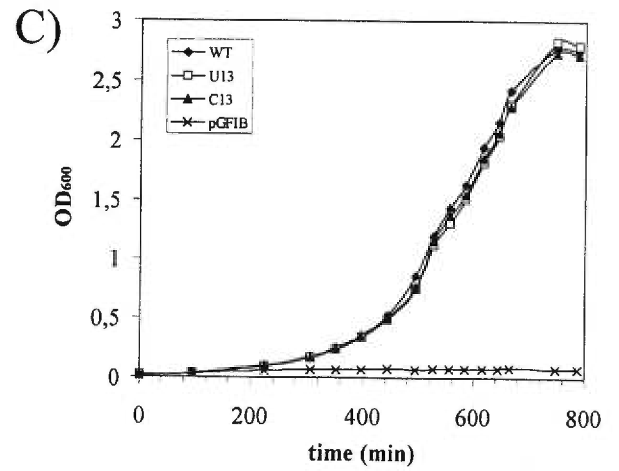
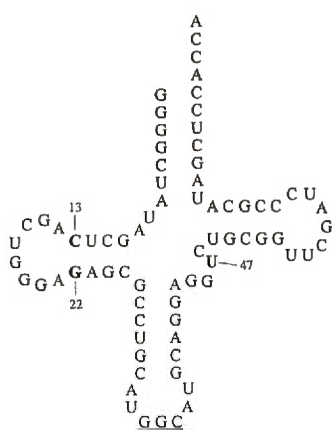

C)
WT
U13
C13
pGFIB
OD600
time (min)

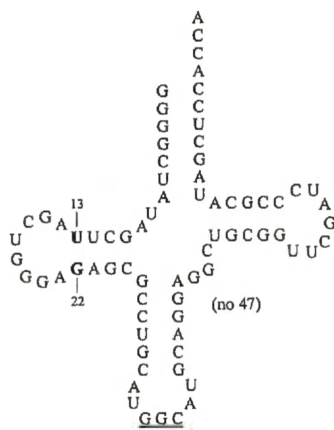
Figure 2. Growth assay with an *E. coli* strain lacking alanine $tRNA_{GGC}$

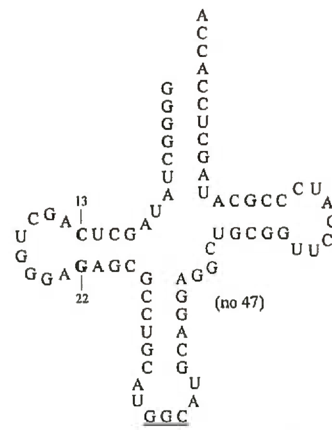
A) The sequence of the three $tRNA_{GGC}$s used in this study, with nts 13, 22 and 47 in bold and the anticodon underlined. The WT_{GGC} has the same sequence as *Escherichia coli* $tRNA^{Ala}_{UGC}$ except for the anticodon loop which is from $tRNA^{Ala}_{GGC}$. $U13_{GGC}$ and $C13_{GGC}$ lack nt 47 and have a U or a C in position 13, respectively. Oligonucleotides corresponding to these sequences were amplified by PCR and cloned in pGFIB.

B) A split-agar plate showing growth of E. coli K45s1ΔArelA cells with a tRNA of (A) or the plasmid pGFIB. The left-hand part of the plate contains minimal medium A with glucose as the carbon source, whereas the right-hand part contains LB medium. Cells were grown to early log phase, and 2 μL of culture were streaked from left to right with a band of sterile paper. The petri dish was then incubated overnight at 37°C. Similar results were obtained with glycerol instead of glucose and at 30°C and 33°C. Consistent results were also obtained with growth curve experiments (data not shown).

A)







WT_{GGC} $U13_{GGC}$ $C13_{GGC}$

B)

WT_{GGC}

$U13_{GGC}$

$C13_{GGC}$

pGFIB

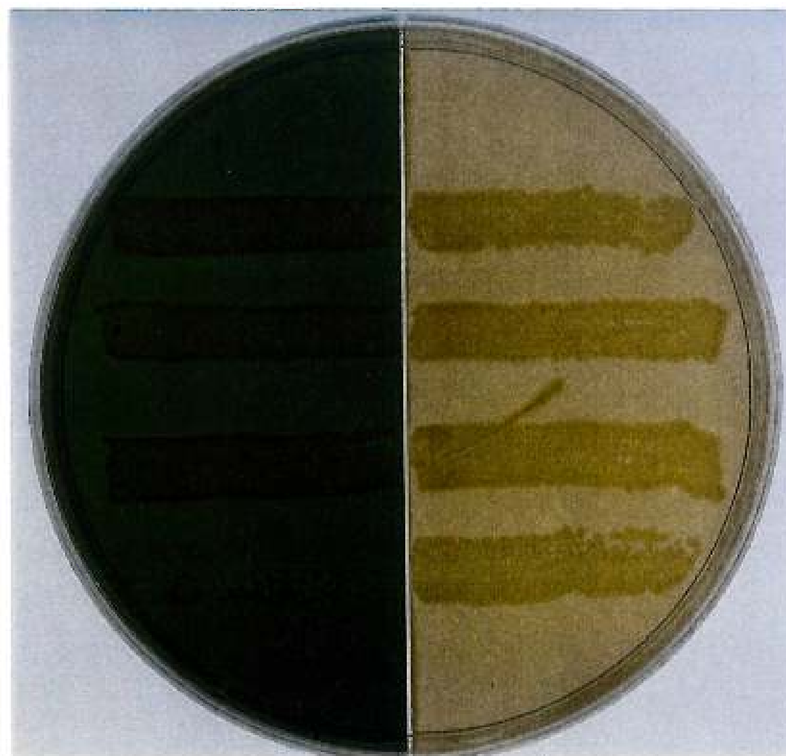


Figure 3. tRNA aminoacylation assay

Total RNA was extracted from cells with tRNAs of Figure 2A, and subjected to polyacrylamide gel electrophoresis in acid conditions followed by Northern blot hybridization with oligonucleotide probes complementary to anticodon stem-loop and, as a control for the amount of RNA, to a part of 5S ribosomal RNA. Prior to electrophoresis, an aliquot of the RNA was deaminoacylated by treatment in basic conditions (lanes "+"). The position of the bands corresponding to charged and uncharged tRNAs is indicated and the percentage of the aminoacylated form is written below the figure. Similar results were obtained with suppressors in the XAC-1 strain.

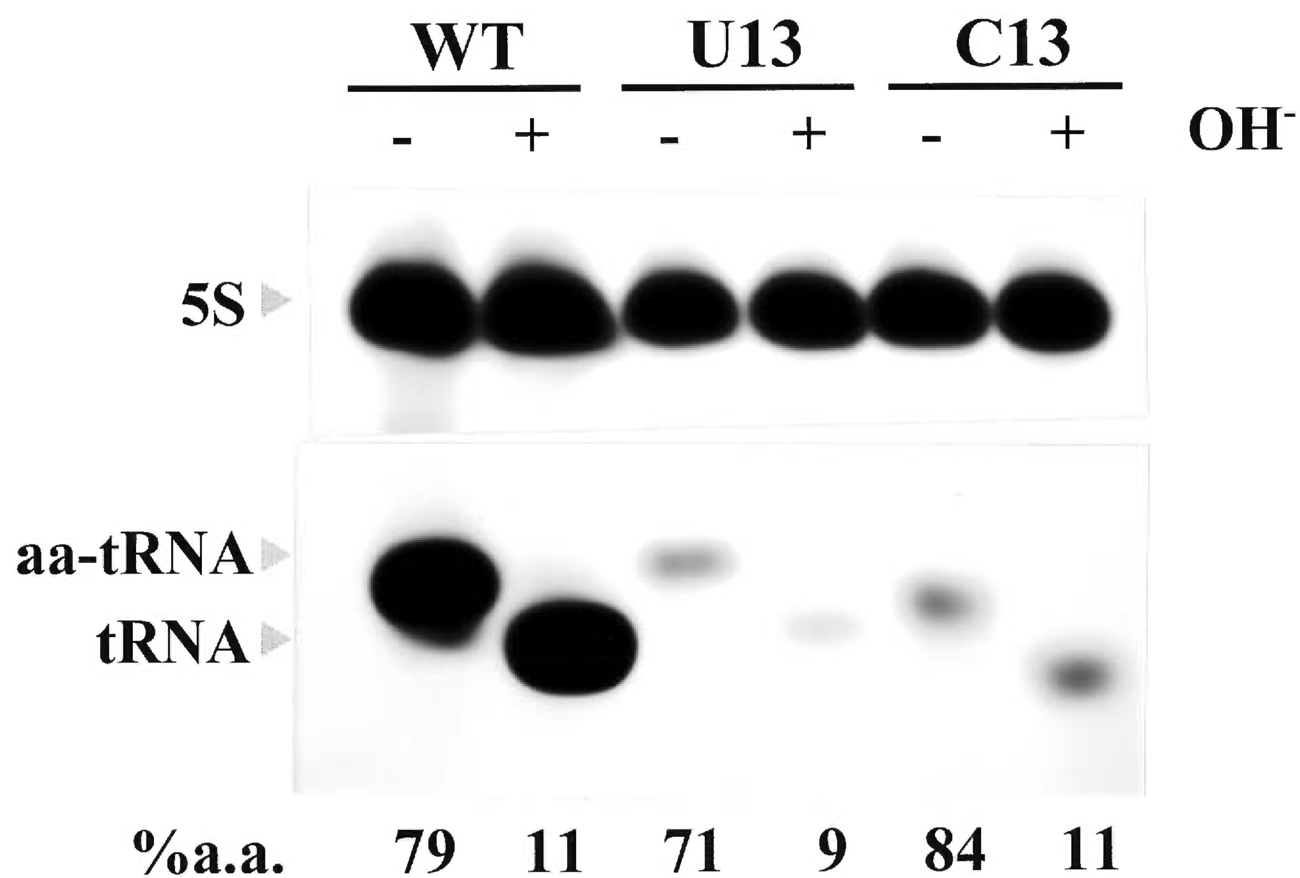
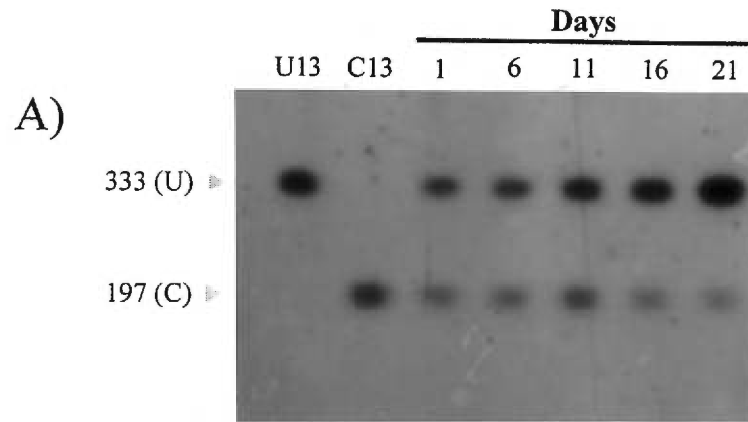

WT
U13
C13
- + - + - + OH-
5S
aa-tRNA
tRNA
%a.a. 79 11 71 9 84 11

Figure 4. Long-term competition between cells with $U13_{am}$ or $C13_{am}$

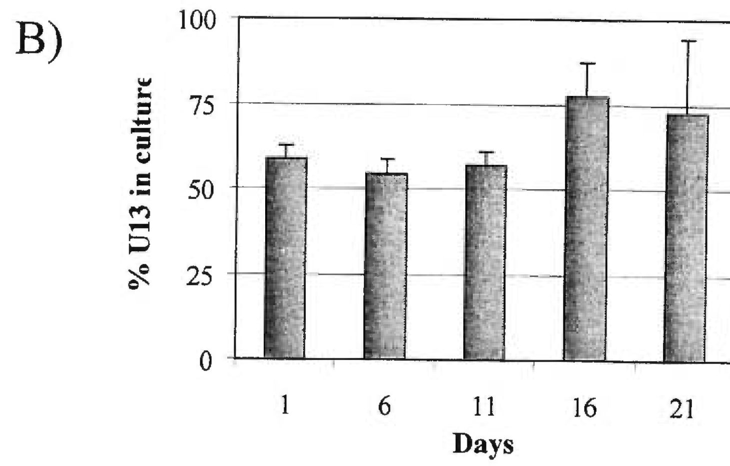
A) One example of competition between cells with $U13_{am}$ and $C13_{am}$. On different days after the first inoculation, plasmid DNA was extracted from cells, digested with PvuII and subjected to agarose gel electrophoresis and Southern blot hybridization with an oligonucleotide probe complementary to anticodon stem-loop. The rationale of the experiment is explained in the text. The lanes "U13" and "C13" contain DNA from cells with only either $U13_{am}$ or $C13_{am}$. The bands corresponding to each clone are shown to the left along with their approximate size.

B) Cumulative results for 24 competitions done in parallel. For different days after the first inoculation, the intensity of the bands derived from $U13_{am}$ and $C13_{am}$ was quantified and the proportion of $U13_{am}$ was calculated. The mean of the values for the 24 competitions is plotted for the selected days of competitions. Details and statistical analyses on these competitions are presented in the text.

A)



B)



ARTICLE 5

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

(Suppression de codons ambre chez *Escherichia coli* par des ARN de transfert atypiques de type mitochondrial)

Bourdeau, V., Steinberg, S.V., Ferbeyre, G., Émond, R., Cermakian, N. et Cedergren, R. (1998) *Proceedings of the National Academy of Sciences of the U S A* **95**:1375-1380.

(inclus avec la permission de la National Academy of Sciences of the USA, Copyright 1997)

Les contributions de N.C. à cet article sont:

- Participation aux expériences.
- Participation à la discussion et à la rédaction du manuscrit.

Biochemistry

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

(transfer RNA/suppressor/mitochondrial/translation/evolution)

Véronique Bourdeau*, Sergey V. Steinberg*, Gerardo Ferbeyre, Rémi Emond, Nicolas Cermakian and Robert Cedergren[1]

Département de biochimie, Université de Montréal
Montréal Québec H3C 3J7 CANADA
December 2, 1997

*The contribution of these two authors is to be considered equivalent.
[1]To whom correspondence should be addressed:

Université de Montréal, Tel.: (514) 343-6320
Département de biochimie, Fax: (514) 343-2210
C.P. 6128, succursale Centre-Ville, E-mail: ceder@bch.umontreal.ca
Montréal, Québec, H3C 3J7, CANADA

20 pages of text, 5 pages of figures, 2 pages of tables and 26,487 characters count
Abbreviations: tRNA, transfer RNA; mt, mitochondria; $tRNA^{Ala}_{su+}$, alanine suppressor transfer RNA; X-Gal, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside; DHFR, dihydrofolate reductase; AC, aminoacceptor; D, D-stem; AN, anticodon; T, T-stem.
Data deposition: GenBank accession numbers AF003201 to AF003233.

ABSTRACT

The "cloverleaf" base pairing pattern was established as the structural paradigm of active transfer RNA (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 five to ten base pairs in the anticodon stem rather than the canonical six. 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 seven, eight, nine and ten 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 which actively suppress amber codons *in vivo*.

INTRODUCTION

The vast majority of transfer RNAs (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 five to ten 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), non-universal genetic codes (6-8) etc. raise the issue that these tRNAs may be simply another anomaly of mitochondria rather than representatives of alternate folding patterns, since they are not found in cytoplasmic translation. We show here, to the contrary, that tRNAs incorporating these unusual features actively suppress amber mutations in *E. coli*, and covariation analysis of their sequences supports the patterns that we have assigned to the unusual mt tRNAs.

