1013

Université de Neuchâtel

Faculté des Sciences

Le génome chloroplastique de Glycine max: Analyse des gènes impliqués dans la structure et la fonction du ribosome

Thèse de doctorat

Jean-Marc von Allmen, 1987

A mes parents

REMERCIEMENTS

Ce travail a été effectué dans le laboratoire de Biochime de l'Université de Neuchâtel, sous la direction de Monsieur le Professeur Erhard Stutz, que nous remercions vivement pour la confiance qu'il nous a accordé, ainsi que pour le soutient et l'intéret qu'il nous a montré tout au long de ce travail.

Nous tenons à remercier M. le professeur R. Mache qui a accepté de juger ce travail, ainsi que M. le Docteur P. Schürmann pour sa disponibilité tout au long de notre travail et pour leur participation au jury de thèse.

Nous n'aimerions pas omettre d'exprimer notre sincère gratitude envers tout ceux qui nous ont aidé dans nos travaux

Mademoiselle Pascale Kaelin qui a participé à ce travail en nous assistant techniquement et qui nous a souvent déchargé d'une partie de notre travail d'assistanat.

M. le Docteur Paul-Etienne Montandon pour les discussions ouvertes et l'assistance tactique qu'il nous a régulièrement apporté.

Tous nos collègues du laboratoire, qui, par leurs discussions, leurs conseils ou simplement leur bonne humeur, ont rendu agréable l'atmosphère dans laquelle nous avons pu effectuer notre travail de recherche.

SUMMARY

The chloroplast genome of Glycine max :
Analysis of genes involved in the structure and the function of ribosome.

The chloroplast *rps12* gene coding for ribosomal protein S12 is divided. Two copies of the *3'-rps12* are localised in the inverted repeats, 1854 bp upstream of *16S* and in the opposite orientation. One copy of the *5'-rps12* is contained in SacD fragment of the large single copy region, at almost 1 kb in front of the SacD-SacF junction. The mRNAs require *cis* and *trans* splicing. The first results obtained by primer extension indicate that a mix of *trans* spliced and precursors RNAs do exist *in vivo*.

The *3'-rps12* and *rps7* are cotranscribed, common 1.1 and 1.6 kb mRNAs are demonstrated by Northern experiments. The *ndhB* gene following the *rps7* is also transcribed, without indication of co-transcription. The intron 2, spliting the *3'-rps12* is processed, leading to a 530 nucleotides RNA shown by Northern hybridization.

The *5'-rps12* is located almost 1 kb in front of *rpl20*. A DNA probe, internal of either *5'-rps12* or *rpl20*, hybridizes against 1.1 kb mRNAs. A 1.1 kb mRNA is proposed to be the result of *cis* and *trans* splicing.

The processing of *rps12* (or the mRNA contents) seems to be a function of the state of the leaves. Quickly growing leaves are not showing big RNAs (of 4.2, 3.5 and 3.1 kb) present in more stable leaves.

trnV, *16S*, *trnI*, *trnA* and the beginning of *23S* are sequenced. Regions of 16S rRNA linked to a function of the ribosome are discussed. The introns dividing *trnI* and *trnA* are compared to those of tobacco and maize and found more similar to maize.

The *psbB* is localised in SacD fragment and the *ndhB* is partly sequenced. Two ORFs are discussed, one of them (ORF 143) being included in the intron 2 and ending in frame with the exon 3 of *rps12*.

TABLE DES MATIERES

1. INTRODUCTION 1

2. MATERIEL ET METHODES 3

2.1 **Les méthodes spécifiques**
2.1.1 Conditions de culture de la plante
2.1.2 Production et isolation de l'ADN-ct
2.1.3 Production et isolation de l'ARNct 4

2.2 **Méthodes d'isolation, d'analyse et de purification des ADN** 5
2.2.1 Caractérisation et séparation des acides nucléiques
2.2.1.1 Electrophorèse 6
2.2.1.2 Elution des acides nucléiques à partir d'un gel
2.2.2 Les systèmes hôte-vecteur
2.2.2.1 Les souches bactériennes 7
2.2.2.2 Milieux de croissance des bactéries et antibiotiques
2.2.2.3 Isolation de plasmides en grande quantités 8
2.2.3 L'introduction de vecteurs dans E.coli
2.2.4 L'identification des recombinants 9
2.2.4.1 Sélection par les antibiotiques
2.2.4.2 Sélection de par l'activité de la β-galactosidase
2.2.5 L'analyse des recombinants
2.2.5.1 Méthode de Hoggness
2.2.5.2 Mini-lysats 10
2.2.5.3 Méthode de Southern
2.2.5.4 Hybridations homologues et hétérologues
2.2.5.5 Construction d'une carte de restriction 11
2.2.5.6 Sous-clonage
2.2.6 Le séquençage selon Sanger 12
2.2.6.1 Préparation de la matrice
2.2.6.2 Les solutions de travail
2.2.6.3 La réaction de séquence 13
2.2.7 Le Marquage des ADN 14
2.2.7.1 Nick-translation
2.2.7.2 Remplissage à la Klenow en présence d'αP^{32}dATP 15
2.2.7.3 Marquage spécifique d'un seul brin d'ADN

2.3 **Méthodes de séparation et d'analyse des ARN** 16
2.3.1 Les méthodes d'analyses
2.3.1.1 Dénaturation des ARN au glyoxal et séparation
2.3.1.2 Northern et électroélution
2.3.2 Séquençage de l'ARN 17
2.3.2.1 Le séquençage par "primer extension" Méthode de Goldenberg et Hauser, (1983).

2.4 **Matériel, enzymes et solutions particulières** 18
2.4.1 Le Matériel particulier
2.4.2 Les enzymes
2.4.3 Le matériel radioactif
2.4.4 Les tampons
2.4.5 Les produits chimiques 19

2.5 **L'outil informatique :** 20
2.5.1 Le BASIC structuré de VAX/VMS
2.5.2 Le programme ANALYSE
2.5.3 Le programme HEID 23

3. RESULTATS ET DISCUSSION 24

3.1 Identification et clonage des gènes rps7 et rps12
3.1.1 La localisation des gènes
3.1.2 Le clonage des fragment d'ADNct intéressants 26
3.1.3 Vérification que les deux IR contiennent une même information
3.1.4 Identification incontestable des gènes rps7 et rps12 29
3.1.5 Cartographie classique du fragment HindIII de 9 kb (clone HD906) 31
3.1.6 Clonage du 5'-rps12 34
3.1.7 Cartographie du 5'-rps12 35
3.1.8 Séquençage 39
3.1.9 Recherche des promoteurs 48

3.2 **Les ARN messagers** 49
3.2.1 Hybridations du type Northern 50
3.2.2 Distribution des ARN messagers en fonction de l'état physiologique de la plante. 52
3.2.3 Détermination du site de splicing par "Primer extension" 55
3.2.4 transcription constitutive ou induite ? 59

3.3 **Résultats corollaires** 61
3.3.1 le gène psbB
3.3.2 Le gène ndhB
3.3.3 Les gènes trnV, trnI, trnA
3.3.4 "Gènes" des ORF 63
3.3.5 Le gène rpl20

3.4 **DISCUSSION I : Quelques aspects du cis- et du trans- splicing** 65
3.4.1 Cas des introns du rps12
3.4.2 Etude de la complémentarité de la limite 5' des introns du groupe II avec une séquence de la partie 3' de l'intron. 67
3.4.3 Les introns internes aux ARN-t Ile et Ala 72
3.4.4 Les consensus 75

3.5 **DISCUSSION II : Quelques corrélations existant entre Le ribosome et sa fonction** 76
3.5.1 Comparaison des protéines S7 et S12 de différents organismes. 77
3.5.2 Rapports structure-fonction 81
3.5.3 Comparaison de la structure des opérons contenant le rps12
3.5.4 L'ARN 16S : structure et fonction 83

4. CONCLUSION 87

5. BIBLIOGRAPHIE 89

6. ANNEXES 99

6.1 Mode d'emploi détaillé des programmes avec TABLE DES MATIERES particulière A0

ABREVIATIONS

Unités physiques

g, mg, µg	gramme, milligramme, microgramme
l, ml, µl	litre, millilitre, microlitre
M, mM	molaire, millimolaire
mA	milliampère
V	volt
Kb	mille paire de base
°C	degré centigrade

Produits chimiques

ATA	acide aurintricarboxylique
BSA	albumine de sérum de boeuf
EDTA	sel disodique de l'acide éthylène diamine tétra-acétique
EtBr	Bromure d'éthidium
DTT	dithio-threitol
PEG	polyéthylène glycol
SDS	dodécyl-sulfate de sodium
SSC	solution standard de citrate
SSPE	solution standard de phosphate, EDTA
TA	tampon acétate
TBE	tampon borate, EDTA
TE	tampon Tris, EDTA
Tris	Tris hydroxyméthyl-aminométhane

Molécules biologiques

ADN	acide désoxyribonucléique
ADN-ct	acide désoxyribonucléique chloroplastique
ARN	acide ribonucléique
ARN-m	ARN messager
ARN-t	ARN de transfert
DNase	désoxyribonucléase
IR	région répétées inverses
ORF	"open reading frame" région pouvant coder pour une protéine de fonction indéterminée
RNase	ribonucléase
RuBP	ribulose bisphosphate

Nomenclature des gènes chloroplastiques

selon Hallick et Bottomley (1983)

1. INTRODUCTION

La source d'énergie la plus importante à disposition du monde vivant est l'énergie solaire. La plupart des organismes dépendent de cette source d'énergie mais sont incapable d'utiliser directement la lumière, l'Homme en fait partie. Nous dépendons étroitement des plantes car elles transforment l'énergie des photons en une source assimilable, nous livrant des sucres que nous sommes capables de brûler et, par la même occasion, nous fournissent un certain nombre de vitamines que nous sommes incapable de synthétiser.

A l'intérieur de la cellule végétale se trouve un organite spécifique, le chloroplaste, qui est responsable de la photosynthèse. Chaque chloroplaste contient 10 à 60 copies identiques de son information génétique propre : l'ADN chloroplastique (Bedbrook et Kolodner, 1979). Récemment, l'ordre des 155844 nucléotides de l'ADN chloroplastique de tabac, *Nicotiana tabacum*, à été déterminé (Shinozaki et al., 1986) et parallèlement, un autre groupe (Ohyama et al. 1986) a séquencé l'ADNct total d'une Hépatique : *Marchantia polymarpha*.

Ces travaux permettent de vérifier quelles sont les protéines dont la synthèse est susceptible d'être dirigée par le chloroplaste et quelles sont les protéines dont la synthèse doit être dictée par le génome nucléaire. En effet, si comme on le pense généralement, le chloroplaste et la mitochondrie sont d'origine endosymbiotique (Schwartz et Dayhoff, 1978), ces organites contiennent beaucoup moins d'informations qu'une bactérie actuelle et n'ont qu'une autonomie partielle.

Une partie des protéines localisée finalement dans le chloroplaste sont codées par le noyau (Ellis, 1981). Les gènes nucléaires sont transcrits, traduits par les polysomes cytoplasmiques et les précurseurs des protéines sont transportés dans leur compartiment propre. Les protéines destinées au chloroplaste, par exemple, sont exportées dans le chloroplaste parce que leurs précurseurs possèdent un peptide de transit spécifique. (Karlin-Neumann et Tobin, 1986).

Le soja, que nous étudions dans ce travail, ne fait pas exception : Par exemple, l'enzyme responsable de la fixation du gaz carbonique lors du processus de la photosynthèse, la ribulose 1,5-bis-phosphate carboxylase est formée de l'assemblage de petites sous-unités d'origine nucléaire (Kawashimia et Wildman, 1972) et de grandes sous-unités d'origine chloroplastiques (Singh et al., 1985). La formation et la fonction du chloroplaste résultent donc de l'interaction complexe existant entre le système génétique nucléaire et celui du chloroplaste.

La participation de l'ADN chloroplastique à la formation du chloroplaste est en cours d'étude. Les protéines et les ARN codés sont déjà connus pour *Marchantia* et le tabac.
Nous trouvons, codés par l'ADN chloroplastique de ces plantes :

Les ARN ribosomiques 16S, 23S, 4.5S, 5S
Les sous-unité α, β, β' de l'ARN polymémase
Les protéines ribosomiques L2, L14, L16, L20, L21, L22, L23, L33, S2, S3, S4, S7, S8, S11, S12, S14, S15, S18, S19, L36 (du gène *secX*)
Le facteur d'initiation 1 de la synthèse des protéines
Des protéines impliquées dans la photosynthèse, tel la grande sous-unité de la RuBP carboxylase, 4 protéines faisant partie des centres réactionnels P700 et P680, les protéines P32 (gène *psbA*), D2 (gène *psbD*) et G (gène *psbG*), les cytochromes b559, f et b6, les sous-unités α, β, ε, I, III et IV de l'ATPase
Des gènes identifiés grâce à leur similitude avec des séquences connues, tel les 6 sous-unités de la NADH deshydogénase similaires à des protéines mitochondriales de mammifères, 3 protéines ayant une structure 4Fe-4S de type Ferredoxine, une protéine liant probablement l'ATP et ressemblant à une perméase d'*E.coli*
Une trentaine d'ARN de transfert différents
Une quarantaine de polypeptides possibles, mais non encore identifiés

Il est intéressant de noter que le chloroplaste, bien que capable d'opérer la réplication de son ADN, la transcription et la traduction, n'a en fait gardé le contrôle que de quelques constituants. Les études concernant les interactions entre le noyau et le chloroplaste ne sont encore qu'à leurs balbutiements.

Nous nous sommes plus particulièrement intéressé au plastoribosomes. Ceux-ci ont une origine mixte (nucléaire + chloroplastique) puisque 20 protéines seulement sur les 53 nécessaires (Wittmann, 1983; Markmann-Mulisch et al., en soumission) au bon fonctionnement d'un ribosome de type procaryotique sont d'origine chloroplastique.

Nous avons choisi d'étudier les gènes chloroplastiques de soja codant pour les protéines ribosomiques S7 et S12 (ainsi nommées par leur similarité avec les 7ème et 12ème protéines en taille de la petite sous-unité du ribosome d'*Escherichia coli*). Ces gènes sont intéressant à plus d'un titre :

- Ils sont codés par le chloroplaste alors que la majorité des gènes ribosomiques sont d'origine nucléaire.

- Ils donnent certaines indications phylogénétique.

- Les protéines impliquées ont un rôle important, sur le plan de l'assemblage et de la fonction du ribosome.

- Le gène *rps12* (protéine S12) apporte une dimension nouvelle à la complexité des étapes de maturation nécessaires à l'obtention d'un ARN messager mature.

La notion classique du "gène", qui était définie comme "une unité biologique de l'hérédité" à l'époque de Mendel, doit subir dans nos esprits une sérieuse modification.

Nous espérons que ce travail permettra de contribuer à cette modification tout en apportant quelques éléments supplémentaires sur la structure du ribosome et sa fonction.

2. MATERIEL ET METHODES

La première condition à remplir pour étudier l'ADN et les ARN chloroplastiques de soja est d'obtenir des plantes dont on puisse repérer l'état physiologique de façon à obtenir des résultats reproductibles et de réussir à isoler des chloroplastes intacts de façon répétitive. A cette fin, nous avons dû affiner les méthodes existantes.

2.1 *Les méthodes spécifiques*

Les méthodes suivantes ont été mise au point particulièrement pour le soja, qui pose certains problèmes de par la dureté de ses feuilles si elles sont trop agées et de la difficulté que l'on a à briser les cellules sans briser les chloroplastes. La présence de polyphénols et d'oxydases endogènes indésirables (Biotto, communication personnelle, 1987) ne facilite pas l'obtention du matériel.

2.1.1 Conditions de culture de la plante

Les graines de soja (*Glycine max*, variété Maple Arrow) nous ont été aimablement fournies par la maison Nestec S.A. Les graines sont tout d'abord imbibées une nuit dans de l'eau courante et à température ambiante, puis sont semées dans des bacs contenant de la vermiculite. Les plantes sont cultivées en chambre de culture sous des conditions contrôlées, avec une photopériode de 16 heures de lumière à 2.5 $mW/cm^2 xS$, à une température de 25°C et une humidité relative de 70%. Nous les arrosons régulièrement, mais nous avons finalement abandonné la solution nutritive de Siegenthaler et Depéry (1976), qui accélère la croissance, mais semble faire produire par la plante des ADN difficiles à couper par enzymes de restriction.

2.1.2 Production et isolation de l'ADN-ct

Les feuilles de soja sont cueillies 16 jours après que les graines aient été mises à gonfler. Physiologiquement, on reconnaît ce stade car il coïncide au moment où la plante a 4 feuilles. Les deux premières sont bien dépliées et vertes, alors que les deux feuilles suivantes sont en train de se déplier et en pleine croissance. Les feuilles sont cueillies, puis stoquées 48 heures à 4°C et à l'obscurité, afin d'éliminer les réserves d'amidons.

Dans une isolation, 100-200 g de feuilles sont homogénéisées par plusieurs broyages brefs (3 à 5 secondes) dans un broyeur bricolé dont les lames ont été remplacées par des lames de scalpel. Les feuilles sont broyées à 4°C dans 400 ml de tampon A (200 mM Tris-HCl, 200 mM EDTA, 300 mM D-mannitol, 20 mM β-mercaptoéthanol, pH 9.0. L'homogénat est filtré à travers 6 couches de gaze et centrifugé à 150xg durant 1 minute, pour éliminer les débris végétaux, une partie des cellules entières et des noyaux. Le surnageant est récolté et centrifugé à 1500xg pendant 5 minutes. Le culot de chloroplastes est resuspendu délicatement dans 6 ml de tampon B (50 mM Tris-HCl, 20 mM EDTA, 300 mM saccharose, pH 8.0) et déposé ensuite sur un gradient de 35 ml contenant deux couches de saccharose 30 et 50%, préparés dans du tampon C

(50 mM Tris-HCl, 20 mM EDTA, pH 8.0). Les tubes sont centrifugés pendant 30 minutes à 1500xg, la fraction contenant les chloroplastes est diluée dans environ 200 ml de tampon B, puis centrifugée 5 minutes à 1500xg. Le culot est resuspendu dans 5 ml de tampon C et les choroplastes sont lysés en ajoutant 1 ml de tampon C contenant 10% de sarcosinate de sodium. Après 5 minutes d'incubation à 0°C, la lyse des chloroplastes est contrôlée au microscope et si la préparation est correcte, le mélange est centrifugé pendant 5 minutes à 1500xg. Le surnageant est traité à la RNase A (50 μg/ml) pendant 1 h à température ambiante, puis à la protéinase K (200 μg/ml) pendant 30 minutes à 37°C. Nous ajoutons alors 0.78 g/ml de CsCl et après 1 h 30 à 0°C, le mélange est centrifugé pendant 30 minutes à 16000xg pour précipiter et éliminer une partie des résidus polypeptidiques. Le surnageant est déposé sur un coussin d'1.5 ml de CsCl (1.009 g/ml CsCl, 200 μg/ml EtBr). Les tubes sont centrifugés pendant 16 h à 130'000xg et à la température de 20°C, dans un rotor de type Ti 65. La bande d'ADN est récoltée sous lumière UV (366 nm) et l'EtBr est éliminé par 5-6 extractions à l'isopentylalcool (saturé avec 10 mM EDTA, pH 8.0). L'ADN est ensuite dialysé contre 3 x 1 litre de 0.1 xSSC pendant 28 h à 4°C, et finalement concentré par précipitation à l'éthanol pendant 30 minutes à -70°C. Le culot est lavé à l'éthanol 80%, puis, après séchage sous vide, il est resuspendu dans 50 μl de tampon TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.5).

A partir de 150 g de feuilles, nous obtenons envirom 50 μg d'ADN chloroplastique.

Remarque : Lorsque l'ADN chloroplastique est peu ou pas digérable par enzymes de restrictions, nous avons utilisé un inhibiteur des protéases, qui restitue une bonne digestion (James, 1978).

2.1.3 Production et isolation de l'ARNct

Le principal problème rencontré lors de l'isolation des ARNct est qu'il est impossible de cueillir les feuilles beaucoup avant l'isolation, ni de laisser s'épuiser les réserves d'amidon de la plante. Les ARN sont des produits dont la durée de vie peut être relativement courte (exemple connu des bactéries), ainsi nous cueillons les feuilles immédiatement avant de commencer l'isolation des ARN.

Pour isoler les chloroplastes, nous avons utilisé dans un premier temps une adaptation des méthodes de Mills et Joy (1980) et de Kalberer *et al.* (1967). 50 g de feuilles sont broyés dans du milieu de broyage MB (25 mM MOPS-KOH, 350 mM sorbitol, 2 mM EDTA, 3 mM β-mecaptoéthanol, pH 7.6). L'homogénat est filtré à travers 6 couches de gaze, les débris cellulaires sont sédimentés en centrifugeant pendant 1 minute à 150xg et le surnageant est chargé sur une couche de MB, 40% Percoll. Le tout est centrifugé pendant 5 minutes à 2500xg et la fraction contenant les chloroplastes est diluée dans du tampon MB puis sédimentée en 3 minutes à 2500xg. Les ARNct sont alors isolés selon Sagher *et al.* (1976) avec 30 minutes de traitement à la DNase (Worthington, ribonuclease free).

De cette manière, 50 g de feuilles délivrent environ 50-80 μg d'ARN chloroplastiques.

Dans un deuxième temps, et en vue d'améliorer les rendements, nous avons adapté la méthode de Hildebrand *et al.* (en soumission) qui a l'avantage de protéger les ARN des effets des ribonucléases au moyen de l'ATA (acide aurintricarboxylique). Par ce protocole, 200 g de feuilles de soja sont homogénéisés dans le broyeur à lames de scalpel, dans 450 ml de tampon de broyage (TB) en train de dégeler (50 mM Tris-HCl, 350 mM Mannitol, 100 mM EDTA, 5 mM β-mercaptoéthanol, 0.1% BSA, 1 mM ATA, pH 8.0. L'homogénat est centrifugé à 150xg durant 1 minute pour éliminer les débris végétaux. Les chloroplastes contenus dans le surnageant sont centrifugés pendant 5 minutes à 1500xg, resuspendus à l'aide d'un pinceau doux, lavés dans 50 ml puis dans 25 ml de tampon TB. Ils sont ensuite resuspendus vigoureusement dans 10 ml de tampon de lyse (2% Triton X-100, 50 mM Tris-HCl, 50 mM NaCl, 1 mM ATA, pH 7.6), et gardés pendant 10 minutes à 0°C. Les chloroplastes lysés sont extraits 2x avec un demi volume de phénol-eau et 2x avec de l'éther, pour être finalement précipités à l'éthanol. L'ATA est retiré par chromatographie, en passant l'échantillon sur une colonne de 5 ml de Sephadex G-25 avec 100 mM acétate de sodium, 0.1% SDS comme tampon d'élution. L'ADN peut ensuite être digéré par la DNase I (Worthington, ribonuclease free) selon la méthode habituelle.

Ainsi, 200 g de feuilles délivrent 800 µg d'ARN selon la densité optique à 260 nm, mais la quantité réelle d'ARNct doit être plus proche de 350 µg, à en juger par les signaux d'hybridations observés.

2.2 Méthodes d'isolation, d'analyse et de purification des ADN

Une fois le matériel biologique préparé, il nous reste à l'analyser. A cette fin, il existe toute une batterie de méthodes qui ont été mises au point dans tous les domaines utilisant la biologie moléculaire. Beaucoup de ces méthodes proviennent des systèmes bactériens ou animaux.

2.2.1 Caractérisation et séparation des acides nucléiques

Chaque année, il s'ajoute une série d'enzymes de restiction sur le marché. Que ce soit à des fins préparatives ou analytiques, les enzymes de restrictions sont pour nous des alliés utiles, puisqu'ils reconnaissent des séquences de 4 à 8 nucléotides et coupent la molécule d'ADN en ces endroits spécifiques. L'emploi des enzymes de restrictions et les conditions de digestion sont décrites par nos fournisseurs (Boehringer Mannheim et Biolabs New England), et dépendent principalement des organismes à partir desquels les enzymes ont été isolés. La séparation des fragments générés se fait le plus souvent au moyen de l'électrophorèse sur gels d'agarose pour les plus grands fragments ou sur gels de polyacrylamide lorsque les fragments sont plus petits que 1000 paires de bases.

2.2.1.1 Caractérisation et séparation des acides nucléiques : Electrophorèse

Les conditions de migration des acides nucléiques dépendent de la force ionique du tampon dans lequel l'électrophorèse a lieu, ainsi que de la concentration et du type de gel utilisé.

Nous avons suivi la méthode décrite par Maniatis *et al.* (1982), en choisissant le tampon TA (40 mM Tris-acide acétique, 20 mM acétate de sodium, 1 mM EDTA, pH 7.8), pour séparer lentement et de façon optimale les grands fragments d'ADN (gels d'une nuit à 38V). Lors des séparations rapides, nous utilisons de préférence du TBE (130 mM Tris, 90 mM acide borique, 25 mM EDTA, pH 8.8).

ADN et ARN sont mis en évidences au bromure d'éthidium, selon Sharp *et al.* (1973).

2.2.1.2 Caractérisation et séparation des acides nucléiques : Elution des acides nucléiques à partir d'un gel

Il est important de pouvoir éluer des fragments d'ADN afin de les utiliser par exemple, dans le clonage ou la cartographie. Nous avons utilisé diverses méthodes, en fonction de la taille des fragments et de la quantité de matériel à récupérer.

Nous avons d'abord essayé la méthode du low-melt agarose, qui évite de perdre les fragments relativements petits que l'on doit éluer à partir de gels d'agarose. Décrite par Weislander (1976), cette méthode consiste à utiliser de l'agarose à bas point de fusion, et d'extraire l'ADN après avoir liquéfié l'agarose à 65°C.

Dans la majorité des cas, l'électroélution est la méthode de choix. Les bandes d'ADN intéressantes sont découpées du gel et l'électroélution est faites dans un sac à dialyse fermé ainsi que le décrit Maniatis *et al.* (1982). Cela à l'avantage d'éviter les pertes de matériel au maximum. Il suffit d'effectuer ensuite deux extractions avec du phénol saturé avec 10 mM Tris, d'une extraction à l'éther et d'une précipitation à l'éthanol, pour se débarasser des polysaccharides sulfatés qui, sinon, inhiberaient la ligase et la plupart des enzymes de restriction.

La troisième méthode, l'élution selon Maxam, est régulièrement utilisée lorsque le matériel est de très petite taille (plus petit que 400 nucléotides) et qu'il faut le sortir d'un gel de polyacrylamide (Maxam et Gilbert, 1977).

2.2.2 Les systèmes hôte-vecteur

Une fois le matériel de base en notre possession, et comme les quantités d'ADN obtenables à partir des plantes sont petites, il est indispensable de pouvoir intégrer dans une bactérie les fragments d'ADN que nous désirons analyser de plus près. Une bactérie transformée nous produit un grand nombre de copies du matériel qui nous intéresse. il ne nous reste plus qu'à séparer cet ADN de l'information génétique propre à la bactérie-hôte.

Nous avons utilisé deux types de vecteurs pour cloner des fragments d'ADN étrangers et les propager dans *Escherichia coli* :

Les plasmides **pBR322**, contenant un ADN double brin, décrit par Sutcliffe (1978) et intégré dans la souche HB101 (Boyer et Roulland-Dussoix, 1969).

pEMBL18, contenant un ADN double ou simple brin, selon le mode de culture (Dente *et al.*, 1983)

Le bactériophage **M13**, décrit par Denhardt *et al.* (1978) et décrit par Messing et Vieira, (1982) dans les formes mp8 et mp9 que nous avons utilisées. C'est un virus dont l'ADN est simple brin qui infecte la souche JM103 d'*E.coli*, principalement utilisé lors du séqençage selon Sanger *et al.* (1980).

2.2.2.1 Les systèmes hôte-vecteur : Les souches bactériennes

Nous conservons les souches bactériennes à -20 ou à -70°C dans leurs milieux de culture complémentés de 15% glycérol afin d'éviter la disparition ou l'apparition de marqueurs génétiques et de limiter les risques de contaminations croisées.

La pureté des souches est régulièrement vérifiée par étalement polygonal (Maniatis *et al.*, 1982).

2.2.2.2 Les systèmes hôte-vecteur : Milieux de croissance des bactéries et antibiotiques

Selon le but poursuivi, nous avons cultivé nos bactéries, soit sur des milieux complexes, (milieux LB, 2xTY), que l'on peut rendre sélectifs pour les bactéries transformées en ajoutant 20 µg/ml d'antibiotiques (ampicilline, tétracycline), soit sur un milieu synthétique complexe dont tous les paramètres sont maîtrisés (Minimal agar).

Le milieu L (10 g/l tryptone, 5g/l extrait de levure, 85 mM NaCl pH 7,5) est utilisé avec *Escherichia coli* souche HB101, le milieu 2xTY (16 g/l tryptone, 10 g/l extrait de levure, 170 mM NaCl), avec la souche JM103. Le milieu Minimal agar (15 g/l bactoagar, 28 mM KH_2PO_4, 60 mM K_2HPO_4, 7.5 mM sulfate d'ammonium, 1.7 mM citrate de sodium, 0.8 mM sulphate de magnésium, 25 µM Vitamine B1, 10 mM glucose) est employé comme milieu solide pour la souche JM103 de *E.coli* K12. Cette souche comporte une délétion du gène *pro* qui permet la sélection de l'épisome sur un milieu glucose/minimum. L'épisome étant nécessaire pour l'infection par le phage M13 qui est mâle spécifique, il est important d'inoculer un milieu riche directement à partir d'une culture sur Minimal agar.

Lorsqu'une amplification des plasmides est nécessaire, du chloramphénicol est ajouté en cours de culture selon Clewell (1972)

2.2.2.3 Les systèmes hôte-vecteur : Isolation de plasmides en grande quantités

Les plasmides utilisés en Biologie moléculaire étant des ADN extra-chromosomiques, circulaires et de tailles relativement petites, il est possible de les isoler de l'ADN bactérien de taille nettement supérieure, qui est linéarisé parce que brisé au cours de nos manipulations.