MATERIAL AND METHODS

STRAINS

Three *E. coli* strains have been used: Top10 (F⁻ *mrc*A Δ(*mrr-hds*RMS-*mcr*BC) F80*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L *end*A1 *nup*G) from Invitrogen (San Diego, CA); XAC-1 (F' $lacI_{373}lacZ_{m118\ am}$ *pro*B^{+}/ F⁻ Δ(*lac-pro*B)$_{XIII}$ *nal*A *rif* *arg*E_{am} *ara*) (9); and XAC/A16 (Δ*lacpro*B *nal*A *rif* *arg*E_{am} / F' *lac*Iq *amber*-Z fusion *pro*B) containing the pDa3am plasmid (10).

CONSTRUCTION OF THE COMBINATORIAL LIBRARY

The template oligonucleotide coding for the combinatorial tRNA library and two flanking primers with restriction enzyme sites for *EcoR*I and *Pst*I (sequences on request) were synthesized by General Synthesis and Diagnostics (Toronto). These were then PCR-amplified: 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, using Vent DNA polymerase (2U), 100 pmol of each primer, 200 µM of dNTPs and 100 ng of the template. All enzymes were from New England Biolabs, Boston. The double-stranded DNA obtained was then digested with *Pst*I and *EcoR*I, purified on Sephadex G-50 column and cloned into the pGFIB-I plasmid, which was predigested with the same enzymes and dephosphorylated with calf intestine phosphatase 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 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 of ampicilin and 20 μg/mL of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and left to grow overnight at 37°C. Blue colonies developed during a 24-hour 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).

tRNA levels and aminoacyl-tRNA levels

The tRNA and aminoacyl-tRNA levels were measured 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 using 2xYT broth containing 50 μg/mL of 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 0.3M sodium acetate (pH 5.2) and 1mM EDTA and extracted with an equal volume of phenol equilibrated with 0.3M 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 two volumes of ethanol. The pellet from centrifugation was washed with 0.5 mL volume of 70% ethanol/10 mM sodium acetate (pH 5.2) and resuspended in 40 mL of 10 mM sodium acetate (pH 5.2) and 1mM EDTA. The concentration of the RNA isolations is determined according to the absorbance of a 1:1000 dilution. Samples of 2 μg to 4 μg of each RNA preparation were distributed in two tubes, one of which was maintained at 4°C, while 1.5 μL of 0.5M 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 composed of 0.1M sodium acetate (pH 5.2), 8M 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 volts at 4°C. After electrophoresis, RNA was transfered 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 to 53 in the *E. coli* 5S sequence).

Aminoacylation levels were obtained from the quantity of aminoacyl-tRNA and uncharged tRNA in lanes which 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.

IDENTIFICATION OF THE CHARGED AMINOACID

The identification of the amino acid charged by 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, St. Louis, MO) 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 BioRad Econopac Q column with a gradient from 0 to 1M KCl. The fractions with DHFR were pooled and desalted on Centricon-30 (Amicon, Beverley, MA). DHFR was microsequenced in the laboratory of Dr. C. Lazure (Institut de recherche clinique de Montréal) after 15% SDS-PAGE and transfer to a PVDF membrane (Amersham).

RESULTS

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. Since 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* $tRNA^{Ala}_{UGC}$ 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* $tRNA^{Ala}_{UGC}$ gene to incorporate the mt characteristics:

1) The G26-A44 base pair was changed to G26-C44 in the designed gene. Since 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 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) (9).

CLONES SELECTION AND CHARACTERISATION

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 $1.3x10^5$. The amplification product was cloned into the pGFIB-I plasmid (20) 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 in order to ensure that the suppressor phenotype was not a result of a host mutation. Fig. 3*A* 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 β-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 $tRNA^{Ala}_{su+}$ control. Certain tRNAs 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. 3*B* 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 $tRNA^{Ala}_{su+}$. Analysis of these data showed that aminoacylation of all tRNAs was between 70% and 90% ± 5% (Table 2). Although generally the charging level of $tRNA^{Ala}_{su+}$ was greater than the variant tRNAs, these differences were within the experimental error. The steady state level for all tRNAs was also calculated 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, since suppression is saturable due to the negative effect of too much read through of stop codons. On the other hand, the gel mobility of the charged and non-charged tRNAs are consistent with the predicted length indicating that nuclease processing of the precursor tRNA was not disrupted by their unusual structure.

The identity of the amino acid inserted at the site of the amber codon was determined using the pDa3am plasmid which contains an amber mutation at the third codon of the encoded dihydrofolate reductase (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 since they were associated with high β-galactosidase activity and represented vastly different structural

groups (see below). 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.

Sequence Analysis

In all, 22 out of the 33 tRNA gene sequences had seven 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 which 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 out of 13, respectively. In all nine Type 7-2 tRNAs, the three D-stem base pairs are Watson-Crick base pairs. Since an insufficient number of unusual mt tRNA sequences are 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.

Discussion

Since mitochondria 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, mitochondria-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 steady 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, since otherwise they would interfere with normal termination of translation. Also, the acid gel experiment shown in Fig. 3*B* demonstrates that aminoacylation and maturation of the 5' and 3' termini is normal.

The fact that these mitochondria-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 cytoplastic protein synthesis system. We think not since: i) the origin of mitochondria is rooted in the α-purple bacteria, 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 mitochondria clearly demonstrates several independent origins for this class of tRNAs (15). Therefore, this structure must have originated several times during the evolution of cyctoplasmic translation, but then systematically rejected.

Our ability to prepare active tRNAs having unusual folding motifs in light of previous attempts (22, 23) (Bourdeau *et al.*, unpublished results) could be due to 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.

ACKNOWLEDGEMENTS

The authors 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 his 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 FCAR (Québec) scholarship to N. Cermakian. R. Cedergren is the Richard Ivey Fellow of the Canadian Institute of Advanced Research.

REFERENCES

1. Steinberg, S. & Cedergren, R. (1994) *Nature 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,* **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 U S A*, **82**, 1401-1405.
6. Barrell, B.G., Bankier, A.T. & Drouin, J. (1979) *Nature,* **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 U S A*, **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., Fritsh, E.F. & Maniatis, T. (1989) *Molecular cloning, a laboraotry manual, 2nd edition.* Cold Spring Harbor Laboratory Press (Cold Spring Harbor, New-York).
12. Vashney, U., Lee, C.-P. & RajBhandary, 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*, **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 Laboratory (Cold Spring Harbor, New York).
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.

Table 1. β-galactosidase activities and gene sequences of suppressor tRNAs.

β-galactosidase activity was determined using the method of Miller (21) in the presence of 50 μg/ml of ampicilin, and the use of chloroform and sodium dodecylsulfate (SDS). The activity is presented as a percentage of the activity of the control $tRNA^{Ala}_{su+}$. 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. Non-canonical base pairs are in italics. These sequence data have been submitted to the Genbank database under accession numbers AF003201 to AF003233. AC refers to the aminoacceptor stem; D, to the D-stem; AN, to the anticodon stem; 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.

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

* structural type defined in Fig. 4. † obtained by averaging β-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 $tRNA^{Ala}_{su+}$ control (Fig. 3b). These values were obtained from three experiments.

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; T, to the T loop and stem.

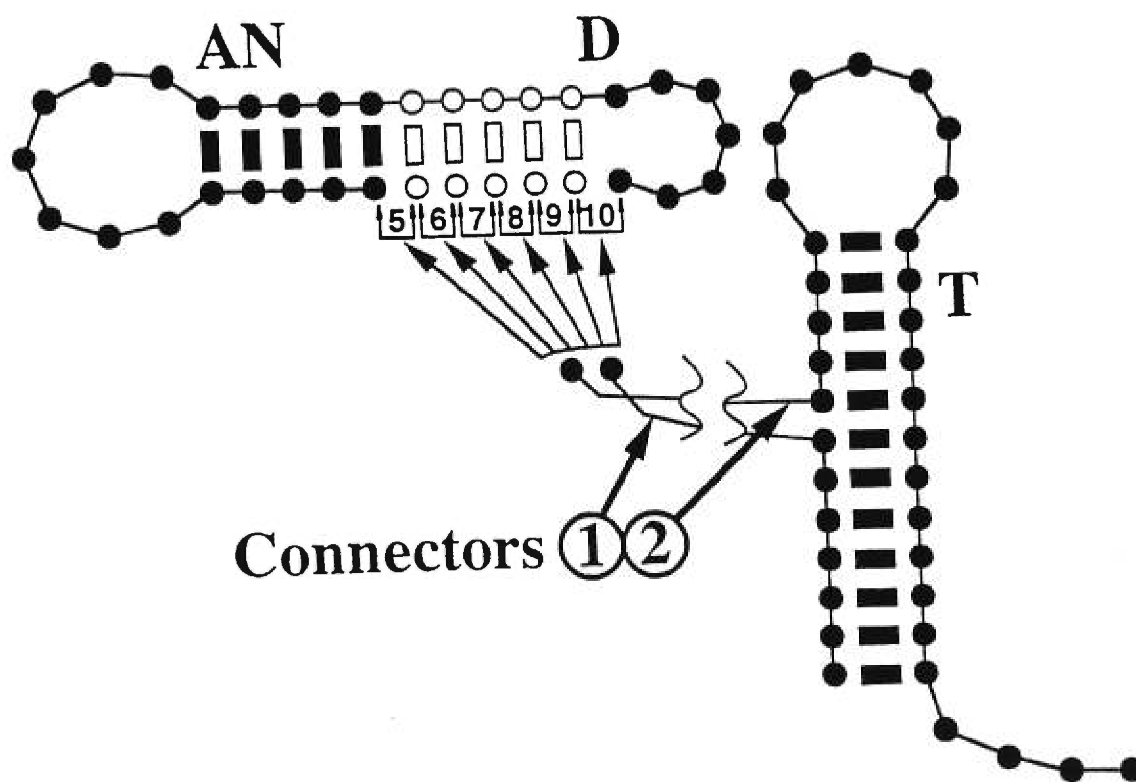

AN
D
5
6
7
8
9
10
T
Connectors
1
2

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. T's are used here, since the gene sequence is given.

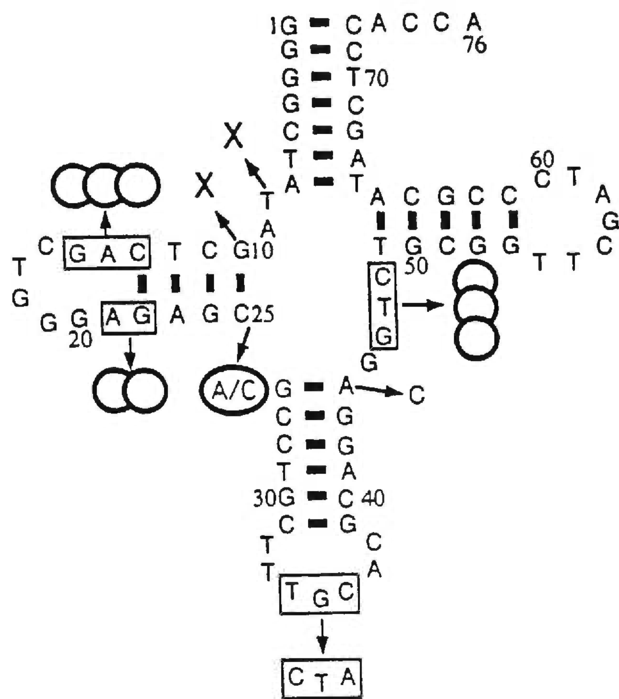
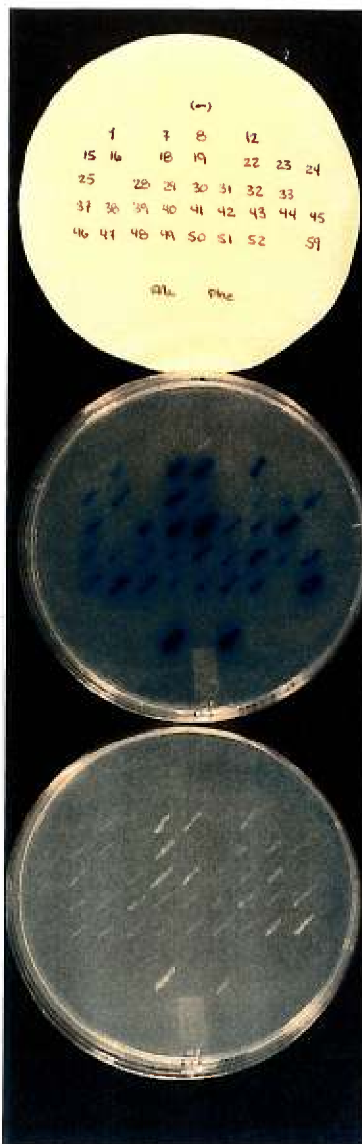

1G C A C C A
76
G C
G T70
G C
C G
X
T A
60
C T
X
A T A C G C C A
T
A G
C G A C T C G10 T G C G G T T C
T
50
G
C
T
G G A G A G C25
20
G
G A C
C G
C G
T A
30G C 40
C G
T C
T A
T G C
C T A

Figure 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 (50μg/ml), without arginine and with or without X-gal (20μg/ml). Ala represents the plasmid containing the $tRNA^{Ala}_{su+}$ 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) 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 non aminoacylated species. The values for each experiment was normalized to the aminoacylated level of the control $tRNA^{Ala}_{su+}$. The levels of expression were determined in the same way and normalized to the level of the 5S RNA in each lane.

A)



B)

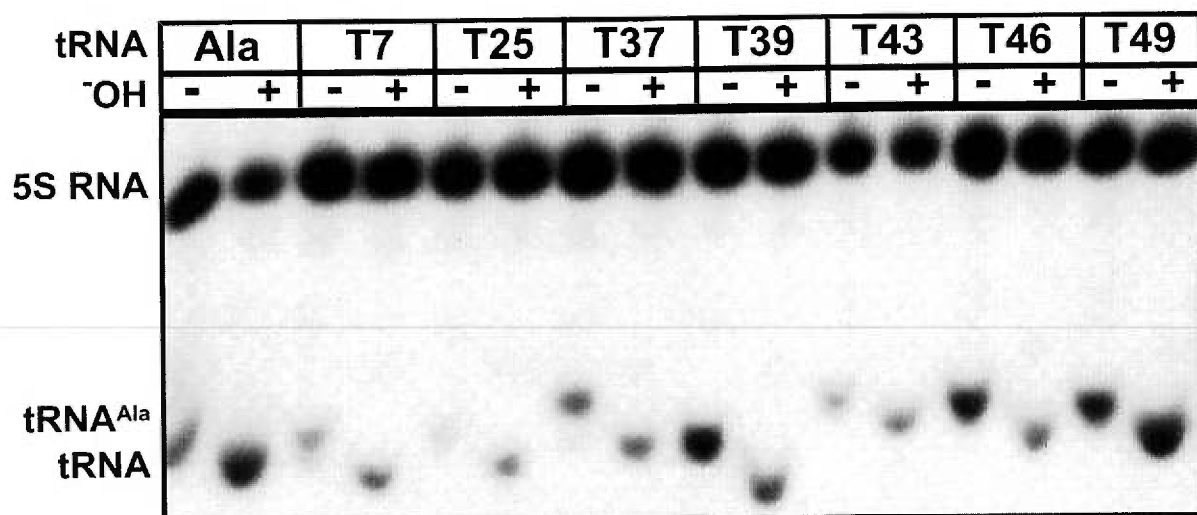
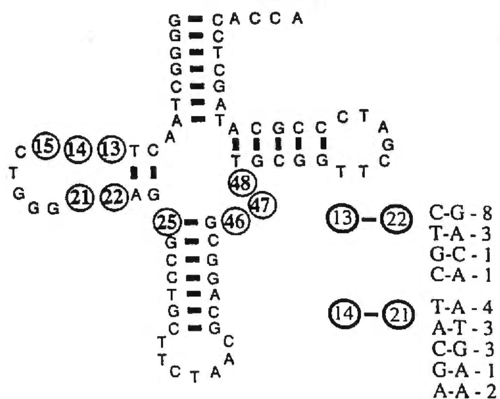
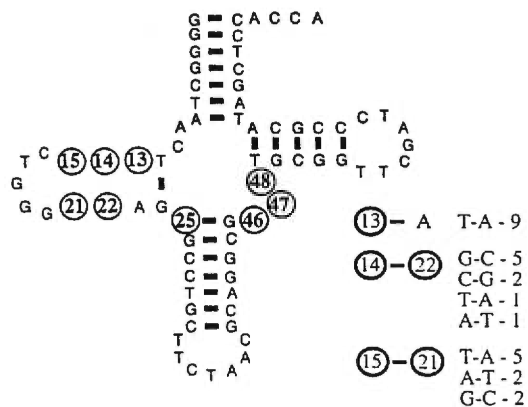


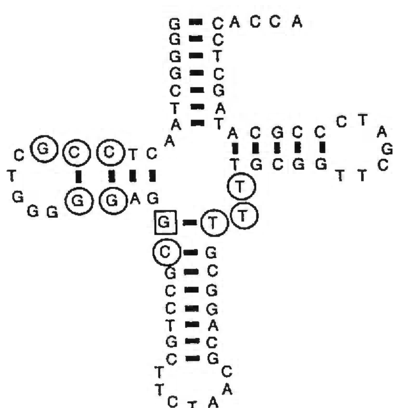
Figure 4. Secondary structure of isolated tRNAs. Optimal secondary structure pattern for the type 7-1, 7-2, 8, 9 and 10 tRNAs which were isolated in these experiments. Type 7-1 and 7-2 are shown with the covariation data from the D stem of these tRNAs. None of these sequences can be folded into the canonical cloverleaf structure.



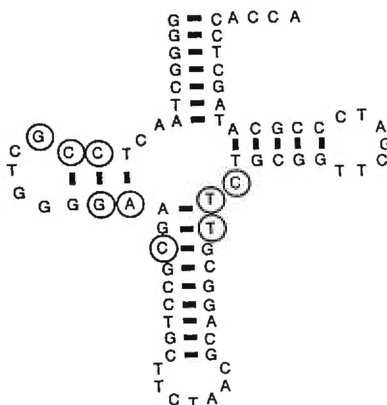
Type 7-1



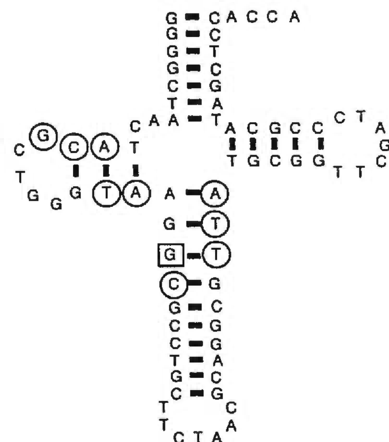
Type 7-2



Type 8
T7



Type 9
T45



Type 10
T37

DISCUSSION

1. La "Fresque de la Mitochondrie"

Vers la fin du XIXè siècle, alors que la recherche sur les mitochondries en était à ses balbutiements, alors que le nom "mitochondrie" n'était qu'une proposition parmi des dizaines, le peintre norvégien Edvard Munch (1863-1944) peignait les tableaux qu'il regrouperait ensuite dans sa "Fresque de la Vie", composée de quatre étapes: "Éveil de l'Amour", "Épanouissement et Déclin de l'Amour", "Angoisse de Vivre" et "Mort" (Bischoff, 1993). De nouvelles études publiées au cours de la dernière année viennent préciser l'histoire de la mitochondrie, de sa naissance survenue il y a un ou deux milliards d'années, à son déclin et même sa disparition apparente chez des organismes actuels. Je discuterai dans les prochaines pages de ces nouvelles études dans le contexte de mes résultats, et je ferai un retour sur mes résultats à la lumière de ces nouvelles découvertes.

1.1 "Éveil de la Mitochondrie"

Nous avons vu dans l'introduction que la mitochondrie a pour origine un événement d'endosymbiose entre une eubactérie et un hôte eucaryote primitif. Cependant, les mitochondries connues sont bien différentes de leurs ancêtres bactériens, en particulier pour ce qui est de leur génome. Pour tenter d'éclaircir les différentes étapes de l'histoire tumultueuse des mitochondries, un groupe de chercheurs canadiens, faisant

partie de l'Institut Canadien de Recherche Avancée, a mis sur pieds un projet de séquençage de génomes mitochondriaux. Ce projet est jusqu'à maintenant très fructueux, puisque plus de vingt génomes mitochondriaux provenant d'eucaryotes les plus diversifiés possible ont été séquencés et analysés, entraînant des découvertes importantes pour la connaissance de ces organites et de l'évolution des eucaryotes (Palmer, 1997b; voir aussi le site: http://megasun.bch.umontreal.ca/ogmp). Le génome mitochondrial le plus intéressant parmi tous ceux séquencés jusqu'ici par ce groupe est sans contredit celui du protiste hétérotrophe flagellé *Reclinomonas americana*, organisme qui a divergé très tôt des autres eucaryotes (Lang et al., 1997). Cet article récent fait état des caractéristiques les plus frappantes de cet ADN mitochondrial (ADNmt). Le génome mitochondrial de *R. americana* est celui qui jusqu'ici présente le plus de caractéristiques eubactériennes. Cette découverte vient confirmer l'hypothèse endosymbiotique, en fournissant en quelque sorte un "chaînon manquant" entre les gros génomes bactériens et les génomes très dérivés des mitochondries d'organismes mieux connus.

Parmi les caractéristiques de l'ADNmt de *R. americana*, on note les faits suivants (Lang et al., 1997; Palmer, 1997b):

• Il contient 92 gènes, soit 25 de plus que le plus riche ADNmt connu jusqu'ici.

• Parmi les 62 gènes codant pour des protéines, 18 n'avaient jamais encore été trouvés dans un génome mitochondrial.

• Cet ADNmt contient des gènes de types encore jamais trouvés dans les mitochondries: *rpoA, B, C, D* (codant pour les sous-unités d'une ARN polymérase à plusieurs sous-unités [msRNAP]), *tufA* (pour le facteur de traduction EF-Tu), *secY* (pour une protéine de voie de sécrétion) et *cox11* (pour une protéine d'assemblage de la cytochrome oxydase).

• L'analyse des séquences d'ARNr et d'ARNm suggère qu'une interaction de type Shine-Dalgarno est possible dans le système de traduction des mitochondries de *R. americana*.

• Plusieurs regroupements de gènes de l'ADNmt de *R. americana* sont très semblables à des opérons bactériens. Un exemple est l'opéron *str* suivi des opérons S10, *spc* et α, comme chez les eubactéries.

1.2 "ÉPANOUISSEMENT ET DÉCLIN DE LA MITOCHONDRIE"

On pourrait dire que la mitochondrie est un laboratoire d'évolution. Tout y est essayé, expérimenté, parfois de façon excessive, et c'est ce qui fait de cet organite un sujet d'étude passionnant et plein de surprises. Palmer (1997b) a pour sa part surnommé la mitochondrie "nature's molecular playground" pour les mêmes raisons. Comment se fait cette évolution effrénée? Eh bien, tous les moyens semblent bons! Pour s'en convaincre, on n'a qu'à se rapporter aux exemples donnés dans l'Introduction et les Chapitres de cette thèse. Et la liste est bien plus longue, puisqu'on n'a pas abordé le sujet de la maturation post-transcriptionnelle des ARN, particulièrement complexe et variée chez les mitochondries, avec les diverses formes d'édition et d'épissage que peuvent subir les transcrits dans ces organites (Palmer, 1997b).

On en arrive au "Pourquoi?" de cette évolution tumultueuse. Nous voilà devant une grande question! Je cède la plume à Stephen Jay Gould pour quelques lignes:

> « Si les animaux étaient façonnés de façon idéale, chacun de leurs organes ne faisant qu'une seule chose à la perfection, alors, il ne pourrait pas y avoir d'évolution, car rien ne pourrait changer (sans perdre une fonction vitale au cours des étapes de transition), et la vie s'arrêterait rapidement dès que le milieu se modifierait, ces organismes ne pouvant y répondre.
>
> Mais les lois structurales, plus profondes que la sélection naturelle elle-même, garantissent que les traits complexes regorgent obligatoirement de possibilités multiples — et l'évolution se trouve dotée de la flexibilité requise grâce au désordre, à la redondance, et à l'absence d'un ajustement parfait. (...) nous avons ici affaire à (...) quelque chose de si général que cela doit s'appliquer à tout. » (Gould, 1996)

Gould parle ici des animaux et de leurs organes, mais comme il le fait remarquer, on peut appliquer les mêmes principes à différents niveaux d'évolution. Dans le cas des mitochondries, il faut chercher ce qui en fait des "laboratoires d'évolution" dans leur statut "d'endosymbiote permanent" (certains de ces points s'appliqueraient d'ailleurs aussi bien aux chloroplastes). Nous allons voir que ces causes rejoignent tout à fait les idées de Gould:

1) Même le génome mitochondrial de *R. americana*, bien que relativement riche avec ses 62 gènes codant pour des protéines, est bien loin des bactéries. Par exemple, le génome de *Mycoplasma genitalium*, un des plus petits génomes bactériens, contient tout de même près de 500 gènes codant pour des protéines (Fraser et al., 1995). Vu que *R. americana* a divergé assez tôt des autres eucaryotes, l'explication la plus simple est que le génome mitochondrial a très tôt subi une perte massive de gènes, soit par transfert vers le génome nucléaire, soit par perte pure et simple (Palmer, 1997b). Une telle décimation dans un court laps de temps a pu se faire de façon plutôt brouillonne, désordonnée, ce qui aurait poussé ce système à improviser dans certaines situations. Par exemple, c'est probablement peu de temps après la divergence de *R. americana* que les autres eucaryotes ont adopté une ARN polymérase à une sous unité (ssRNAP) pour transcrire les gènes mitochondriaux; cette ssRNAP aurait acquis le monopole de la transcription mitochondriale suite à une perte irrémédiable des gènes pour la msRNAP (Chapitre 1).

2) Comme on l'a vu dans l'Introduction et au Chapitre 3, il est probable que la machinerie d'expression des gènes mitochondriaux, en particulier pour ce qui est de la traduction, soit devenue imparfaite, suboptimale, peut-être à cause d'un plus petit nombre de gènes à prendre en charge (Hasegawa et al., 1984; Benne et Sloof, 1987) et du fait que des mutations dans cette machinerie n'entraînent pas dans les mitochondries une cascade d'erreurs, comme ce doit être le cas dans le cytoplasme (Hasegawa et al., 1984). Les mitochondries se seraient donc affranchies de certaines contraintes des cellules. Quoi de plus favorable à une évolution rapide? Un exemple est le cas du nombre réduit d'ARNt dans les mitochondries animales, avec des règles

de *wobble* beaucoup plus relâchées: ceci permet de toute évidence une dérive plus rapide de l'ADN de ces mitochondries (Gray, 1989; Osawa et al., 1992). Un autre exemple est celui des ARNt mitochondriaux atypiques qui ont par contre été soigneusement évités dans les cytoplasmes (voir Introduction et Chapitre 3). Même les protéines régulatrices mitochondriales codées par le génome nucléaire évoluent vite, comme l'avaient prédit Hasegawa et al. (1984); cela a par exemple été démontré pour mtTFB (Carrodeguas et al., 1996).

3) Une autre notion de Gould qui s'applique certainement à l'évolution des mitochondries est la multiplicité. Multiplicité de gènes, premièrement: plusieurs gènes proto-mitochondriaux avaient un homologue nucléaire. On connaît certains cas où la même protéine est utilisée à la fois dans le cytoplasme et les mitochondries. C'est par exemple le cas d'enzymes de modification d'ARNt (Wolfe et al., 1994) et d'aminoacyl-ARNt synthétases (Natsoulis et al., 1986). Certains gènes codant pour des protéines mitochondriales pourraient même provenir d'autres sources (plastides ou endosymbiotes temporaires) (Gray, 1992). Deuxièmement, on peut avoir multiplicité de fonctions pour une même protéine. Un exemple déjà discuté dans cette thèse est le cas de mtTFA, qui sert à la fois de protéine structurale pour le génome mitochondrial et d'activateur transcriptionnel, en plus d'être impliqué dans l'initiation de la réplication de l'ADNmt. Le simple fait que les trois mtTFA étudiés jusqu'ici (Parisi et al., 1993; Antoshechkin et Bogenhagen, 1995) ont une importance relative différente de ces activités montre bien le potentiel de plasticité que permet une telle multiplicité de fonctions. Troisièmement, le simple fait qu'il y a un grand nombre de mitochondries dans une cellule pour accomplir un rôle global à l'intérieur de celle-ci peut favoriser une évolution rapide. Le fait d'avoir plusieurs mitochondries permet une additivité de leur fonction: l'union fait la force. De plus, une mutation même défavorable peut survenir et éventuellement être fixée sans coupure brusque et baisse d'efficacité soudaine, puisque les autres mitochondries de la cellule peuvent prendre la relève durant la transition; une situation opposée est par contre retrouvée dans le noyau, où seulement une ou deux copies d'un gène donné sont habituellement présentes, ce qui confère un effet important à toute mutation.

On peut cependant aborder le problème sous un autre angle: les mitochondries peuvent être vues comme une petite population de micro-organismes qui se propagent dans le cytoplasme (Kurland, 1992; Andersson et Kurland, 1995; Lynch, 1997). À partir de là, deux forces peuvent pousser la mitochondrie à réduire le plus possible son génome. La première est la compétition entre les mitochondries dans cette population, compétition qui favorise celles dont le taux de croissance sera le plus élevé (Kurland, 1992). Ainsi, des mitochondries ayant un génome plus petit à répliquer, moins de gènes à exprimer ou toute autre caractéristique augmentant leur vitesse de croissance et de division seront avantagées. La deuxième force impliquée est appelée “le cliquet de Muller” (*Muller's ratchet*; Lynch, 1996, 1997). Selon ceci, de petites populations peuvent subir des fluctuations statistiques. Même des génomes défectifs peuvent ainsi supplanter un génome normal; ceci arrive d'ailleurs dans le cas de certains mutants respiratoires de la levure, et dans le cas de mutations mitochondriales humaines qui sont retenues et causent des maladies (Kurland, 1992). Puisque pour les délétions, la probabilité de revenir en arrière est à peu près nulle, la population tendra nécessairement vers de plus petits génomes. Lynch (1996, 1997) a comparé le taux d'évolution de la séquence d'ARNt mitochondriaux et nucléaires, et en a conclu que le génome mitochondrial subit les effets du “cliquet de Muller”. Cette perte peut être compensée par transfert de séquences vers le noyau (Kurland, 1992). En fait, les deux mêmes forces influent sur l'évolution des génomes de bactéries parasites (Andersson et Kurland, 1995).