Nous avons préféré une méthode d'isolation (Rütti, 1982) où les bactéries sont lysées au SDS (selon Godson et Vapnek 1973), les débris cellulaires précipités selon Hirt, (1967) et où l'ADN bactérien est ensuite éliminé par une dénaturation à l'alcali suivie d'une extraction phénol en milieu à haute force ionique (débarassant de tout l'ADN bactérien, qui dans ces conditions est simple brin).

La méthode de purification des ADN circulaires par centrifugation à l'équilibre dans un gradient de chlorure de césium, est nettement plus longue, (Maniatis *et al.*, 1982) et n'a été utilisée que lorsqu'il fallait un matériel de particulièrement bonne qualité.

2.2.3 L'introduction de vecteurs dans *E.coli*

Pour préparer de l'ADN double brin en grande quantité, nous avons cloné nos fragments d'ADN dans les extrémités cohésives des sites HinDIII et BamHI du plasmide pBR322. L'ADN plasmidial est digéré, déphosphorylé (Chaconas et van de Sande, 1980), puis ligué avec les fragments désirés (Weiss *et al.*, 1968). La transformation des cellules de la souche HB101 (Mandel et Higa, 1970) est augmentée en gardant les cellules traitées au $CaCl_2$ pendant quelques heures dans de la glace, comme c'est conseillé par Dagert et Ehrlich (1979).

Le clonage des phages de type M13 dans *E. coli* est compliqué par le fait que seule la forme réplicative double brin de l'ADN des bactériophages est utilisable pour la digestion, la ligation et la transformation des bactéries. La forme réplicative est obtenue en récoltant des cellules infectées par le phage M13 mp8 ou mp9 qui sont récoltées alors que les cellules sont encore en phase de croissance exponentielle. L'ADN est isolé par la méthode décrite au §(2.2.2.3). Le vecteur est digéré par HincII (site aux extrémités nettes), déphosphorylé, purifié par électrophorèse sur gel d'agarose, et ainsi prêt pour la ligation (Messing *et al.* 1981). Les extrémités des fragments à cloner sont remplies avec l'ADN polymérase I, fragment de Klenov lorsque c'est nécessaire. Les ligations contiennent 25 ng de vecteur et 3 à 6 fois plus d'insertion en terme de molarité. L'ADN ligase du bactériophage T4 est utilisée selon les données de Boehringer (Mannheim) et les ligations incubées à 14°C pendant la nuit dans la plupart des cas et à 0°C, si des insertions de plus de 1000 paires de bases sont attendues. La tranformation au $CaCl_2$ suit le protocole d'Amersham, (1983).

2.2.4 L'identification des recombinants

Les vecteurs que nous utilisons possèdent un ou plusieurs marqueurs, qui permettent l'identification rapide des transformants et de repérer les plasmides ou les phages contenant une insertion de ceux qui n'en ont pas.

2.2.4.1 L'identification des recombinants : Sélection par les antibiotiques

Outre son origine de réplication, le plasmide pBR322 contient les gènes *bla* et *tet*, codant respectivement pour une β-lactamase clivant l'ampicilline et une protéine qui modifie la membrane bactérienne et empêche le transport de la tétracycline dans la cellule. Les sites HinDIII et BamHI interrompant le gène bla, un simple test de croissances sur des milieux contenant l'un de ces antibiotiques permet de déterminer s'il y a eu insertion d'un vecteur (résistance à la tétracycline) et si ce vecteur a intégré un fragment d'ADN étranger (sensibilité à l'ampicilline).

2.2.4.2 L'identification des recombinants : Sélection de par l'activité de la β-galactosidase

Dans le cas des vecteurs M13, un test de complémentation permet de distinguer facilement les plaques des recombinants de celles du phage original. Les phages de type M13 contiennent une partie de l'opéron lactose de *E. coli* comprenant l'opérateur, le promoteur et les 145 acides aminés N-terminaux de la β-galactosidase (α-peptide). La bactérie ne possédant qu'une β-galactosidase comprenant une délétion des acides aminés 11 à 41, ne produit pas un enzyme fonctionnel. Lorsque la β-galactosidase est induite par de l'IPTG (isopropyl-β-thio-galactopyranoside), les enzymes de la bactérie et du phage se complémentent, produisant une β-galactosidase fonctionnelle, capable d'hydrolyser un analogue du lactose, le BCIG (5-bromo-4-chloro-3-indoyl-β-galactoside), en un colorant bleu. Les plaque blanches contiennent donc des phages avec insertions, les plaques bleues appartenant aux phages de type "sauvage".

2.2.5 L'analyse des recombinants

Les méthodes précédentes, nous permettent de savoir qu'il y a eu insertion, mais cela n'est généralement pas suffisant. Nous nous trouvons souvent devant plusieurs centaines de recombinants et seuls ceux contenant un gène particulier nous intéressent.

2.2.5.1 L'analyse des recombinants : Méthode de Hoggness

Cette méthode d'isolation simultanée des ADN d'une centaine de colonies de bactéries ayant poussé sur un filtre nitrocellulose (Grunstein et Hogness, 1975) à l'avantage de permettre de tester rapidement la présence d'un gène par hybridation §(2.2.5.4). Nous avons utilisé une adaptation de la méthode originale (Maniatis *et al.* 1982), qui évite l'achat de l'appareil particulier utilisé par les précédents auteurs pour imbiber leurs filtres nitocellulose, le remplaçant par du papier filtre.

2.2.5.2 L'analyse des recombinants : Mini-lysats

La méthode précédente a le désavantage de ne pas différencier un signal obtenu par hybridation contre l'insertion, du signal obtenu contre le vecteur ou l'ADN bactérien total, il est généralement utile de vérifier les clones qui semblent positifs par la méthode de Birnboim et Doly, (1979). Le procédé consiste en une rapide extraction des plasmides en conditions alcalines, qui est généralement suivie par une analyse des clones par restriction, d'un transfert selon Southern §(2.2.5.3) et d'une hybridation contre une sonde spécifique §(2.2.5.4).

2.2.5.3 L'analyse des recombinants : Méthode de Southern

Nous utilisons la méthode de Southern, (1975), pour la localisation de séquences particulières parmi un certain nombre de fragments de restrictions qui ont été séparés par électrophorèse. L'ADN est dénaturé, transféré sur papier nitrocellulose, puis hybridé.

2.2.5.4 L'analyse des recombinants : Hybridations homologues et hétérologues

La dénaturation d'un ADN marqué radioactivement, suivie d'une renaturation d'un double brin avec des ADN non marqués et simples brins, permet de repérer les molécules hybrides formées. La température de renaturation optimale dépendra du pourcentage en bases guanine + cytosine de l'ADN, de la force ionique du tampon, de la quantité de formamide présente, ainsi que de l'homologie existant entre les deux séquences à hybrider. Pour le détail, se référer au §(2.5), où les relations mathématiques sont intégrées à la partie informatique de ce travail. La préparation des sondes est décrite dans le §(2.2.7).

Tenant compte des données théoriques, nous avons utilisé quatre tampons d'hybridations. En début de thèse, nous avons utilisé les conditions suivantes :

Dans le cas d'hybridations hétérologues, lorsque la sonde provient d'un autre organisme que celui étudié, le tampon contient 15% formamide, 5xSSPE, 5xDenhardt, 200 μg/ml d'ADN de thymus de veau simple brin et 0.1% SDS. L'hybridation est optimale à 37°C et en présence de 100-200 ng/ml de sonde 10^5-10^6 cpm/ml

Dans le cas d'hybridations homologues, lorsque la sonde est spécifique au fragment d'ADN ou d'ARN étudié, le tampon contient 50% formamide, 5xSSPE, 5xDenhardt, 200 μg/ml d'ADN de thymus de veau simple brin et 0.1% SDS. L'hybridation est optimale à 42°C et en présence de 100-200 ng/ml de sonde 10^5-10^6 cpm/ml

Nous avons ensuite préféré les conditions d'hybridations conseillées par la maison NEN (1984), qui donnent de bien meilleurs rapports signal/bruit de fond :

Selon le type d'hybridation, le tampon contient 15% ou 50% formamide, 1M NaCl, 1% SDS, et plus de 100 μg/ml d'ADN de thymus de veau. L'hybridation à lieu dans les mêmes conditions de températures (à 37 ou à 42°C) que précédemment, en présence d'au maximum 10 ng/ml de sonde (10^5 dpm/ml).

Après 5-6 heures de préhybridation (sans la sonde), 16-24 heures d'hybridation, les filtres sont lavés avec agitation constante :

2 x 5 minutes dans 2xSSC,
2 x 30 minutes à 65°C dans 2xSSC, 1% SDS
2 x 30 minutes à température ambiante dans du 0.1xSSC

2.2.5.5 L'analyse des recombinants : Construction d'une carte de restriction

La cartographie détaillée d'un fragment de restriction, est obtenue à partir de plusieurs méthodes. Le choix de la technique à employer dépend de la taille de l'ADN à cartographier.

Nous avons utilisé les digestions simples et multiples avec des enzymes coupant peu, lorsque le fragment à localiser faisait partie de l'ADN chloroplastique total. L'hybridation spécifique d'un fragment face à une série de digestions, nous donne alors la localisation de ce fragment par rapport aux autres. Une série d'expériences, permet de reconstituer une carte complète de la région étudiée.

Pour cartographier des fragments plus courts (10 Kb dans notre cas), nous avons eu recours à la digestion partielle (Smith et Birnstiel, 1976), La méthode consiste à faire le marquage terminal d'une seule des extrémités d'un fragment linéaire d'ADN et d'ajuster les conditions de digestions de façon à obtenir un clivage par molécule. Il suffit ensuite d'observer la taille des sous-fragments produits, pour pouvoir dresser la carte du fragment initial, le fragment le plus court étant le fragment le plus proche de l'extrémité marquée.

Parfois, nous avons procédé à des digestions séquentielles, en éluant par électroélution un fragment dont on désirait une carte plus détaillée : Adaptation de la méthode d'Herrmann *et al.* (1980)

2.2.5.6 L'analyse des recombinants : Sous-clonage

Il est parfois utile de cloner des sous-fragments d'ADN afin d'en faciliter l'analyse. Le clonage des fragments relativement grands (plus grands que 1 Kb) est effectué dans pBR322. Si des fragments plus petits sont attendus, il est possible de les intégrer au hasard dans le bactériophage M13 et de les séquencer. Dans ce cas, il faut rester attentif au fait que si les insertions ont plus de 1 Kb, elles ont tendance à subir des délétions durant la propagation des phages.

2.2.6 Le séquençage selon Sanger

L'analyse la plus fine possible au niveau de l'ADN est le séquençage, qui permet de déterminer la structure chimique de la molécule. Nous avons utilisé la méthode de Sanger *et al.* (1980). Cette méthode se base sur la faculté qu'a le fragment Klenov de l'ADN polymérase I d'incorporer des dideoxynucléotides, lorsqu'il synthétise une copie d'un ADN monocaténaire hybridé à une amorce. Un dideoxyribonucléotide, dont le dideoxyribose est dépourvu de groupe hydroxyle en 3', bloque la synthèse du brin dans lequel il vient d'être incorporé. Quatre réactions sont effectuées en parallèle, chacune en présence de l'un des dideoxynucléotides triphosphates (ddGTP, ddATP, ddCTP, ddTTP) et des quatre deoxyribonucléotides triphospates (dNTP) dont l'un est radioactif (α-P^{32} dATP). Les produits sont séparés par électrophorèse sur gels d'acrylamide dénaturants.

2.2.6.1 Le séquençage selon Sanger : Préparation de la matrice

La préparation des matrices simples brins est très critique pour obtenir des gels de séquences lisibles. Nous avons suivi la méthode de Messing *et al.* (1981) et plus particulièrement le protocole optimalisé d'Amersham, (1983). Les cellules bactériennes sont centrifugées et les phages contenus dans le surnageant sont précipités au PEG (Lis, 1980). Il est critique d'enlever toute trace de PEG. L'ADN est alors extrait au moyen d'une extraction phénol et d'une extraction éther. Il faut éviter à tout prix l'emploi de chloroforme, qui fait obtenir des séquences totalement illisibles. Après précipitation à l'éthanol, une culture d'une nuit de 1.4 ml de bactériophage donne de 100 à 200 ng d'ADN simple brin qui sont resuspendus dans un volume final de 50 μl de tampon Tris-HCl 10 mM, EDTA 0.1 mM, pH 8.

2.2.6.2 Le séquençage selon Sanger : Les solutions de travail

Les solutions stocks de deoxyribonucléotides triphosphates (dNTPs) et de dideoxyribonucléotides triphospates (ddNTPs) sont préparées dans de l'eau distillée, neutralisées à l'aide de Tris base et ajustées à 10 mM. Nous gardons ces solutions à -70°C.

Nous préparons les solutions de dGTP, dCTP, dTTP suivantes dans 5 mM Tris-HCl, 0.1 mM EDTA, pH 7.5. En première approximation, les solutions de travail de ddGTP et ddCTP sont à 0.3 mM, le ddATP à 0.1 mM et le ddGTP à 0.5 mM. La solution de dATP non-radioactif (solution de chasse) est à 0.25 M. Nous donnons dans un tableau, la composition des solutions de mélange pour chaque réaction de séquence.

Sol. STOCK	Mél. dGTP	Mél. dATP	Mél. dCTP	Mél. dTTP
0.5 mM dGTP	1 µl	20 µl	20 µl	20 µl
0.5 mM dCTP	20 µl	1 µl	20 µl	20 µl
0.5 mM dTTP	20 µl	20 µl	1 µl	20 µl
Tampon TE	5 µl	5 µl	5 µl	5 µl
Eau bidist.	454 µl	454 µl	454 µl	435 µl
où TE = 50 mM Tris-HCl, 1 mM EDTA, pH 8				

2.2.6.3 Le séquençage selon Sanger : La réaction de séquence

Nous avons utilisé une simplification des méthodes publiées, en procédant à nos réactions de séquences dans des tubes Eppendorf et non dans des capillaires scellés. Les clones sont d'abord triés en analysant la réaction de séquence en présence de ddT (ddTTP). Généralement, 24 clones sont analysés simultanéments. Par ADN, nous utilisons 1 μCi d'αP^{32} dATP (0.1 μl), 0.3 ng de pentadecamère "sequencing primer" pour l'amorce (0.3 μl), 0.3 μl de tampon 10x Hin DTT (110 mM Tris-HCl, 100 mM $MgCl_2$, 500 mM NaCl, 10 mM DTT, pH 7.4) et 0.8 μl d'eau. A 1.5 μl de ce mélange sont ajoutés 3-6 ng de matrice d'ADN (1.5 μl) qui sont déposés sur le bord du tube et brièvement centrifugé. L'hybridation entre l'amorce et la matrice est effectuée à 55°C pendant 5 minutes, puis 10 minutes à température ambiante. Pour chaque réaction, 2 μl d'un mélange dTTP:ddTTP (1:1) sont ajoutés ainsi que 0.2 UI de Klenov (2 μl). La synthèse du deuxième brin d'ADN est effectuée à température ambiante pendant 15 minute, le dATP de chasse (2 μl) est ajouté, puis la réaction est continuée pendant 15 minutes supplémentaires avant d'être stoppée par l'addition de 4.5 μl de colorant-formamide (3 μg/ml xylène cyanol FF, 3 μg/ml bleu de bromophénol, 20 mM EDTA dans de la formamide désionisée). L'ADN est dénaturé à 100°C pendant 3 minutes et chargé sur un gel dénaturant 6% acrylamide, 8M urée (400 x 200 x 0.3 mm), (Sanger et Coulson, 1978). Nous faisons migrer les gels à 30 mA/gel pendant 2 h30 à 3 h, en surveillant que la température n'excède pas 55°C. Les gels sont fixés pendant 5 minutes dans une solution 20% éthanol, 10% acide acétique), transférés sur papier 3MM, séchés par chauffage et sous vide pendant 2 heures, puis une autoradiographie est faite durant la nuit sur film KODAK XAR-5.

Lorsque les empreintes des T de tous les clones sont analysées, La réaction de séquence proprement dites est faites. Les seules différences consistent en la préparation du tampon d'hybridation et en le changement des rapports dNTP:ddNTP. Pour chaque clone analysé, nous utilisons 5 μCi d'αP^{32} dATP (0.5 μl), 1 ng de pentadecamère "sequencing primer" pour l'amorce (1 μl), 1 μl de tampon 10x Hin DTT et 2.5 μl d'eau. A 10-20 ng de matrice d'ADN (5 μl) sont ajoutés 5 μl de ce mélange. Après hybridation, 2 μl de solution sont ajoutés aux différents mélanges de dNTP:ddNTP.

Dans le tableau suivant, nous indiquons le type de gel utilisé, la duré des électrophorèses et les conditions particulières des mélanges dNTP:ddNTP. Pour lire des séquences de plus de 350 nucléotides, un gel de huit heures est indispensable.

gel acrylamide	durée	dG:ddG	dA:ddA	dC:ddC	dT:ddT
6%	3-6 h	2:1	2:1	4:1	2:1
5%	8 h	3:1	3:1	5:1	3:1

Il est à noter que nous avons parfois utilisé des gels avec gradient de tampon (Biggin *et al.*, 1983), qui ont l'avantage de permettre de supprimer le gel de six heures, lorsque les insertions ne sont pas trop longues. Par contre, il faut que la technique du coulage des gels soit vraiment au point, pour ne pas observer de différences locales de migrations.

2.2.7 Le Marquage des ADN

Différents types de marquages sont utilisés pour fabriquer des sondes et ainsi diminuer la quantité minimale d'ADN ou d'ARN nécessaire pour pouvoir être détecté. Nous avons évité la technique du marquage terminal par la kinase (Maxam et Gilbert, 1977), qui nous demandait l'achat de γP^{32}-dATP, peu utilisé sinon dans le laboratoire.

2.2.7.1 Le Marquage des ADN : Nick-translation

Cette méthode (Rigby *et al.*, 1977) est basée sur l'utilisation de l'ADN polymérase I d'*Escherichia coli*, qui possède 3 activités enzymatiques distinctes (Kelley et Stump, 1979) : Une activité polymérasique (5'→ 3'), une activité exonucléasique (3'→ 5') qui est bloquée par l'activité de polymérisation, et une activité exonucléasique (5'→ 3') qui combinées, permettent la réparation d'un ADN à partir d'un trou ("nick") dans la molécule. Si le mélange réactionnel contient un nucléotide marqué (αP^{32} dATP), alors l'ADN réparé sera une copie radioactive du matériel initial.

Remarques : La qualité des préparations de l'enzyme ayant augmenté, il est maintenant nécessaire d'ajouter 0.6 mUI de DNase I (Worthington, ribonuclease free) par réaction.
Ce type de sonde doit être dénaturé à 100°C pendant 3 minutes, avant d'être utilisé.

2.2.7.2 Le Marquage des ADN :
Remplissage à la Klenow en présence d'αP^{32} dATP

Lorsque les ADN ont été digérés par des enzymes de restriction donnant des extémités 5' protubérantes, il est parfois possible de faire un marquage terminal en remplissant les extrémités à l'aide de l'ADN polymérase I, fragment de Klenov. A cela, deux conditions, l'emploi d'αP^{32} dATP et la possibilité d'intégration de ce dATP dans le site de restriction en question. L'avantage de cette méthode est qu'il est possible de jouer sur les sites de restrictions des deux extrémités du fragment et d'obtenir le marquage terminal d'une seule extrémité.

2.2.7.3 Le Marquage des ADN :
Marquage spécifique d'un seul brin d'ADN

Il existe plusieurs méthodes de fabrication de sondes à partir de la matrice simple brin que représente un clone M13. Hu et Messing, (1982) utilisent une amorce universelle "hybrid probe primer", complémentaire à la région 5' au site de clonage multiple du M13 mp7, pour initier la synthèse d'un ADN marqué complémentaires au séquences du M13 en aval de l'insertion. Nous avons testé l'utilisation de cette sonde partiellement double brin, mais la fixation parfois non spécifique de la sonde diminue l'efficacité de la méthode. Boyer, (1986) rend économique cette technique de sandwich, effectuant une hybridation en deux temps. De l'ADN froid d'un clone M13 contenant une insertion, puis de l'ADN marqué de la forme réplicative de M13 sont séquentiellement hybridés, délivrant une sonde très chaude.

Nous avons mis au point une technique personnelle de marquage, qui si elle donne des sondes moins chaudes, donne néanmoins des sondes nettement plus spécifiques, ne contenant pas de longues séquences simple-brins de M13 qui peuvent hybrider avec les ADN ou les ARN à tester. Nous isolons une sonde marquées en plusieurs points et séparée de la séquence complémentaire, froide, ce qui améliore l'efficacité d'hybridation.

Cette méthode est la suivante :

20-40 ng d'ADN provenant d'un clone M13, contenant une insertion "RNA-like" (9 μl), 10 μCi d'αP^{32} dATP (1 μl), 3 ng de pentadecamère "sequencing primer" pour l'amorce (3 μl) sont ajoutés dans un tampon contenant (11 mM Tris-HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT, pH 7.4, 200 μM dGTP, 200 μM dCTP et 200 μM dTTP). L'amorce est hybridée pendant 5 minutes à 55°C, puis 10 minutes à température ambiante. 1 à 2 UI de Klenov sont ajoutées. Le tube est incubé pendant 20 minutes à température ambiante et lorsque l'insertion est de taille supérieure à 150 nucléotides, 2 μl de solution de dATP de chasse §(2.2.6.2) sont ajoutés avant d'incuber le tube pendant 20 minutes supplémentaires à 37°C. La réaction est stoppée par une incubation de 5 minutes à 65°C et l'ADN est précipité à l'éthanol à -70°C pendant 25 minutes. Après 10 minutes de centrifugation dans une centrifugeuse Eppendorf, les sondes sont resuspendues dans du tampon de digestion EcoRI (M13 mp8) ou HindIII (M13 mp9) et digérées par les enzymes correspondants. Les sondes sont ensuite dénaturées et purifiées sur gels d'acrylamide 4 à 6%, urée 8M et où la bande spécifique est repérée après une autoradiographie de 15 minutes avec écran intensifiant. L'élution du fragment est faite selon Maxam et Gilbert, (1977) et le fragment est finalement resuspendu dans 20 μl de TE.

La séquence complémentaire à notre sonde étant l'ADN complet de M13, nous nous en débarassons lors de la purification du fragment sur gel dénaturant : Ainsi toute molécule compétitrice à l'hybridation souhaitée est éliminée.

2.3 Méthodes de séparation et d'analyse des ARN

Une fois que la molécule d'ADN est décodée, l'étape suivante est d'analyser quels ARN messagers sont transcrits, à quel moment et comment. Pour répondre à ces questions, il faut s'appuyer sur un certain nombre de méthodes relativement délicates, puisque le matériel de base est l'ARN chloroplastique, qui mis à part sa rareté, est sensible au RNAses. Tous le matériel utilisé (verrerie, tampons, etc.) a donc toujours été stérilisé, que se soit à l'étuve à 160°C, à l'autoclave à 120°C ou au DEP (diéthyl pyrocarbonate).

2.3.1 Les méthodes d'analyses

2.3.1.1 Les méthodes d'analyses : Dénaturation des ARN au glyoxal et séparation

Lorsque les échantillons à séparer sont des ARN, nous les dénaturons généralement au glyoxal et les séparons sur gel d'agarose 1.2% dans du tampon 10 mM phosphate pH 6.6, avec circulation continue du tampon, selon la méthode de Mc Master et Carmichael (1977). La calibration des gels d'ARN est obtenue en chargeant sur le gel 2 échantillons d'ARN ribosomique obtenus à partir d'une préparation de ribosomes d'*Escherichia coli* :
L'un est dénaturé au glyoxal et sert de référence de poids moléculaire, l'autre est un contrôle non dénaturé qui permet de se rendre compte si la dénaturation au glyoxal est effective.

2.3.1.2 Les méthodes d'analyse : Northern et électroélution

Les ARN sont transférés sur papier nitrocellulose par Northern (Thomas, 1980) ou sur nylon (GeneScreen™) par électroélution (NEN, 1984). Nous avons rapidement préféré la seconde méthode, qui permet des hybridation avec un bruit de fond minime.

2.3.2 séquençage de l'ARN

Plusieurs moyens peuvent être utilisés pour obtenir la séquence d'un ARN :
Si l'ARN est de faible taille et abondant, le séquençage directe de l'ARN par méthode chimique est possible, après marquage terminal des molécules (Peattie, 1979), (Waldmann *et al.*, 1987). Nous avons écarté cette méthode car les messagers n'étaient pas assez abondants.

Le séquençage d'un cDNA cloné est une méthode de choix lorsque une copie de l'ARN messager a été obtenue par l'emploi de la trancriptase inverse (Maniatis *et al.*, 1982). Nous aurions pu choisir de monter une banque de cDNA, mais l'absence de queue de poly A chez les messagers chloroplastiques, et les manipulations relativement compliquées et longues rendaient assez aléatoire l'obtention d'un résultat dans le temps imparti.

Nous avons donc choisi une méthode de séquençage directe de l'ARN utilisant l'hybridation d'une amorce radioactive sur le messager, puis la fabrication d'un cDNA à partir de cette amorce, cela en présence de dideoxynucléotides, selon le même principe que pour le séquençage de l'ADN selon Sanger :

2.3.2.1 Le séquençage par "primer extension" Méthode de Goldenberg et Hauser, (1983).

8 μl d'amorce simple brin préparée selon notre méthode personnelle du §(2.2.7.3), mais obtenue par digestion du site de restriction unique et interne à l'insertion (SnaBI) ou digéré par RsaI, auquel on ajoute 13.2 μl d'eau bidistillée est dénaturée pendant 2 minutes à 100°C. Ensuite : 10 μg d'ARN chloroplastique et 5.6 μl de tampon 10xRT (500 mM Tris-HCl, 20 mM MgCl2, 840 mM KCl, 10 mM DTT, pH 8.3 pH ajusté à 37°C) sont ajoutés. L'hybridation est faites 5 minutes à 65°C, puis le bain est réglé sur 55°C et laissé à refroidir pendant 55 minutes. Une incubation de 90 minutes à 37°C, puis de 30 minutes à 25°C, termine cette étape.

La réaction avec la transcriptase inverse est initiée dans du tampon 1xRT auquel est ajouté un supplément de 20 mM KCl, les dNTP et les ddNTP sont ajoutés selon (Goldenberg et Hauser, 1983). Chaque réaction est faite en présence de 6.5 UI d'enzyme, d'abord 5 minutes à 25°C puis, 15 minutes à 37°C.
La solution est amenée à une concentration de 125 μM dNTP, puis incubée pendant 15 minutes à 37°C. La réaction est arrêtée en ajoutant 10 mM EDTA. Les ARN sont hydrolysés pendant 3 minutes à 100°C en présence de 100 mM NaOH. Le pH est neutralisé par l'addition d'une même quantité d'HCl et les cDNA sont précipités avec 3 volumes d'isopropanol, en présence de 500 mM acétate d'amonium et de 1 ug/ul d'ARN-t, pendant 20 minutes à -80°C. Après centrifugation, le culot est lavé avec de l'éthanol 100%, resuspendu dans 2 μl de tampon d'électrophorèse (TBE) 2x et 2 μl de solution colorant-formamide (3 μg/ml xylène cyanol FF, 3 μg/ml bleu de bromophénol, 20 mM EDTA dans de la formamide désionisée). Le tout est ensuite chargé sur gel de séquence et manipulé comme tel, §(2.2.6.3). Pour obtenir un signal suffisant, le gel est exposé pendant une semaine à -70°C, avec un écran intensifiant.

2.4 Matériel, enzymes et solutions particulières

2.4.1 Le Matériel particulier

- Les écrans intensifiants "Lightning-Plus" proviennent de Cronex, Du Pont.

- Le broyeur à lames de scalpels est un broyeur standard, pour lequel les lames classiques ont été remplacées par 2 lames de scalpel tournant horizontalement et 2 lames faisant un angle de 45° ayant été renforcées par des lames cassées.

2.4.2 Les enzymes

- Les enzymes de restriction :

AluI, BamHI, BglII, ClaI, EcoRI, HaeIII, HinCII, HinDIII, HinFI, HpaI, HpaII, KpnI, PstI, PvuI, RsaI, SmaI, SacI, SalI, Sau3AI, SnaBI, TaqI et XhoI ont été utilisés au cours de ce travail. Tous proviennent de Boehringer Mannheim ou lorsqu'ils n'étaient pas au catalogue, de Biolabs New England.

- La T4 DNA ligase,
- La DNA polymérase I sans endonucléases,
- La DNA polymérase I, fragment de Klenov,
  proviennent de Boehringer Mannheim.

- La DNAse est garantie sans RNase par Worthington

- La reverse transcriptase a été achetée chez Stehlin, puis suite à des problèmes de fiabilité, chez Boehringer Mannheim.

- La RNase A (de pancréas de boeuf, Sigma) est traitée dix minutes à 100°C, ce qui élimine tout risque de contamination par quelqu'enzyme que ce soit.

2.4.3 Le matériel radioactif

l'αP^{32} dATP provient de Amersham

2.4.4 Les tampons

Denhardt (1x) : 0.2% BSA (sérum albumine de boeuf)
0.2% PVP (polyvinylpyrrolidon)
0.2% Ficoll

SSC (1x) : 150 mM NaCl
15 mM citrate de sodium

SSPE (1x) : 180 mM NaCl
1 mM EDTA, pH 7.7
10 mM Tampon phosphate de sodium, pH 7.7

TA (1x) : 400 mM Tris
200 mM acétate de sodium
10 mM EDTA
pH 7.8

TBE (1x) : 130 mM Tris
90 mM acide borique
25 mM EDTA
pH 8.8

2.4.5 Les produits chimiques

La formamide : est toujours utilisée sous forme désionisée. Nous désionisons la formamide en lui additionnant une résine échangeuse d'ion (Amberlite MB1, Serva, Heidelberg), puis en la filtrant. Nous stoquons la formamide désionnisée à -20°C

Le glyoxal : subit le même traitement que la formamide. Les filtrations doivent être particulièrement soignées. Le glyoxal est fractionné en très petites quantités (env. 20 ul) et stoqué à -70°C. Les tubes dégelés et ouvert une fois sont jetés.

L'urée : utilisée pour les gels de séquence doit être pure. L'urée Merk est de pureté suffisante. L'urée ultra pure de Schwarz Mann Inc., Spring Valley est un réel luxe (trop chère), mais d'excellente qualité.