Ainsi, selon ce point de vue, la tendance à avoir un génome toujours plus petit entraînerait les mitochondries dans la voie des innovations. Cela expliquerait par exemple la petite taille de certains ARNt et ARNr mitochondriaux: même si leur efficacité ou leur aptitude à remplir leurs fonctions est amoindrie, ceci est contrebalancé par l'avantage d'un taux de croissance plus grand de la mitochondrie (Andersson et Kurland, 1995). Par ailleurs, les escargots ont un type d'édition de leur ARNt mitochondriaux qui permet d'avoir des gènes d'ARNt qui se chevauchent; or,

les escargots ont parmi les plus petits ADNmt de tous les animaux (Yokobori et Pääbo, 1995).

Cependant, cette vision des choses, bien que sensée, n'explique pas tout: entre autres, que certains génomes de plantes ont une taille avoisinant celle des génomes bactériens (Gray, 1989). Par ailleurs, il est difficile de faire rentrer dans un tel cadre les longues séquences apparemment sans rôle et même les nombreux introns qu'on retrouve dans les mitochondries de nombreux eucaryotes (Gray, 1989).

Peu importe la raison exacte du déclin des génomes mitochondriaux et de l'épanouissement des expériences évolutives de ces organites, une question demeure: pourquoi les chloroplastes ressemblent-ils plus aux bactéries que les mitochondries? En fait, ce problème ne persiste que si on croit que les organites ont cessé d'évoluer. Mais, comme tout être vivant ou système biologique, les organites sont très loin d'être statiques, évolutivement parlant. Les mitochondries, les chloroplastes et leurs génomes *sont en évolution active*! En fait, des expériences révèlent que la tendance vers les petits génomes continue à agir dans les génomes mitochondriaux de champignons et d'animaux (voir ci-haut; Andersson et Kurland, 1995). De plus, des gènes en cours de transfert des mitochondries et des chloroplastes vers le noyau ont en quelque sorte été "pris la main dans le sac": on a pu trouver des étapes intermédiaires successives de transfert chez des plantes très rapprochées (Baldauf et Palmer, 1990; Nugent et Palmer, 1991; Brennicke et al., 1993).

Vu sous cet angle, on peut maintenant émettre l'hypothèse que les chloroplastes ont moins de caractères dérivés parce qu'ils sont le fruit d'une (ou des) symbiose plus récente. L'illustration la plus frappante touche directement au sujet principal de cette thèse: l'évolution des ARN polymérases. Comme nous l'avons mentionné, l'ADN mitochondrial de *R. americana* contient tous les gènes nécessaires pour produire une msRNAP de type bactérien (Lang et al., 1997). On ne sait toujours pas si cet organisme a un gène nucléaire pour une ARN polymérase à une sous-unité (ssRNAP). Par ailleurs, les mitochondries plus dérivées (presque toutes les autres) ne semblent

avoir qu'une seule mtRNAP, qui est une ssRNAP codée par un gène nucléaire (voir le Chapitre 1). Tel que souligné plus haut, tôt dans l'évolution de la mitochondrie, la msRNAP héritée de l'ancêtre eubactérien a dû être remplacée par une ssRNAP. Les chloroplastes, pour leur part, ne semblaient avoir qu'une msRNAP semblable à celle des cyanobactéries, et codée par le génome chloroplastique. Mais au cours des dernières années, des expériences ont suggéré que ces organites avaient peut-être une autre RNAP, codée dans le noyau (Hess et al., 1993; Lerbs-Mache, 1993). Ceci a été confirmé tout récemment par la publication de la séquence de deux ADNc d'*Arabidopsis thaliana*, tous deux codant pour une ssRNAP, l'une étant ciblée à la mitochondrie, l'autre au chloroplaste (Hedtke et al., 1997). Ces deux ssRNAP branchent ensemble dans les phylogénies, ce qui suggère que la cp-ssRNAP (ARN polymérase chloroplastique à une sous-unité) origine d'une duplication du gène de la mt-ssRNAP (Hedtke et al., 1997). Les deux organites n'ont donc pas adopté indépendamment une ssRNAP. Ces résultats suggèrent que les chloroplastes peuvent renvoyer une image intermédiaire entre l'endosymbiote bactérien et les systèmes très dérivés de la plupart des mitochondries actuelles. La cp-msRNAP (type bactérien) sera-t-elle un jour remplacée par la cp-ssRNAP, comme dans les mitochondries?

Attention, cependant: les chloroplastes ne sont pas les mitochondries! Les voies d'évolution que ces organites empruntent ont été et devraient être différentes (sauf pour certains thèmes généraux), premièrement parce que leurs rôles sont différents, mais aussi parce que l'évolution n'emprunte pas deux fois le même chemin, si l'on suit une autre des idées si chères à S.J. Gould, la *contingence* (Gould, 1991).

1.3 "Angoisse de Vivre, ou plutôt de Respirer"

Étant donné les forces puissantes guidant la réduction du génome mitochondrial, quelle est la limite? Est-ce que les mitochondries se dirigent vers la perte pure et simple de leur génome? Bien qu'on ne puisse écarter une telle éventualité, encore aucune mitochondrie sans génome n'a été trouvée. Le plus petit ADNmt connu, celui de *P. falciparum*, contient les gènes des deux grands ARNr, et *cob*, *cox1* et *cox3* (Wilson et

Williamson, 1997), les quatre premiers étant d'ailleurs présents dans tous les ADNmt séquencés jusqu'ici (Palmer, 1997b).

Certaines hypothèses ont été avancées pour expliquer que certains gènes n'ont jamais été transférés au noyau. Von Heijne (1986) a suggéré que certaines protéines étaient trop hydrophobes pour qu'il leur soit possible d'être importées dans les mitochondries *in vivo* sans éviter d'être détournées ailleurs (comme au réticulum endoplasmique). Selon Jacobs (1991), la rétention de l'ADNmt résulterait d'un mécanisme actif, semblable au système agressif de rétention de plasmide chez les bactéries. Dans le cas des grands ARNr, on pourrait proposer que leur grande taille empêche leur transport à travers la double membrane mitochondriale, contrairement à de petits ARN.

Quoi qu'il en soit, la perte totale du génome mitochondrial pourrait bien signifier la perte de la fonction de phosphorylation oxydative. Ceci est appuyé par quatre études phylogénétiques récentes (Bui et al., 1996; Germot et al., 1996; Horner et al., 1996; Roger et al., 1996). Les hydrogénosomes sont des organites présentant des caractéristiques en commun avec les mitochondries, mais qui n'ont pas de génome et ne produisent pas l'ATP par phosphorylation oxydative, mais plutôt de façon anaérobique (Sogin, 1997; voir la section 2.3 de l'Introduction). Le fait que des phylogénies faites avec des séquences de protéines chaperonnes regroupent ces deux types d'organites suggère que les hydrogénosomes dérivent des mitochondries, ou du moins que ces deux organites ont une origine évolutive commune (Palmer, 1997a). De plus, le fait que les hydrogénosomes soient retrouvés dans des groupes eucaryotes non reliés (c'est-à-dire qu'ils sont polyphylétiques) et coïncident avec une absence de mitochondrie (Bui et al., 1996; Germot et al., 1996), peut plus facilement être expliqué par une réduction extrême des mitochondries que par une acquisition indépendante d'organites semblables par des organismes diversifiés.

1.4 "MORT (OU DISPARITION) DE LA MITOCHONDRIE"

Les expériences sur les hydrogénosomes rapportées ci-haut ont d'autres implications peut-être encore plus fondamentales. Trois groupes de protistes situés à la base de l'arbre eucaryote, les diplomonades, les microsporidies et les trichomonades, n'ont pas de mitochondries (Sogin, 1997). Les trichomonades ont des hydrogénosomes dérivant probablement des mitochondries. De plus, le diplomonade *Giardia lamblia* possède apparemment des protéines pouvant provenir d'un endosymbiote mitochondrial (Roger et al., 1996). Ces observations suggèrent que les eucaryotes sans mitochondrie en ont en fait déjà eu mais les ont perdues (Sogin, 1997). Ce point referme la boucle, puisqu'il ramène à se demander quand a eu lieu l'endosymbiose qui a donné naissance à la mitochondrie. Cet événement pourrait donc avoir eu lieu très tôt dans l'évolution des eucaryotes, avant la divergence des diplomonades, et peut-être peu de temps après la divergence des eucaryotes des archaebactéries, ou la création de l'eucaryote ancestral par fusion de deux procaryotes, selon l'hypothèse que l'on préfère...

Si cela était vérifié, l'histoire évolutive des mtRNAP devrait être réévaluée. Nos résultats, alliés à ceux sur *R. americana* (Lang et al., 1997), suggèrent que la RNAP de l'endosymbiote a rapidement été remplacée par une ssRNAP. Cependant, si les mitochondries peuplaient déjà les cytoplasmes depuis longtemps, l'acquisition des ssRNAP pour jouer le rôle de mtRNAP n'aurait plus à être considérée comme un événement urgent, pressé, puisqu'on sait que *R. americana* a une msRNAP et que l'acquisition de la ssRNAP s'est donc probablement faite subséquemment à la divergence de cet organisme. Cet événement aurait donc eu lieu après l'endosymbiose qui est à l'origine des mitochondries, suite à un événement indépendant de celle-ci. Bien que tout ceci ne demeure que spéculation, on voit bien que la connaissance de l'histoire des mtRNAP gagnera beaucoup de l'étude des organismes sans mitochondries et de l'analyse des génomes mitochondriaux.

2. Implications générales des travaux présentés dans cette thèse

Prenons un peu de recul et nous verrons que la plupart des études présentées dans cette thèse peuvent être vues dans un cadre beaucoup plus grand. Analysons donc maintenant ces résultats en regard de l'évolution des êtres vivants et non plus seulement de l'évolution de la mitochondrie.

2.1 L'origine des ARN polymérases

L'hypothèse du monde à ARN suggère qu'aux commencements de la vie sur Terre, les fonctions de stockage de l'information génétique et de catalyse étaient toutes deux remplies par des molécules d'ARN (Darnell et Doolittle, 1986; Cedergren et Grosjean, 1987). Certains de ces ARN catalytiques avaient une activité de polymérase à ARN. Par la suite, cette activité fut laissée à la charge de protéines, comme la majorité des fonctions catalytiques. Comme il est peu probable qu'une ARN polymérase à plusieurs sous-unités soit apparue d'un seul coup, il est plus sage de proposer que les premières ARN polymérases étaient plus simples, avec une seule constituante. Cependant, la discussion du deuxième article du Chapitre 1 met en évidence que les ssRNAP du type T7 n'étaient pas ces ssRNAP ancestrales; il est même probable qu'elles aient vu le jour bien après les msRNAP. Mais il se pourrait bien que les petites cousines des ssRNAP du type T7 se méritent l'honneur d'être les premières ARN polymérases! On se rappelle en effet que ces ssRNAP font partie d'une grande famille comprenant des ARN et ADN polymérases, dépendantes de l'ARN ou de l'ADN (Delarue et al., 1990). Les premières RNAP protéiques étaient vraisemblablement des RNAP dépendantes de l'ARN (Lazcano et al., 1988), surtout si l'ADN n'est apparu que plus tard. Ensuite, pour des besoins de régulation plus complexe, le système à plusieurs sous-unités a été mis en place. Mais alors pourquoi les ssRNAP du type T7 sont-elles apparues? Certains virus, la plupart en fait, se débrouillent très bien avec la polymérase de l'hôte (Watson et al., 1987). Si la réponse à cette question ne s'est pas effacée des "archives moléculaires" au cours du temps, c'est probablement dans la multitude de génomes

viraux non encore explorés qu'elle se trouve; un exemple frappant est le cas du phage de cyanobactéries LPP-1, qui possède peut-être une ssRNAP (voir deuxième article du Chapitre 1).

2.2 ÉPISSAGE ALTERNATIF, GONADES ET COMMUNICATION MITOCHONDRIE-NOYAU

On m'a posé il y a quelque mois une question très intéressante et à laquelle je réfléchis depuis: comment se fait-il que tellement d'événements d'épissage alternatif spécifiques aux gonades aient lieu? Le cas de mtTFA/tsHMG présenté au Chapitre 2 est loin d'être le seul exemple de protéine somatique constitutive ayant un isoforme spécifique pour les testicules exprimé de cette façon. Toutes sortes de protéines sont dans cette situation: facteurs de transcription tel CREM (Foulkes et Sassone-Corsi, 1996), DNA ligase (Mackey et al., 1997), synthétase d'oxyde nitrique (Wang et al., 1997), protéine liant l'actine (von Bülow et al., 1997) ne sont que quelques exemple parmi tant d'autres. Serait-ce parce que les cellules des gonades ont besoin d'une régulation spécifique et précise de l'expression des gènes et des activités intracellulaires? On pourrait faire la même remarque pour bien des organes et tissus. Cette dernière remarque suggère que si l'on cherchait pour d'autres tissus, on trouverait également tout un groupe d'isoformes spécifiques de protéines somatiques. Une recherche d'articles portant sur des isoformes du système nerveux, un autre tissu qui a "besoin d'une régulation spécifique et précise", m'a permis de remarquer que c'était peut-être tout bêtement la réponse à la question...

Comme nous l'avons vu, la plupart des protéines mitochondriales sont codées par des gènes nucléaires. De plus, presque toutes les macromolécules impliquées dans l'expression et la réplication du génome mitochondrial, ainsi que dans la biogenèse de cet organite, proviennent également de gènes nucléaires. Ces faits fournissent différents niveaux de régulation possible du noyau vers la mitochondrie (Poyton et McEwen, 1996), et d'autres mécanismes sont peut-être impliqués. D'un autre côté, les mécanismes de la régulation en sens inverse, c'est-à-dire de la mitochondrie vers le noyau, sont encore mal connus. Une des hypothèses est que la mtRNAP ou des

facteurs de transcription mitochondriaux seraient actifs dans le noyau, mais ceci n'a pu être appuyé expérimentalement (Marczynski et al., 1989). Des facteurs de transcription nucléaires distincts pourraient par ailleurs répondre à des signaux venant de la mitochondrie. Ce pourrait être le cas de NRF-1 et NRF-2 (*Nuclear Respiratory Factors*), qui modulent l'expression de différentes protéines mitochondriales dans le noyau (Poyton et McEwen, 1996). Le Chapitre 2 de cette thèse présente une autre voie possible de communication noyau-mitochondrie. Comme décrit dans ce chapitre, tsHMG pourrait être impliquée dans l'activation ou la répression de gènes nucléaires importants dans les transformations que subissent les mitochondries dans les spermatides. Cependant, des résultats récents de Larsson et al. (1997) suggèrent que le rôle de cet événement d'épissage alternatif serait plutôt de diminuer la quantité de messagers pouvant coder pour mtTFA; en effet, les spermatides humains ne semblent pas avoir d'isoforme nucléaire tsHMG, mais la protéine mitochondriale est tout de même sous-exprimée, suite à un épissage alternatif produisant des ARNm inutilisables. Mais cela n'explique pas pourquoi les spermatides de souris font un tsHMG nucléaire apparemment en bon état.