2.5 L'outil informatique :

Pour faciliter le travail d'analyse des données récoltées au laboratoire, ou simplement comparer celles-ci avec celles déjà cataloguées dans une base de données telle que l'EMBL (European Molecular Biology Laboratory), il est maintenant impensable de travailler sans l'informatique.

Une partie de notre travail à consisté à mettre au point des utilitaires simplifiant l'analyse des séquences d'ADN et de protéines. Nous avons travaillé alternativement sur des ordinateurs VAX 785 et VAX 750. Ces ordinateurs ont l'avantage d'être construits de manière à ce que l'utilisateur n'ait aucune limitation en ce qui concerne la place en mémoire centrale. L'ordinateur gère lui-même sa mémoire centrale.

2.5.1 Le BASIC structuré de VAX/VMS

Nous avons choisi de programmer en langage BASIC pour les raisons suivantes :

Le langage BASIC dans la version 2.3 (Digital, 1982) implantée sur VAX/VMS est un langage structuré, c'est à dire où il n'est pas nécessaire de numéroter chaque ligne, un langage qui est beaucoup plus lisible que le BASIC standard et où le nombre des embranchements (ordres GOTO) peut être limité au strict minimum. De plus, le BASIC à l'avantage de nécessiter un minimum de déclaration de variables et d'être relativement plus souple à l'utilisation que d'autres langages. Nous avons aussi profité du fait qu'il est possible avec VAX d'importer des ordres de différents langages, et dans notre cas, particulièrement à partir du FORTRAN. Nous avons écarté le FORTRAN pour l'écriture du programme prncipal à cause de ses déclarations et PASCAL parce que si ce langage est très bien adapté aux mathématiques, il est très mal-pratique en ce qui concerne le traitement des chaines de caractères, qui dans notre cas est le problème principal. Le BASIC implanté sur VAX/VMS est compilé, ce qui le rend performant (environ 1'000'000 de boucles contenant une ou deux opérations simples sont effectuées par secondes). Enfin, nous avons adapté un programme à nos besoin, plutôt que de rester dans les limites imposées par l'utilisation de programmes déjà établis.

2.5.2 Le programme ANALYSE

Le programme ANALYSE contient en fait quatre programmes qui fonctionnent avec plusieurs types de fichiers d'entrées. Les fichiers d'ADN peuvent être selon le format d'Heidelberg (EMBL) ou de type quelconque, ne contenant que la séquence sans aucun commentaire (Staden, 1980).

Sous-programme ANALYSE

Celui-ci est à peu près équivalent aux programmes :

SEQLST, TRANSQ, SEARCH	Staden, (1977)
TRANDK, TRNTRP	Staden, (1978)
TRANMT, DIAGON	Staden, (1980)

avec des possibilités d'impressions beaucoup plus étendues, et la compatibilité aux fichiers de l'EMBL, cela même si le fichier n'est pas mémorisable dans une seule variable VAX (>65'535 caractères). Nous avons des options supplémentaires très utiles comme la recherche de "N-track" à l'écran, la transcription automatique des séquences codantes répertoriées dans les fichiers d'ADN de l'EMBL et la fabrication de fichiers de protéines du type standardisé de l'EMBL, utilisables dans le programme HEID mis au point par nous, ou dans le programme PRPLOT décrit par Hopp et Woods, (1981).

Les fichiers d'enzymes de restrictions

peuvent être du type de Staden (1977), ou des deux types utilisés à l'EMBL. Les recherches de séquences sont possible en utilisant le code préconisé par Heidelberg (EMBL), code particulièrement pratique et dont l'emploi a tendance à se généraliser :

R : PURINE (A,G)	Y : PYRIMIDINE (C,T)	S : C ou G
M : A ou C	W : A ou T	K : G ou T
D : A ou G ou T	H : A ou C ou T	V : A ou C ou G
B : C ou G ou T	N : A ou C ou G ou T	

auquel est ajouté le code de Staden (1977) :

- : A ou C ou G ou T

Nous avons délibéremment évité d'utiliser le code utilisé par Staden (1979) lorsque la séquence est incertaine et qu'un certain nombre d'insertion-délétion est probable. Il est préférable d'utiliser le code présenté ici ou de faire des recherches dans le fichier en utilisant l'option "N-track" du programme, qui permet de repérer une séquence spécifique correspondant à une réaction de séquence particulière (celle du T par exemple), donc en ne connaissant que la situation des T.

Sous-programme Calcul_Tm :

Le calcul des températures de dénaturation et de renaturation des ADN en fonction des conditions ioniques est facilité, par l'emploi de différentes formules dispersées dans la littérature et traitant de ce sujet. Ce programme est particulièrement pratique lors du choix des conditions de lavage d'un hybride.

Nous utilisons les formules suivantes :

(1) $Tm = 0.41 \cdot \%(G+C) + 69.37 + 13.3 \cdot \log (X)$
si $0.5 \leq X \leq 2$

$Tm = 0.41 \cdot \%(G+C) + 69.37 + 15.7 \cdot \log (X)$
si $0.125 \leq X < 0.5$

où Tm représente la température de dénaturation de l'ADN
X exprime la concentration en SSC

(Gillis *et al.*, 1970)

(2) $Tor = 1.24 \cdot Tm - 38.8$

où Tor est la température de renaturation optimale

(Gillis *et al.*, 1970)

(3) 1% formamide réduit le Tm de 0.72°C

(Mc Conaughy *et al.*, 1969)

(4) L'augmentation de 1% du nombre de missmatch entre paires de bases descend le Tm du duplex de 1°C

(Bonner *et al.*, 1973)

(5) $(Tm)_{\mu 2} - (Tm)_{\mu 1} = 18.5 \cdot \log (\mu 2/\mu 1)$

ou $\mu 1$ et $\mu 2$ sont les forces ioniques respectives des deux solutions.

(Schleif et Wensink, 1981)

(6) $\mu = 1/2 \cdot \sum Ci \cdot Zi^2$

ou μ represente la force ionique,
Ci la concentration de chaque ion, Zi sa charge

(Weil, 1979)

Sous-programme Calcul_GC :

Basé sur l'observation de Staden et Mc Lachlan, (1982), qui ont remarqué la différence de structure existant entre les régions codantes et non codantes de l'ADN, cette partie du programme permet de chercher pour une séquence donnée, le pourcentage en chacun des nucléotides, et ceci en fonction de la position dans le codon. Le programme permet donc d'évaluer si une ORF est probablement transcrite et traduite ou non.

Cette option est également utilisable pour analyser, par exemple la G-périodicité d'un gène de manière simple (Trifonov, 1987).

Sous-programme "Diagon"

Nous avons mis au point un programme permettant de comparer graphiquement deux séquences de protéines. Le problème posé par un tel programme est qu'il faut l'adapter au système spécifique utilisé. Ici, nous avons dû jongler en fonction du fait que l'imprimante matricielle graphique PRISM 80 travaille avec des informations codées sur 8 bits alors que le terminal que nous possédons (CIT 101, Captronix), ne peut travailler que sur 7 bits. Il s'en est suivi un certain nombre de problèmes de liaisons, que nous avons règlé au mieux.

Vu que nous travaillons avec une matrice de $4 \cdot 10^6$ caractères, il nous a fallu optimaliser le programme au maximum, en travaillant directement au niveau des bits et non des variables, (stockage de 7 points graphiques par caractère, au lieu d'un point par caractère). Nous avons néamoins prévu une sortie de secours du programme, qui crée un fichier où "0" correspond à l'absence de point et "1" à la présence d'un point. Cela permet d'adapter facilement un programme ou d'imprimer les résultats sur n'importe quel type d'imprimante en ne demandant qu'une programmation minimale.

2.5.3 Le programme HEID

Le programme est prévu pour analyser une séquence protéique particulière et la comparer face à toutes les séquences déjà publiées et stoquées dans les fichiers de protéines de l'EMBL (European Molecular Biology Laboratory). Il est également possible de chercher des séquences consensus (sites actifs) ou des domaines communs à plusieurs protéines. La recherche permet d'autoriser une insertion dont la taille variable est déterminée lors de l'analyse. Il est possible de trier les séquences en fonction du pourcentage d'homologie existant entre les protéines comparées dans la zone contiguë au domaine analysé.

Il est également possible de faire l'analyse en utilisant les équivalences suivante :

Table 1	Lys/Arg, Ser/Thr, Asp/Glu, Val/Ile/Leu
Table 2	Cys Lys/Arg/His Ala/Gly/Pro/Ser/Thr Val/Ile/Leu/Met Phe/Trp/Tyr Asp/Asn/Glu/Gln

3. RESULTATS ET DISCUSSION

Après cette présentation des quelques techniques de biologie moléculaire que nous avons largement utilisées au cours de ce travail, nous nous intéresserons au cheminement parcouru et aux résultats obtenus. Nous verrons comment la recherche d'un gène fantôme nous a aiguillé sur le passionnant problème des gènes ribosomiques.

3.1 Identification et clonage des gènes *rps7* et *rps12*

3.1.1 La localisation des gènes

(La carte actuelle de l'ADN chloroplastique de soja est présentée dans la Figure 8, page 39)

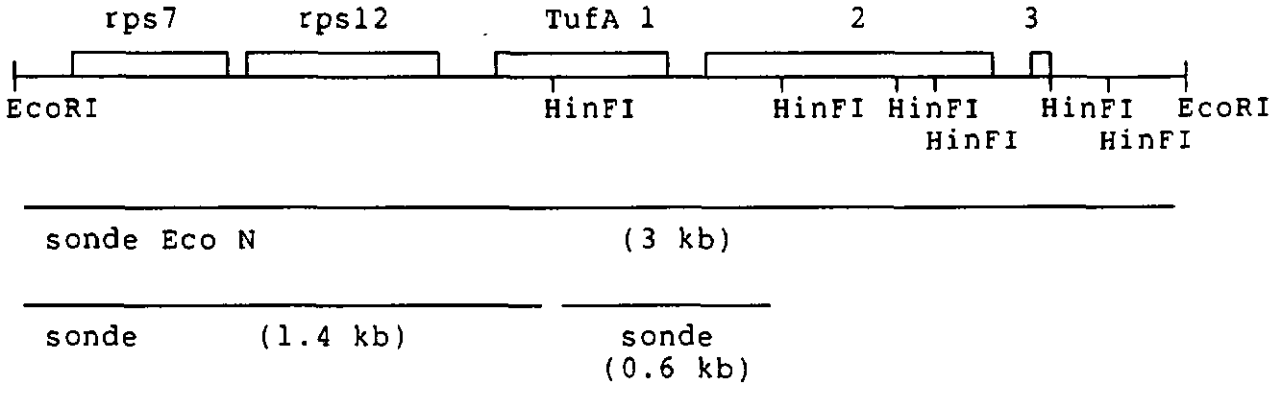
Spielmann (1983) a localisé un fragment faisant partie des régions répétées inverses de l'ADN chloroplastique de soja. Ce fragment BamHI de 1.2 kb, coupé par l'enzyme PstI, chevauche les fragments Pst A : Pst B (IRA), et Pst A : Pst D (IRB). Il pensait alors que le fragment contenait le gène du facteur d'élongation Tu car la sonde qu'il a utilisée contenait ce gène.

Nous avons repris ce travail, en utilisant plusieurs sondes spécifiques internes au fragment EcoN de l'ADN chloroplastique d'*Euglena gracilis* (Montandon et Stutz, 1983). *Euglena gracilis* est un organisme unicellulaire, flagellé, capable de vivre en hétérotrophie. Lorsque une source lumineuse est disponible, cette organisme rejoint le phytoplancton en subissant le verdissement, différenciant des chloroplastes qui lui permettent de vivre en autotrophie. La sonde hétérologue que nous avons utilisée, (le fragment Eco N complet) contient le gène *TufA* codant pour le facteur d'élongation Tu et hybride effectivement avec l'ADN-ct de soja, témoignant de séquences communes à l'Euglène et au soja.

Mais ce gène, que nous désirions en effet localiser, ne se situe pas dans l'ADN chloroplastique des plantes supérieures, comme nous l'avons remarqué dans l'introduction. A titre de vérification, nous avons effectué des hybridations entre l'ADNct de soja et des sous-fragments de notre sonde initiale. Les résultats nous ont permis de voir que le facteur d'élongation Tu n'est pas codé au niveau du chloroplaste.

Pour des raisons didactiques, nous présentons la carte actuelle du fragment d'*Euglena gracilis* et nous nous rappellerons que seuls les 3 exons du *TufA* étaient séquencés et connus au début de ce travail.

Figure 1a : Carte du fragment Eco N d'*Euglena gracilis*



Nous avons ainsi utilisé de l'ADN chloroplastique de soja digéré par différents enzymes de restriction, transféré selon Southern et utilisé les 3 sondes d'Euglène que nous avons préparées par élution d'un gel d'agarose, puis marquées par la méthode de "nick-translation".

Le fragment Eco N et le fragment EcoRI-HinFI de 1.4 kb hybrident dans les conditions d'hybridation hétérologues décrite au §(2.2.5.4), alors que le fragment HinFI de 0.6 kb ne donne aucun signal. Nous avions choisi ce dernier fragment car il contient la partie du gène qui code pour le site de fixation du GTP du facteur d'élongation Tu, partie de la molécule qui est bien conservée entre *E. coli* et l'Euglène pour des raisons de fonctionnalité. Nous pouvons négliger l'influence des 102 nucléotides d'intron qui interrompent le gène, car son faible pourcentage en G+C (14.5%) ne permet pas la formation d'hybrides assez stables pour résister aux conditions de lavage.

Ce résultat permettait de prédire que la région en 5' du *TufA* d'*Euglena gracilis* et le fragment BamHI de 1.2 kb de l'ADN chloroplastique de soja contenait en commun un gène intéressant.

Figure 1b : Hybridation de l'ADN chloroplastique de soja en fonction de la sonde d'Euglène utilisée.

Sonde	Hybridation
Eco N	+
1.4 Kb	+
0.6 kb	-

3.1.2 Le clonage des fragment d'ADNct intéressants

Connaissant la taille des fragments de l'ADN chloroplastique qui nous intéressent (Spielmann, 1983), nous avons isolé 50 μg d'ADNct que nous avons digéré, séparé sur gel d'agarose à bas point de fusion, puis nous avons récupéré le fragment BamHI de 1.2 kb ainsi que le fragment HinDIII de 9.0 kb que nous avons insérés dans pBR322. Nous avons transformé des bactéries *Escherichia coli* de souche HB101, puis découvert et analysé les colonies ayant intégré un plasmide recombinant.

Nous avons nommé les clones BamHI : BH101, BH102, ... , alors que les clones HinDIII sont désignés : HD901, ... , HD906. Les clones que nous avons finalement séquencés sont BH101 et HD906, que nous avons du séquencer parce que les gènes étudiés (rps12 et rps7) n'étaient pas totalement inclus dans le fragment BH101. Plus tard, nous avons sous-cloné les plus grands fragments BamHI internes à HD906 dans pBR322 et les avons nommés BH096 (d'environ 1 kb) et BH406 (d'environ 4 kb). Les sous-fragmemts BamHI de faible tailles (< 1kb) ont été directement clonés dans M13 et séquencés.

Pour mémoire, la nomenclature des fragments est indiquée sur la Figure 5, page 33

3.1.3 Vérification que les deux IR contiennent une même information

Pour savoir si nous avions le droit de ne séquencer qu'un seul fragment BamHI (et un seul fragment HinDIII), alors que l'ADN chloroplastique en contient 2 copies, une dans chacune des régions répétées inverses (IR), il est indispensable de vérifier que l'information contenue dans le fragment BH101, se retrouve dans l'autre IR. Cette vérification est d'autant plus importante que l'un des gènes codé (le *rps12*) contient des introns et qu'il aurait été pensable que nous trouvions un gène non interrompu sur la région non séquencée.

A cette fin, nous avons effectué une hybridation du fragment HD906 marqué par nick-translation contre plusieurs profils de restriction. Nous avons utilisé un ancien filtre contenant de l'ADNct de soja digéré par divers enzymes de restriction et transféré par Spielmann (1983). L'hybridation homologue a été effectuée en utilisant de l'ADN du fragment HD906 que nous avons marqué par "nick translation" (Figure 2).

Figure 2 : Localisation du fragment HD906 contenant le *3'-rps12* et vérification que les deux régions répétées inverses de l'ADN chloroplastique de soja sont similaires.

A : Digestion BamHI
B : Digestion BamHI + PstI
C : Digestion PstI
D : Digestion HinDIII + PstI

Présentation de l'ADN digéré et coloré à l'EtBr en face du résultat de l'hybridation avec le fragment HD906. Les flèches indiquent les bandes ayant hybridé lorsque la digestion est double et que les fragments ne sont pas répertoriés. Les lettres indiques le nom d'un fragment cartographié. (ex : BamHI V correspond au clone BH101 de 1.2 Kb)

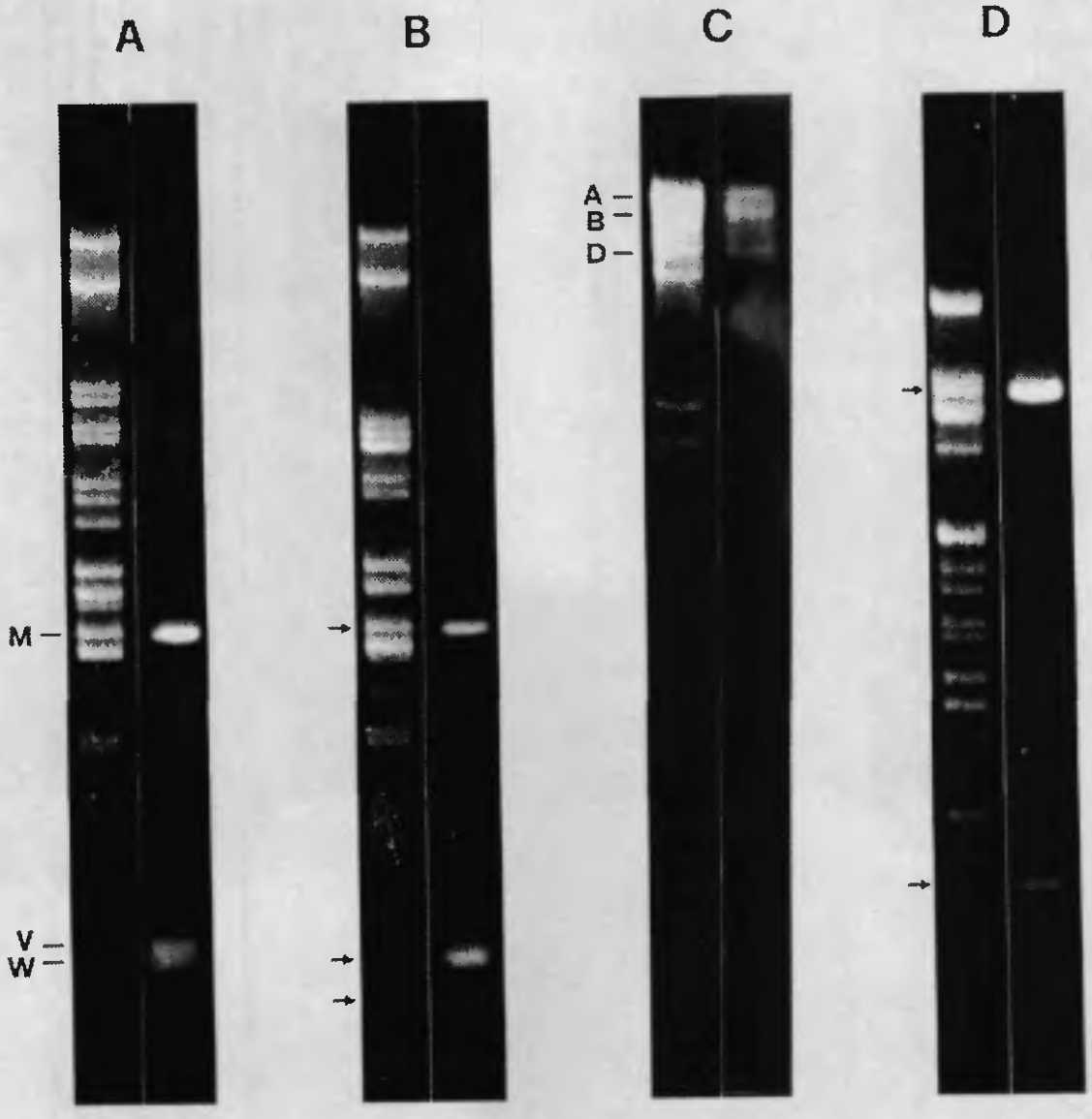
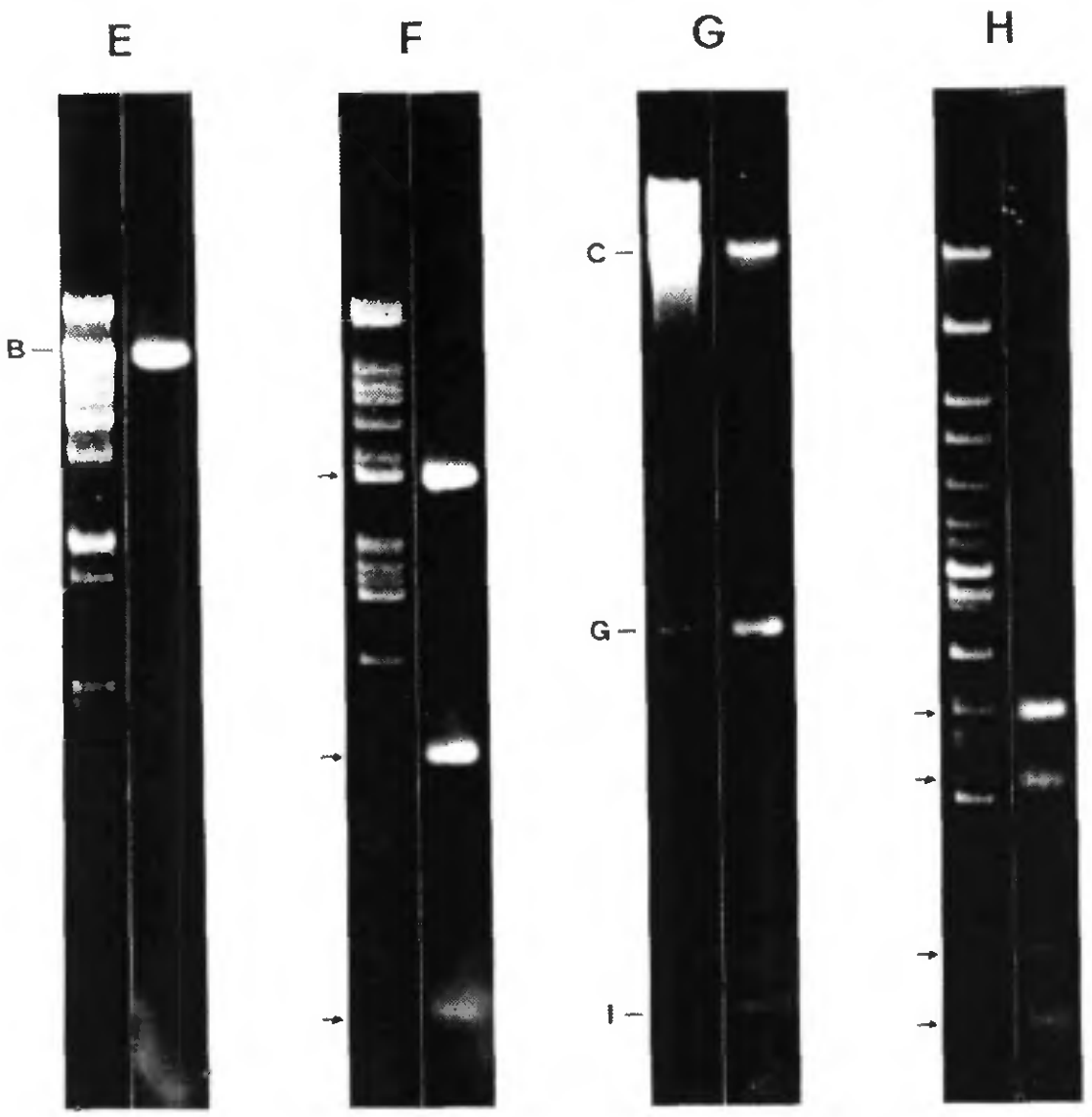


Figure 2 (suite) E : Digestion HinDIII
F : Digestion HinDIII + SmaI
G : Digestion SmaI
H : Digestion SmaI + BamHI



Les résultats de l'hybridation nous permettent de placer le fragment HinDIII sur la carte de restriction de l'ADNct. Nous observons que le fragment HinDIII est le fragment HinD B et qu'il hybride avec les fragments BamHI M, V, W et faiblement avec BamHI J. Le fragment est localisé dans les régions répétées inverses et est plus précisément placé à l'intersection des fragments PstI A et B, D. La digestion par SmaI permet de localiser un fragment SmaI I que nous avons ajouté à la carte de l'ADN chloroplastique entre les fragments SmaI C et G.

Nous pouvons ainsi vérifier la localisation du fragment cloné (HD906) et observer qu'il n'existe pas de différences de tailles significatives entre les deux IR. Seuls des changements ponctuels peuvent éventuellement différencier ces deux régions.

Nous remarquerons également que lorsque nous avons séquencé le clone HD906 (9 kb) et qui contient un fragment similaire au clone BH101 (1.2 kb) provenant d'une expérience de clonage différente, nous n'avons trouvé aucune différence de structure entre les deux fragments. Nous conviendrons pourtant volontiers que nous avions une chance sur deux que les deux fragments considérés proviennent de la même région répétée inverse.

3.1.4 Identification incontestable des gènes *rps7* et *rps12*

Parallèlement à Montandon et Stutz (1984), qui séquençaient la partie en 5' du gène *TufA*, nous avons séquencé le fragment BH101 qui devait contenir des séquences communes. Nous avons comparé les polypeptides suceptibles d'être codés par ces fragments avec les séquences en 5' du *TufA* qui font parties de l'opéron *str* d'*Escherichia coli* (Post et Nomura, 1980). Il était en effet possible qu'il existe une organisation à peu près similaire des 3 génomes. Nous avons aussi cherché dans la base de donnée des protéines (PIR-NBRF) de l'EMBL (European Molecular Biology Laboratory) au moyen du programme HEID que nous avions déjà mis au point, mais dans une version moins complète que la version actuelle.

Figure 3 : Repérage dans la base de donnée de l'EMBL de la protéine S7 (séquence complète).

Comparaison de la protéine chloroplastique S7 de soja avec les fichiers de la base de donnée de l'EMBL :

NBRF Code	Gène	Nbre	a.a	Séquences trouvées							
R3EC1	30S ribosomal protein S1 - Escherichia coli	3	557	LVDA	AAKG	MAEA					
	Position dans la protéine de référence :			37	553	547					
	Position dans notre protéine :			124	127	144					
R3EC7K	30S ribosomal protein S7 - Escherichia coli	12	177	GKKS	RRVG	RVGG	VGGS	GGST	QVPV	VPVE	HRMA
	Position dans la protéine de référence :			33	77	78	79	80	85	86	140
	Position dans notre protéine :			34	78	79	80	81	86	87	142
R3EG7	Ribosomal protein S7 - Euglena gracilis chloroplast	5	155	DPIY	QVPV	VPVE	KKEE	AEAN			
	Position dans la protéine de référence :			14	85	86	136	144			
	Position dans notre protéine :			15	86	87	137	145			

La même protéine (S7), lorsqu'elle est comparée aujourd'hui avec les protéines codée par l'ADNct de tabac

Code	Gene	Nbre	a.a	Séquences trouvées							
100004	B : 100004 99537 (C) ribosomal protein S7 (rps7)	133	155	MSRR	SRRG	RRGT	RGTA	GTAE	KTAK	TAKS	AKSD
	Position dans la protéine de référence :			1	2	3	4	5	10	11	12
	Position dans notre protéine :			1	2	3	4	5	10	11	12
142525	A : 142525 142992 ribosomal protein S7 (rps7)	133	155	MSRR	SRRG	RRGT	RGTA	GTAE	KTAK	TAKS	AKSD
	Position dans la protéine de référence :			1	2	3	4	5	10	11	12
	Position dans notre protéine :			1	2	3	4	5	10	11	12

Légende : 1°) La protéine tracée est la S7. Une bonne similarité existe avec les protéines homologues de *E.coli* et d'Euglène. Remarquons également que la taille des différentes protéines en acides aminés est un critère de sélection (Indiqué sous : a.a). Le nombre de blocs similaires pour les 2 protéines comparées est indiqué sous : Nbre. Seuls les 8 premiers sont explicitement présentés.

2°) Nous trouvons 3 blocs de 4 acides aminés identiques entre la protéine S1 de *E.coli* et la S7 de soja soja, mais la position des différents blocs n'est pas conservée. Nous pouvons donc écarter l'hypothèse que la protéine inconnue de soja soit la S1, même sans regarder la taille des deux protéines.

3°) La protéine S7 est trouvée 2x dans le fichier des protéines codées au niveau du chloroplaste de tabac et les gènes sont très similaires à ceux du soja. (Le Fichier a été construit par le programme ANALYSE et le code correspond au 1er nucléotide codant de la séquence).

4°) En faisant une analyse semblable avec la protéine S12, nous trouvons une similarité plus élevée qu' entre les différentes protéines S7. Nous ne présentons donc pas ces résultats ici, d'autant plus que pour l'analyse de la protéine complète, il faut tenir compte de la place des introns qui interrompent le gène.

Ayant de bonnes évidences que le fragment BH101 contient les gènes *rps7* et *rps12*, nous avons été étonné de trouver que la séquence du *rps12* était interrompue par des introns. C'était d'autant plus déroutant que le groupe travaillant parallèlement sur l'Euglène, n'observait pas d'introns dans le gène *rps12* et que les seuls gènes chloroplastiques codant pour des protéines et connus pour contenir des introns à ce moment là étaient le *rbcL* (Stiegler *et al.* 1982) et le *TufA* (Montandon et Stutz, 1983).

Nous étions les premiers a avoir localisé et séquencé l'unité *3'-rps12* et le gène *rps7* chez une plante supérieure (von Allmen et Stutz 1985).

Nous nous sommes vite rendu à l'évidence que le *5'-rps12* ne se trouvait pas dans le fragment BH101 et qu'il fallait le chercher en amont de la région séquencée ... Ce qui était un réflexe logique en pensant à la définition du gène :

le gène est une unité biologique de l'hérédité

Avec les connaissances actuelles, il est compréhensible que l'unité en question puisse être l'ADN (Avery *et al.*, 1944; Hershey et Chase, 1952). Depuis les travaux sur l'adenovirus (Chow *et al.* 1977; Berget *et al.* 1977), il est facile d'imaginer que l'ARNm soit synthétisé sous la forme d'un précurseur dont la taille est plus grande que le messager final.

le gène est parfois un gène en mosaïque

L'épissage (="splicing") est alors une étape de maturation de certains ARN, dont le but est de joindre des parties distantes de la molécule du précurseur (exons) en éliminant les parties intermédiaires non codantes (introns). Dans ce cas, il existe toujours une unité de transcription qui schématiquement produit un ARN complet, même s'il s'agit d'un précurseur.

Nous avons donc tout d'abord logiquement cherché la partie *5'-rps12* dans le fragmemt HinDIII que nous avions également cloné, d'autant plus qu'aucun autre signal d'hybridation n'avait été observé en utilisant la sonde d'Euglène.

3.1.5 Cartographie classique du fragment HindIII de 9 kb (clone HD906)

L'expérience présentée au §(3.1.3) et qui donne déjà une bonne indication de l'emplacement du fragment HinDIII contenant la partie *3'-rps12*, a été complétée par le placement des sites BamHI internes au fragment.

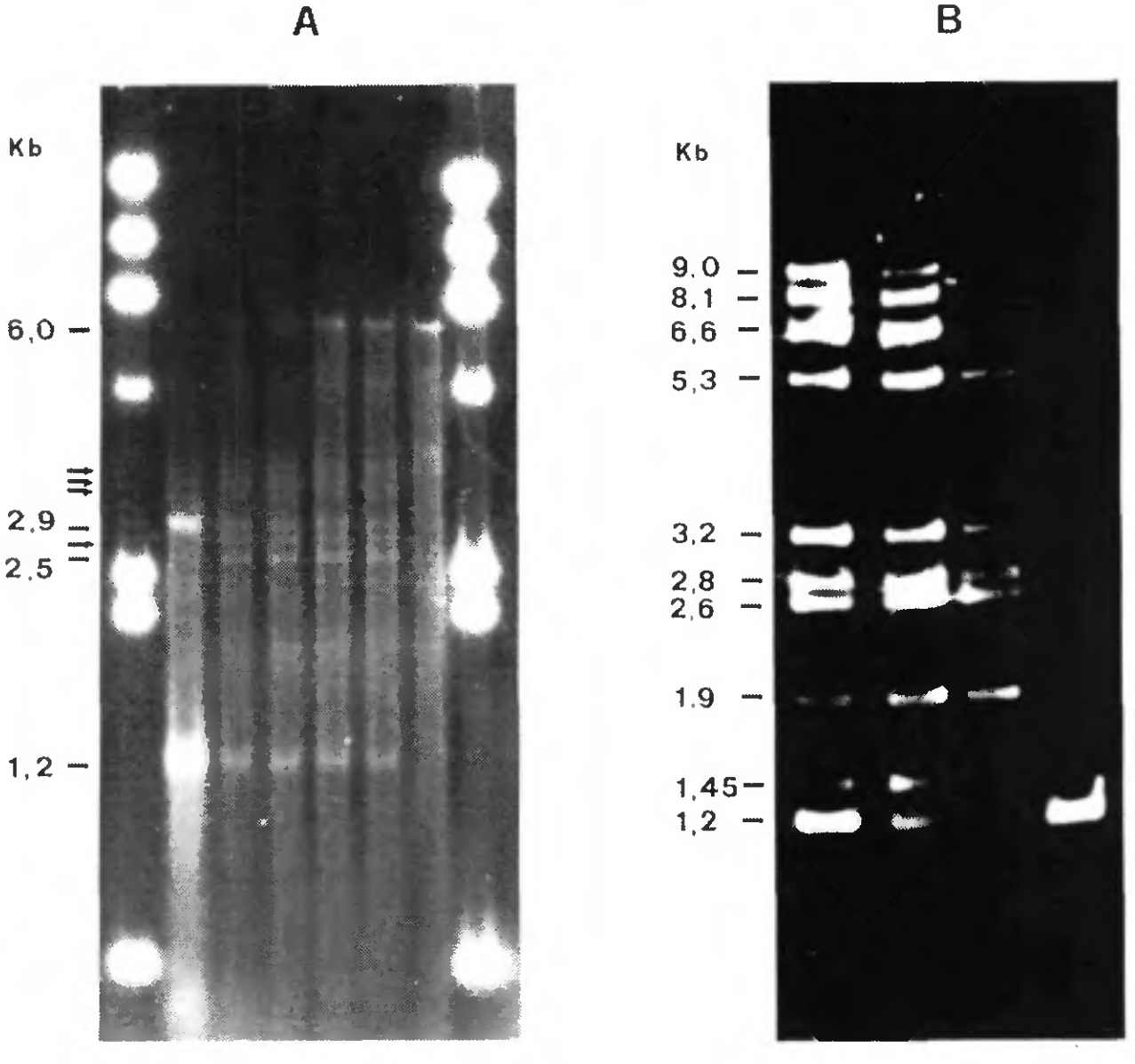
Nous avons localisé les sous-fragments BamHI internes au fragment HinDIII en éluant le sous-fragment HindIII-SmaI (6.0 Kb), en le marquant au moyen d'un remplissage des extémités à l'ADN polymérase I (fragment de Klenov) en présence d'αP^{32} dATP. Nous avons ensuite digéré partiellement le fragment à l'aide d'une unité d'enzyme BamHI pendant des temps variant de 1 minute à 2 heures. Les échantillons ont été séparés sur gel d'agarose 1.2% et autoradiographiés (Figure 4 A).

Nous présentons également la figure d'un gel contenant des digestions partielles du fragment HinDIII par BamHI, transféré puis hybridé contre le fragment BH101, contenant la partie *3'-rps12* du gène (Figure 4 B).

Figure 4 : Digestions partielles BamHI du fragment HinDIII
Etablissement de la carte BamHI

A : Digestions partielles BamHI du fragment HinDIII-SmaI marqué aux extrémités. Dans l'ordre de gauche à droite : λ/HindIII, digestions de 60' à 1', λ/HinDIII

B : Digestions partielles BamHI du fragment HindIII-SmaI hybridé contre BH101 (sonde contenant le *3'-rps12*). Dans l'ordre : Digestions partielles de 5' à 60'

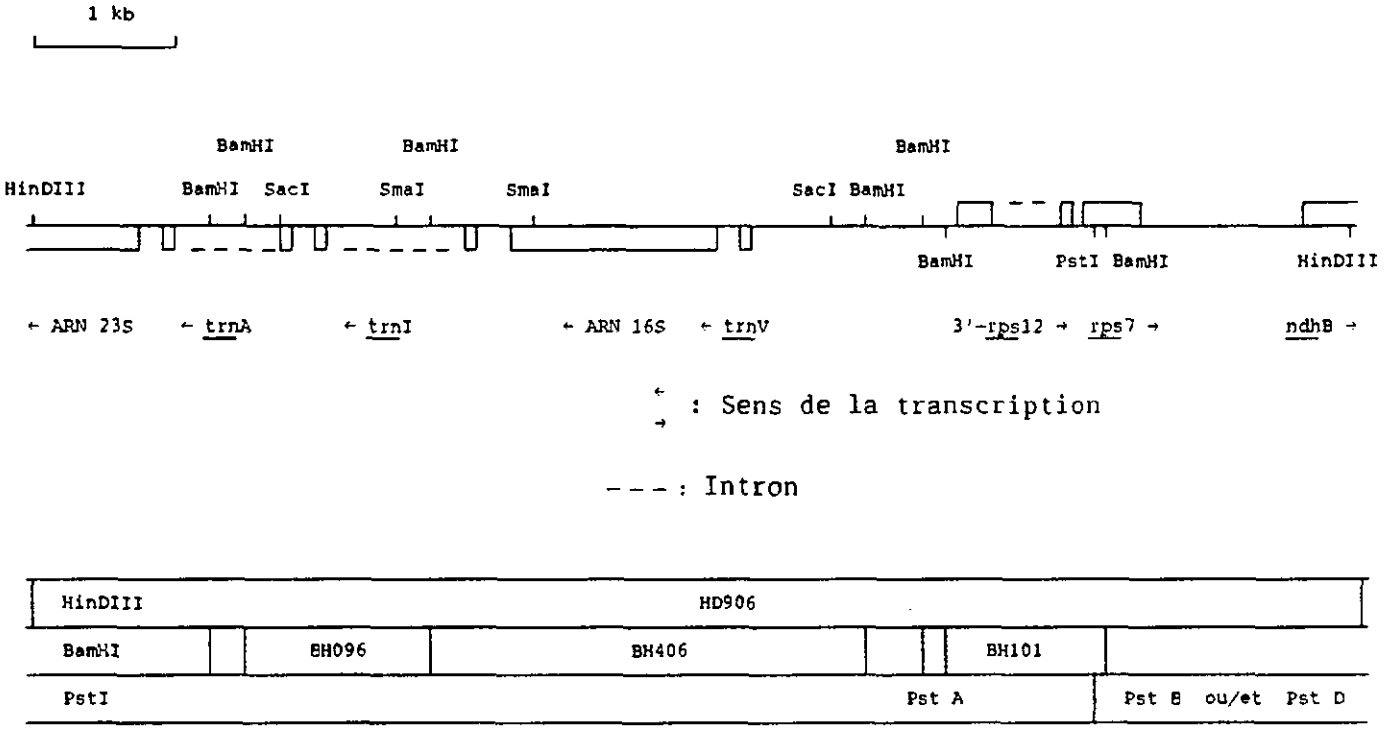


Le marquage du fragment HinDIII-SmaI (6.0 Kb) (Figure 4 A) qui est plus fort du côté HinDIII (1 ATP ajouté) que du côté SmaI (0 ATP) permet de mettre en évidence la capacité 3'-exonucléasique de l'ADN polymérase (fragment de Klenov) qui attaque le brin d'ADN en 3' et le répare par son activité polymérasique. D'autre part, cette figure permet d'identifier à partir du site HindIII des sites BamHI à 1200, 2500, 2700, et 3100 nucléotides.

La Figure 4 (B) permet de terminer la localisation des sites BamHI, à l'exception d'un petit fragment de 237 nucléotides que nous avons placé après une expérience supplémentaire. Ce fragment interne à l'intron du gène *trnA* était en dessous de la résolution du gel car deux fragments sont confondus dans les bandes de 6.6 et de 8.1 Kb.

Ces résultats, ajoutés à ceux de la figure 2, nous ont permis de construire la carte du fragment HD906. L'emplacement des gènes présenté sur la carte suivante a été défini d'après la séquence du fragment dans son ensemble. Nous devons signaler que Singh *et al.* (1984) ont entretemps publié une carte partielle couvrant la région comprise entre le gène *trnV* et la fin de l'ARN 23S qui confirme nos résultats.

Figure 5 : Construction de la carte du fragment HinDIII de 9 kb et emplacements indicatifs des gènes d'après la séquence.



Comme nous pouvons l'observer sur la carte, nous n'avons pas le début du gène *rps12*. Nous avons bien trouvé une ORF susceptible d'être le début du gène (cf. Figure 10, séquence : position 4569), localisée sur le brin anti-sens du gène codant pour l'ARN 16S. Nous avons également étudié de plus près une région que le professeur Edelmann, pendant la préparation du papier de Fromm *et al.* (1986), considérait comme un bon candidat pour le *5'-rps12*. Cette dernière région contenait malheureusement un codon stop dans le cas du soja (cf. Figure 10, séquence : position 5104). Nous avons alors envoyé de l'ADN du clone BH406, contenant ces séquences, au Professeur H. Weissbach du Roche Institute of Molecular Biology de New York pour tester *in vitro* si l'initiation de la traduction des régions concernées était possible en utilisant la méthode préconisée par Cenatiempo *et al.* (1982).

Les résultats ont été doublement négatifs, ce qui donne une indication supplémentaire que la partie *5'-rps12* n'est pas située directement en 5' de l'exon 2.

Déjà à ce momment là, l'idée nous est venue que le début du gène se trouvait peut-être placé ailleurs et pourquoi pas dans une autre direction. Cette idée qui à première vue était tout à fait fantaisiste, et n'avait aucun fondement sur les "lois" bien connues de la transcription, s'est trouvée soudain à la mode, lorsque les travaux sur le *trans* splicing *in vitro* (Konarska *et al.* 1985; Solnick, 1985) et, *in vivo*, chez le Trypanosome ont paru (Sather et Agabian, 1985).

Le *trans* splicing qui d'abord a été considéré comme un artefact du splicing *in vitro*, consiste en la liaison de deux exons situés sur deux molécules d'ARN différentes, pour former un ARN résultant, unique. Notre problème revenait à trouver l'emplacement du premier exon !

Malheureusement, la sonde d'Euglène refusait toujours de signaler le début du gène, à cause de sa relative petite taille et des nombreuses divergences existant dans la partie 5' des deux gènes. La taille et l'organisation des équipes japonaises (nous étions alors seul à travailler sur le soja) fut payant et sur les deux génomes séquencés entièrement, le gène fut trouvé en utilisant des moyens informatiques (Fukuzawa *et al.* 1986; Torazawa *et al.* 1986). Le professeur Edelmann nous a alors envoyé une sonde de tabac et nos avons pêché le gène.

Connaissant les travaux de (Palmer *et al.* 1983) comparant les différents génomes chloroplastiques, et remarquant que la région du *5'-rps12* de tabac devait environ coïncider avec la situation du gène chez le soja, nous pouvions prévoir que le fragment à cloner serait de très grande taille, si nous utilisions un des enzymes de restriction pour lequel la carte de restriction de soja était déjà établie. Nous avons donc cloné au hasard des fragments d'ADN chloroplastiques de soja digérés par HindIII et BamHI, et repéré les clones contenant la région désirée par les méthodes habituelles : Hogness, Mini-lysats, hybridation hétérologue contre une sonde ne contenant que le premier exon du gène *rps12* de tabac.

3.1.6 Clonage du *5'-rps12*

Nous avons directement sous-cloné le fragment HindIII de 4.4 kb dans M13 après digestion avec l'enzyme de restriction AluI. Puis, utilisant le programme ANALYSE que nous avons mis au point, il nous a été possible de reconnaître les clones contenant des fragments du gène *rps12* de soja en comparant les "T-tracks" (réaction de séquence pour mettre en évidence les résidus tymidine) de soja avec la séquence connue de tabac (Shinozaki *et al.* 1986) et disponible parmi les fichiers de l'EMBL. Cette méthode à l'avantage d'être beaucoup plus rapide que la méthode classique, ne demandant pas de connaître la carte du fragment à séquencer, si il existe à disposition la séquence d'un gène homologue...

Remarquons que le repérage par ordinateur est très facile si les régions comparées sont internes à un gène, ou comprises dans les séquences non-codantes des régions répétées inverses, il est par contre beaucoup moins facile de se repérer dans les zones non codantes et uniques du génome, qui ont divergé beaucoup plus au cours de l'évolution. Il est clair que de bons gels et une connaissance des quelques arte-facts possibles au niveau des réactions de séquences facilite le travail.

3.1.7 Cartographie du *5'-rps12*

Pour connaître l'emplacement précis du *5'-rps12* sur la carte de l'ADNct, nous avons ensuite utilisé des sondes simples brins provenant de clones M13 marquées selon notre méthode personnelle §(2.2.7.3). Nous avons procédé à des digestions complètes et partielles d'ADNct de soja et procédé à une hybridation homologue avec les sondes de soja.

Figure 6 : Cartographie du *5'-rps12* en utilisant de l'ADN-ct digéré par différents enzymes de restrictions :

A) Digestion BamHI + digestion HinDIII
B) Digestion BamHI + digestion SacI
C) Digestion HinDIII
D) Digestion SacI

a) ADNct coloré au bromure d'éthidium

b) Marquage simple à l'aide d'une sonde interne au gène *rpl20* externe au fragment HinDIII de 4.4 kb (cf. sonde n°3 décrite au §(3.2.1))

c) Marquage double, avec une sonde interne au *5'-rps12* (sonde n°1 décrite au §(3.2.1)) en plus du marquage (b)

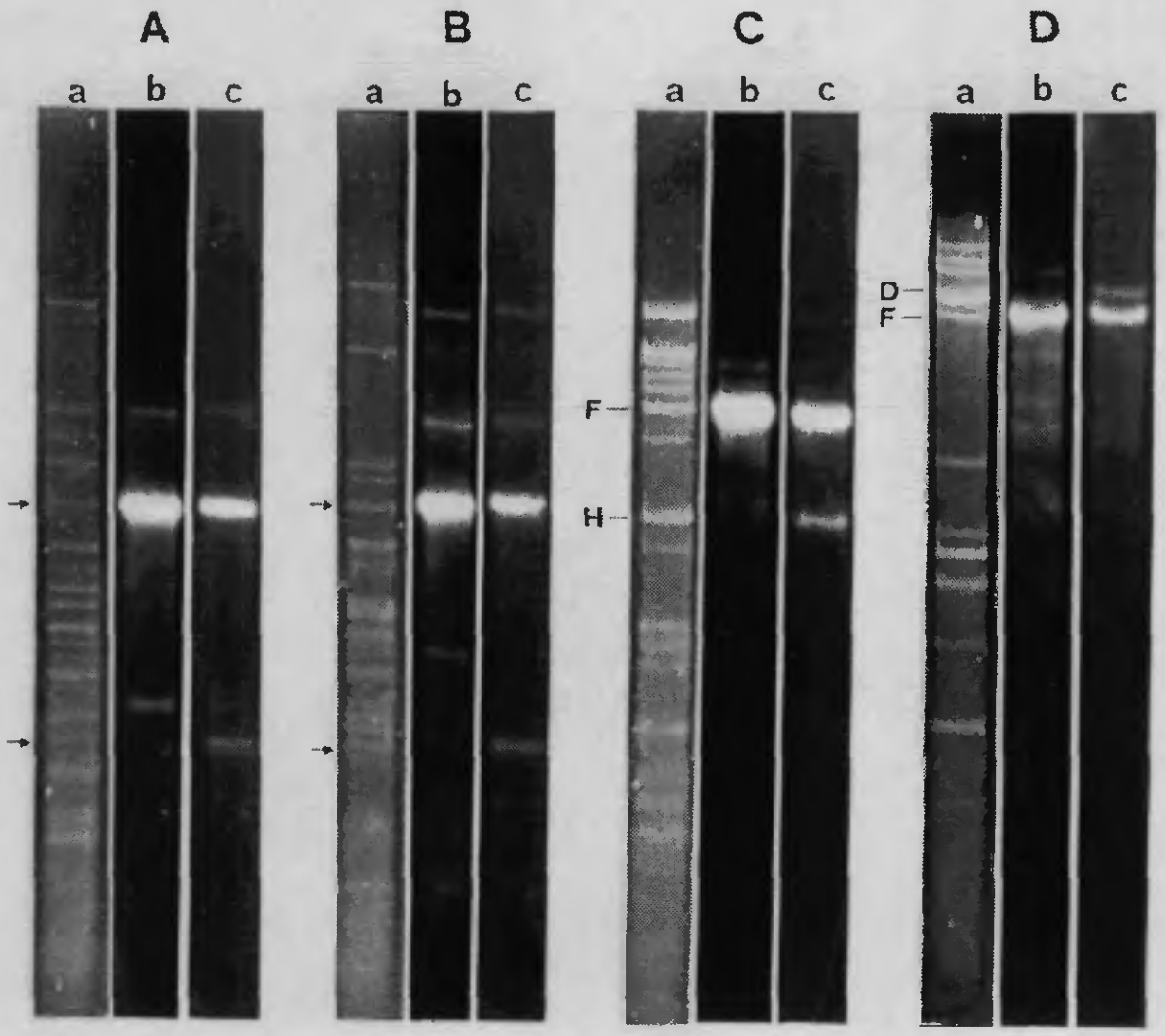
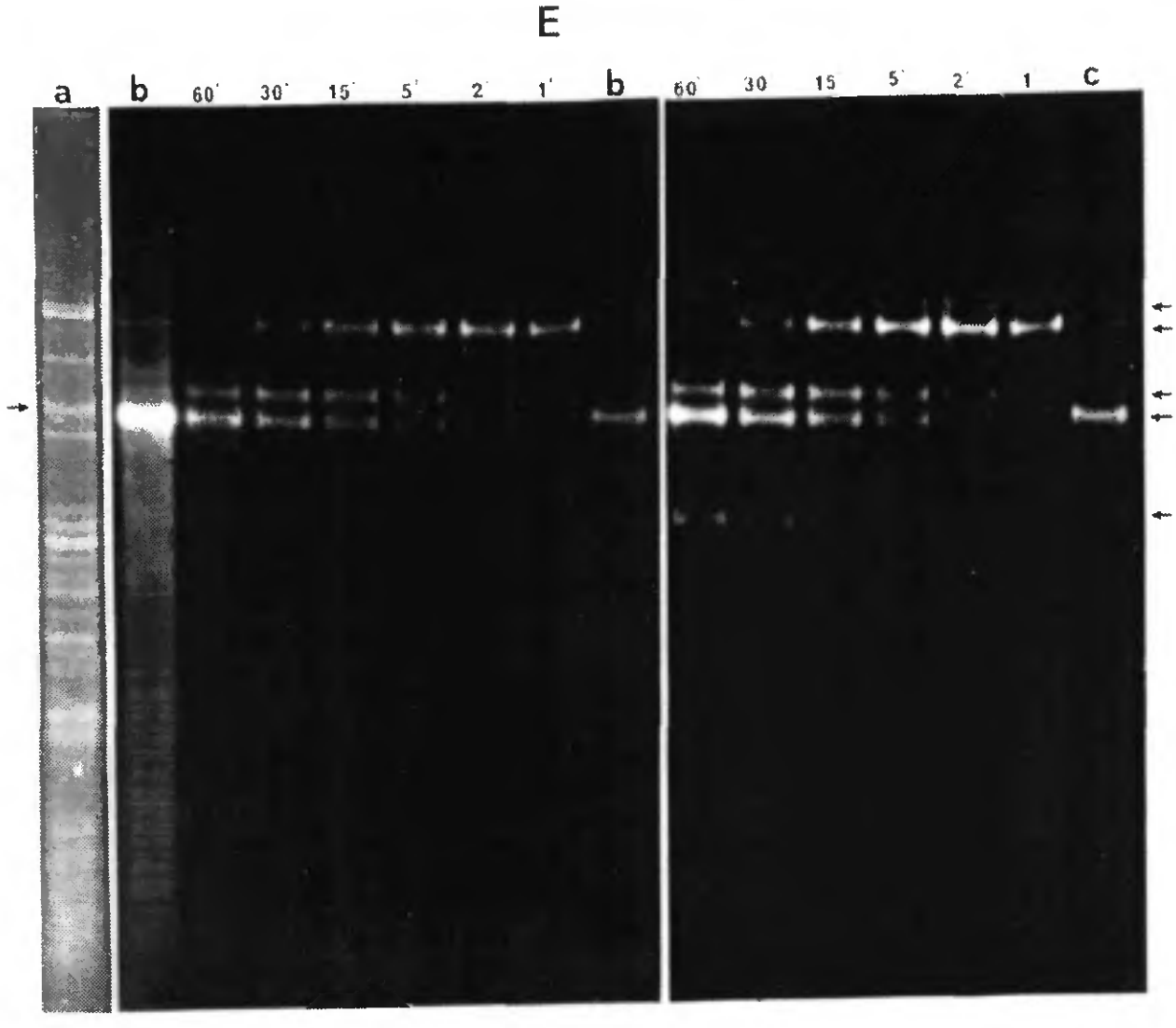


Figure 6 (suite)

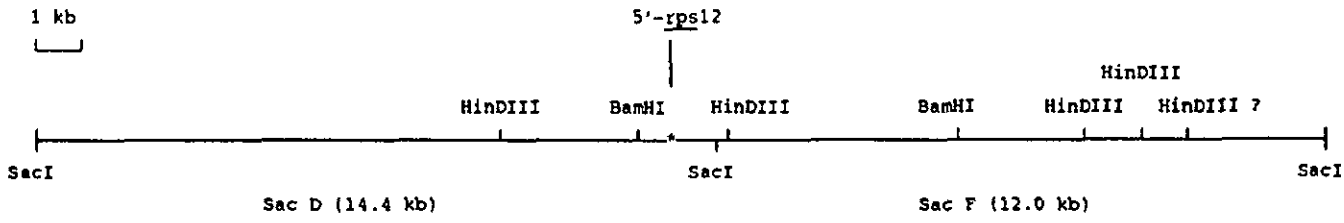
E) Digestion SacI totale + digestion partielle HinDIII.
Une unité d'enzyme a été ajouté pour 5 μg d'ADN chloroplastique. 500 ng d'ADN ont été récoltés après le temps indiqué en minutes sur la figure.

a, b, c identiques à la page précédente

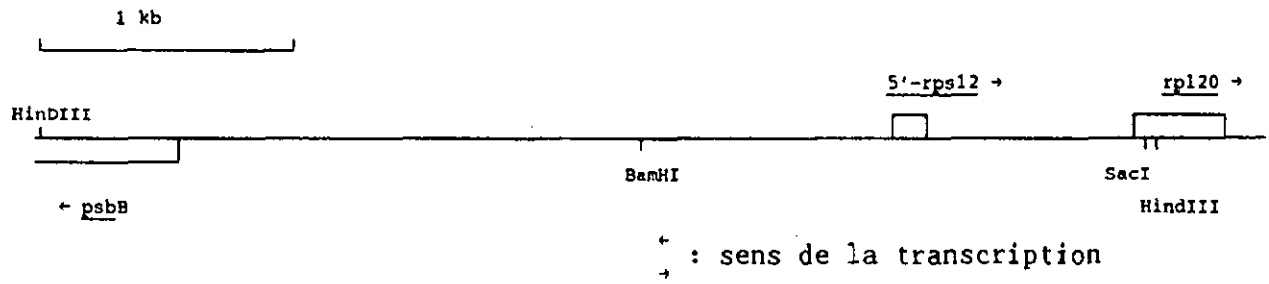


Nous obtenons hybridation de l'exon 1 du *3'-rps12* avec un fragment BamHI de 6.6 kb et un fragment HinDIII de 4.4 kb. Les hybridations montrent clairement que la partie 3' du *rpl20* est située dans le fragment Sac F et donne par la même occasion l'orientation du fragment HinDIII de 4.4 kb. Nous remarquerons que ce dernier fragment est principalement situé dans le Sac D. L'analyse de la séquence démontre que le site SacI se trouve à 40 nucléotides de l'extrémité du fragment. L'exploitation des résultats de la Figure 6 nous permet de construire la carte de la Figure suivante :

Figure 7 : Carte de la région du *5'-rps12* (=exon1)



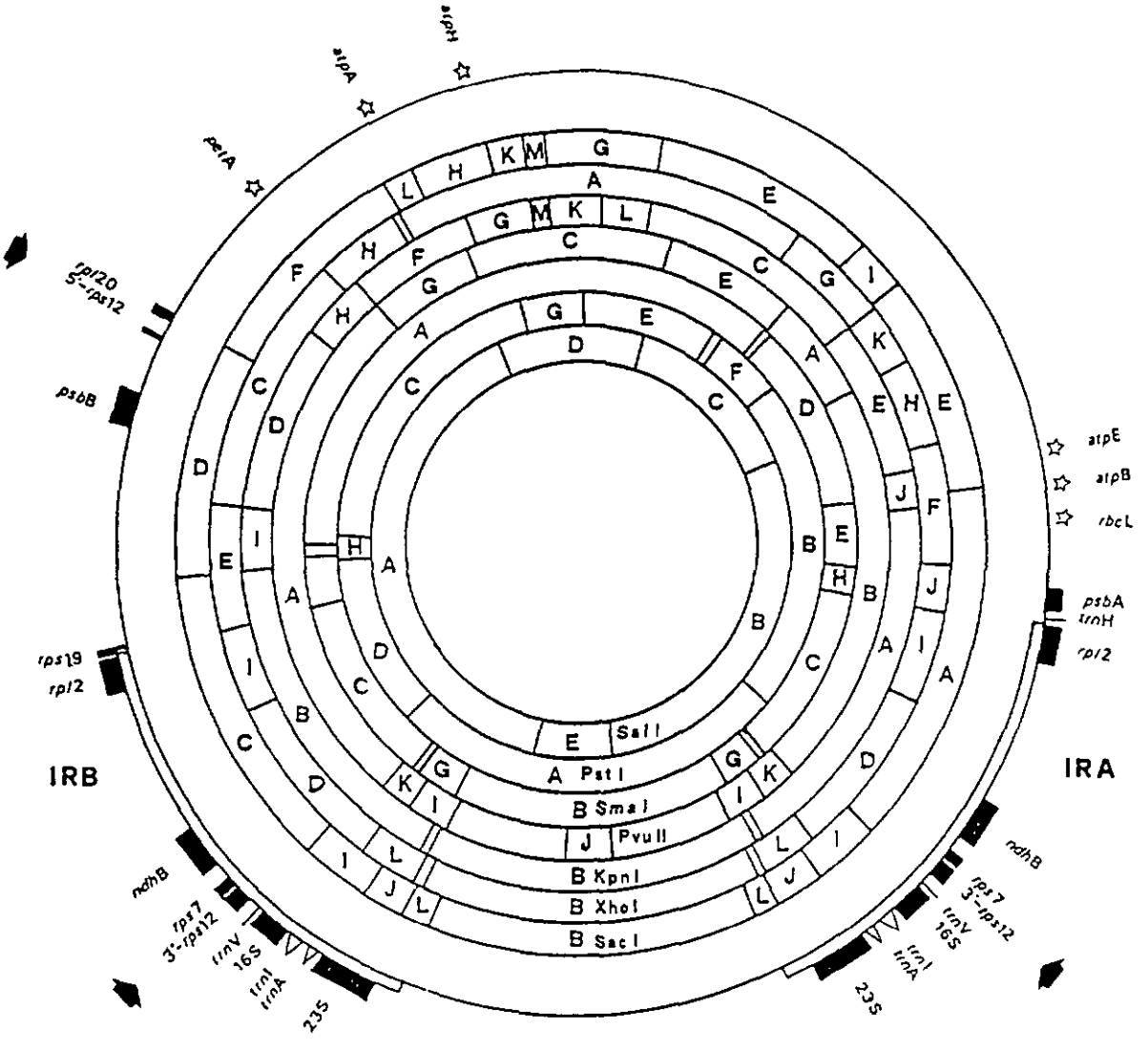
Carte : Détail



Nous pouvons ainsi construire la carte générale de l'ADN chloroplastique, (Figure 8), qui inclu les résultats obtenus jusqu'à ce jour : (Spielmann et Stutz, 1983; Spielmann *et al.* 1983; Spielmann, 1983; Singh *et al.* 1984; Singh *et al.* 1985; von Allmen et Stutz, 1985; von Allmen et Stutz 1987; von Allmen, présent travail.)