2.3 PETIT NE SIGNIFIE PAS NÉCESSAIREMENT SIMPLE, ET SIMPLE, PAS NÉCESSAIREMENT VIEUX; LE CAS DES ARN DE TRANSFERT

Les ARNt sont à première vue de petites molécules (Figure 4 de l'Introduction) et certains pourraient en conclure que leurs caractéristiques, en particulier leur évolution, sont simples. S'il y a une chose que j'ai apprise lors de mes études sur les ARNt, c'est bien qu'il en est tout autrement. Les ARNt interagissent avec une multitude de macromolécules, aussi bien des protéines (ribonucléases spécifiques, enzymes de modification, aminoacyl-ARNt synthétases, facteurs d'élongation) que des ARN (autres ARNt, ARNr, ARNm) ou des ribonucléoprotéines (ribosome, ribonucléase P). Söll (1993) a évalué qu'un ARNt moyen interagit avec 30 à 40 protéines différentes au cours de son cycle de vie! Différents éléments de structure primaire, secondaire ou tertiaire sont des signaux pour la modification (Björk, 1995), l'aminoacylation (McClain, 1995), l'édition (dans le sens de "vérification et correction

d'aminoacylation", Hale et al., 1997; et dans le sens de "transformation de la séquence", Smith et al., 1997), la reconnaissance par EF-Tu (Clark et al., 1995) et bien sûr l'interaction codon-anticodon. Tout ça dans environ 80 nucléotides. Quel chef-d'œuvre d'économie et de subtilité! Donc, les ARNt n'évoluent pas comme toute autre macromolécule; ils subissent peut-être même beaucoup plus de contraintes que les gènes codant pour des protéines (Cedergren et al., 1981; Kumazawa et Nishida, 1993). On a vu au Chapitre 3 que l'évolution n'y va pas de main morte avec les ARNt. Par exemple, dans le premier article, un simple nucléotide cause une infime différence d'activité ou de stabilité et est remplacé dans à peu près tous les ARNt (sauf dans les mitochondries...).

Il reste beaucoup de choses à découvrir sur la régulation et la fonction des ARNt (et des autres ARN). Ce constat est bien illustré par le cas des nucléosides modifiés. Face à l'ampleur du phénomène et à la complexité de l'appareillage cellulaire qui y est consacré (Björk, 1995), il serait difficile de prétendre que la modification des nucléosides dans les ARN n'a aucune fonction. Dans un chapitre de livre présenté en annexe de cette thèse, nous proposons un modèle évolutif pour l'origine de ce phénomène et sa rétention subséquente. Cette dernière est sans doute due à certains avantages évolutifs conférés aux ARN par les nucléosides modifiés. Pourtant, dans bien des cas, les ARN ou protéines nécessaires pour de telles modifications ne sont pas essentiels pour la cellule (Björk, 1995; Smith et Steitz, 1997), et leur délétion peut même être exécutée sans entraîner de phénotype anormal (Smith et Steitz, 1997). Il est intéressant de noter que les ARNt mitochondriaux sont très peu modifiés en comparaison des ARNt cytoplasmiques (Martin, 1995). L'analyse des nucléosides modifiés des ARNt de type mitochondrial exprimés dans une bactérie (Chapitre 3) pourrait permettre de mieux comprendre les particularités des ARNt mitochondriaux, mais aussi le rôle de la modification post-transcriptionelle des ARN en général.

Enfin, beaucoup prennent pour acquis que tout va vers une complexité croissante. Or, l'évolution n'est pas une marche implacable vers le mieux, le plus grand, le plus fort (Gould, 1991). Un bon exemple est le cas des ARNt

mitochondriaux. Ceux-ci présentent souvent des caractéristiques plus simples que leurs analogues cytoplasmiques: moins de nucléotides modifiés, tel que noté ci-haut, moins d'ARNt pour chaque acide aminé, domaines manquants, etc. Or, dans tous les cas, ces simplifications ont nécessairement une origine tardive, puisque la mitochondrie a un ancêtre bactérien avec des ARNt plus complexes, sans compter que plusieurs de ces phénomènes sont survenus plusieurs fois. Donc, l'idée courante en biologie que ce qui est plus simple doit être plus ancien est dans bien des cas erronée, tel qu'illustré par l'étude des ARNt mitochondriaux.

3. CONCLUSION

Les études évolutives sont d'une grande importance et valent largement la peine d'y consacrer temps et efforts, et ce pour différentes raisons. Premièrement, comme nous l'avons vu, parce que ces études nous permettent de mieux comprendre les mécanismes biologiques actuels. Deuxièmement et plus directement, pour la connaissance de l'histoire de la vie, en fait la réponse à la première des trois questions traditionnellement fondamentales: D'où venons-nous? Qui sommes-nous? Où allons-nous? Ces questions n'intéressent pas seulement les scientifiques ou les intellectuels. En effet, j'ai fini par remarquer que contrairement à ce que beaucoup pensent, la plupart des gens dans la population en général ne se préoccupent pas seulement de la recherche qui mènera à des médicaments ou autres applications concrètes. Ils sont également fascinés par des sujets tels que l'évolution des êtres vivants.

Les études sur l'évolution ont gagné beaucoup de puissance grâce aux techniques de biologie moléculaire, et surtout grâce aux nombreuses séquences de gènes et de protéines maintenant disponibles. Leur nombre continue à augmenter de façon exponentielle et des dizaines de génomes sont en train d'être séquencés au complet. Parmi ceux-ci, on nous promet la séquence du génome humain pour environ 2005. Le *Voyageur imprudent* de Barjavel et le héros de *La machine à explorer le temps* de H.G. Wells ont voyagé jusqu'à quelques dizaines ou centaines de milliers d'années dans le

futur. Les expériences d'évolution moléculaire nous font voyager des *milliards* d'années dans le passé! N'est-ce pas au moins aussi fascinant, d'autant plus que ce n'est pas de la fiction?...

BIBLIOGRAPHIE

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. et Watson, J.D. (1994) *Molecular Biology of the Cell*, 3è édition, Garland Publishing Inc., New York, Londres.

Andersson, S.G. et Kurland, C.G. (1995) Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* **73**:775-787.

Antoshechkin, I. et Bogenhagen, D.F. (1995) Distinct roles for two purified factors in transcription of *Xenopus* mitochondrial DNA. *Mol. Cell. Biol.* **15**:7032-7042.

Archambault, J. et Friesen, J.D. (1993) Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol. Rev.* **57**:703-724.

Baldauf, S.L. et Palmer, J.D. (1990) Evolutionary transfer of the chloroplast *tufA* gene to the nucleus. *Nature* **344**:262-265.

Baldauf, S.L., Palmer, J.D. et Doolittle, W.F. (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**:7749-7754.

Barjavel, R. (1944) *Le voyageur imprudent*. Éditions Gallimard (Folio), Paris.

Barrell, B.G., Bankier, A.T. et Drouin, J. (1979) A different genetic code in human mitochondria. *Nature* **282**:189-194.

Benne, R. et Sloof, P. (1987) Evolution of the mitochondrial protein synthetic machinery. *Biosystems* **21**:51-68.

Bischoff, U. (1993) *Munch*. Benedikt Taschen Verlag GmbH, Köln.

Björk, G.R. (1995) Biosynthesis and function of modified nucleosides. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

Boer, P.H. et Gray, M.W. (1988) Scrambled ribosomal RNA gene pieces in *Chlamydomonas reinhardtii* mitochondrial DNA. *Cell* **55**:399-411.

Bogenhagen, D.F. (1996) Interaction of mtTFB and mtRNA polymerase at core promoters for transcription of *Xenopus laevis* mtDNA. *J. Biol. Chem.* **271**:12036-12041.

Bogenhagen, D.F. et Insdorf, N.F. (1988) Purification of Xenopus laevis mitochondrial RNA polymerase and identification of a dissociable factor required for specific transcription. *Mol. Cell. Biol.* **8**:2910-2916.

Bonner, G., Patra, D., Lafer, E.M. et Sousa, R. (1992) Mutations in T7 RNA polymerase that support the proposal for a common polymerase active site structure. *EMBO J.* **11**:3767-3775.

Brennicke, A., Grohmann, L., Hiesel, R., Knoop, V. et Schuster, W. (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. *FEBS Lett.* **325**:140-145.

Brown, J.R. et Doolittle, W.F. (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**:2441-2445.

Broyles, S.S. et Moss, B. (1986) Homology between RNA polymerases of poxviruses, prokaryotes, and eukaryotes: nucleotide sequence and transcriptional analysis of vaccinia virus genes encoding 147-kDa and 22-kDa subunits. *Proc. Natl. Acad. Sci. USA* **83**:3141-3145.

Bryant, D.A. (1992) Puzzles of chloroplast ancestry. *Curr. Biol.* **2**:240-242.

Bui, E.T., Bradley, P.J. et Johnson, P.J. (1996) A common evolutionary origin for mitochondria and hydrogenosomes. *Proc. Natl. Acad. Sci. USA* **93**:9651-9656.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D. et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**:1058-1073.

Carrodeguas, J.A., Yun, S., Shadel, G.S., Clayton, D.A. et Bogenhagen, D.F. (1996) Functional conservation of yeast mtTFB despite extensive sequence divergence. *Gene Expr.* **6**:219-230.

Cavalier-Smith, T. (1993) Kingdom protozoa and its 18 phyla. *Microbiol. Rev.* **57**:953-994.

Cedergren, R. et Grosjean, H. (1987) On the primacy of primordial RNA. *Biosystems* **20**:175-180.

Cedergren, R. et Miramontes, P. (1996) The puzzling origin of the genetic code. *Trends Biochem. Sci.* **21**:199-200.

Cedergren, R., Gray, M.W., Abel, Y. et Sankoff, D. (1988) The evolutionary relationships among known life forms. *J. Mol. Evol.* **28**:98-112.

Cedergren, R.J., Sankoff, D., LaRue, B. et Grosjean, H. (1981) The evolving tRNA molecule. *CRC Crit. Rev. Biochem.* **11**:35-104.

Chen, B., Kubelik, A.R., Mohr, S. et Breitenberger, C.A. (1996) Cloning and characterization of the *Neurospora crassa cyt-5* gene. A nuclear-coded mitochondrial RNA polymerase with a polyglutamine repeat. *J. Biol. Chem.* **271**:6537-6544.

Clark, B.F.C., Kjeldgaard, Barciszewski, J. et Sprinzl, M. (1995) Recognition of aminoacyl-tRNAs by protein elongation factors. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

Clayton, D.A. (1991) Replication and transcription of vertebrate mitochondrial DNA. *Annu. Rev. Cell Biol.* **7**:453-478.

Clayton, R.A., White, O., Ketchum, K.A. et Venter, J.C. (1997) The first genome from the third domain of life. *Nature* **387**:459-462.

Conway Morris, S. (1997) Defusing the cambrian 'explosion'? *Curr. Biol.* **7**:R71-R74.

Cousineau, B., Leclerc, F. et Cedergren, R. (1997) On the origin of protein synthesis factors: a gene duplication/fusion model. *J. Mol. Evol.* **45**:661-670.

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

Dairaghi, D.J., Shadel, G.S. et Clayton, D.A. (1995) Addition of a 29 residue carboxyl-terminal tail converts a simple HMG box-containing protein into a transcriptional activator. *J. Mol. Biol.* **249**:11-28.

Darnell, J.E. et Doolittle, W.F. (1986) Speculations on the early course of evolution. *Proc. Natl. Acad. Sci. USA* **83**:1271-1275.

Darwin, C. (1962) *The origin of species* (texte de la 6è édition, 1872), Collier Books, Macmillan Publishing Co., New York.

Delarue, M., Poch, O., Tordo, N., Moras, D. et Argos, P. (1990) An attempt to unify the structure of polymerases. *Protein Eng.* **3**:461-467.

Delwiche, C.F., Kuhsel, M. et Palmer, J.D. (1995) Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.* **4**:110-128.

D'Erchia, A.M., Gissi, C., Pesole, G., Saccone, C. et Arnason, U. (1996) The guinea-pig is not a rodent. *Nature* **381**:597-600.

Dirheimer, G., Keith, G., Dumas, P. et Westhof, E. (1995) Primary, secondary, and tertiary structure of tRNAs. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

Doolittle, R.F. (1995) Of archae and eo: what's in a name? *Proc. Natl. Acad. Sci. USA* **92**:2421-2423.

Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. et Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**:470-477.

Douglas, S.E., Murphy, C.A., Spencer, D.F. et Gray, M.W. (1991) Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* **350**:148-151.

Dover, G.A. (1988) rDNA world falling to pieces. *Nature* **336**:623-624.

Fedonkin, M.A. et Waggoner B.M. (1997) The late precambrian fossil *Kimberella* is a mollusc-like bilaterian organism. *Nature* **388**:868-871.

Fisher, R.P. et Clayton, D.A. (1985) A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the

heavy- and light-strand promoters dissected and reconstituted in vitro. *J. Biol. Chem.* **260**:11330-11338.

Fisher, R.P. et Clayton, D.A. (1988) Purification and characterization of human mitochondrial transcription factor 1. *Mol. Cell. Biol.* **8**:3496-3509.

Fisher, R.P., Parisi, M.A. et Clayton, D.A. (1989) Flexible recognition of rapidly evolving promoter sequences by mitochondrial transcription factor 1. *Genes. Dev.* **3**:2202-2217.

Fisher, R.P., Topper, J.N. et Clayton, D.A. (1987) Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. *Cell* **50**:247-258.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**:496-512.

Foulkes, N.S. et Sassone-Corsi, P. (1996) Transcription factors coupled to the cAMP-signalling pathway. *Biochim. Biophys. Acta* **1288**:F101-F121.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A. et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**:397-403.

Germot, A., Philippe, H. et Le Guyader, H. (1996) Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proc. Natl. Acad. Sci. USA* **93**:14614-14617.

Golding, B. (1996) When was life's first branch point? *Curr. Biol.* **6**:679-682.

Gould, S.J. (1991) *La vie est belle*. Éditions du Seuil, Paris. (Traduit de l'américain par Marcel Blanc. Titre original: *Wonderful life*, W.W. Norton and Co., 1989)

Gould, S.J. (1996) Des paroles en l'air, p.118-131. In *Comme les huit doigts de la main*, Éditions du Seuil, Paris. (Traduit de l'américain par Marcel Blanc. Titre original: *Eight little piggies*, W.W. Norton and Co., 1993)

Graur, D. (1993) Towards a molecular resolution of the ordinal phylogeny of the eutherian mammals. *FEBS Lett.* **325**:152-159.

Graur, D., Duret, L. et Gouy, M. (1996) Phylogenetic position of the order Lagomorpha. *Nature* **379**:333-335.

Gray, M.W. (1989) Origin and evolution of mitochondrial DNA. *Annu. Rev. Cell Biol.* **5**:25-50.

Gray, M.W. (1992) The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* **141**:233-357.

Gray, M.W. (1993) Origin and evolution of organelle genomes. *Curr. Opin. Genet. Dev.* **3**:884-890.

Griffiths, A.J.F. (1995) Natural plasmids of filamentous fungi. *Microbiol. Rev.* **59**:673-685.

Grosschedl, R., Giese, K. et Pagel, J. (1994) HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. *Trends Genet.* **10**:94-100.

Gupta, R.S. et Golding, G.B. (1993) Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* **37**:573-582.

Gupta, R.S. et Golding, G.B. (1996) The origin of the eukaryotic cell. *Trends Biochem. Sci.* **21**:166-171.

Hale, S.P., Auld, D.S., Schmidt, E. et Schimmel, P. (1997) Discrete determinants in transfer RNA for editing and aminoacylation. *Science* **276**:1250-1252.

Hasegawa, M., Yano, T. et Miyata, T. (1984) Evolutionary implications of error amplification in the self-replicating and protein-synthesizing machinery. *J. Mol. Evol.* **20**:77-85.

Hedtke, B., Börner, T. et Weihe, A. (1997) Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*. *Science* **277**:809-811.

Helmann, J.D. et Chamberlin, M.J. (1988) Structure and function of bacterial sigma factors. *Annu. Rev. Biochem.* **57**:839-872.

Henze, K., Badr, A., Wettern, M., Cerff, R. et Martin, W. (1995) A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proc. Natl. Acad. Sci. USA* **92**:9122-9126.

Hess, W.R., Prombona, A., Fieder, B., Subramanian, A.R. et Börner, T. (1993) Chloroplast *rps15* and the *rpoB/C1/C2* gene cluster are strongly transcribed in ribosome-deficient plastids: evidence for a functioning non-chloroplast-encoded RNA polymerase. *EMBO J.* **12**:563-571.