Figure 8 : Carte générale de l'ADN chloroplastique de soja

Les étoiles indiquent un gène localisé par hybridation. Les rectangles, une localisation précise par séquençage.



Nous remarquerons qu'il existe une différence majeure entre la carte du tabac et celle du soja, qui consiste en l'inversion de la région comprise entre le *psbA* et le *petA*.

3.1.8 Séquençage

La région que nous avons séquencé étant relativement grande, nous avons eu le temps d'affiner notre stratégie du séquençage.

Lorsque la région n'a jamais été séquencée pour un autre organisme, ce qui était le cas pour la région du *3'-rps12* comprise entre les gènes *trnV* et *ndhB*, nous avons procédé ainsi :

- Nous avons choisi des enzymes de restriction qui coupent peu la séquence, pour obtenir rapidement l'ordre des fragments plus petits et gagner en efficacité.

Nous avons séquencé toute cette région obligatoirement sur les deux brins de la molécule. Il est ainsi possible de passer à travers certaines régions qui posent quelques problèmes à l'ADN polymérase I, fragment de Klenov, comme les séquences très riches en guanosine.

Si la région est connue pour un autre organisme et que la séquence du gène homologue est publiée :

- Alors une technique nouvelle de séquençage est disponible, en utilisant le programme ANALYSE que nous avons mis au point. Ce programme permet de trouver, par exemple, l'emplacement des clones de soja correspondant à l'emplacement que le clone occupe dans la séquence de tabac. Visuellement les T-tracks des clones de soja peuvent être comparés au T-tracks de tabac. Nous conseillons ainsi de travailler avec un fichier de résultats qui contient la séquence de tabac en lettres minuscules et de remplacer au fur et à mesure la séquence de tabac par la séquence de soja, mais en majuscules cette fois. Il est ainsi possible de gagner beaucoup de temps dans les analyses.

Figure 9 : Exemple d'analyse de T-tracks avec notre programme.

EXEMPLE D'UTILISATION : Recherche de N-tracks

MODE D'EMPLOI :

POUR RECHERCHER UN N-TRACK, IL SUFFIT DE TAPER LA SEQUENCE
EN REMPLACANT LES ESPACES VIDES PAR UN <.>
ex : .TTTT.TT.T..TT. recherche les T-TRACKS correspondants
ainsi que les A-TRACKS .AA...A.AA.AAAA.

SEQUENCE ? .TTT..TTT..

ANALYSE DE LA SEQUENCE DONNEE

GCCCCATCTATCCTCCTGAGGATAAGTTTGGTTTCAAACCCCGGTTCAAACAGGAGAAGTA 2634

ATTCTACAAGAATTTTTATGAAAGCCTTTCATTTGCCTCTCTTCGATGGAAGTTTTATTTT 8108

ANALYSE DE LA SEQUENCE COMPLEMENTAIRE

GCTTCTTTTTGGGTGTGAAGCAGTGTCAAACTAAACTACCCAACAAGCATTAGCTCTCCCT 3068

TATCACCGGCAGTCTGCTCAGGGTTCCAAACTAAATGTTGGCAACTAAACACGAGGGTTGC 3495

Ainsi, pour construire la séquence du gène *trnI'* au *23S*, nous avons utilisé les séquences du tabac publiées par Tohdoh *et al.* (1981); Tohdoh et Sugiura (1982); Takaiwa et Sugiura (1982)a; Takaiwa et Sugiura (1982)b et avons séquencé toute la région.

Figure 10 : séquence complète du fragment HinDIII de soja contenant le *3'-rps12*.

Séquence du clone HD.906 de l'ADN chloroplastique de soja

```
AAGCTTCATCCTGGTCATGGACAGATCACCCAGGTTCGGGTCCATAAGCAGTGACATTTGCCCTGTGAAGACTCGCTTTC 80
TTCGAAGTAGGACCAGTACCTGTCTAGTGGGTCCAAGCCCAGGTATTCGTCACTGTAAACGGGACACTTCTGAGCGAAAG
HinDIII

GCTACGGCTCCGGTGGTTTCCCTTAACCAAGCCACTGCCTATGAGTCGCCGGCTCATTCTTCAACAGGCACGCGGTCAGA 160
CGATGCCGAGGCCACCAAAGGGAATTGGTTCGGTGACGGATACTCAGCGGCCGAGTAAGAAGTTGTCCGTGCGCCAGTCT

GTCCCGAGCCTCCTCCCACTGCTTGGGAGCTTACGGTTTCATGTTCTATTTCACTCCCCGATGGGGTTCTTTTCACCCTT 240
CAGGGCTCGGAGGAGGGTGACGAACCCTCGAATGCCAAAGTACAAGATAAAGTGAGGGGCTACCCCAAGAAAAGTGGGAA

CCCTCACGGTACTACTTCACTATCGGTCACCCAGGAGTATTTAGCCTTGCAAGGTGGTCCTTGCTGATTcacacgggatt 320
GGGAGTGCCATGATGAAGTGATAGCCAGTGGGTCCTCATAAATCGGAACGTTCCACCAGGAACGACTAAgtgtgccctaa

ccacgtgccccatgctactcgggtcagagcataagctagtgatgctttcggctactgGACTCTCGCCATCTAGGGTGCAG 400
ggtgcacggggtacgatgagcccagtctcgtattcgatcactacgaaagccgatgacCTGAGAGCGGTAGATCCCACGTC

CATTCTCTGCTCTTCGCCTAGCAGCACGACACTTGTATAGCTCTCCCACAACCCGTTTTCACGGTTTAGGCTGCTCCCAT 480
GTAAGAGACGAGAAGCGGATCGTCGTGCTGTGAACATATCGAGAGGGTGTTGGGCAAAAGTGCCAAATCCGACGAGGGTA

TCGTCGCCGCTACTACGGGAATCgcttttgctttcttttcctctggctcactaagatgtttcagttcgccaggttgtctc 560
AGCAGCGGCGATGATGCCCTTAGcgaaaacgaaagaaaaggagaccgagtgattctacaaagtcaagcggtccaacagag

ttgcctgcccatggattcagcagcagttcgaaaggttgccctattcgggaatctccggatctatgcttattttcaactcc 640
aacggacgggtacctaagtcgtcgtcaagctttccaacgggataagcccttagaggcctagatacgaataaaagttgagg

ccgaagcatttcgtcgattactacgcccttcctcgtctctgggtgcctaggtatccaccgtaagcctttcctcgtttgaa 720
ggcttcgtaaagcagctaatgatgcgggaaggagcagagacccacggatccataggtggcattcggaaaggagcaaactt

cctcgcccttcacttttaaggctatgccatcctaaggtgctgctaaatggatggatcttatcaacgtccatgaatgataa 800
gg   <ARN-r 16S

aatcatagatcgaaccgccGAATCGGAAAAATGGAAGTGCTATCATATAGCTTTGTATCGGACTAAGTTCACGAGTTGGA 880
                                                                    ACCT

GATAAGCGGACTCGAACCGCTGACATCCGCCACAGGGTAAACCACCGCCTCTCAGGCCCGACTGATTCTACCATAGAGGC 960
CTATTCGCCTGAGCTTGGCGACTGTAGGCGGT   <Exon2 ARN-t Ala
      890        900        910

CAACGATAGACAATAACTCCCCCCCGAACACAGCTCACAACTTTCATCGTACTGTGCTCTCCAAAGAGCAACTCTTCTCA 1040
AAATCTCAAAAGGTGCTGAGTTGGAATCCCATTCTTGTGGTTCCGGAGATTCCAGCTACAGGAACCGGGagctctcccct 1120
ttttccgcccgactctttgatcttaactttatgaatgctggttttatgaacgagtgattgcccttcGGATCCGTCGTCCA 1200
                                                                    BamHI

ACCTGAGAGCGGATAGTAATGCGTTCCGTTTTTGAACAGGGTTCTATGGTCGGTCCGCGACCCCTGGATGCCGACGTCCT 1280
TGGGGTGATCTCGTAGTTCCTACGGGGTGGAGACGATGGGGTCGGTCCATGGATTCTTTTTCCTTTTGCCACATTTCGCT 1360
CAAAGGGTTGAAGGGAGATAGTGCATCAACTGTTCGCAAGGGCCAGCTTGATCCTCTTCCCCAGGATCCCAAATGAGGAA 1440
                                                                    BamHI
```

```
ACCCTAGGAGAGCCACCGACTCCAACTACCGTCCATGTACGATCCATACTAGATCTGACCAACTGCCCATCCTACCTCCT 1520
CTACGTTCTTGACAGCCCGTCTTTGTCTCAGTAGAGTCTTTCAGTGGCATGTTTCGGTCCTCTTTCCCATTACTTAGAAA 1600
AAGTGAGCCACCGGTTCAGGTACAAGATACTATCATTACCGCCTGGACAATTAGACATCCAACCCGTAATCGCAACGACC 1680

CAATTGCAAGAGCGGAGCTCTACCAACTGAGCTATATCCCCCCGAGCCGAGTGGAGCATGCATGAAGGAGTCAAGTCAGT 1760
    ACGTTCTCGCCTCGAGATGGTTGACTCGATATAGGGG   <Exon1 ARN-t Ala
      1690      1700      1710      1720
              SacI

TTTCTATTCTTTTCCTTGACGCAGCTGGGCCAACCTGGACTTGAACCAGAGACCTCTCCCGCGAAGTAAATCATCGCACC 1840
                          ACCCGGTTGGACCTGAACTTGGTCTCTGGAGAGGGC   <Exon2 ARN-t Ile
                           1790      1800      1810      1820

TACAGTCCAACCAATTGGGAGAGAATCAATAGATTCCTTTTCGGGAGCGATTCATCCTTCCCGAACGCAGCATACAACTA 1920
TCCGTTGTACTGCGCTCTCCAAGTGTGCTTGTTCCCCCTTCTTCTTTACCATGGCAAGTCTTTGTGAAATAACTCCGATG 2000
AGAAGAAAAAGGAAGGCATTAAGAGACCCTCCTGGCACAACCCTAGACACTCTAAGATCCTTTTTCAAACCTGCTCCCAC 2080
CATTTCGAGTCAAGAGATAGATAAATAGACACATCTCATTGCACTGATCGGGGGCGTTCGTAGTGACTGAGGGGGTCAAA 2160
GACCAAGAAGTGAGTTATTTATCAGCCAAGCATTCTATTCTTCTTATGGCTAGATCCAATCTCCTGGTCCCTGTGGAAAG 2240
GAAAAAGAATTTTTCACGTTCTTCCTTTCGGGAAGGAAGGATTAGTAAAATCCTATTGATTGCAATTTTCTCCAGACCCC 2320
                                                                        SmaI

GGGAAAAGCATGAAAAAAATTCTCGAATGGTACGAACCCACCGTCACCCCAGAATGAAAGGGGCGATCTCGTAGTTCTTG 2400
GTCTGTGAATACGTTGTTAGGTGCTCCGTTTTTTTCCATTGAGGCCAAACCTAAACCTGTGCTCGAGAGATAACTGTCCA 2480
TATACTGATAAGGGATGTATGAATTCTCGAGAAGAGAGGAGCCGTGGTGGTCCCCCCCGGATCGCCCGGATCACGGATCC 2560
                    EcoRI                                                   BamHI

CACGAGTGAAGAGAAAGTTGGATCTACATTGGATCTCACCTGAATCGCCCCATCTATCCTCCTGAGGATAAGTTTGGTTT 2640
CAAACCCCGGTTCAAACAGGAGAAGTACGCCATGCTAATGTGCCTTGGATGATCCACATCTCAGGGTCAGGCGCCGATGA 2720

GCACATTAAACTATCCATGTGGCTGAGAGCCTCACAGTCCAGGCACAACGACGCAATTATCAGGGGCGCGCTCTACCACT 2800
                                                        AATAGTCCCCGCGCGAGATGGTGA
                                                        2780      2790      2800

GAGCTAATAGCCCGTCGTGCGGCTCCCGCGGGGCCCGCTATGCCAAAAGCAAGAGAAACCCCATCCCTCTCTTTCCTTTT 2880
CTCGATTATCGGG   <Exon1 ARN-t Ile
      2810

TTCACCCCCATCCATGTCGCCCCCGTGTGGCGACATGAGACAAAAGTAGCTCCTATCAACTCGTTCCGACCTAGGATAAT 2960
AAGCTCATGAGTTTAGTCTTACTTCACCATCGAGAAACGAAAGAAGACTTCCATCTCCAAATTTGACTCAGACATAACTC 3040

GCTTCTTTTTGGGTGTGAAGCAGTGTCAAACTAAACTACCCAACAAGCATTAGCTCTCCCTGAAAAGGAGGTGATCCAGC 3120
                                                            TTCCTCCACTAGGTCG
                                                             3110      3120

CGCACCTTCCAGTACGGCTACCTTGTTACGACTTCACTCCAGTCACTAGCCCCTTCGGCATCCCCCTCTTGCGGTTAAGG 3200
GCGTGGAAGGTCATGCCGATGGAACAATGCTGAAGTGAGGTCAGTGATCGGGGAAGCCGTAGGGGGAGAACGCCAATTCC

TAACGACTTCGGCATGGCCAGCTCCCATAGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACGAATTCACCGCCGTATGG 3280
ATTGCTGAAGCCGTACCGGTCGAGGGTATCACACTGCCCGCCACACATGTTCCGGGCCCTTGCTTAAGTGGCGGCATACC
                                                        SmaI      EcoRI

CTGACCGGCGATTACTAGCGATTCCGGCTTCATGCAGGCGAGTTGCAGCCTACAATCCGAACTGAGGACGGGTTTTTGGA 3360
GACTGGCCGCTAATGATCGCTAAGGCCGAAGTACGTCCGCTCAACGTCGGATGTTAGGCTTGACTCCTGCCCAAAAACCT

GTTAGCTCACCTCGCGGGATCGCGATCCTTTGTCCATTGTAGCACGTGTGTCGCCCAGGGCATAAGGGGCATGATGACTT 3440
CAATCGAGTGGAGCGCCCTAGCGCTAGGAAACAGGTAACATCGTGCACACAGCGGGTCCCGTATTCCCCGTACTACTGAA

CGACGTCATCCTCACCTTCCTCCGGCTTATCACCGGCAGTCTGCTCAGGGTTCCAAACTAAATGTTGGCAACTAAACACG 3520
GCTGCAGTAGGAGTGGAAGGAGGCCGAATAGTGGCCGTCAGACGAGTCCCAAGGTTTGATTTACAACCGTTGATTTGTGC
```

```
AGGGTTGCGCTCGTTGCGGGACTTAACCCAACACCTTACGGCACGAGCTGACGACAGCCATGCACCACCTGTGTCCGCGT 3600
TCCCAACGCGAGCAACGCCCTGAATTGGGTTGTGGAATGCCGTGCTCGACTGCTGTCGGTACGTGGTGGACACAGGCGCA

TCCCGAAGCACCCCTCTCTTTCAAGAGGATTCGCGGCATGTCAAGCCCTGGTAAGGTTCTTCGCTTTGCATCGAATTAAA 3680
AGGGCTTCGTGGGGAGAGAAAGTTCTCCTAAGCGCCGTACAGTTCGGGACCATTCCAAGAAGCGAAACGTAGCTTAATTT

CCACATGCTCCACCGCTTGTGCGGCCCCCGTCAATTCCTTTGAGTTTCATTCTTGCGAACGTACTCCCCAGGCGGGATAC 3760
GGTGTACGAGGTGGCGAACACGCCGGGGGCAGTTAAGGAAACTCAAAGTAAGAACGCTTGCATGAGGGGTCCGCCCTATG

TTAACGCGTTAGTACAGCATTGCACGGGTCGATACGCACAGCGCCTAGTATCCATCGTTTACGGCTAGGACTACTGGGGT 3840
AATTGCGCAATCATGTCGTAACGTGCCCAGCTATGCGTGTCGCGGATCATAGGTAGCAAATGCCGATCCTGATGACCCCA

ATCTAATCCCATTCGCTCCCCTAAGTTTCGTCTCTCAGTGTCAGTGTCGGCCCAGCAGAGTGCTTTCGCCGTTGGTGTTC 3920
TAGATTAGGGTAAGCGAGGGGATTCAAAGCAGAGAGTCACAGTCACAGCCGGGTCGTCTCACGAAAGCGGCAACCACAAG

TTTCCGATCTCTACGCATTTCACCGCTCCACCGGAAATTCCCTCTGCCCCTACCGTACTCCAGCTTGGTAGTTTCCACCG 4000
AAAGGCTAGAGATGCGTAAAGTGGCGAGGTGGCCTTTAAGGGAGACGGGGATGGCATGAGGTCGAACCATCAAAGGTGGC

CCTGTCCAGGGTTGAGCCCTGGGATTTGACGGCGAACTTAAAAAGCCACCTACAGACCTTTCGCCCAATCATTCCGGATA 4080
GGACAGGTCCCAACTCGGGACCCTAAACTGCCGCTTGAATTTTTCGGTGGATGTCTGGAAAGCGGGTTAGTAAGGCCTAT

ACGCTTGCATCCTCTGTCTTACCGCGGCCGCTGGCACAGAGTTAGCCGATGCTTATTCCCCGGATCACCGTCATTGCTTC 4160
TGCGAACGTAGGAGACAGAATGGCGCCGGCGACCGTGTCTCAATCGGCTACGAATAAGGGGCCTAGTGGCAGTAACGAAG

TTCTCCGGGAAAAGAAGTTCATGACCGTAGGCCTTCTACCTTCACGCGGCATTGCTCCGTCAGCTTTCGCCATTGCGGAA 4240
AAGAGGCCCTTTTCTTCAAGTACTGGCATCCGGAAGATGGAAGTGCGCCGTAACGAGGCAGTCGAAAGCGGTAACGCCTT

AATTCCCCACTGCTGCCTCCCGTAAGAGTCTGGGCCTCGTCTCAGTCCCAGTGTGGCTGATCATCTCTCGGACCAGCTAC 4320
TTAAGGGGTGACGACGGAGGGCATTCTCAGACCCGGAGCAGAGTCAGGGTCACACCGACTAGTAGAGAGCCTGGTCGATG

TGATCATCGCCTTGGTAAGCTATGCCTCACCAACTAGCTAATCAGACGCGAGCCCCTCCTCGGGCGGATTCTCCTTTTGC 4400
ACTAGTAGCGGAACCATTCGATACGGAGTGGTTGATCGATTAGTCTGCGCTCGGGGAGGAGCCCGCCTAAGAGGAAAACG

TCCTCAGCCTACGGGGTATTAGCAGCCGTTTCCAGCTGTTGTTCCCCTCCCAAGGGTAGGTTCTTACGCGTTACTCACCC 4480
AGGAGTCGGATGCCCCATAATCGTCGGCAAAGGTCGACAACAAGGGGAGGGTTCCCATCCAAGAATGCGCAATGAGTGGG

GTCCGCCACTGGAAACACCACTTCCCGTCCGACTTGCATGTGTAAGGCATGCCGCCAGCGTTCATCCTGAGCCAGGATCG 4560
CAGGCGGTGACCTTTGTGGTGAAGGGCAGGCTGAACGTACACATTCCGTACGGCGGTCGCAAGTAGGACTCGGTCCTAGC

AACTCTCCATGAGATTCCTAGTTGCATTACTTATAGCTTCCTTGTTCGTAGACAAAGCTGATTCGGAATTCTCTTCCATT 4640
TTGAGAGGTACTCT   <ARN-r 16S                                          EcoRI
      4570

CCAAGGCATAACTTGTATCCAGGCGCTTCATATTCGACTGGAGTTCGCTCCCAGAAATATAGCCCCCCCACCCCTCGTGT 4720

AATCCCACGAGCCTCTTATCCATTCTCAATTATTATTCGATCACGGCATACAAATAGAAAAACTTACATTGGGTTTAGGG 4800
                                                                    ATCCC

ATAATCAGGCTCGAACTGATAACTTCCACCACGTCAAGGTGACACTCTACCGCTGAGTTATATCCCTTGCCTCCATCGAG 4880
TATTAGTCCGAGCTTGACTATTGAAGGTGGTGCAGTTCCACTGTGAGATGGCGACTCAATATAGGGA   <ARN-t Val
                                                    4860      4870      4880

AAATAGAACTGACTAATCCTAAGGCAAAGGGTCAAGAAACTCAACGCCACTATTCTTAAACAACTTCTTGGAGTTAGGCT 4960
TTCTTTTCACACTATTACGGATACGAAAATTATGGGAAAAATTGGATTCCATTGTCAACTGCTCTATCGGAAATAGGATT 5040
GACCACCGATTCCGAAGTAACTGGAGTTAGATCTCAATTTTCCATTTCAATTCAAGAGTTCTTATGTGTTTCCACGCCCC 5120
TTCGAGACCCCGAAAAAATGGACAAATTGCTTTTCTTAGGAACACATATTCGTTACCACAAAAAGGATAATGGTAACCCC 5200
ACCATTAACTACTTCAATTTAATAGTAATAAAAATACATGCCCTACCAAGACAGAATTTGTAATTTGCTATACTCTTGCC 5280

                                                  M  R  V  Q  P  H  C  I  A  R
TAGCAGGCAAAGATTTACCTCCGTGGAAAGGATGATTCATTCAGATCGACATGAGAGTCCAACCACATTGCATTGCCAGA 5360
      5290      5300      5310      5320      5330
                                              ORF 163
```

```
 T  H  V  V  Y  L  K  K  V  E  P  P  C  F  S  H  V  T  I  L  L  P  L  S  S  L  S
ACCCATGTTGTATATTTGAAAAAGGTTGAACCTCCTTGCTTCTCTCATGTTACAATCCTCTTGCCGCTGAGCTCCCTTTC 5440
                                                                      SacI

  P  R  S  T  E  T  K  C  K  T  G  A  N  S  S  S  R  K  K  G  L  A  E  P  R  S
TCCTCGGTCCACAGAGACAAAATGTAAGACAGGTGCCAACAGTTCATCACGAAAGAAAGGACTAGCTGAGCCGAGATCAC 5520

L  T  K  T  N  L  I  L  I  E  N  T  N  R  I  E  I  E  L  S  F  H  I  D  I  P  S
TAACTAAAACGAATTTAATACTAATCGAAAATACTAATAGAATAGAAATAGAACTGTCTTTTCATATAGATATCCCTTCA 5600

 T  Q  H  R  S  S  K  G  S  R  R  P  T  N  A  R  K  P  G  S  F  R  K  W  I  P  I
ACACAACATAGGTCATCGAAAGGATCTCGAAGACCCACCAACGCACGAAAGCCGGGATCTTTCAGAAAATGGATTCCTAT 5680

  R  R  V  H  N  R  M  D  K  L  T  L  T  R  Q  F  G  I  Q  F  G  S  L  G  G  I
TCGAAGGGTGCATAACCGCATGGATAAGCTCACACTAACCCGTCAATTTGGGATCCAATTCGGTTCCTTAGGAGGTATCG 5760
                                                    BamHI

G  K  G  F  W  N  V  I  I  S  I  H  T  D  T  E  E  R  V  L  Y  *
GGAAGGGATTTTGGAATGTAATAATATCGATTCATACAGATACAGAAGAAAGGGTTCTCTATTGATTCAAACGCTCTACC 5840
      5770      5780      5790      5800      5810      5820

TACGAGATAGGGATAGAGAAAGAGGAAAAAACCGAGGATTTCACATAGTACTTTTGGTCGAAAAATCGATCTGATTTTTT 5920
TCGTACTTTTCGCTCAATGATGAAATGGGTCAGATTCTACAGGACCAAACCTTGTGGGAGTTAAGGAATGATGGAAGAAA 6000
AGAAATAAAAAAAAAGTAAATAAAAATGCAGTAGAAGAGTCCTGATTCCAAACGAAAAAATTCAAACTTGATTGAAAAGGA 6080

TCTTTCTGATTATTGAAGAATGGGACAAAAGGATTGATCGCGAAAAAGATTTCTTGTTCTTATTAGAAGATCGTGATTGG 6160
                                                                      BamHI

ATCCACATATGTTTGGTAAAAAGAAAGAAAAATATTCTTCTTTGAGAATAATAAAAAAGGAAAGTGTTCAATTGGAACAT 6240
GAAAACGTGACAGAATTGGTCCTAGTTACTCTTCGGAACGGAATAGAAGAAGGGGAGAGATTCTCGAACCGAACTAGGAA 6320

ACGGATCCAATGACTTCAAAAGAATTGAACGAGGAGCCATATGAAGTGAAAATCTCATGTACGGTTCTGTAGAGTGGCAG 6400
      6330      6340      6350      6360      6370      6380      6390      6400
  BamHI                           Structure (B5) Intron du groupe II

                                T  I  T  P  K  K  P  N  S  A  L  R  K  V  A  R  V
TAAGGGTGACTTATCTGTCAACTTTTCCACTATTACACCCAAAAAACCAAACTCTGCTTTACGTAAAGTTGCCAGAGTAC 6480
      6410      6420      6430                                SnaBI
                             S12-exon2>

R  L  T  S  G  F  E  T  T  A  Y  I  P  G  I  G  H  N  L  Q  E  H  S  V  V  L  V
GCTTAACCTCTGGATTTGAAATTACTGCTTATATACCCGGTATTGGCCATAATTTACAAGAACATTCTGTAGTCTTAGTA 6560
                                  HpaII   HaeIII

 R  G  G  R  V  K  D  L  P  G  V  R  Y  H  I  V  R  G  T  L  D  A  V  G  V  K  D
AGAGGGGGGAGGGTTAAAGATTTACCCGGTGTGAGATATCACATTGTTCGAGGAACCCTAGATGCTGTCGGAGTAAAGGA 6640
                               Cis_splicing

  R  Q  Q  G  R  S  S  A  L  *
TCGTCAACAAGGGCGTTCTAGTGCGTTGTAGATTCTTATCCAAGACTTGTATCATTTGAAAATGCCATGTGAATCGCTAG 6720
      6650      6660      6670
                    GTGCGNNNYR - Intron 2
                    Cis_splicing

                                                          M  R  Q  K  I
AAACATGTAAAGTGTGTGGCTAACCCAATAACGAAAGTTTCGTAAGGGAACTGGAGCAGGCTACCATGAGACAAAAGATT 6800
                                                    6780      6790      6800
                                                          ORF 143

 F  F  L  K  K  F  D  S  E  L  R  Y  V  Q  G  P  I  W  K  E  F  Q  R  F  S  L  T
TTCTTTCTAAAGAAATTCGATTCGGAACTACGATATGTCCAAGGTCCAATATGGAAAGAATTTCAGAGGTTTTCCTTGAC 6880
```

```
  L  S  V  S  T  N  N  S  K  C  L  D  F  F  R  I  G  P  S  Q  I  A  M  I  R  S
TTTGTCCGTGTCAACAAACAATTCGAAATGCCTCGACTTTTTTAGAATAGGTCCGAGTCAAATAGCAATGATTCGAAGCA 6960

  T  L  H  Y  F  G  N  P  R  T  Q  S  Y  G  Y  V  K  Y  R  I  S  N  P  S  R  K  R
CTTTACACTATTTTGGAAACCCAAGGACTCAATCGTATGGATATGTAAAATACAGGATTTCCAATCCTAGCAGGAAAAGG 7040

  G  G  T  D  T  Q  F  Q  V  S  K  Q  N  S  I  L  D  I  I  D  T  Y  R  I  L  W  K
GGGGGAACGGATACTCAATTTCAAGTGAGTAAACAGAATTCCATACTAGATATCATAGATACATATAGAATTCTGTGGAA 7120
                                    EcoRI                                 EcoRI      Struc-

                                                                                 K  Y  G
  A  V  F  D  E  S  R  M  Y  G  L  E  G  D  L  S  Y  L  S  R  S  T  L  Q  Y  G
AGCCGTATTCGATGAAAGTCGTATGTACGGCTTGGAGGGAGATCTTTCATATCTTTCGAGATCCACCCTACAATATGGGG 7200
      7130      7140      7150      7160      7170      7180      7190      7200
ture (B5) Intron du groupe II                                                S12-exon3>

  A  K  K  P  K  *  M  I  L  A  L  I  K  V  I  K  K  R  I  P  P  F  T  L  M  S  R
CCAAAAAGCCAAAATAAATGATTTTAGCCCTTATAAAAGTTATAAAAAAACGGATTCCTCCTTTCACGCTCATGTCACGG 7280
      7210      7220      7230      7240      7250      7260      7270      7280
            ***                                                          S7

  R  G  T  A  E  E  K  T  A  K  S  D  P  I  Y  R  N  R  L  V  N  M  L  V  N  R  I
CGAGGTACTGCAGAAGAAAAAACCGCAAAATCCGATCCAATTTATCGTAATCGATTAGTTAACATGTTGGTTAACCGTAT 7360
       PstI

  L  K  H  G  K  K  S  L  A  Y  Q  I  I  Y  R  A  M  K  K  I  Q  Q  K  T  E  T
TCTGAAACACGGAAAAAAAATCATTGGCTTATCAAATTATCTATCGAGCTATGAAAAAGATTCAACAAAAGACAGAAACAA 7440

  N  P  L  S  V  L  R  Q  A  I  R  G  V  T  P  D  I  A  V  K  A  R  R  V  G  G  S
ATCCACTATCTGTTTTACGTCAAGCAATACGTGGAGTAACTCCCGATATAGCAGTAAAAGCAAGACGCGTAGGCGGATCA 7520

  T  H  Q  V  P  V  E  I  G  S  T  Q  G  K  A  L  A  I  R  W  L  L  G  A  S  R  K
ACTCATCAAGTTCCCGTTGAAATAGGATCCACACAAGGAAAAGCACTTGCCATTCGTTGGTTATTGGGGGCATCCCGAAA 7600
                        BamHI

  R  P  G  R  N  M  A  F  K  L  S  S  E  L  V  D  A  A  K  G  S  G  D  A  I  R
ACGTCCGGGTCGAAATATGGCTTTCAAATTAAGTTCCGAATTAGTGGATGCTGCCAAAGGTAGTGGCGATGCCATACGCA 7680
    HpaII

  K  K  E  E  T  H  R  M  A  E  A  N  R  A  F  A  H  F  R  *
AAAAGGAAGAGACTCATAGAATGGCAGAGGCAAATAGAGCTTTTGCACATTTTCGTTAATCCATGAACAGGGATTCTATA 7760
      7690      7700      7710      7720      7730      7740

TAGACACATAGACCCATGATCCTACATCTCGATCGAAAAAGAATCGAAATTGATGAATGTATCTCTCGAAACAAACGAAA 7840
                              PvuI

AGGAAACGAAAAACAAAAACATAAATCATGGATCAATTAAGCCTCTCGGGACTTGCTTAAGAATGAGAAAGAGGAATCTT 7920
ATGTAAATACCACAGAATAAGGTTTGGTCCTATTCATGGGGATTCTGTAAATATTCCATTCCAAAAAGAGAAAAAATTCG 8000
AAACAATTGGGATTTTTTGGAGATTGGATGCAGTTACTAATTCATGATCTGACATGTACAGAATGAAAACTTCATTCTCG 8080

                  M  K  A  F  H  L  P  L  F  D  G  S  F  I  F  P  E  C  I  L  I
ATTCTACAAGAATTTTTATGAAAGCCTTTCATTTGCCTCTCTTCGATGGAAGTTTTATTTTCCCAGAATGTATCCTAATT 8160
      8090      8100
                > ndhB Premier Méthionine possible

  F  G  L  I  L  L  L  M  I  D  S  T  S  D  Q  K  D  I  S  W  F  Y  F  I  S  S  T
TTTGGCCTAATTCTTCTTCTGATGATTGATTCAACCTCTGATCAAAAAGATATATCTTGGTTCTATTTTATCTCTTCAAC 8240

  S  L  V  M  S  I  T  A  L  L  F  R  W  R  E  E  P  M  I  S  F  S  G  N  F  Q
AAGTTTAGTAATGAGCATAACGGCCCTATTGTTCCGATGGAGAGAAGAACCTATGATTAGCTTTTCGGGAAATTTCCAAA 8320
```

```
T  N  N  F  N  E  I  F  Q  F  L  I  L  L  C  S  T  L  C  I  P  L  S  V  E  Y  I
CGAACAATTTCAACGAAATCTTTCAATTTCTTATTTTACTATGTTCAACTCTATGTATTCCTCTATCCGTAGAGTACATT 8400

 E  C  T  E  M  A  I  T  E  F  L  L  F  I  L  T  A  T  L  G  G  M  F  L  C  G  A
GAATGTACAGAAATGGCTATAACAGAGTTTCTGTTATTCATATTAACAGCTACTCTAGGAGGAATGTTTTTATGCGGCGC 8480
      8410  > ndhB début du gène proposé (Shinozaki et al. 1986)

  N  D  L  I  T  I  F  V  A  L  E  C  F  S  L  C  S  Y  L  L  S  G  Y  T  K  K
TAACGATTTAATAACTATCTTTGTAGCTCTAGAATGTTTCAGTTTATGTTCCTATCTACTATCTGGATATACCAAGAAAG 8560

D  V  R  S  N  E  A  T  T  K  Y  L  L  M  G  G  A  S  S  S  I  L  V  H  G  F  S
ATGTACGGTCTAATGAGGCTACTACGAAATATTTACTCATGGGTGGGGCAAGCTCTTCTATTCTGGTTCATGGTTTCTCT 8640

 W  L  Y  G  S  S  G  G  E  I  E  L  Q  E  I  V  N  G  L  I  N  T  Q  M  Y  N  S
TGGCTATATGGTTCATCCGGGGGAGAGATCGAGCTTCAAGAAATAGTGAACGGTCTTATCAATACACAAATGTATAACTC 8720

  P  G  I  L  I  A  L  L  F  I  T  V  G  I  G  F  K  L
CCCAGGAATTTTAATTGCACTTTTATTCATCACTGTAGGAATTGGGTTCAAGCTT 8775
                                                HinDIII
```