Hillis, D.M. (1997) Biology recapitulates phylogeny. *Science* **276**:218-219.

Horner, D.S., Hirt, R.P., Kilvington, S., Lloyd, D. et Embley, T.M. (1996) Molecular data suggest an early acquisition of the mitochondrion endosymbiont. *Proc. R. Soc. Lond. B Biol. Sci.* **263**:1053-1059.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. et Miyata, T. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86**:9355-9359.

Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. et Miyata, T. (1991) Evolution of RNA polymerases and branching patterns of the three major groups of Archaebacteria. *J. Mol. Evol.* **32**:70-78.

Jacobs, H.T. (1991) Structural similarities between a mitochondrially encoded polypeptide and a family of prokaryotic respiratory toxins involved in plasmid maintenance suggest a novel mechanism for the evolutionary maintenance of mitochondrial DNA. *J. Mol. Evol.* **32**:333-339.

Jaehning, J.A. (1991) Sigma factor relatives in eukaryotes. *Science* **253**:859.

Jaehning, J.A. (1993) Mitochondrial transcription: is a pattern emerging? *Mol. Microbiol.* **8**:1-4.

Jang, S.H. et Jaehning, J.A. (1991) The yeast mitochondrial RNA polymerase specificity factor, MTF1, is similar to bacterial sigma factors. *J. Biol. Chem.* **266**:22671-22677.

Kempken, F., Hermanns, J. et Osiewacz, H.D. (1992) Evolution of linear plasmids. *J. Mol. Evol.* **35**:502-513.

Klenk, H.P. et Doolittle, W.F. (1994) Archaea and eukaryotes versus bacteria? *Curr. Biol.* **4**:920-922.

Klenk, H.P. et Zillig, W. (1994) DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J. Mol. Evol.* **38**:420-432.

Kuhn, T.S. (1983) *La structure des révolutions scientifiques*. Flammarion, Paris. (Traduit de l'américain par Laure Meyer. Titre original: *The structure of scientific revolutions*, University of Chicago Press, 1970)

Kumazawa, Y. et Nishida, M. (1993) Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. *J. Mol. Evol.* **37**:380-398.

Kurland, C.G. (1992) Evolution of mitochondrial genomes and the genetic code. *Bioessays* **14**:709-714.

Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M. et Gray, M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**:493-497.

Langer, D., Hain, J., Thuriaux, P. et Zillig, W. (1995) Transcription in archaea: similarity to that in eucarya. *Proc. Natl. Acad. Sci. USA* **92**:5768-5772.

Larsson, N.G., Oldfors, A., Garman, J.D., Barsh, G.S. et Clayton, D.A. (1997) Down-regulation of mitochondrial transcription factor A during spermatogenesis in humans. *Hum. Mol. Genet.* **6**:185-191.

Lazcano, A., Fastag, J., Gariglio, P., Ramírez, C. et Oró, J. (1988) On the early evolution of RNA polymerase. *J. Mol. Evol.* **27**:365-376.

Lehninger, A.L. (1964) *The mitochondrion: molecular basis of structure and function.* W.A. Benjamin Inc., New York, Amsterdam.

Lerbs-Mache, S. (1993) The 110-kDa polypeptide of spinach plastid DNA-dependent RNA polymerase: single-subunit enzyme or catalytic core of multimeric enzyme complexes? *Proc. Natl. Acad. Sci. USA* **90**:5509-5513.

Li, W.-H. (1997) *Molecular evolution.* Sinauer Associates Inc., Sunderland MA.

Liljas, A. (1996) Imprinting through molecular mimicry. *Curr. Biol.* **6**:247-249.

Lisowsky, T. et Michaelis, G. (1988) A nuclear gene essential for mitochondrial replication suppresses a defect of mitochondrial transcription in Saccharomyces cerevisiae. *Mol. Gen. Genet.* **214**:218-223.

Lynch, M. (1996) Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol. Biol. Evol.* **13**:209-220.

Lynch, M. (1997) Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Mol. Biol. Evol.* **14**:914-925.

Mackey, Z.B., Ramos, W., Levin, D.S., Walter, C.A., McCarrey, J.R. et Tomkinson, A.E. (1997) An alternative splicing event which occurs in mouse pachytene spermatocytes generates a form of DNA ligase III with distinct biochemical

properties that may function in meiotic recombination. *Mol. Cell. Biol.* **17**:989-998.

Maizels, N. et Weiner, A.M. (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl. Acad. Sci. USA* **91**:6729-6734.

Mangus, D.A., Jang, S.H. et Jaehning, J.A. (1994) Release of the yeast mitochondrial RNA polymerase specificity factor from transcription complexes. *J. Biol. Chem.* **269**:26568-26574.

Marczynski, G.T., Schultz, P.W. et Jaehning, J.A. (1989) Use of yeast nuclear DNA sequences to define the mitochondrial RNA polymerase promoter in vitro. *Mol. Cell. Biol.* **9**:3193-3202.

Margulis, L. (1981) *Symbiosis in cell evolution.* W.H. Freeman and Co., San Francisco.

Margulis, L. (1988) *Five kingdoms: an illustrated guide to the phyla of life on earth.* W.H. Freeman and Co., New York.

Margulis, L. (1996) Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc. Natl. Acad. Sci. USA* **93**:1071-1076.

Marsh, T.L., Reich, C.I., Whitelock, R.B. et Olsen, G.J. (1994) Transcription factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes. *Proc. Natl. Acad. Sci. USA* **91**:4180-4184.

Martin, N.C. (1995) Organellar tRNAs: biosynthesis and function. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

Masters, B.S., Stohl, L.L. et Clayton, D.A. (1987) Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* **51**:89-99.

McClain W.H. (1995) The tRNA identity problem: past, present, and future. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

McGraw, N.J., Bailey, J.N., Cleaves, G.R., Dembinski, D.R., Gocke, C.R., Joliffe, L.K., MacWright, R.S. et McAllister, W.T. (1985) Sequence and analysis of the gene for bacteriophage T3 RNA polymerase. *Nucleic Acids Res.* **13**:6753-6766.

Meinhardt, F., Kempken, F., Kamper, J. et Esser, K. (1990) Linear plasmids among eukaryotes: fundamentals and application. *Curr. Genet.* **17**:89-95.

Munn, E.A. (1974) *The structure of mitochondria.* Academic Press, London, New York.

Natsoulis, G., Hilger, F. et Fink, G.R. (1986) The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of *S. cerevisiae*. *Cell* **46**:235-243.

Noller, H.F., Hoffarth, V. et Zimniak, L. (1992) Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256**:1416-1419.

Nugent, J.M. et Palmer, J.D. (1991) RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. *Cell* **66**:473-481.

Osawa, S., Jukes, T.H., Watanabe, K. et Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**:229-264.

Osumi-Davis, P.A., de Aguilera, M.C., Woody, R.W. et Woody, A.Y.M. (1992) Asp537, Asp812 are essential and Lys631, His811 are catalytically significant in bacteriophage T7 RNA polymerase activity. *J. Mol. Biol.* **226**:37-45.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**:734-740.

Palenik, B. (1992) Polymerase evolution and organism evolution. *Curr. Opin. Genet. Dev.* **2**:931-936.

Palmer, J.D. (1997a) Organelle genomes: going, going, gone! *Science* **275**:790-791.

Palmer, J.D. (1997b) The mitochondrion that time forgot. *Nature* **387**:454-455.

Parisi, M.A. et Clayton, D.A. (1991) Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. *Science* **252**:965-969.

Parisi, M.A., Xu, B. et Clayton, D.A. (1993) A human mitochondrial transcriptional activator can functionally replace a yeast mitochondrial HMG-box protein both in vivo and in vitro. *Mol. Cell. Biol.* **13**:1951-1961.

Pennisi, E. (1997) Microbial genomes come tumbling in. *Science* **277**:1433.

Poyton, R.O. et McEwen, J.E. (1996) Crosstalk between nuclear and mitochondrial genomes. *Annu. Rev. Biochem.* **65**:563-607.

Ridley, M. (1993) *Evolution.* Blackwell Scientific Publications Inc., Boston.

Roeder, R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21**:327-335.

Roger, A.J. et Brown, J.R. (1996) A chimeric origin for eukaryotes re-examined. *Trends Biochem. Sci.* **21**:370-371.

Roger, A.J., Clark, C.G. et Doolittle, W.F. (1996) A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis. Proc. Natl. Acad. Sci. USA* **93**:14618-14622.

Rowland, G.C. et Glass, R.E. (1990) Conservation of RNA polymerase. *Bioessays* **12**:343-346.

Russo, F.D. et Silhavy, T.J. (1992) Alpha: the Cinderella subunit of RNA polymerase. *J. Biol. Chem.* **267**:14515-14518.

Schimmel, P. et Ribas de Pouplana, L. (1995) Transfer RNA: from minihelix to genetic code. *Cell* **81**:983-986.

Schnare, M.N. et Gray, M.W. (1990) Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. *J. Mol. Biol.* **215**:73-83.

Shadel, G.S. et Clayton, D.A. (1993) Mitochondrial transcription initiation. Variation and conservation. *J. Biol. Chem.* **268**:16083-16086.

Shadel, G.S. et Clayton, D.A. (1995) A *Saccharomyces cerevisiae* mitochondrial transcription factor, sc-mtTFB, shares features with sigma factors but is functionally distinct. *Mol. Cell. Biol.* **15**:2101-2108.

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

Smith, H.C., Gott, J.M. et Hanson, M.R. (1997) A guide to RNA editing. *RNA* **3**:1105-1123.

Sogin, M.L. (1997) Energy-producing symbionts in early eukaryotes? *Curr. Biol.* **7**:R315-R317.

Söll, D. (1993) Transfer RNA: an RNA for all seasons. In *The RNA World*, Gesteland et Atkins, ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Sousa, R. (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem. Sci.* **21**:186-190.

Sousa, R., Chung, Y.J., Rose, J.P. et Wang, B.C. (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature* **364**:593-599.

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 N^2-dimethylguanosine presence and alternate tRNA conformers. *RNA* **1**:886-891.

Steinberg, S. et Ioudovitch, A. (1996) A role for the bulged nucleotide 47 in the facilitation of tertiary interactions in the tRNA structure. *RNA* **2**:84-87.

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

Stewart, C.B. (1993) The powers and pitfalls of parsimony. *Nature* **361**:603-607.

Swofford, D.L. et Olsen, G.J. (1990) Phylogeny reconstruction. In *Molecular Systematics* (chap. 11), Hillis et Moritz, ed., Sinauer, Sunderland, MA.

Szostak, J.W. (1992) In vitro genetics. *Trends Biochem. Sci.* **17**:89-93.

Szostak, J.W. et Ellington, A.D. (1993) In vitro selection of functional RNA sequences. In *The RNA World*, Gesteland et Atkins, ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Tracy, R.L. et Stern, D.B. (1995) Mitochondrial transcription initiation: promoter structures and RNA polymerases. *Curr. Genet.* **28**:205-216.

Von Bülow, M., Rackwitz, H.R., Zimbelmann, R. et Franke, W.W. (1997) CP beta3, a novel isoform of an actin-binding protein, is a component of the cytoskeletal calyx of the mammalian sperm head. *Exp. Cell Res.* **233**:216-224.

Von Heijne, G. (1986) Why mitochondria need a genome. *FEBS Lett.* **198**:1-4.

Wang, Y., Goligorsky, M.S., Lin, M., Wilcox, J.N. et Marsden, P.A. (1997) A novel, testis-specific mRNA transcript encoding an NH_2-terminal truncated nitric-oxide synthase. *J. Biol. Chem.* **272**:11392-11401.

Watanabe, K. et Osawa, S. (1995) tRNA sequences and variations of the genetic code. In *tRNA: Structure, Biosynthesis and Function*, Söll and RajBhandary, ed., American Society for Microbiology, Washington DC.

Watson, J.D., Hopkins, N.H., Roberts, J.W., Steitz, J.A. et Weiner, A.M. (1987) *Molecular biology of the gene*, 4è édition. The Benjamin/Cummings Publishing Company Inc., Menlo Park CA.

Wells, H.G. (1895) *La machine à explorer le temps*. Éditions Gallimard (Folio Junior), Paris. (Traduit de l'anglais par Henry D. Davray. Titre original: *The time machine*)

Wilson, R.J.M. et Williamson, D.H. (1997) Extrachromosomal DNA in the Apicomplexa. *Microbiol. Mol. Biol. Rev.* **61**:1-16.

Woese, C.R. (1967) *The genetic code: the molecular basis for genetic expression.* Harper and Row, New York.

Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* **51**:221-271.

Woese, C.R. et Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**:5088-5090.

Woese, C.R., Kandler, O. et Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**:4576-4579.

Wolfe, C.L., Lou, Y.C., Hopper, A.K. et Martin, N.C. (1994) Interplay of heterogeneous transcriptional start sites and translational selection of AUGs dictate the production of mitochondrial and cytosolic/nuclear tRNA nucleotidyltransferase from the same gene in yeast. *J. Biol. Chem.* **269**:13361-13366.

Wray, G.A., Levinton, J.S. et Shapiro L.H. (1996) Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**:568-573.

Xu, B. et Clayton, D.A. (1992) Assignment of a yeast protein necessary for mitochondrial transcription initiation. *Nucleic Acids Res.* **20**:1053-1059.

Yokobori, S. et Pääbo, S. (1995) Transfer RNA editing in land snail mitochondria. *Proc. Natl. Acad. Sci. USA* **92**:10432-10435.

Zillig, W., Palm, P. et Klenk, H.P. (1992) A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor. In *The Origin and Evolution of Prokaryotic and Eukaryotic Cells*, Hartman et Matsuno, ed., World Scientific Publishing Co., New Jersey.

Zillig, W., Palm, P., Klenk, H.P., Langer D., Hüdepohl, U., Hain, J., Lanzendörfer, M. et Holz, I. (1993) Transcription in Archaea. In *Biochemistry of Archaea*, Kates, Kushner et Matheson, ed., Elsevier Science Publishers B.V., Amsterdam.

ANNEXE

Chapitre de livre

Modified nucleosides always were: an evolutionary model

CHAPITRE DE LIVRE

Modified nucleosides always were: an evolutionary model

(Les nucléosides modifiés existent depuis toujours; un modèle évolutif)

Cermakian, N. et Cedergren, R. (1997) In *Modification and editing of RNA : the alteration of RNA structure and function*, H. Grosjean et R. Benne (Ed.), Chapitre 29, American Society for Microbiology Press, Washington DC, Sous presse.

(inclus avec la permission de l'American Society for Microbiology)

Modified nucleosides always were: an evolutionary model

Nicolas Cermakian and Robert Cedergren*

Département de biochimie

Université de Montréal

Montréal (Québec)

CANADA H3C 3J7

Phone: (514) 343-6320

Fax: (514) 343-2210

E-mail: ceder@bch.umontreal.ca

* Corresponding author. Send proofs to:

Robert Cedergren

Département de biochimie (D-334)

Université de Montréal

2900, boul. Edouard-Montpetit

Montréal (Québec)

Canada H3T 1J4

September 12, 1997

THE EVOLUTIONARY FRAMEWORK

The RNA World hypothesis of the origin of cellular life on earth is a good example of how interpreting a process or a structure, in this case RNA catalysis, within an evolutionary framework can have profound consequences for subsequent directions in biology. The origin of introns is an even better example for our purposes, since it has been approached in different ways: hypotheses generated by phylogenetic analyses of present-day introns have co-existed, not always very happily, with chemical/biological rationalizations of intron origins. Since we will use both phylogenetic arguments and chemical rationalizations in analyzing the origin of modified nucleotides in biological systems, the differences between them must be clear. Phylogenetic analyses start at present-day and make use of the organismal distribution of traits to infer the past, that is, properties of the most recent common ancestor. Chemical and biological rationalizations of the past are based on the extrapolation of certain molecular properties, which can be logically linked to a process or a structure present during the period of emerging life forms. These extrapolations require neither intermediates nor ancestors, the plausibility being judged solely on the coherence of the proposal.

The differences between these two approaches are brought into focus when one considers that intron/splicing is a phenomenon that can be invoked in an undeniably rational manner to explain the formation of modern genes (Gilbert, 1978). On the other hand, no one has been able to demonstrate a wide enough distribution of introns to support hypotheses of a prebiotic origin. The controversy surrounding the origin of introns is thus fundamentally due to a conflict between the inferences of the two methodologies.

In general, when one speaks of a modified nucleoside, minor nucleosides are meant, even though this certainly does not mean minor in the sense of function (Chap 1 by Lane). Under this definition, the spectrum of "modified" nucleosides is expanded out of the realm of post-transcriptional modifications of RNA and leads to the consideration of the origin of nucleosides and their derivatives in a more global

manner. Within our definition, it seems perfectly clear that modified nucleosides did not have an origin any different from that of the "basic nucleosides" or the plethora of other small organic compounds that were found in the primordial soup. Indeed, it was proposed some time ago that the numerous nucleotide-derived cofactors and prosthetic groups had an ancient origin and could even be viewed as relics of the early involvement of RNA molecules in catalysis of various cellular reactions (Lazcano, 1994a). These hypotheses suggest as well that during the emergence of the key cellular components, these nucleoside components in RNA were already present. We believe therefore, that the issue is not the origin of modified nucleotides - they always were - but rather how and why they remain in RNA and DNA molecules. We do, however, bow to consensus opinion in this book and use the term modified nucleoside.

THE PHYLOGENETICS OF MODIFIED NUCLEOSIDES

The tried and true methodology in evolutionary studies is the determination of the organismal distribution of a given trait. Then using the principle of parsimony, the origins of a property can generally be ascribed to the most recent ancestor of the two most distant organisms possessing the trait. In the case of the modified nucleotides, we are interested not only in the organism in which it is found but also the RNA species and the position occupied by the modified nucleoside.

Modified nucleosides are found in ribosomal RNAs, messenger RNAs (eukaryotic and viral), small nuclear RNAs, and other small RNAs (Appendix I). Thus we can infer that modified nucleosides are present in most known RNA species. As can be seen in Appendix I, most of the modified nucleosides that have been compiled and studied are found in tRNAs. Here, we evaluate the distribution of modified nucleosides in certain RNA species of the three domains of life (archaea, bacteria, eukaryotes).

From the phylogenetic distribution of modified nucleosides in tRNAs, more than half (60 out of 96) are found only in either eukaryotes, bacteria or archaea, whereas 22 are found in two: 7 are common to both eukaryotes and archaea, 7 to bacteria and archaea, and 8 to eucaryotes and bacteria. However, these data do not allow us to infer the presence of modified nucleotides before the divergence of the three domains.

Fourteen modifications are present in all three domains (fifteen, including queuosine derivatives) and of these several are found at comparable positions in tRNAs. Table 1 is derived from the information provided in Appendices I and V, and Grosjean et al. (1995b). First, most of the conserved modified nucleosides are in or near the anticodon loop (positions 34, 37-39). They are among the simplest modifications with the exception of queuosine (Q) and the threonylated residue t^6A37. In fact, these conserved modifications, that is pseudouridine (Ψ) and methylated nucleotides, are the most frequent modifications in tRNAs and in other RNAs as well (Appendices I and V). Another striking point is that two of the modified nucleosides of Table 1, m^1G37 and t^6A37 are nearly ubiquitous in tRNAs reading codons beginning by a C or an A, respectively (Björk, 1995). Therefore based on the distribution of modified nucleosides, those of Table 1 would have been present in tRNA at the point at which the eukaryote/archaea branch diverged from the bacterial lineage (Fig. 1). We may assume that the enzymatic systems required to modify the nucleosides were present as well at this point in time, although the precise mechanisms of their formation may well have differed from the modern enzymatic processes for their biosynthesis. Note that these modifications infered to have been present before the divergence of the three domains are likely to form a minimal set, since certain modifications could have been lost in one domain or the other. The distribution of modified nucleosides showing presence in two of the three domains support this conjecture.

It is interesting that two of the most widespread tRNA modifications of Table 1, pseudouridylation and 2'-O-methylation, are also the most frequent modifications in ribosomal RNAs. In fact, Ψ and Cm are among the only four modified nucleosides that

are found in ribosomal RNAs from all three organismal domains (Table 2). These observations strengthen our point that these modifications are very ancient. At the very least therefore, there is strong phylogenetic evidence for the early appearance of modified nucleosides.

THE DEEP PAST

The RNA World hypothesis states that at some point in early evolution, living organisms relied entirely on RNA for both the storage of genetic information and catalysis; they would have been devoid of encoded proteins and DNA (Gilbert, 1986; Darnell and Doolittle, 1986; Cedergren and Grosjean, 1987). The intertwining of the RNA World and RNA catalysis has led to novel hypotheses, experiments and discoveries in almost every biological process that involves RNA. The study of RNA processing and translation, as well as in vitro selection of RNA molecules with a wide range of catalytic activities, provides additional support for the notion of an all-RNA World (Lazcano, 1994b). RNA catalysis was such a perfect solution to the genotype/phenotype dichotomy in origin of life theories that practically all modern scenarios are based on some form of RNA catalysis. As a result, many underpinnings of this hypothesis have unfortunately managed to escape detailed critical evaluation.

Among the issues that we believe to be worthy of thought is the purity of the RNA in the RNA World. The emergence of the RNA World would have taken place in a molecular soup composed of massive numbers of small chemicals. Even though molecular complexity can be a precondition to the origin of life in some theories (Kauffman, 1993), chemically this mixture is a nightmare (Joyce and Orgel, 1993). How do the poorly represented purines and pyrimidines in this diluted soup get together with the poorly represented ribose to form nucleosides? More importantly, how could the chemistry of these precursors distinguish so well between the "non-modified" and the "modified" bases and riboses? In summary, how does a self-replicating, presumably polymeric, RNA arise *de novo* from a complex mixture of

nucleotides and derivatives thereof (Joyce and Orgel 1993). It seems much more plausible that primordial “RNA” was more a representative of the variety of nucleoside-like compounds that existed in the soup than a pure ribonucleotide species.

A possible criticism to the above scenario would be if modified nucleotides somehow blocked the catalytic activity of RNA upon which the emergence of an RNA World depended. However, work in our laboratory has shown that mixed ribo- and deoxyribo-oligonucleotides can perform catalytic functions (Perreault et al., 1990; Yang et al., 1992; Chartrand et al., 1995). Based on this work, it was proposed that such mixed polymers could have preceded the RNA World or served as a bridge between the RNA World and the present-day DNA-protein World. This possibility was presented in some depth by Bussière and Perreault (1995). Since those early experiments, many nucleic acid catalysts have been found, either by using modified nucleosides in chemical synthetic schemes (Usman et al., 1996) or, in the case of catalytic DNAs, by in vitro selection (Cuenoud and Szostak, 1995). Rather than thinking that the idea of mixed polymers is ill-founded, we consider it surprising that such little critical thought has been given to the state of RNA produced in the primordial soup.

At the present time, the mechanism of abiotic polymerization of RNA (here, we use RNA to signify any polymer involving a sugar, a nitrogen base and a phosphate residue) or their precursors is controversial (Joyce and Orgel, 1993; Szathmáry, 1997); therefore, we can only imagine how modified nucleosides might have become part of the early polymers. First, nucleic acids could have been produced by condensation of nucleotides (as they are today, albeit in an indirect manner), or by copolymerization of nucleosides and phosphates. Alternate schemes would involve addition of a nitrogen base to a sugar-phosphate backbone, or less likely insertion of a base or a nucleoside into a preformed polyphosphate backbone. These three possible, but non-exhaustive, pathways relate in different ways to how modified nucleosides could be incorporated during synthesis of the polymer.

The condensation reaction of nucleosides takes place at the 3' and 5' positions of the sugar and thus unless some intramolecular cyclic intermediate would be formed, this reaction could not discriminate with regard to the substituents on the nitrogen base of the nucleotide. If this condensation were templated by a complementary strand, discrimination would be possible, but only in the case where the nitrogen base could not form the proper hydrogen bonds with the template strand (Ferris, et al., 1996). Clearly this property would tend to eliminate some modified (and unmodified) nucleotides, but many would still be good candidates. For example, no effective discrimination could be imagined between uracil or thymine derivatives. Modifications on the sugar residue such as 2'-O-methylation of a ribose should not affect condensation either, unless the 2'-OH was the primary nucleophile. Therefore, a condensation mechanism could only be partially discriminatory if a template strand was involved in polymerization.

The alternate synthetic pathway, the addition of a nitrogen base to a sugar/phosphate backbone is subject to the reactivity of the ring nitrogen involved in formation of the glycosidic bond. The synthesis of a nucleoside would involve the nucleophilic attack of the ring nitrogen (either N3 or N9) on the C1' carbonyl carbon, so the better the nucleophile, the more likely the reaction. Normally, methyl groups would be considered electron donating and therefore depending on their ring position, would provide a higher nucleophilicity to the ring nitrogen and therefore a greater probability of reaction. Electron withdrawing substituents on the ring would lower the nucleophilicity. This second way of incorporating bases (eventually modified bases) into RNA can be illustrated by current mechanisms involved in some tRNA modifications. For instance, G is replaced by queuine (Q) at wobble position 34 in many tRNAs, by a mechanism known as *transglycosylation*, in which the base is replaced without cleavage of the phosphodiester backbone (Chap 9 by Romier et al.). Although Q is a hypermodified nucleotide, the transglycosylation reaction *per se* is rather simple; indeed, the enzyme that performs the base transfer does not require any additional energy source (Chap 9 by Romier et al.). Also significant here is the discovery that abasic hammerhead ribozymes can be rescued by the addition of

exogenous bases, which restore the activity by inserting in the cavity created by the absence of the normal base (Peracchi et al., 1996). Finally, the insertion of a preformed nucleoside to a polymeric phosphate should not be sensitive to the structure of the nitrogen base for the same reasons as we invoked above for the condensation reaction.

The incorporation of modified nucleosides into nucleic acid polymers is therefore quite plausible, and the use of a particular nucleoside may depend only on its concentration in the primordial soup. Again, considering the chemistry of nucleosides, it would seem that the moderately modified nucleosides would be present in the highest concentrations in the primordial soup along with the standard nucleosides (Fig. 1). Such modified nucleosides would likely include: the sugar or base methylated nucleosides; inosine (I), which can be obtained easily by deamination of A (Chap 19 by Emeson; Grosjean et al., 1995a); pseudouridine (Ψ), which could theoretically be generated from U via an isomerization pathway; and possibly the aminoacylated nucleosides; but not the hypermodified nucleotides generally found in the anticodon loop of tRNAs.

A MODEL OF THE EMERGENCE OF MODIFIED NUCLEOSIDES

Up to this point, we have made a case for the presence and the incorporation of modified nucleosides into RNA under prebiotic conditions. In parallel, the phylogenetic evidence suggests that some simple modified nucleosides existed at the point when the major organismal groups diverged. The question is now: are we able to provide a reasonable scenario to bridge the one or two billion-year gap between the primordial soup and the divergence of major living forms? Fortunately, the use of the phylogenetic data not only provides evidence for the early existence of modified nucleosides in RNA molecules, but also allows us to focus on only those modifications that appear to be the oldest.

At the beginning, a minimum of distinctive chemical characteristics (see above) would have promoted the more or less random incorporation of nucleosides and modified nucleosides into RNA. In an evolutionary sense, this situation is acceptable but hardly ideal, since the information on the position and the nature of the modification would be lost at each generation in the absence of some way to store this information. Therefore, rather early in the emergence of more sophisticated systems, the position and the nature of the modified nucleoside would have to be encoded. A hint as to how this could have happened has been provided by recent experimental results showing that modifications of ribosomal RNAs are addressed or guided by certain small nucleolar RNAs (Ofengand and Fournier, Chapter 12; Bachellerie and Cavaille, Chapter 13). Intriguingly, the modifications that we have predicted to be most prevalent in the primordial soup and demonstrated to be the most widespread from phylogenetic analysis, are also those that have an RNA involvement in their synthesis. It must be kept in mind, however, that RNA-guided rRNA modification has only been found in eukaryotes up until now.

We propose therefore that at some early stage, small RNA molecules were used to direct the positions at which modified nucleosides were to be present. At first, it could be imagined that a modified nucleoside was incorporated during synthesis of the RNA, thanks to the presence of small RNA molecules that added new discriminatory properties to the template copying mechanism. Later, the modifications would have been products of post-transcriptional modifications. This eventuality would be required when the exceedingly more discriminatory protein polymerases took over the role of RNA synthesis. Here again the known chemistry of the modifications is in agreement with the interpretation that we make. The modern synthesis of the methylated nucleosides is assured by a methyl transferase and S-adenosylmethionine, the methyl donor (Appendix III). S-Adenosylmethionine, the product of a simple addition of an amino acid to adenosine (i.e., a modified nucleoside), likely existed at the time when cellular systems were emerging. The presence of an RNA to direct the positioning of the methyl group raises the possibility that an RNA once had a methyl transferase activity in this process. Pseudouridine is simply an isomer of uridine and

could also be the product of an RNA-catalyzed reaction. The possibility also exists that these guide RNAs might have been internal. Perhaps the complex present-day recognition of modification sites in tRNA (for example, Grosjean et al., 1996) is an example of how cis-guidance could occur, with the guide RNA being intramolecular or within an intron (Grosjean et al., 1997). Cis-guidance has been proposed in the case of modification of bacterial rRNA (Chap 12 by Ofengand and Fournier). An interesting parallel can be made here with RNA editing (see also below): in trypanosomatid mitochondria, insertion and deletion of U is mediated by guide RNA acting in trans through base pairing (Chap 21 by Hajduk); moreover, it has been proposed that editing could be in some cases mediated by cis-acting guide RNA, for example in tRNAs (Chap 16 by Price and Gray) or in glutamate neuroreceptor editing (Herbert, 1996). These RNA-mediated RNA modification systems would then have given way to the presumed, modern protein-enzyme modification system.

The driving force for the presence of modified nucleosides could have been in rendering superior properties to the RNA. First, they could confer thermal or pH stability on RNA molecules, which would otherwise have been rather fragile and reactive (Conrad et al., 1995). Such a stability advantage has also been discussed in the context of mixed RNA-DNA polymers (Yang et al. , 1992; Bussière and Perreault, 1995). Second, the presence of a variety of chemical groups on modified nucleosides could have expanded the functionality of early ribozymes, by modulating the recognition of other nucleic acids, or by allowing a greater range of catalytic activities. An example of the modulation of nucleic acid interactions is the influence of modified nucleosides in the evolution of the genetic code, through altering codon-anticodon interaction (Cedergren et al., 1986; Osawa et al., 1992). Enlargement of the range of catalytic activities is illustrated by the case of Ψ, which is much more chemically versatile than U (Chap 1 by Lane; Chap 12 by Ofengand and Fournier; Lane et al., 1995; Ofengand and Bakin, 1997). Third, modified nucleosides could modify RNA base pairing as in the case of m^2_2G (Steinberg and Cedergren, 1996) and m^1A (Helm et al., 1997) which may prevent alternative base pairing in tRNAs. Furthermore, they could be involved in regulation of RNA activity by modulating the turnover of these

molecules; in present-day RNAs, modified nucleosides can be recognition signals for specific nucleases (Masaki et al., 1997). Finally, several hydrophobic isopentenyl-containing modified nucleosides known as cytokinins, apart from their occurrence in tRNAs, are present in a free state in a variety of organisms where they exert regulatory functions (Björk, 1995). Whatever the evolutionary mechanism by which these new properties arise, once established, it may be impossible to go back.