Légende : Les régions en minuscules sont des séquences de tabac, mis à part celle commençant à la position (1110) qui est tirée du maïs pour des raisons de similarités dans la région concernée. Nous n'avons pas cherché à remplir les trous de cette séquence, le groupe du professeur Pillay y travaillant.

Un travail identique a été fait pour la région du *5'-rps12*. Cette fois ci la séquence provenait directement du fichier CHNTXX.DAT; de l'EMBL, qui est la séquence chloroplastique de tabac complète, annoncée par Shinozaki *et al.* (1986). Pour nous éviter de devoir changer les formats très standardisés des fichiers de l'EMBL, nous avons alors modifié le programme ANALYSE pour que ce type de fichier soit directement utilisable. Une mémoire de l'ordinateur Vax ne pouvant contenir que 65535 caractères, nous avons dû moduler le type d'entrée du programme en fonction du type de fichier. Quelques problèmes annexes comme la modification instantanée des "U" des fichiers d'ARN en "T" sont traités. Pour avoir une idée des problèmes qui peuvent se poser dans ce domaine, nous donnerons l'exemple suivant :

Les fichiers de l'EMBL contiennent 60 nucléotides par ligne. Un fichier contenant la séquence d'un ADNct contient de 120'000 à 160'000 nucléotides. Si nous voulons analyser une séquence à partir de la position 65000 par exemple, nous devons éviter de charger en mémoire les 65000 premiers caractères de la séquence. Nous calculons donc le nombre de ligne de 60 caractères à éviter et calculons la position du premier nucléotide à mettre en mémoire de la première ligne effectivement intéressante pour nous. La résolution de ce problème nous conduit alors au petit calcul suivant :

$$POS_1 = 60 \cdot \left(\frac{F_INF - 1}{60} - INT\left(\frac{F_INF - 1}{60}\right)\right) + 1$$

POS_1 : Position n°1 à mettre en mémoire
F_INF : Position inférieure dans le fichier à analyser
INT( ) : Partie entière de ( )

Ce calcul qui est absolument juste lorqu'on calcul à la main peut donner quelques surprises ! Si nous remplaçons F_INF par 60'000. L'ordinateur va trouver POS_1 = 59.9998 : Comme le fichier ne peut pas être coupé à une fraction de caractère, c'est le 59ème caractère et non le 60ème qui sera le premier nucléotide de la chaine analysée. Nous avons donc dû tester les erreurs d'arrondi avant de faire le découpage du fichier. Sinon, le repérage des gènes aurait été décalé d'une position dans les quelques cas où l'erreur d'arrondi due au manque de précision de l'ordinateur donne une valeur légèrement inférieure à la position réelle.

D'autres problèmes typiquement informatiques ont été résolus durant la période de conception des programmes. Nous avons entre autre multiplié par 10 la vitesse d'exécution du sous-programme DIAGON, par la mise au point d'algorithmes personnels.

L'analyse du *5'-rps12* à été grandement facilitée par l'emploi des programmes. Nous avons séquencé en partie le fragment HinDIII de 4.4 kb, mettant en évidence la présence du gène *psbB* codant pour une protéine faisant partie du centre réactionnel P680 du photosystème II, le gène *rpl20*, dont il nous manque malheureusement encore les 9 premiers codons et qui porte l'information pour la protéine ribosomique L20 (Figure 24, page 64) et finalement la partie *5'-rps12*.

Figure 11 : Séquence partielle du fragment HinDIII de 4.4 kb et du fragment BamHI de 6.6 kb. Séquence du *5'-rps12*.

```
GGTTATTTAATATTATTTTTGCCCATTTTTTTATTTTCTTATTTCTATTTATCCCTTTTACCTTTATTTGGAATAAGCAT 80
AATCTTTACCAATATAGGATAAATCATTATCAAAATCTTTATCAAGATAAGGGAAATACTTATTAAGTTATAAAATCAGG 160
GTAATGATCCACCAACCCGCTAGTTCTTTTTATGAGGCACAAACGGGACCATTTATCCTGGAAGCGGAAGAGCTGCTGAA 240
                                                                        AluI

AATGCGTGAAACTATCACAAGGGTTTATGTACAAAGAACGGGCAAACCTTTATGGGTTATATCCGAAGACATGGAAAGGG 320
AATGTTTTTATGTCAGCAGCAGAAGCCCAAGCTCACGGAATTGTAGATTCTTGTAGCCGTTGAATAAAAAATAAGTGGCA 400
                          AluI

AGGATTATTCTAAAAAAATACAGGATTTGAAATTTCATTTTTGAATCTTCTTACTTTCATTTTAATATAATATCTCTATT 480
AGTTAATCTATTAATATCTATTAGTTAATCTATTAATATAAGTAATATAGAATCTAAGTAATTCACGGTTAGGATAGATC 560
                                                                    BglII

                                    M  P  T  M  K  Q  L  I  R  N  T  R  Q  P  I
TAAACCAGCTCATTATATATATATTCACAATTCATATGCCAACTATGAAACAACTTATTAGAAACACAAGACAGCCAATC 640
      570       580       590       600       610       620       630       640
     AluI

 R  N  V  T  K  S  P  A  L  R  G  C  P  Q  R  R  G  T  C  T  R  V  Y  V  R  L  V
CGAAACGTCACAAAATCCCCAGCTCTTCGGGGATGCCCTCAGCGCCGAGGAACGTGTACCAGGGTGTATGTGCGACTCGT 720
      650       660       670       680       690       700       710       720
                  AluI                                                 Trans splicing

  Q  I  I  Y  *
TCAGATCATATACTAAGAAGAAAAAACCAATTTCATTTTTTCAGTATTGTTGATCGGTTAAGATAGTGTGAACTGTAGTT 800

TTCCCATTAGACTATCTTAACTTATATATATACATTCTTTACATTCTTCTATCTTGTATCAAATTTATGGTTTCGATTGG 880
```

Nous avons ainsi le gène *rps12* complet et un problème nouveau : Savoir si le gène est exprimé et comment il est possible qu'un "intron" d'une taille supérieure à la moitié de l'ADN chloroplastique puisse être supprimer. Remarquons que le problème est plus aiguë pour *Marchantia*, qui possède une seule copie du *3'-rps12* et sur le brin complémentaire à la partie 5' du gène (Fukuzawa *et al.* 1986). Nous avons donc déjà cherché les promoteurs, avant de nous intéresser à la transcription proprement dites.

3.1.9 Recherche des promoteurs

Une fois la séquence établie, commence le travail d'analyse des résultats découlant de la séquence. La recherche des promoteurs est l'une des étapes de cette analyse. Deux approches sont possibles. L'approche classique consiste à chercher les séquences similaires aux consensus des promoteurs d'*E. coli* situés en "-35" et en "-10" du site d'initiation de la transcription (Harley et Reynolds, 1987). Une méthode indépendante et complémentaire (Tung et Harvey, 1987), tenant compte de la présence de structures dont les angles d'hélicité ("helix twist angle") sont de 36.2°, 37.8°, 26.4°, 37.8°, 36.2° en succession et avec une tolérence de $S^2 \leq 5.0\ deg^2$, a été testée sur nos séquences. La La succession de ces angles pouvant indiquer la région -55 à +5 des promoteurs de la bactérie *E. coli* selon les auteurs.

L'analyse des angles d'hélicité que nous avons faite, en cherchant toutes les séquences possédant les angles corrects et énumérées par Tung et Harvey (1987) nous a paru décevante, dans le sens où trop de sites possibles sont observés, même à l'intérieur des gènes. Nous pouvons difficilement mettre en relation cette analyse avec les promoteurs, d'autant plus que nous n'avons pas testé expérimentalement la position réelle des sites de début de transcription. Du moins, si les promoteurs éventuels que nous localisons au moyen de leur ressemblance avec les promoteurs d'*Escherichia coli* sont fonctionnels, il se pourrait que l'ARN polymérase du chloroplaste pourrait peut être avoir une spécificité un peu différente que celle d'*Escherichia coli*.

Nous nous sommes alors contenté de l'approche classique et avons cherché les séquences caractéristiques "-35" et "-10". Dans la figure suivante nous proposons des promoteurs pour les unités de transcriptions du *trnI* de l'opéron ribosomique et du *3'-rps12*. Nous les comparons avec d'autres promoteurs énumérés dans la littérature. Remarquons qu'il est connu depuis longtemps chez le maïs (dont la séquence est très proche du soja) qu'il existe des sites d'initiation de la transcription situés en 5' de l'ARN 16S et du *trnV*. Il existe chez cette plante une co-transcription de (*trnV*), *16S*, *trnI*, *trnA*, *23S*, *4.5S* et *5S* (Strittmatter et Kössel, 1984; Strittmatter *et al.* 1985).

Figure 12 : Comparaison des promoteurs présomptifs avec quelques séquences publiées.

Organisme	Gène	"-35"	"-10"	Distance du gène	Réf.
E.coli	consensus	TTGACA NNNNNNNNNNNNNNNNN	TATAAT		1
E.coli	trn Tyr	AACACTTTACA GCGGCGCGTCATTTGA	TATGATGCGCCCCG	41	2
G.max	trnV,p2	TGGCGTTGAGT TTCTTGACCCTTTGCCT	TAGGATTAGTCAGT	27	-
Z.mays	trnV,p2	TGGCCTTGCGT TTCTCGCCCCTTTGCCT	TAGGATTCGTTAAT	31	3
N.tabacum	trnV	TGGCGTTGAGT TTCTCGACCCTTTGACT	TAGGATTAGTCAGT	31	4
G.max	trnV,p1	AGCAGTTGACA ATGGAATCCAATTTTTCC	CATAATTTTCGTAT	151	-
E.coli	rrnA,p1	CTCCTTTTTGT CAGGCCGGAATAACTCC	TATAATGCGCCACC'A	300	5
E.coli	rrnA,p2	AATGCTTGACT CTGTAGCGGGAAGGCG	TATTATGCACACC'C	181	5
S.oleracea	rrn,p1	CTATTTTGACT TGCTCCCCCGCCGTGA	TTGAATGAGAATGAATA'	175	6
S.oleracea	rrn,p2	TGGGATTGACG TGAGGGGGTAGGGATGGC	TATATTTCTGGG'A	152	6
E.coli	str	ATTTCTTGACA CCTTTTCGGCATCGCCC	TAAAATTCGGC'G	64	7
G.max, p1	3'-rps12	ATTTCTTGTTC TTATTAGAAGATCGTGA	TTGGATCCACATAT	277	-
G.max, p2	3'-rps12	TATGTTTGGTA AAAAGAAAGAAAAATAT	TCTTCTTTGAGAAT	238	-
Z.mays	rps4	TTTGCTAGATA CAAAGCATAAAAATAAA	TAATATATATACAA'A	190	8

(1) Siebenlist et al. (1980)
(2) Berman et Landy (1979)
(3) Schwarz et al. (1981)
(4) Tohdoh et al. (1981)
(5) Lindahl et Zengel (1986)
(6) Lescure et al. (1985)
(7) Post et Nomura (1980)
(8) Russell et Bogorad (1987)

Nous avons ainsi un catalogue de séquence qui pourraient avoir le rôle de promoteurs pour les gènes étudiés.

Une des questions suivantes à résoudre est de tester si les différents gènes localisés sont transcrits ou non. En effet, il ne suffit pas qu'un ADN porte une information pour qu'il soit transcrit en un ARN messager. Nous avons donc parallèlement commencé l'étude de la transcription des gènes *rps7*, *rps12*, *rpl20* et *ndhB*.

3.2 Les ARN messagers

Si l'isolation des ADN chloroplastiques de soja pose déjà parfois de sérieux problèmes, car certaines préparations ont une fâcheuse tendance à ne pas se laisser digérer, les préparations d'ARN sont peut-être encore plus fastidieuses car nous avons dû tout d'abord, préparer une isolation d'ARNct (50 ug d'ARN) pour 1 à 2 Northern §(2.3.1.2). Nous comprendrons aisément que dans ces conditions, la préparation du matériel de base prenne énormément de temps.

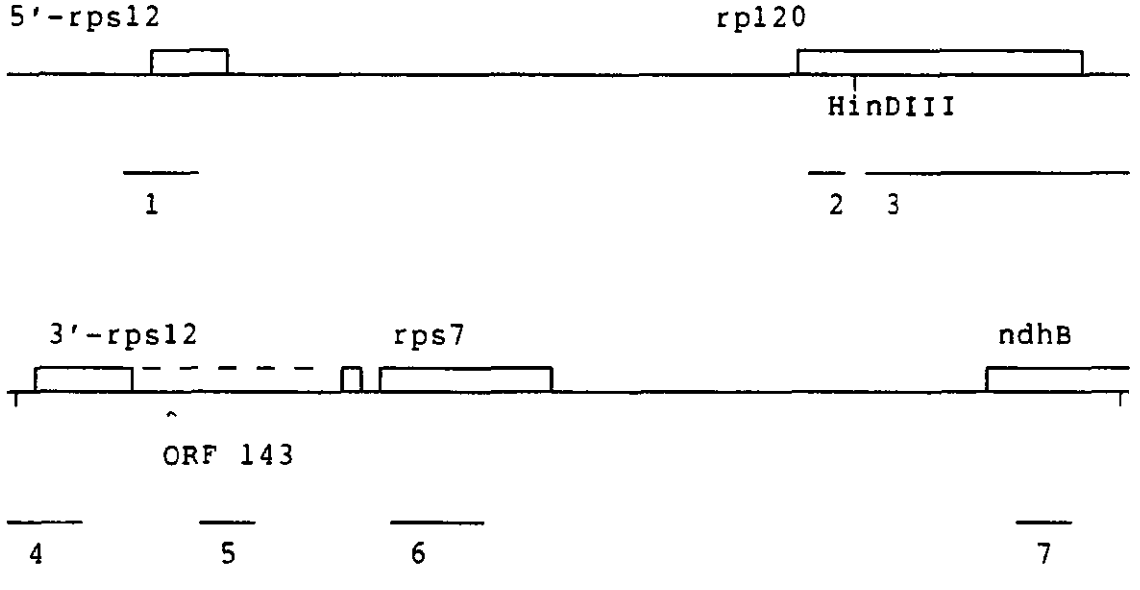
3.2.1 Hybridations du type Northern

Pour mettre en évidence les messagers, nous avons utilisé des sondes simples brins préparées selon la méthode décrite au §(2.2.7.3). Une série de clones simples brins intégrés dans le vecteur M13 ont été utilisés comme sonde. Les clones suivants qui nous avaient servi à déterminer la séquence, nous ont servi de sondes :

(1) HinD_4.4, sous-fragmemt Alu (5)	1er exon du *rps12*
(2) HinD_4.4, Alu (17)	*rpl20*
(3) BamH_6.6, HindIII (9)	*rpl20*
(4) BH101, HaeIII/HpaII (27)	2ème exon du *rps12*
(5) BH101, HinFI (46)	Intron 2, (ORF 143)
(6) BH101, Sau3A (69)	*rps7*
(7) PK, Alu (4)	*ndhB*

Nous indiquons dans la Figure 13, la situation exacte des sondes par rapport aux gènes. Rappelons que les sondes n'ont ici qu'un seul brin marqué et que le brin complémentaire froid a été éliminé lors de la préparation de la sonde.

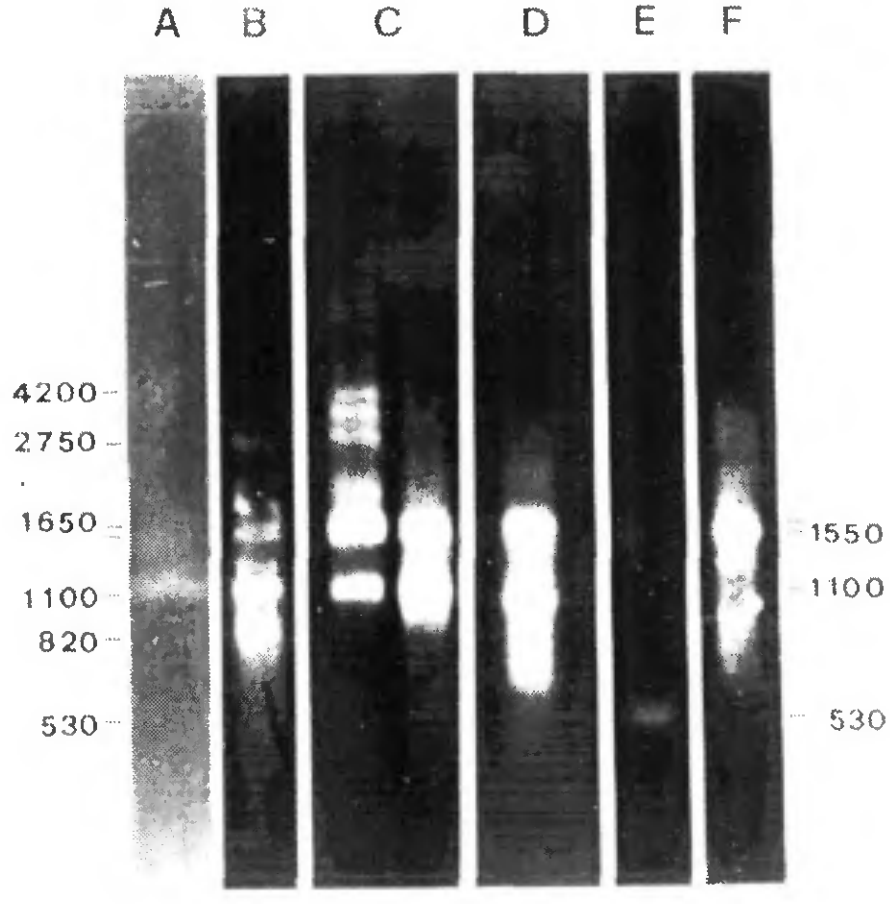
Figure 13 : Description et localisation des sondes



Nous avons utilisé les sondes 1,2,4,5,6 et 7 pour tenter de d'analyser la situation plutôt compliquée qui existe au niveau de la transcription du soja. La Figure 14 présente les résultats obtenus.

Figure 14 : Hybridations de différentes sondes contre les ARN chloroplastiques. Les tailles sont indiquées par référence aux tailles connues des ARN 16S et 23S d'*Escherichia coli*

A : sonde (2), interne au *rpl20*
B : sonde (1), interne à l'exon 1 du *rps12*
C : sonde (4), interne au *rps12*, ARN-m récoltés des jours différents
D : sonde (6), interne au *rps7*
E : sonde (5), interne à l'intron 2 du *rps12*
F : sonde (7), interne au *ndhB*



Nous obtenons les résultats suivants :

1°) Chaque sonde hybride généralement avec plusieurs ARN messagers. Il en résulte soit que la maturation des ARN est particulièrement complexe et que nous observions des produits intermédiaires de cette maturation, et/ou qu'il existe des ARN avec plusieurs extrémités 5' ou 3' possibles. Les résultats présentés étant insuffisant pour prendre position.

2°) La sonde interne à l'exon 2 du *rps12* hybride avec des messagers relativement abondants dont les tailles (1100 et 1650 nucléotides) sont similaires au produit de l'hybridation avec le *rps7*. Ce résultat suggère fortement qu'il existe une co-transcription du *3'-rps12* avec le *rps7*.

3°) L'intron 2 du *rps12* (530 nucléotides) contenant l'ORF est enlevé avec une très grande efficacité. Il a fallu un temps d'exposition de 3 semaines avec écran intensifiant pour mettre en évidence autre chose que la bande principale de 530 nucléotides et correspondant à l'intron coupé.

4°) L'exon 1 est en commun avec une série de messagers. Nous observons des messagers de 2750, 2100, 1900, 1500, 1300, 1100 et 830 nucléotides. Le messager mature, contenant les exons 1+2+3 (seule bande forte en commun avec les hybridations contre l'exon II) à une taille de 1100 nucléotides. Le *trans* splicing semble donc d'une efficacité réduite.

5°) la sonde interne au *ndhB* allume des bandes dont la taille ne correspond pas aux transcrits des gènes *rps7* et *rps12*. Il existe probablement une autre unité de transcription pour ce gène. Les ARN-m principaux observés ont 1300 et 1550 nucléotides.

6°) Un messager de 1100 nucléotides hybride avec le *rpl20*. Ce résultat peut suggérer une co-transcription avec l'un des précurseurs du *5'-rps12*, ou une correspondance due au hasard entre la taille de ce précurseur et la taille du messager final.

Une rapide comparaison avec les résultats obtenus avec le tabac (Koller *et al.* 1987; Zaita *et al.* 1987) donne une bonne correspondance entre les deux plantes. Les analyses au microscope électronique indiquent que la partie *3'-rps12* et le *rps7* sont cotranscrits et qu'il existe des précurseurs de grandes tailles (2.7 Kb) contenant la partie *5'-rps12* et le *rpl20*. Nous observons également une taille légèrement supérieure de certains transcrits de tabac, mais nous reviendrons sur ce point plus tard.

Le résultat de l'hybridation avec l'exon 2 du *rps12* pose en effet le problème de vérifier la qualité des ARN en fonction du moment où les feuilles sont cueillies. Nous avons constaté une nette différence dans le contenu en ARN messagers, selon le jour de la cueillette des feuilles.

3.2.2 Distribution des ARN messagers en fonction de l'état physiologique de la plante.

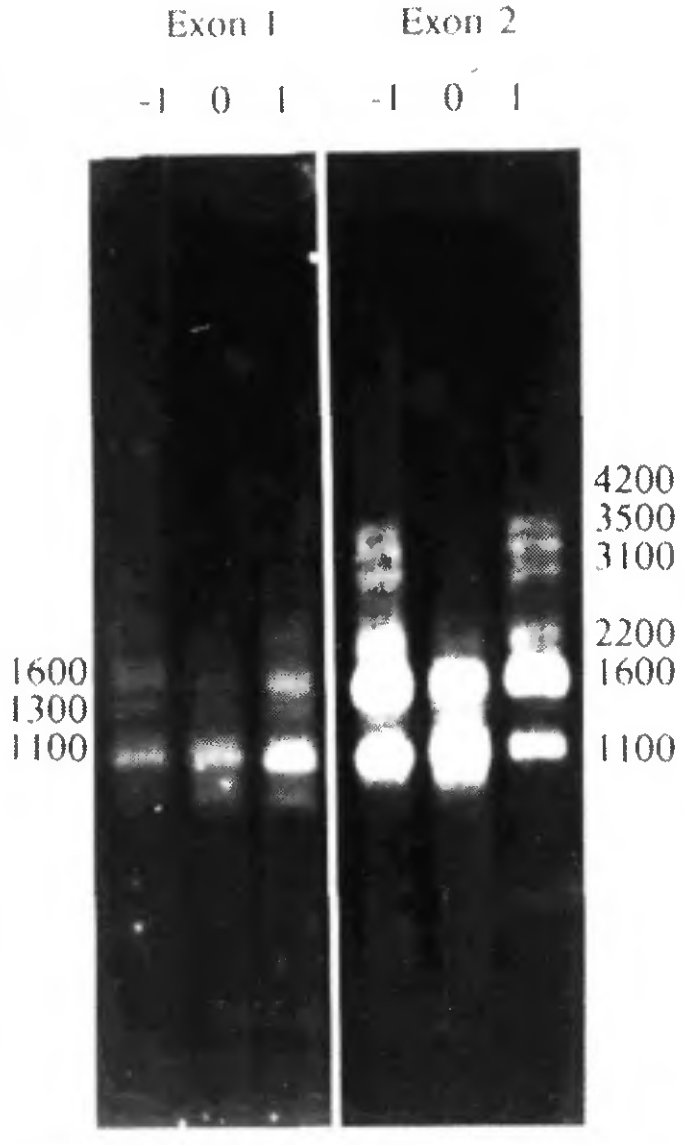
Nous nous sommes ainsi intéressé à la transcription en fonction du temps pendant lequel le soja a poussé. Une telle étude est rendue possible par l'emploi du phytotron dans lequel des conditions contrôlées permettent d'obtenir des états de croissances reproductibles. Nous avons récolté des feuilles, les 15ème, 16ème et 17ème jours de croissance à partir du jour où les graines sont mises à gonfler.

Nous présentons dans la Figure 15 une photo des feuilles, qui permet de reconnaître les différents stades visuellement. La lumière fournie par les néons pouvant varier légèrement et l'état des feuilles pouvant varier lui aussi quelque peu, nous parlerons des jours (J-1), (J) et (J+1). Le point de repère que nous avons choisi est le jour (J), où les 2 premières feuilles sont largement dépliées et les deux feuilles suivantes sont en pleine croissance : En train de se déplier et en début de verdissement.

Figure 15 : Photographie des stades : jour J-1, J et J+1
Nous remarquerons la grande différence de taille des feuilles à 24 heures d'écart.

L'étude de la transcription en fonction du jour de la récolte des feuilles nous fournit des résultats intéressants dans le sens où le contenu en ARN précurseurs et matures varie. Nous avons chargé 50 μg d'ARN chloroplastique dénaturé au glyoxal par poche d'un gel 1.2% agarose, effectué l'électrophorèse et transféré l'ARN par Northern. Les résultats en fonction du jour de la récolte des feuilles (-1, 0 ou 1) sont présentés dans la Figure 16. Il est particulièrement intéressant de noter que notre analyse est non seulement qualitative, mais aussi semi-quantitative. Tous les ARN ont été isolés par la même méthode d'isolation des ARN et une quantité identique d'ARN chloroplastique a été chargée sur le gel. Cela nous permet de tirer plus de renseignements de cette expérience.

Figure 16 : Transcription en fonction de l'état physiologique. Hybridation de 50 μg d'ARN chloroplastique de soja récolté le jours (-1), (0) et (1) transféré sur filtre nitrocellulose contre l'exon 1 (sonde 1) et l'exon 2 (sonde 4) du *rps12*.



Nous constatons que le splicing semble avoir une efficacité maximale le jour J (0) et que les feuilles une fois constituées, le *trans* splicing redevient moins efficace. Alors que les observations de Deng et Gruissem (1987) indiquent que la régulation de l'initiation de la transcription a souvent (mais pas toujours) un rôle mineur dans le contrôle de l'expression des gènes chloroplastiques, il n'est pas impossible que dans le cas du soja, il existe un phénomène intéressant de régulation du processus de maturation des ARN qu'il faudrait étudier de plus près. Nous observons la disparition de deux précurseurs de l'exon 1 (de 1650 et 1300 nucléotides) le jour J alors que, simultanément, cinq précurseurs de la partie 3' du gène (4200, 3500, 3100 et 2200 nucléotides) disparaissent ou (1650 nucléotides) diminuent probablement au profit du messager de 1100 nucléotides. En effet, la bande de 1100 nucléotides est visiblement plus forte le jour J, en ce qui concerne son hybridation face à l'exon 2 du *rps12* malgré le fait que la quantité totale d'ARN chloroplastique est identique à celle des autres jours.

La comparaison de nos résultats avec les données connues concernant la transcription du *rps12* de tabac (Zaita *et al.* 1987) nous donne quelques indications supplémentaires. Les ARN messagers hybridant avec les exons 1 et 2 mis en évidence pour le tabac ont des tailles (2.4 et 1.5 Kb) légèrement plus grandes que les ARN-m de soja (1.1 et 1.65 Kb). Un messager de 1.1 Kb est mis en évidence par la sonde interne à l'exon 1 chez le tabac, qui pourrait correspondre à la bande de 830 nucléotides que nous observons. Les profiles d'hybridations de tabac ressemblent à la situation chez le soja lorsque les feuilles sont cueillies le jour (J) et les précurseurs en dessus de 2.4 Kb ne sont observés que par microscopie électronique (Koller *et al.* 1987).

L'analyse des ARN-m chloroplastiques de soja suggère que l'un des ARN a une relativement bonne évidence d'être un ARN-m mature possédant les exons 1+2+3 et qu'il s'agit du messager de 1100 nucléotides. La bande observée lors de l'hybridation (avec la sonde interne à l'exon 2) se renforçant semble-t-il au dépend de l'ARN de 1650 nucléotides.

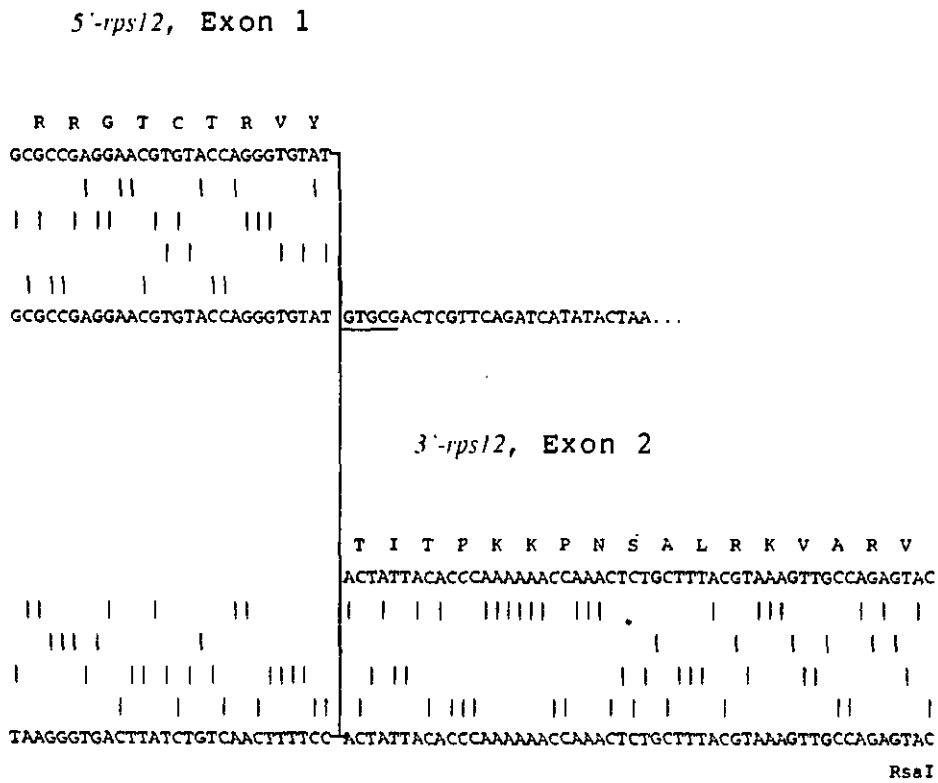
3.2.3 Détermination du site de splicing par "Primer extension"

Une méthode de choix pour déterminer la séquence au niveau du site de splicing, est la méthode du §(2.3.2.1). Nous l'avons appliquée en utilisant des ARN d'origines chloroplastique. Cette préparation des ARN, permet d'éviter la présence de la plupart des ARN d'origine nucléaire. Il n'est malheureusment pas possible de séparer les messagers chloroplastiques des ARN ribosomiques et de transferts. En effet, il n'existe pas de queue de poly A au niveau des messagers chloroplastiques comme c'est le cas avec les messagers nucléaires. Cet handicap demande de jouer sur les concentrations d'ARN et de transcriptase inverse présents dans la réaction.

Nous avons obtenu un résultat mitigé, lorsque nous avons utilisé des ARN chloroplastiques totaux et une amorce RsaI-HaeIII isolée par nos soins et selon la méthode de préparation des sondes qui nous est habituelle. L'amorce a été choisie interne à l'exon 2. La séquence de cet exon est lisible, mais, dès que la séquence devrait appartenir à l'exon 1, la séquence devient illisible (Figure 17). Comparant les artefacts obtenus avec ce que nous devrions obtenir si un mélange d'ARN messagers précurseurs et matures est présent, nous obtenons une correspondance relativement bonne entre la séquence théorique du mélange et ce que nous avons séquencé par primer extension. Remarquons que les fortes concentrations d'ARN identifiées par Northern lors des hybridations contre l'exon 2 du *rps12*, (messagers de 1650 et 1100 nucléotides) plaident pour l'hypothèse du mélange d'ARN précurseurs et matures. Le signal d'hybridation obtenu en utilisant la sonde interne au premier exon est beaucoup plus faible que celui des messagers contenant l'exon 2.

Nous présentons dans la Figure 17 le résultat de la "primer extension" face à une superposition des réactions de séquence que nous devrions observer en présence d'un mélange d'ARN ayant subit et n'ayant pas subit le *trans* splicing.

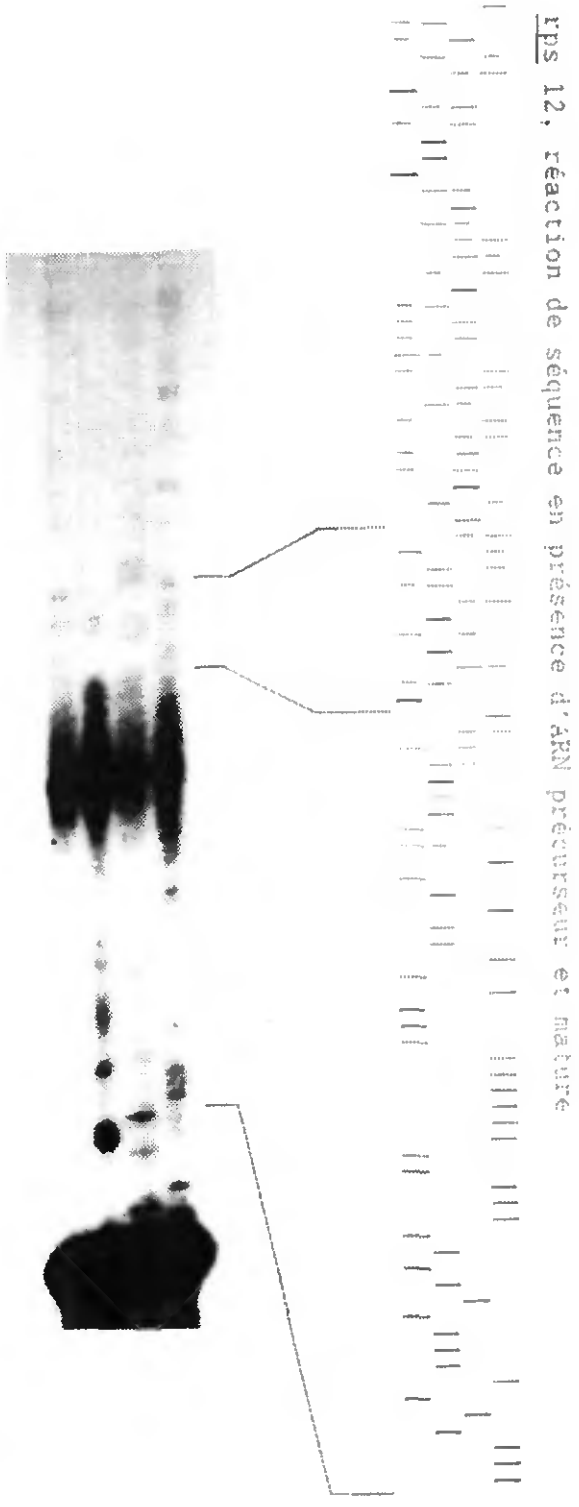
En effet, deux séquences distinctes peuvent être situées en amont de l'exon 2 du rpsl2 (*3'-rps12*) :



Si un mélange de molécules d'ARN précurseur et mature sont séquencées, nous devons alors observer une superposition des deux séquences.

Nous obtenons expérimentalement les résultats suivants :

Figure 17 : Séquence par "primer extension" des messagers du *rps12* à partir du site RsaI et en direction de la région 5' du gène. Comparaison avec une réaction théorique en présence de précurseurs et d'ARN matures.



Si nous avions le temps et le matériel pour faire des expériences supplémentaires, nous suggérerions de fractionner les ARNct en fonction de leurs taille. Nous effectuerions une primer extension, en utilisant une sonde simple brin, synthétisée chimiquement et complémentaire à une séquence plus proche du site de splicing. En fabriquant les cDNA complémentaires au fractions contenant les ARN de 1100 et de 1650 nucléotides, nous pourrions vérifier si effectivement le précurseur ayant subi le *trans* splicing est l'ARN de 1100 nucléotides et si l'ARN de 1650 est un précurseur de l'unité de transcription du *3'-rps12*.

Une telle expérience, si elle semble évidente, pose toute fois un certain nombre de problèmes, de par la quantité d'ARN dont il faut disposer. Les rendements des isolations d'ARN chloroplastiques sont faibles et nous calculons que pour obtenir un résultat, en tenant compte des perte encourues au cours de la purification de tels ARN, il faut environ l'équivalent de 8 phytotrons pleins pour une expérience de "primer extension". Ainsi, dans les conditions d'équipement actuelles du laboratoire, la récolte des feuilles de soja prend déjà environ 16 semaines... Et il faudrait compter 16 isolations d'ARN à partir de 200 g de feuilles. De plus, il faudrait éluer le matériel à partir d'un gel dénaturant (Reijnders *et al.* 1973), en prenant des bandes de largeur minimale de manière à éviter d'avoir trop d'ARN indésirables, tels les ARN ribosomiques et les ARN de transfert.

En effet, la transcriptase inverse peut faire l'élongation d'un cDNA à partir d'une amorce de type ARN, ce qui permet la synthèse d'un cDNA à partir du repliement d'un ARN sur lui-même, lorsque celui-ci à une structure secondaire adéquate. Une autre possibilité serait d'étudier le *trans* splicing sur un matériel plus facile à isoler, tel le tabac.

Au niveau du soja, nous sommes convaincu que les ARN du *rps12* subissent du splicing et du *trans* splicing. Nous pourrions étudier par hybridation quels sont les messagers ayant subi le splicing au niveau de l'intron 2 et de l'intron 1 (transon) en construisant artificiellement des sondes contenant 7 nucléotides de part et d'autre des sites de splicing et en les hybridant contre les ARN. Nous pourrions recommencer les expériences de "primer extension" ou nous lancer dans la construction d'une banque de cDNA en utilisant une amorce synthétique et interne au *rps12*.

Toutes ces expériences et d'autres encore pourraient nous aider à comprendre le "comment" du splicing. Remarquons que les résultats sur le tabac (Zaita *et al.*, 1987; Koller *et al.* 1987) semblent prouver que le *trans* splicing *in vivo* des messagers du *rps12* est effectif. Dans le cas du tabac, la réaction de primer extension à donné une séquence lisible, ce qui tends à prouver que l'adjonction de l'exon 1 aux exons 2+3 est éventuellement plus efficace avec le tabac dans les conditions particulières de culture de Zaita *et al.* (1987). Remarquons que les résultats de Koller *et al.* (1987) font plutôt penser qu'il existe simultanément des ARN précurseurs et matures.

La seule véritable évidence qu'un *trans* splicing à lieu pour lier l'exon 1 à l'exon 2 est donné par (Koller *et al.* 1987) Plusieurs méchanismes sont théoriquement possibles pour obtenir un ARN *rps12* fonctionnel.

La première possibilité serait qu'il existe un transcrit primaire contenant les 3 exons du *rps12*, mais cela implique la formation d'un précurseur d'environ 130 Kb et que l'intron 1 ainsi formé serait codant et/ou un ARN anti-sens pour beaucoup de gènes. Un précurseur de cette taille à peu de chance d'être mis en évidence avec les techniques d'isolation des ARN actuelles. Mais un tel méchanisme impliquerait que le gène *rps12* de *Marchantia polymorpha* est non fonctionnel, car la partie *5'-rps12* n'est pas codée par le même brin d'ADN que la partie 3' du gène (Fukuzawa *et al.* 1986).

La seconde possibilité serait une *trans* addition par ligation de l'extrémité 3' du transcrit de l'exon 1 avec l'extrémité 5' du transcrit de l'exon 2. Un tel mécanisme serait alors suivi par un (cis-) splicing conventionnel (Milhausen *et al.* 1984; Campbell *et al.* 1984). Mais Koller *et al.* (1987) n'a pas observé au microscope électronique de molécules hybridant à la fois avec des sondes en 3' de l'exon 1 et en 5' de l'exon 2.

La possibilité qui semble donc avoir le plus de chance d'être réelle est le *trans* splicing. Un splicing où les exons liés proviennent de deux molécules d'ARN différentes et ajoutés en *trans*.

Nous préciserons quant à nous, certaines évidences structurales permettant de mieux délimiter les introns et ferons certaines observations à propos des gènes chloroplastiques que nous avons séquencés dans la discussion sur le splicing §(3.4).

3.2.4 transcription constitutive ou induite ?

Un élément supplémentaire concernant la transcription est de déterminer si les gènes des protéines séquencées sont exprimés de manière constitutive ou induis par la lumière. Il est certain que le moyen le plus direct de déterminer de quel type est un gène est de tester la quantité des ARN produits avant et après exposition de la plante à la lumière.

Il est vrai pourtant que l'analyse de la structures de l'ADN pour quelques gènes dont la transcription était connue a permi à Gingrich et Hallick (1985) de formuler l'hypothèse suivante :

L'usage des codons semble être différent si le gène est constitutif ou induit par la lumière. Le rapport (T+C)/(A+G) en troisième position du codon est plus élevé dans le cas des gènes étudiés et qui sont induits par la lumière que pour les gènes exprimés de manière constitutive.

Voulant tester le résultat d'une analyse opérée selon les critères de ces auteurs et conscient que les résultats obtenus sont hypothétiques, nous avons cherché les rapports (T+C)/(A+G) pour quelques gènes de soja. Nous indiquons également les rapports trouvés pour les gènes correspondants d'Euglène et qui étaient initialement étudiés par Gingrich et Hallick (1985).

Figure 18 : Usage des codons et mode d'expression

Ac. aminé	Induits par la lumière				Expression constitutive				?	
	psbA E.g		psbA G.m		TufA-rps12 rps7 E.g		rps12-rps7 ndhB G.m		ORF 143 G.m	
	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G
Leu (CT-)	11	1	6	8	7	9	8	10	3	5
Ile (AT-)	18	2	30	2	32	33	21	9	5	5
Ser (TC-)	13	4	14	3	10	15	18	4	5	6
Pro (CC-)	12	8	12	3	15	18	6	8	1	4
Thr (AC-)	19	14	15	1	16	37	15	11	5	3
Ala (GC-)	35	9	27	8	18	29	19	12	2	1
Arg (CG-)	24	1	10	2	12	28	13	11	1	2
Gly (GG-)	38	12	25	8	11	39	13	24	3	6
Total	170	51	139	35	121	208	113	89	25	32
T/C:A/G	3.3		3.9		0.58		1.2		0.7	

Exprimé selon Gingrich et Hallick, (1985)
psbA d'*Euglena gracilis*, selon Karabin *et al.* (1984)
psbA de *Glycine max* (soja), selon Spielmann et Stutz (1983)
TufA, *rps7*, *rps12* d'*Euglena gracilis*, selon Montandon et Stutz, (1983), (1984)

Nous remarquerons que les résultats concernant les gènes *rps7*, *rps12* et la partie 5' séquencée du *ndhB* (du premier méthionine annoncé pour le tabac au site HinDIII) ont été groupés. Ces séquences étant relativement courtes, nous avons voulu éviter une trop grande variation des rapports observés. En effet, le poids statistique d'une analyse portant sur 399 codons est plus élevé que si la même analyse est faite sur une centaine de codons seulement. Constatons que les rapports sont similaires lorsque chacun des gènes est étudié individuellement.

La Figure 18 montre que les gènes *rps7*, *rps12* et *ndhB* de soja ont des scores comparables aux gènes constitutifs d'Euglène. Nous observons toutefois que le rapport (T+C)/(A+G) des gènes de soja semble en général un peu plus élevé. A priori, le modèle de Gingrich et Hallick (1985) semble donc être utilisable avec le soja également. Nous aurions ainsi une indication du type d'expression des gènes *rps7*, *rps12* et *ndhB*.

L'ORF étudiée et qui est interne à l'intron 2 du *rps12* à un rapport plus bas que le *rps12*. Nous ne pouvons pas décider, dans la connaissance actuelle que nous avons des gènes de soja, si cette ORF est transcrite et traduite constitutivement où si elle n'est pas traduite. Nous reviendrons plus tard sur ce point particulier.

3.3 Résultats corollaires

Au cours de la recherche de l'exon 1 du *rps12*, nous avons rencontré un certain nombre de gènes. Les ayant localisés et souvent séquencés, nous indiquons les résultats obtenus.

3.3.1 le gène *psbB*

Le gène *psbB* a été repéré en séquençant le fragment HinDIII de 4.4 kb dans lequel est situé la partie 5' du gène *rps12* et dans l'orientation inverse. le *psbB* code pour une protéine chloroplastique associée au centre réactionnel P680 du photosystème II.

3.3.2 Le gène *ndhB*

Ce gène est situé en 3' du *rps7*. Il a été mis en relation avec un gène d'origine mitochondriale (Shinozaki *et al.* 1986). La protéine codée ferait partie d'une NADH deshydrogénase d'origine chloroplastique. La séquence de la protéine est donnée au §(3.1.8) dès le premier acide aminé méthionine trouvé dans la séquence de soja (position 8098).

Nous avons en effet remarqué que la protéine codée peut être beaucoup plus longue que ce que nous propose le groupe japonais (Shinozaki *et al.* 1986), qui situe le premier méthionine traduit à l'endroit équivalent à notre position 8413. Le choix de ce méthionine ayant été décidé par similarité avec le gène codant pour la NADH deshydrogénase mithocondriale humaine (Chomyn *et al.* 1985).

3.3.3 Les gènes *trnV, trnI, trnA*

Les ARN de transfert trouvés dans l'opéron ribosomique sont les ARN-t^{Val}, ARN-t^{Ile} et ARN-t^{Ala}. Nous présentons les séquences avec leurs structures en forme de trèfle caractéristiques dans les Figures suivantes en comparant chaque fois la séquence avec les ARN-t correspondant de tabac et de maïs. Nous rappellerons que la séquence terminale "CCA" n'est pas codée au niveau de l'ADN, mais ajoutée postérieurement à la transcription.

Figure 19 : *Glycine max* : ARN-t^{Val} (GAC)

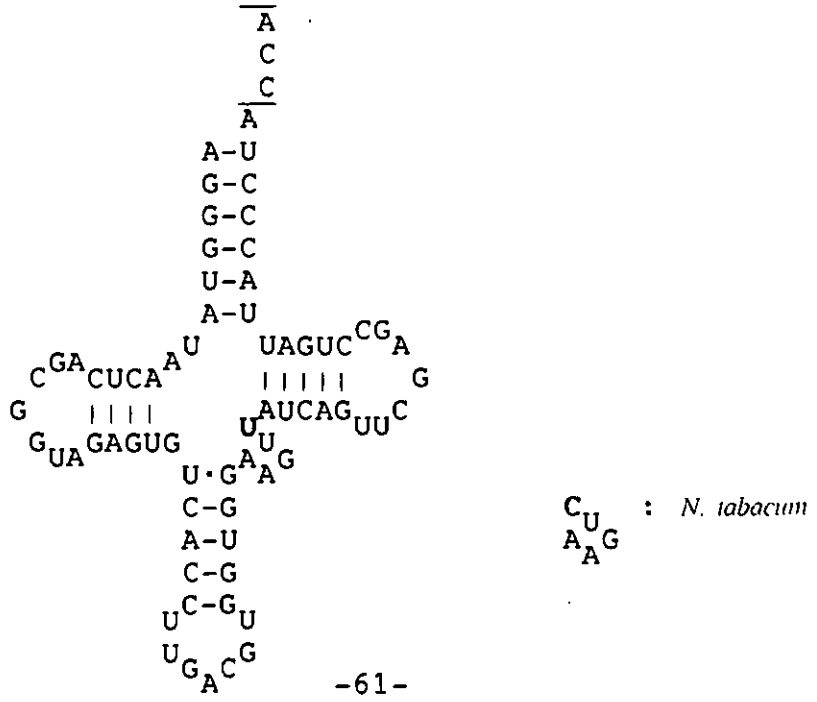


Figure 20 : *Glycine max* : ARN-t ^{Ile} (GAU) (après splicing)

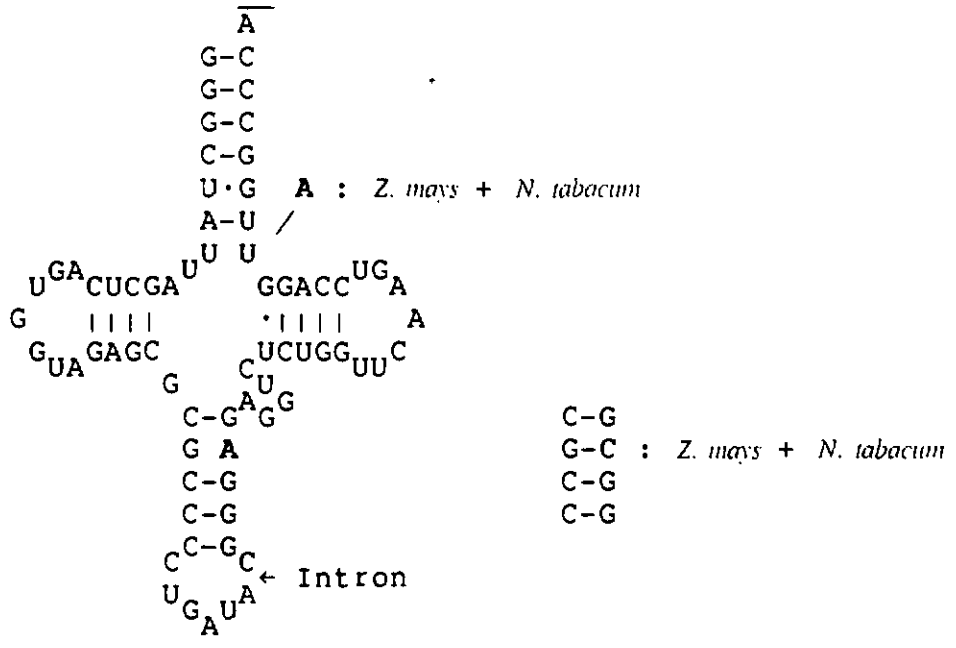
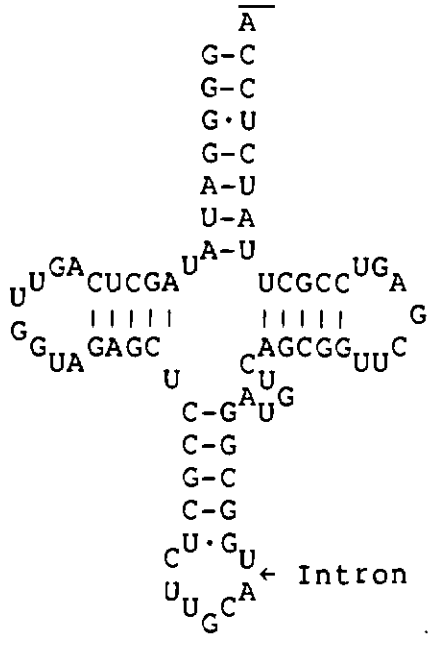


Figure 21 : *Glycine max* : ARN-t ^{Ala} (UGC)



Nous constatons une fois encore, la grande similarité qui existe entre les différents génomes chloroplastiques. La structure primaire des ARN de transfert est très conservée et seules des micro-différences existent entre les différentes séquences. Nous remarquerons également que la position des introns interrompant les ARN-t Ile et Ala est identique. Nous discuterons plus tard des particularités observées dans les introns de ces ARN de transfert pour le maïs, le tabac et le soja.

3.3.4 "Gènes" des ORF

Nous avons trouvé deux séquences ouvertes à la traduction. Le principal problème dans un tel cas est de savoir si une protéine est réellement fabriquée à partir de ce message ou si le hasard seul à présidé au codage de l'ORF. Nous avons premièrement trouvé une ORF de 163 acides aminés, située en 5' du *3'-rps12* (position 5331) et codée par le même brin. Dans la même région de l'ADN chloroplastique de tabac, a été localisé l'ORF 70B. Le manque de correspondance entre les deux séquences indique peut être que l'ORF 70B de tabac et l'ORF 163 de soja sont des pseudogènes et démontre une fois de plus qu'il existe des différences de contenu au niveau des ADN chloroplastiques.

Figure 22 : ORF 163

```
         10        20        30        40        50
MRVQPHCIARTHVVYLKKVEPPCFSHVTILLPLSSLSPRSTETKCKTGAN
SSSRKKGLAEPRSLTKTNLILIENTNRIEIELSFHIDIPSTQHRSSKGSR
RPTNARKPGSFRKWIPIRRVHNRMDKLTLTRQFGIQFGSLGGIGKGFWNV
IISIHTDTEERVLY*
```

La seconde région ouverte à la traduction est particulièrement intéressante car elle se trouve intégrée dans l'intron 2 du *rps12* et en phase avec le 3ème exon du *rps12*, nous avons alors pensé aux maturases internes aux introns mitochondriaux, qui codent pour les enzymes responsables du splicing. Malheureusement, il n'existe pas d'homologie notable entre l'ORF 143 et les maturases connues. Nous donnons tout de même la séquence de cette protéine possible, qui commence à la position 6786 du fragment Hd906.

Figure 23 : ORF 143

```
         10        20        30        40        50
MRQKIFFLKKFDSELRYVQGPIWKEFQRFSLTLSVSTNNSKCLDFFRIGP
SQIAMIRSTLHYFGNPRTQSYGYVKYRISNPSRKRGGTDTQFQVSKQNSI
LDIIDTYRILWKAVFDESRMYGLEGDLSYLSRSTLQYGAKKPK*
```

Nous devons signaler que l'usage des codons est relativement différent de la normale (Figure 35) et que la chance que l'ORF 143 soit traduite est minime. Nous discutons ce point particulier après l'analyse de la G-périodicité (Figure 35).

3.3.5 Le gène *rpl20*

Le gène *rpl20* code pour la protéine ribosomique L20. Cette protéine est l'une des protéines impliquée dans l'assemblage de la sous-unité 50S du ribosome et se fixe directement sur l'ARN ribosomique 23S. Nous donnons la séquence partielle de la protéine car il nous manque encore 9 codons qui n'étaient pas présents dans les clones séquencés. Nous avons d'ailleurs constaté avec surprise que la partie 3' du gène était beaucoup moins bien conservée que ne l'est le *rps7*.

La séquence ayant été vérifiée avec soin, nous pouvons nous poser la question de savoir si ce gène de soja est fonctionnel ou non. Une seconde possibilité envisageable est que le clone M13 contenant la région terminale du gène ait une délétion car le clone considéré est un clone de plus de 1000 nucléotides. Constatons toutefois que la similitude existant entre le gène de tabac et celui de soja plaide pour une même fonction.