The research on editing and on nucleoside modification that has been carried out during the last decade, and which is reviewed in detail in other chapters in this book, brought to light various analogies between these two phenomena. One striking example is the use of guide RNAs both in rRNA modification (Chap 12 by Ofengand and Fournier; Chap 13 by Bachellerie and Cavaille) and in some editing reactions (Chap 21 by Hajduk). Another point of convergence of editing and modification studies is the deamination of bases from A to I and from C to U, which are generated via similar mechanisms (Chap 19 by Emeson; Chap 20 by Carter). In fact, in some cases, it is not clear whether we should speak of editing or modification (Chap 16 by Price and Gray; Covello and Gray, 1993), especially in the case of the production of I by deamination of A: its occurrence is called editing when the mRNA for some glutamate receptors (GluR) is concerned (Herbert, 1996) but base modification in the case of tRNA (Appendix I). Also, U formed by an editing phenomenon involving enzymatic deamination of an encoded C can be viewed as a modified nucleoside (see Appendix I).

These parallels between RNA editing and modification prompted us to examine the former through the same evolutionary approach as for the latter. In fact, an evolutionary scenario for the origin and retention of editing in RNA molecules parallels that of nucleoside modifications. Editing might have first occurred in an undirected and random way through spontaneous chemical base transformations. We have already discussed, for example, the likelihood of base deamination in the primordial soup. Thereafter, editing of RNA molecules might have been selected for some advantage, for example the capacity of forming more stable double-stranded

structures, as in the case of tRNA editing (Chap 16 by Price and Gray), or assisting early, less discriminatory RNA polymerases (Conrad et al., 1995). This would have favored the establishment of a way to specifically edit specific sites in RNA, for example through guide RNAs in trans (Chap 21 by Hajduk) or in cis (Chap 16 by Price and Gray).

In spite of this rationalization of what might have happened in the deep past, there is no phylogenetic evidence that RNA editing is very ancient. In fact, there is general agreement that all present-day manifestations of editing are likely to have arisen recently, in particular lineages of eukaryotes (Chap 17 by Brennicke; Chap 21 by Hajduk; Covello and Gray, 1993; Scott, 1995; Weiner and Maizels, 1990). This hypothesis is based on the fact that phenomena grouped under the banner "editing" are diverse in effects and molecular mechanisms, and are each restricted to a group of organisms or subcellular compartments: for instance, RNA-guided U insertion/deletion in trypanosome mitochondria (Chap 21 by Hajduk); C-to-U and U-to-C transition by deamination in plant organelles (Freyer et al., 1997; Simpson, 1990; Yu and Schuster, 1995); intron-guided A-to-I deamination in GluR mRNA (Chap 19 by Emeson); or editing by various mechanisms in the amino acid acceptor stem of tRNA of different lineages (Chap 16 by Price and Gray). On the other hand, no editing mechanism is known to be widespread, and editing is apparently absent in archaea and bacteria (Appendix II).

This situation is in sharp contrast to nucleotide modifications, for which, as we showed above, both chemical rationalization and phylogenetic evidence point toward an ancient, prebiotic origin. The situation regarding what we know about the origin of editing is rather analogous to the case of introns. Whereas RNA splicing fit well in the RNA World hypothesis, and can help to explain many observation pertaining to the early evolution of RNA and protein domains, no phylogenetic evidence, based on the known distribution of introns, clearly supports an early appearence of these entities (Hurst and McVean, 1996).

CONCLUSION

We have proposed a step-wise chronology for the early origin and the maintenance of modified nucleosides in RNA. A similar model could be established for RNA editing, which is similar to RNA modification in many of its present-day manifestations; however, contrary to the case of modifications, no phylogenetic evidence supports an early occurrence of editing. Our model involves first a random incorporation of modified nucleosides which later gave way to an RNA-guided modification system. In the case of tRNAs, there is no evidence for present-day RNA guiding, although such a system could be imagined early in the evolution of replicating cells. In the case of the rRNAs, aspects of the RNA-guided system are still present as evidenced by the startling results demonstrating a role for the small nucleolar RNAs in the maturation of ribosomal RNA. The current evidence on the enzymes involved in RNA-guided modification leaves the door open to speculation on the possibility that RNA may have a greater role in this process than simply guiding modification. Could it be that remnants of the modification system of the RNA World still exist?

ACKNOWLEDGEMENTS

We are particularly grateful to Michael Gray for a critical reading of this chapter. We thank the Natural Science and Engineering Research Council of Canada for support. RC is Richard Ivey Professor of the Canadian Institute for Advanced Research. NC has a predoctoral fellowship from the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

REFERENCES

Björk, G.R. 1995. Biosynthesis and function of modified nucleosides, p. 165-205. *In* D. Söll, U. RajBhandary (ed.), *tRNA: structure, biosynthesis and function*, American Society for Microbiology, Washington, D.C.

Bussière, F. and J.-P. Perreault. 1995. On the road to a DNA-protein world. *RNA* 1:451-452.

Cedergren, R. and H. Grosjean. 1987. On the primacy of primordial RNA. *Biosystems* 20:175-180.

Cedergren, R., H. Grosjean and B. LaRue. 1986. Primordial reading of genetic information. *BioSystems* 19:259-266.

Chartrand, P., S.C. Harvey, G. Ferbeyre, N. Usman, and R. Cedergren. 1995. An oligodeoxyribonucleotide that supports catalytic activity in the hammerhead ribozyme domain. *Nucleic Acids Res* 23:4092-4096.

Conrad, F., A. Hanne, R.K. Gaur and G. Krupp. 1995. Enzymatic synthesis of 2'-modified nucleic acids: identification of important phosphate and ribose moieties in Rnase P substrates. *Nucleic Acids Res* 23:1845-1853.

Covello, P.S. and M.W. Gray. 1993. On the evolution of RNA editing. *Trends Genet* 9:265-268.

Cuenoud, B. and J.W. Szostak. 1995. A DNA metalloenzyme with DNA ligase activity. *Nature* 375:611-614.

Darnell, J.E. and W.F. Doolittle. 1986. Speculations on the early course of evolution. *Proc Natl Acad Sci U S A* 83:1271-1275.

Ferris, J.P., A.R. Hill, R. Liu and L.E. Orgel. 1996. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* 381:59-61.

Freyer, R., M. Kiefer-Meyer and H. Kössel. 1997. Occurence of plastid RNA editing in all major lineages of land plants. *Proc Natl Acad Sci U S A* 94:6285-6290.

Gilbert, W. 1978. Why genes in pieces? *Nature* 271:501.

Gilbert, W. 1986. The RNA World. *Nature* 319:618.

Grosjean, H., F. Constantinesco, D. Foiret and N. Benachenhou. 1995a. A novel enzymatic pathway leading to 1-methylinosine modification in *Haloferax volcanii* tRNA. *Nucleic Acids Res* 23:4312-4319.

Grosjean, H., M. Sprinzl and S. Steinberg. 1995. Posttranscriptionally modified nucleosides in transfer RNA: their locations and frequencies. *Biochimie* 77:139-141.

Grosjean, H., J. Edqvist, K.B. Stråby and R.Giégé. 1996. Enzymatic formation of modified nucleosides in tRNA: dependence on tRNA architecture. *J. Mol. Biol.* 255:67-85.

Grosjean, H., Z. Szweykowska-Kulinska, Y. Motorin, F. Fasiolo, and G. Simos. 1997. Intron-dependent enzymatic formation of modified nucleosides in eukaryotic tRNAs: A review. *Biochimie* 79:293-302.

Helm, M., H. Brulé, F. Degoul, C. Cepanec, J.P. Leroux, R. Giegé and C. Florentz. 1997. The presence of a modified nucleotide is required for the cloverleaf folding of a human mitochondrial tRNA. Oral presentation 3-13 of the 17th International tRNA Workshop, May 10-15, Kazusa Akademia Center, Japan.

Herbert, A. 1996. RNA editing, introns and evolution. *Trends Genet* 12:6-9.

Hurst, L.D. and G.T. McVean. 1996. A difficult phase for intron-early. *Curr Biol* 6:533-536.

Joyce, G.F. and L.E. Orgel. 1993. Prospects for understanding the origin of the RNA World, p. 1-25. *In* R.F. Gesteland, J.F. Atkins (ed.), *The RNA World*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Kauffman, S.A. 1993. The origins of order. Self-organization and selection in evolution. Oxford University Press, New York, Oxford.

Lane, B.G., J. Ofengand and M.W. Gray. 1995. Pseudouridine and O2'-methylated nucleosides. Significance of their selective occurrence in rRNA domains that function in ribosome-catalyzed synthesis of the peptide bonds in proteins. *Biochimie* 77:7-15.

Lazcano, A. 1994a. The transition from nonliving to living, p. 61-69. *In* S. Bengtson (ed.), *Early life on earth*, Columbia University Press, New York.

Lazcano, A. 1994b. The RNA World, its predecessors, and its descendants, p.70-80. *In* S. Bengtson (ed.), *Early life on earth*, Columbia University Press, New York.

Limbach, P.A., P.F. Crain and J.A. McCloskey. 1994. Summary: the modified nucleosides of RNA. *Nucleic Acids Res* 22:2183-2196.

Mazaki, H., T. Ogawa, K. Tomita, T. Ueda, K. Watanabe, and T. Uozumi. 1997. Colicin E5; a novel ribonuclease targeting a specific group of tRNAs. Oral presentation 2-1 of the 17th International tRNA Workshop, May 10-15, Kazusa Akademia Center, Japan.

Ofengand, J. and A. Bakin. 1997. Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J Mol Biol* 266:246-268.

Osawa, S., T.H. Jukes, K. Watanabe and A. Muto. 1992. Recent evidence for evolution of the genetic code. *Microbiol Rev* 56:229-264.

Peracchi, A., L. Beigelman, N. Usman, D. Herschlag. 1996. Rescue of abasic hammerhead ribozymes by exogenous addition of specific bases. *Proc Natl Acad Sci U S A* 93:11522-11527.

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

Scott, J. 1995. A place in the world for RNA editing. *Cell* 81:833-836.

Simpson, L. 1990. RNA editing - a novel genetic phenomenon? *Science* 250:512-513.

Szathmáry, E. 1997. The first Two billion years. *Nature* 387:662-663

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

Usman, N., L. Beigelman, and J.A. McSwiggen. 1996. Hammerhead ribozyme engineering. *Curr Opin Struct Biol* 6:527-533.

Weiner A.M. and N. Maizels. 1990. RNA editing: guided but not templated? *Cell* 61:917-920.

Yang, J.H., N. Usman, P. Chartrand, and R. Cedergren. 1992. Minimum ribonucleotide requirement for catalysis by the RNA hammerhead domain. *Biochemistry* 31:5005-5009.

Yu, W. and W. Schuster. 1995. Evidence for a site-specific cytidine deamination reaction involved in C to U RNA editing of plant mitochondria. *J Biol Chem* 270:18227-18233.

Table 1. Modified nucleosides found at corresponding positions in tRNAs from eukaryotes, bacteria and archaea [a]

Position in tRNA	Modified nucleotide[b]
13	Ψ [c]
34	Cm
34	Q [d]
37	t^6A
37	m^1G [e]
38	Ψ
39	Ψ [f]
55	Ψ [g]
58	m^1A [c]

[a] Data taken in Appendices I and V, and Grosjean et al. (1995b). [b] Abbreviations as in Appendix I. [c] Not found in chloroplasts. [d] There are different variants of Q depending on base substituents (see Appendix I). [e] Modification found in more than 80% of tRNAs with a G37. [f] Modification found in more than 80% of tRNAs with a U39. [g] Modification found in more than 90% of tRNAs with a U55.

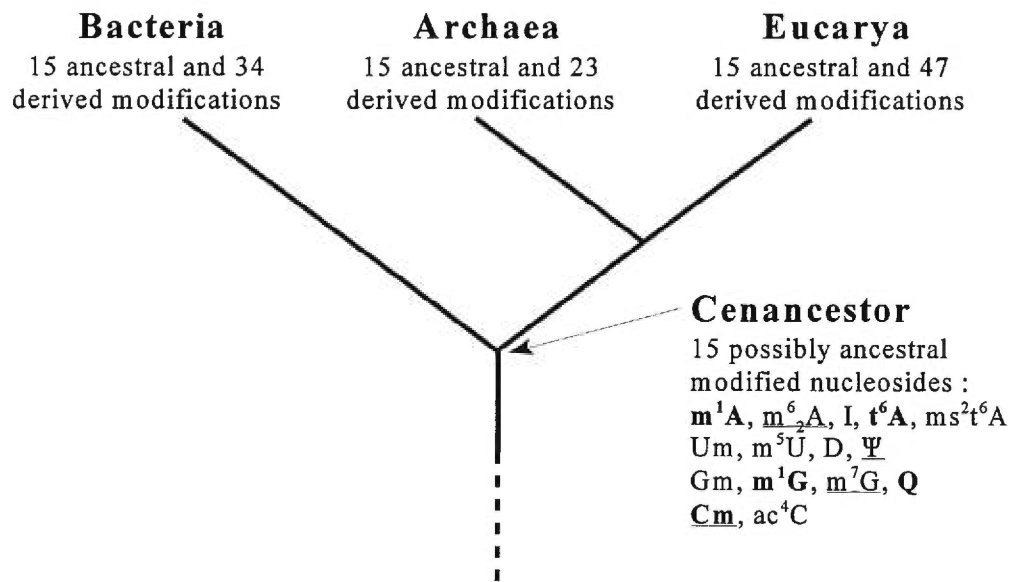
Table 2. Conserved modified nucleotides in rRNAs from eukaryotes, bacteria and archaea [a]

Modified nucleotide [b]	rRNA species [c]		
	archaea	bacteria	eucaryote
$m^6{}_2A$	16S	16S/23S	18S
Cm	5S	23S	5.8S/17S/18S/28S
m^7G	16S	? [d]	5S/17S/18S/26S/28S
Ψ	16S	16S/23S	5S/5.8S/18S/28S

[a] Data taken from Limbach et al. (1994). [b] Abbreviations as in Appendix I. [c] Refers to the rRNAs containing the modified nucleotide. [d] The precise position of the modified nucleotide has not been determined.

Figure 1. Nucleoside modifications likely to have been occurring in the cenancestor and in the RNA World.

The upper part of the figure presents the results of the phylogenetic analysis of modified nucleotides. The data are from Table 1 and Appendix I. The modified nucleosides found at corresponding positions in tRNAs from the three domains (Table 1) are in bold. The modified nucleosides found in rRNAs of all three domains are underlined (Table 2). Derived modifications are considered to have occured later in evolution. The lower part of the figure refers to the discussion about the prebiotic world and RNA World. See the text for more details.


Bacteria
15 ancestral and 34
derived modifications
Archaea
15 ancestral and 23
derived modifications
Eucarya
15 ancestral and 47
derived modifications
Cenancestor
15 possibly ancestral
modified nucleosides :
m1A, m62A, I, t6A, ms2t6A
Um, m5U, D, Ψ
Gm, m1G, m7G, Q
Cm, ac4C
Likely modified nucleosides in the
prebiotic world and the RNA World :
- Ψ
- Methylated nucleosides
- I
- Aminoacylated nucleosides