Figure 24 : Séquence partielle du *rpl20* et de la protéine L20
Nous soulignons la présence des sites SacI et HinDIII qui interrompent le gène.

```
N.t  M  T  R  I  K  R  G  Y  I  *  *  R  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
G.m                            A  R  K  R  R  T  K  I  R  L  F  A  S  S  F  R  G  A
     atgaccagaattaaacggggatatatAGCTCGGAAAAGAAGGACAAAAATTCGTTTATTTGCATCAAGTTTTCGAGGAGC
           10        20        30        40        50        60        70        80

N.t   *  *  *  *  *  *  *  *  *  *  *  *  *  R  *  *  *  *  *  *  *  *  *  *  *  *
G.m   H  S  R  L  T  R  T  I  T  Q  Q  K  I  K  A  L  V  S  A  H  R  D  R  D  R  K
     TCATTCAAGACTTACTCGAACTATTACTCAACAGAAAATAAAAGCTTTGGTTTCGGCTCATCGGGATAGAGATAGGAAAA
           90        100       110       120       130       140       150       160

N.t *  *  *  *  *  R  *  *  *  T  *  *  *  *  *  *  *.*  Δ  E  R  G  *  S  *  S  *  *
G.mK  R  D  F  R  G  L  W  I  S  R  I  N  A  V  I  R  G  N  I  K  V  Y  Y  I  Y  S
     AAAGAGATTTTCGTGGTTTATGGATTAGTCGAATAAATGCAGTAATTCGCGGGAATATTAAAGTATATTATATTTATAGT
           170       180       190       200       210       220       230       240

N.t  R  *  I  H  D  L  *  K  R  *  L  *  *  *  *  *  *  L  *  *  *  *  *  S  N  R  *
G.m  N  L  V  Y  S  L  Y  T  G  Q  W  L  L  N  R  K  I  V  A  Q  I  A  I  L  K  Q  N
     AATTTAGTATATAGTCTATACACGGGGCAATGGCTTCTTAATCGTAAAATAGTTGCACAAATAGCTATATTAAAACAGAA
           250       260       270       280       290       300       310       320

N.t   *  *  Y  *  *  S  *  E  *  *  *  E  V  W  K  E  S  T  R  I  I  *
G.m   C  L  F  M  I  A  N  D  I  I  K  T  K  N  P  L  R  L  Y  S  G  K  V  *
     TTGTCTTTTTATGATTGCCAATGATATCATAAAAACCAAAAATCCCCTTAGACTTTACTCCGGGAAGGTATAGATTGTCT
           330       340       350       360       370       380       390       400
```

La L20 nous intéresse par le fait qu'elle fait partie des protéines directement utiles à l'assemblage du ribosome et qui ont été démontrée en interaction avec l'ARN 23S. Il est intéressant de noter que parmi les protéines utiles dans l'assemblage du ribosome et en interaction avec l'ARN 16S : Les S4, S7, S8 et S15 sont codée par le chloroplaste alors que seule la S17 semble être d'origine nucléaire.

Au niveau de l'ARN 23S et impliqué dans l'assemblage de la sous-unité 50S du ribosome, les protéine L2, L20 et L23 sont d'origine chloroplastique tandis que les autres protéines de première importances (L1, L3, L4, L9 et L24) sont codées semble-t-il par le noyau.

3.4 DISCUSSION I : Quelques aspects du *cis-* et du *trans-* splicing

L'étude de la maturation des ARN confronte à une controverse les chercheurs qui cherchent à entrer dans le domaine du splicing. Les définitions ne sont pas encore toutes standardisées. C'est pourquoi, nous ferons référence aux définitions que pour notre part nous acceptons.

Les types de splicing bien définis sont :

Le type eucaryotique (Breathnach et Chambon, 1981; Shapiro et Senapathy, 1987), qui est utilisé pour le processing des ARNm synthétisés (par la RNA polymérase II) d'origine nucléaire :

$$AG\ 'GT^{A}_{G}AGT\cdots\cdots Intron\cdots\cdots {}^{TT}_{CC}TT\ {}^{TTTTTT}_{CCCCCC}NCAG'\ G$$

le groupe I (Michel et Dujon, 1983; Burke *et al.* 1987) répond initialement à la présence des séquences hautement conservées suivantes dont l'ordre 5' à 3' est conservé, mais où la distance qui les sépare peut varier :

····YTCANNGACTA···· et ····TNANGANATAGTC····

Nous observerons que la présence d'une structure secondaire associée aux introns du groupe I peut être plus importante que la structure primaire (Burke *et al.* 1987).

les introns du groupe II (Michel et Dujon, 1983), parmi lesquels se rangeront les introns que nous avons séquencés, sont caractérisés par une structure secondaire (Figure 26) caractéristique et un consensus de début d'intron bien conservé. Leurs structures devraient répondre au schéma suivant :

```
      |=|  |=|     |=| |=|   |=|  |=|
'GTGCGY-RGAU-ACGGAG-RAG-TGGAG-RGGG-RA'
```

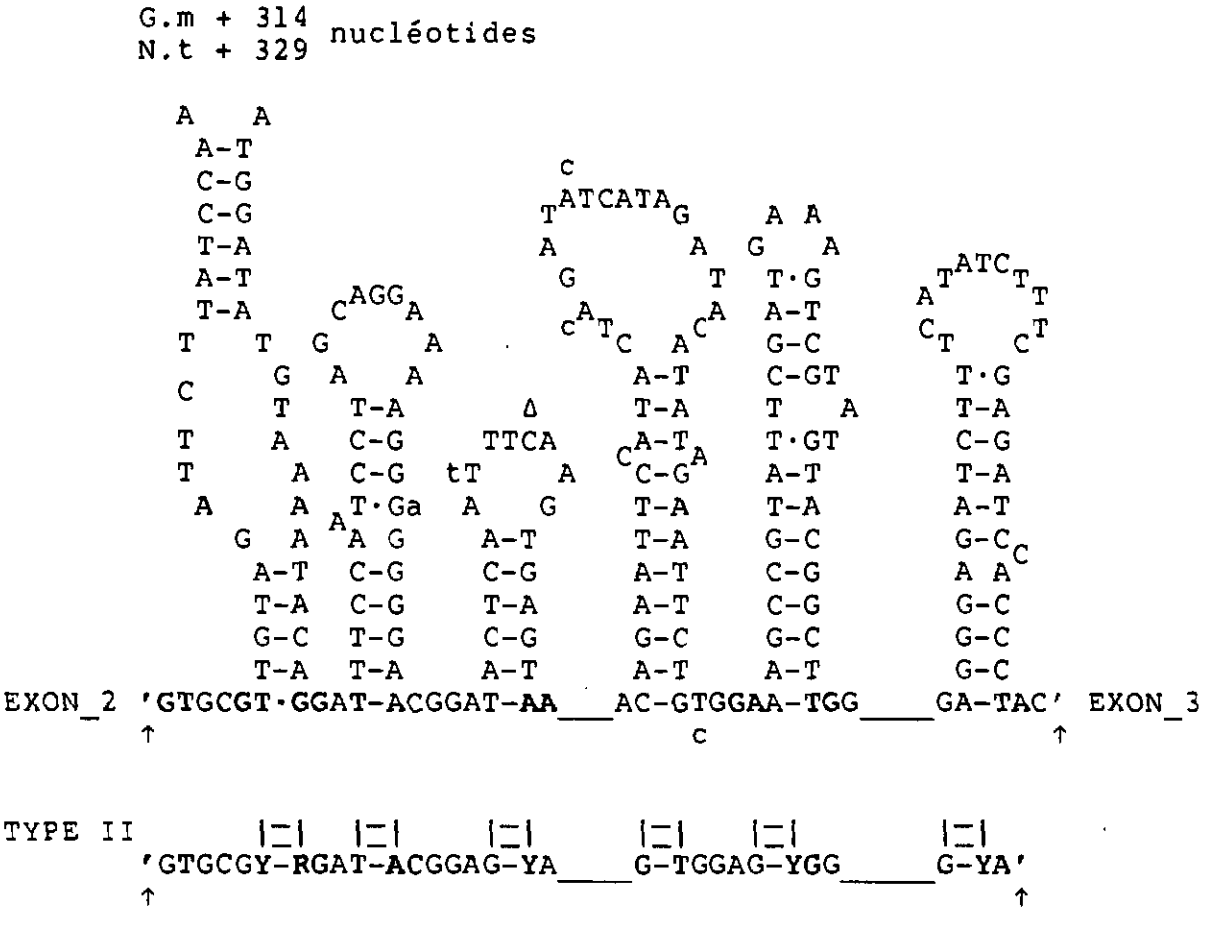
Nous n'entrerons pas dans la définition du groupe III, car il n'existe pas encore unanimité quant à sa définition. Ce groupe à été utilisé dans des cas totalement différents et sans rapports les uns avec les autres ! (Yamada et Shimaji, 1987; Shinozaki *et al.* 1986b). Ce dernier article parlant d'ailleurs des introns du *rps12* et les classant dans le "groupe III", mais cette classification est prématurée.

3.4.1 Cas des introns du *rps12*

Le premier critère qui puisse permettre de repérer la limite entre les parties codantes et non codantes d'un gène est de voir comment il est possible de placer les limites de l'intron de manière à ce que la protéine codée ait une similarité maximale avec une protéine homologue (une protéine du même gène, mais appartenant à un autre organisme).

Le second critère est de chercher les consensus connus qui marquent généralement le début ou la fin d'un intron. Cherchant un intron chloroplastique, nous nous sommes d'abord référé au consensus 5'-GYGYG (Koller *et al.* 1984) définissant la limite 5' de l'intron. Ce consensus est compatible avec le début d'un intron du groupe II 5'-GTGCGY tel qu'il était déterminé. Cela nous a poussé à chercher une structure secondaire spécifique de ces introns.

Figure 25 : Structure d'intron du groupe II pour l'intron 2 du *rps12* de soja (*Glycine max*) et de tabac (*Nicotiana tabacum*)



Légende : La séquence du *rps12* de soja (G.m) est donnée en lettres majuscules. Les nucléotides différents dans le cas de la séquence de tabac (N.t) sont ajoutées en lettres minuscules (Fromm *et al.* 1986).
Le rapport avec le consensus général des introns du groupe II (Michel et Dujon, 1983) est donné.

_____ : représente 0 nucléotides

Nous pouvons d'ailleurs vérifier que les limites d'introns observées ont de bonnes chances d'être justifiées en comparant les séquences des protéines S12 d'*Escherichia coli*, d'*Euglena gracilis*, de *Nicotiana tabacum* et de *Glycine max*, mais nous en parlerons lors de la discussion II axée sur le ribosome §(4.5).

Nous nous sommes alors intéressé de plus près à la structure des introns de type eucaryotique et du groupe I. L'étude du splicing est beaucoup plus avancée pour ces deux groupes et nous pouvons peut être ainsi mettre en évidence des caractéristiques passées inaperçues. En partant d'articles de revues, (Cech, 1983; Padgett *et al.* 1986) nous nous sommes intéressé aux différentes séquences importantes pour le splicing que ce soit dans les cas du groupe I, II, ou de type nucléaire. Le splicing de tous les types d'introns, bien que possédant des méchanismes différents, semble être des processus que l'on puisse mettre en relation (Cech, 1986). Nous avons alors particulièrement cherché si les introns chloroplastiques du groupe II présentaient des structures où la bordure en 5' de l'intron était complémentaire à une séquence de la partie 3' du même intron.

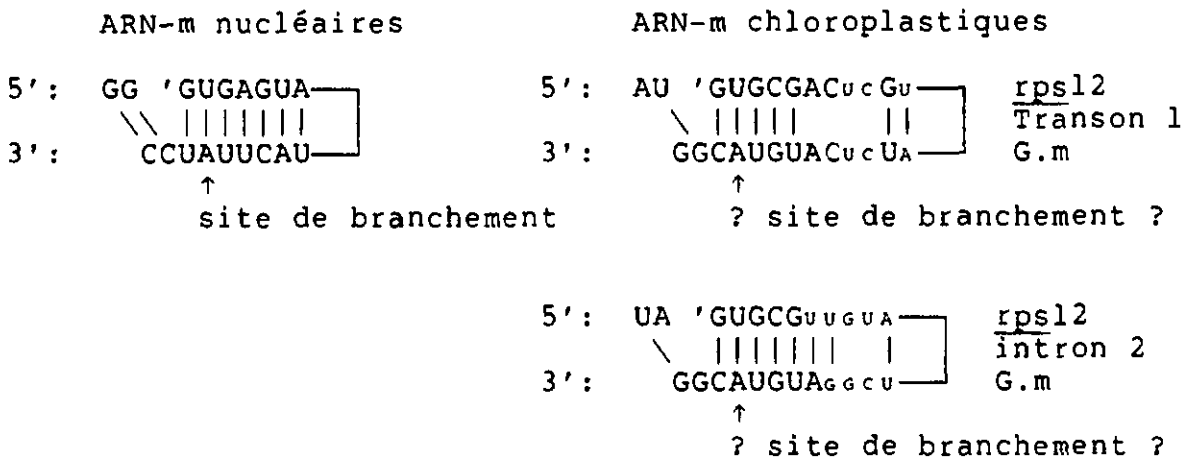
3.4.2 Etude de la complémentarité de la limite 5' des introns du groupe II avec une séquence de la partie 3' de l'intron.

Dans le cas du splicing de type nucléaire, il existe un site de branchement situé entre 16 et 64 nucléotides en amont du site de splicing en 3' de l'intron (Teem *et al.* 1984).
La région entourant le site proprement dit est bien conservée et complémentaire aux premiers nucléotides de la partie 5' de l'intron (Padgett *et al.* 1986).

GG'GUGAGUA——————————UACUUAUCC————CAG'C

Nous avons alors cherché si une structure de même type, mais complémentaire à la séquence des introns du groupe II était observable. Nous nous sommes fixé les limites approximatives de 16 à 64 nucléotides en amont du site de splicing en 3' comme distances optimales pour trouver une telle structure et avons étudié une série d'introns chloroplastiques présentés plus loin (Figure 27). L'analyse a révélé qu'il existait effectivement complémentarité. Nous avons en même temps remarqué une légère particularité des transons chloroplastiques connus vis à vis des introns.

Figure 26 : Comparaison des séquences très conservées liées à la formation de lariat dans les introns des messagers nucléaires (Padgett *et al.* 1986) avec une structure possible dans le cas des transons chloroplastiques connus et d'une série d'introns



Distance: 20-40 bases avant le site 3' de l'intron
Limites : (16-64)

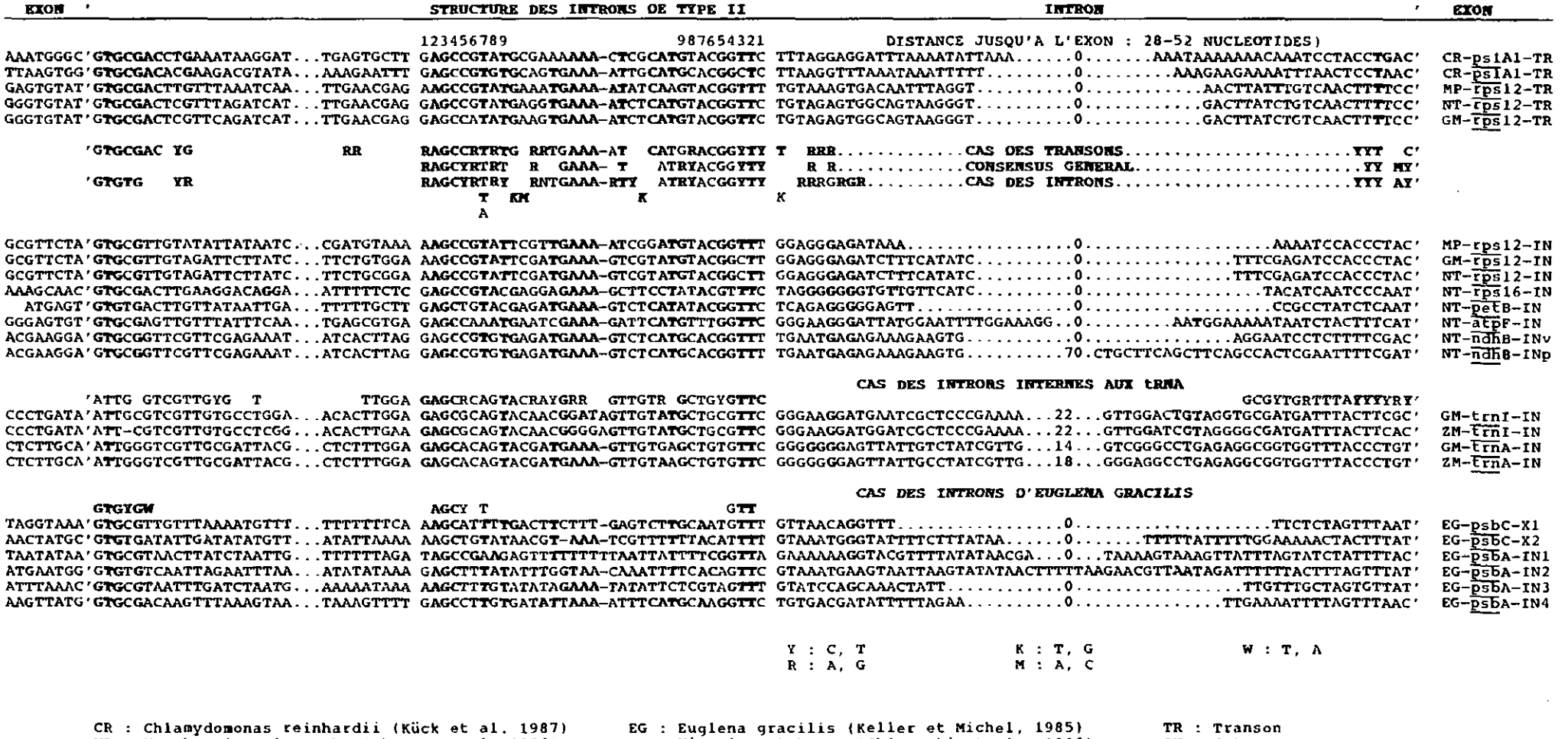
29-55 bases avant le site 3' de l'intron
Limites : inconnues
Recherche sur 12 gènes de plantes supérieures (Figure 27)

Nous avons ainsi retrouvé des séquences très conservées et faisant partie de la boucle n°5 (avant dernière structure secondaire) du schéma pour les introns du groupe II de Michel et Dujon (1983). Il est intéressant de noter que cette boucle a non seulement une structure secondaire, mais aussi une structure primaire très conservée, ce qui peut éventuellement indiquer qu'une fonction plus particulière peut lui être impartie.

Nous formulons alors une première réserve : La structure que nous observons possède effectivement une complémentarité avec la partie 5' de l'intron, mais elle possède déjà une structure secondaire relativement stable puisqu'elle est comprise dans la boucle n°5 des introns du groupe II. Il est alors difficile de savoir si l'interaction montrée (Figure 26) entre la partie 5' et la boucle n°5 est possible *in vivo*.

Cette première objection théorique ne permet pas, pourtant, de rejeter ni l'hypothèse d'un appariement, ni celle de la formation d'un site de branchement. Van der Veen *et al.* (1987) proposent en effet un mécanisme à peu près similaire pour un intron mitochondrial du groupe II de levure, impliquant une séquence comprenant la fin de l'exon et le début de l'intron. L'interaction proposée met en jeu une zone également appariée dans le modèle de Michel et Dujon (1983), cette fois-ci la boucle n°6. L'appariement ne serait donc pas plus impossible avec une séquence interne à la boucle n°5 qu'avec la boucle n°6. L'ARN étudié par ces auteurs est d'ailleurs, dans le cas précis, un ARN subissant le self-splicing, donc un ARN possédant toutes les qualités requises pour supprimer lui-même son intron.

Figure 27 : **ANALYSE DES INTRONS DE TYPE II INTERROMPANT LES GENES CHLOROPLASTIQUES DE PROTEINES ET DE TRNA : QUELQUES REGLES OBSERVEES**



La conservation des séquences et la conservation des distances de ces séquences par rapport à l'extrémité 3' de l'intron donnent à notre avis une indication qu'une fonction importante pourrait appartenir au consensus observé. Les observations de Koller *et al.* (1985), (1987) ajoutent que des ARN circulaires correspondant à la taille des introns du groupe II existent *in vivo* et sont visibles au microscope électronique. Il est évident que la microscopie ne permet pourtant pas de distinguer si l'ARN est refermé sur lui-même par la présence d'une complémentarité de ses nucléotides, ou s'il forme un lariat.

Le lariat est une structure circularisée d'un ARN possédant un site de branchement et fermé sur lui même au moyen d'une liaison covalente. Les comparaisons que nous avons effectuées, montrent que certaines caractéristiques des séquences internes aux introns chloroplastiques du groupe II sont étonnament proches de ce que l'on attendrait du domaine entourant les sites de branchement de type nucléaire. Nous pouvons penser que le site de branchement proposé dans la Figure 26 est soit fonctionnel, soit cryptique. La conservation de ces séquences en ces positions particulières apporte de toute manière un élément intéressant. Les introns chloroplastiques du groupe II ont certaines caractéristiques structurales qui peuvent les rapprocher phylogénétiquement des introns de type nucléaire.

Cette analyse des introns, nous a d'autre part apporté des renseignements supplémentaires concernant les consensus de début, de boucle 5 et de fin d'introns des gènes en mosaïque subissant du *trans* et du *cis* splicing. Nous avons voulu vérifier si les transons chloroplastiques connus sont différemment délimités que les introns. Nous présentons les résultats dans la Figure suivante (Fig. 27), où nous avons indiqué la complémentarité des nucléotides par la numérotation :

(1:1), (2:2), ... , (9:9).

La première constatation que nous avons faite est que la boucle No 5 interne aux ARN messagers des introns de type II se situe entre 26 et 52 nucléotides en 5' de l'extrémité 3' de l'intron. Nous avons observé une seule exception, pour le gène *ndhB* (NT-ndhB-INp) (Shinozaki *et al.* 1986), ou la distance observée est de 121 nucléotides, ce qui est même plus grand que la distance existant dans le cas des introns internes aux ARN de transfert. Nous avons alors analysé la séquence de tabac dans la région du *ndhB* et avons observé qu'à la distance "réglementaire", il existe une possibilité de splicing plus convaincante. Nous avons appris entre temps (Shinozaki, communication personnelle) qu'il existe une erreur dans la localisation de cet intron. Nous n'avons pas réussi à obtenir de précisions quant à son emplacement exact, mais nous pourrons bientôt vérifier dans la littérature si notre supposition (NT-ndhB-INv) est exacte.

Figure 28 :

Sites de splicing du gène ndhB de tabac

```
   D  V  Y  E  G '
...GACGTATACGAAGGA'GTGCGGTTCG....Intron.... GAGCCGTGTGAGATGAAAGTCTCATGCACGGTTT T
                  ↑

                  V  R  N  P  L  F  D ' S  D  S  P  T  P  V  V  A  F  L  S  V  T
GAATGAGAGAAAGAAGTGAGGAATCCTCTTTTCGAC'TCTGACTCTCCCACTCCAGTCGTTGCTTTTCTTTCTGTTACTT
                                     └ notre proposition

S  K  V  A  A  S  A  S  A  T  R  I  F  D ' I  P  F  Y  F  S  S  N  E  W  H  L
CGAAAGTAGCTGCTTCAGCTTCAGCCACTCGAATTTTCGAT'ATTCCTTTTTATTTCTCATCAAACGAATGGCATCT...
                                        └ proposition publiée
```

Nous avons aussi effectué l'analyse des introns d'*Euglena gracilis* et des introns internes aux ARN-t Ile et Ala. Ces résultats sont intéressants, car ils permettent d'apprécier la divergence qui peut exister au niveau des séquences impliquées dans un splicing correct des introns et donnent une idée de l'évolution parallèle qu'ont subi ces séquences. Il est frappant que chez les plantes (Euglène excepté), les distances et les consensus soient aussi bien conservés.

Ces observations militent en faveur d'une fonction à garder. Nous envisageons ici, la formation de lariat. Nous pensons qu'il faudrait tester cette hypothèse, peut-être en effectuant ant une "primer extension" qui utilise une amorce synthétique hybridant avec une séquence en 3' de l'adénine mentionné (Figure 26). Il devrait y avoir un arrêt de la transcriptase inverse au site de branchement. Nous proposerions d'utiliser l'intron 2 du *rps12*, qui devrait se prêter à une telle analyse, puisqu'il est detecté, même au niveau des Northern et que la "maturation de l'intron" semble complète.

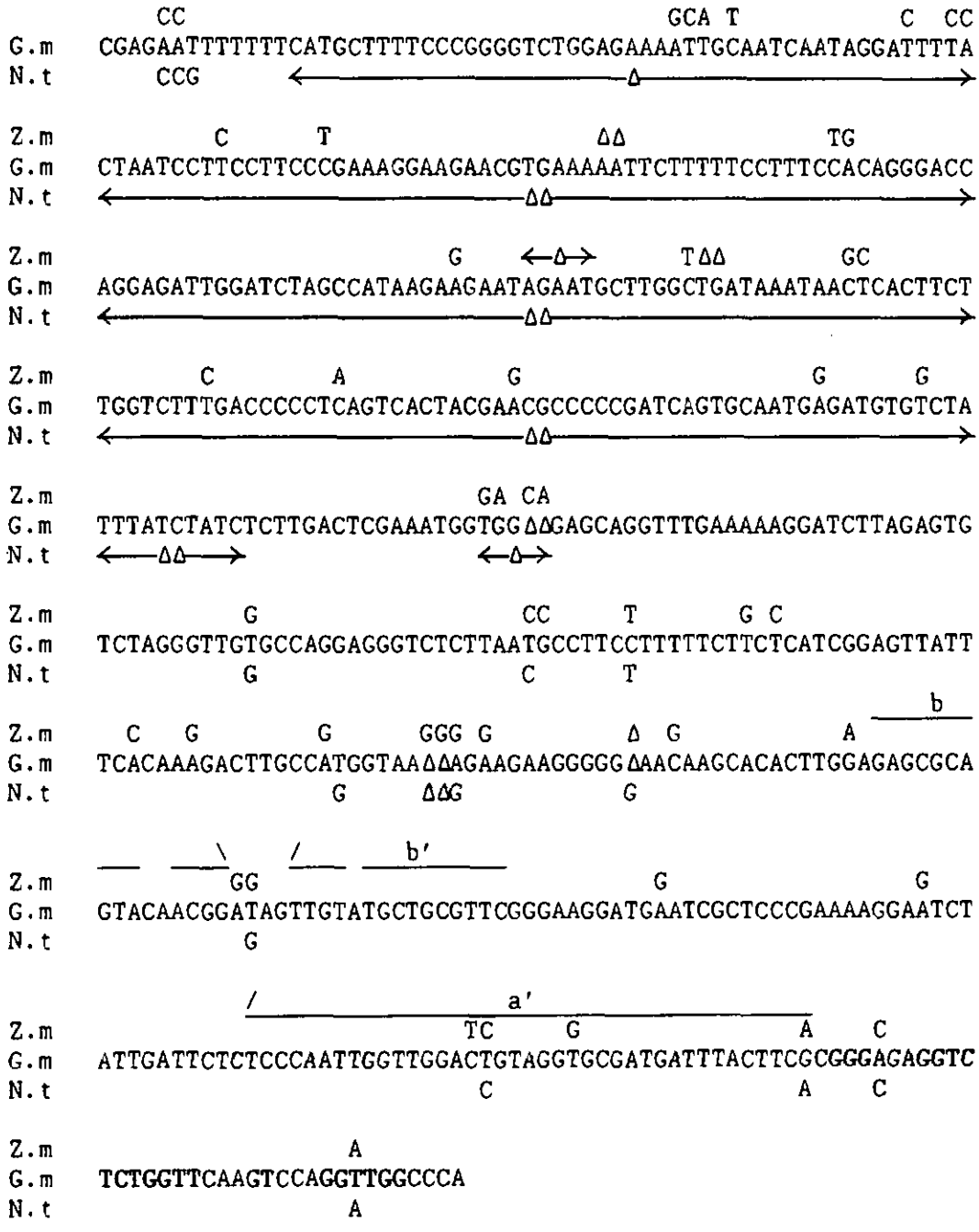
3.4.3 Les introns internes aux ARN-t Ile et Ala

Nous avons remarqué que les séquences des introns sont très bien conservées entre les différentes plantes supérieures. La Figure 25, déjà, nous a démontré que les différences existant entre les introns 2 du *rps12* de soja et de tabac sont insignifiantes, si ce n'est que nous n'observons la présence d'une ORF (région ouverte à la traduction) de 143 acides aminés que chez le soja. (cf. résultats corollaires §(3.3)). Mais la similarité entre les différents introns ne s'arrête pas à ce seul exemple. Nous avons comparé les introns qui interrompent les gènes *trnI* et *trnA* que nous avons séquencés et qui codent pour les ARN-t Ile et Ala, avec les gènes homologues de maïs (Koch *et al.* 1981) et de tabac (Takaiwa et Sugiura, 1982a).

Figure 29 : Comparaison des introns interrompant les ARN de transferts Ile et Ala de *Zea mays*, *Glycine max*. et de *Nicotiana tabacum*. Il existe complémentarité entre les région (a:a') et (b:b'). La structure (b:b') correspond à la boucle 5 des introns du groupe II.

29a) ARN-t Ile

```
                                                                    a
                                                                    ____
Z.m                                                      Δ          G C
G.m   GGGCTATTAGCTCAGTGGTAGAGCGCGCCCCTGATAATTGCGTCGTTGTGCCTGGACTGT
N.t                                                                 G

          \
      ____
Z.m       G                    C              A            GA
G.m   GAGGΔCTCTCAGCCACATGGATAGTTTAATGTGCTCATCGGCGCCTGACCCTGAGATGTG
N.t       G                    C

Z.m                                      C            T   A
G.m   GATCATCCAAGGCACATTAGCATGGCGTACTTCTCCTGTTTGAACCGGGGTTTGAAACCA
N.t                                      C        C

Z.m       AAAC  C                                                C         G
G.m   AACΔΔΔΔTTATCCTCAGGAGGATAGATGGGGCGATTCAGGTGAGATCCAATGTAGATCCA
N.t       ←ΔΔ→ CC                              G

Z.m           A                  ←—Δ—→     G           GG     G
G.m   ACTTTCTCTTCACTCGTGGGATCCGTGATCCGGGCGATCCGGGGGGGAΔCCACCACGGCT
N.t           A                  ←—Δ—→     G           AΔ

Z.m                        C                  G         C
G.m   CCTCTCTTCTCGAGAATTCATACATCCCTTATCAGTATATGGACAGTTATCTCTCGAGCA
N.t                        C                  G         C

Z.m           G     C T          G     T←—Δ—→                  CATC
G.m   CAGGTTTAGGTTTGGCCTCAATGGAAAAAAAACΔΔΔGGAGCACCTAACAACGTAΔΔTTCA
N.t             ←——Δ——→      G     AAAAT                    CATC

Z.m                            A                          A   A
G.m   CAGACCAAGAACTACGAGATCGCCCCTTTCATTCTGGGGTGACGGTGGGTTCGTACCATT
N.t                                                       A   A
```



Zea mays : (Koch *et al.* 1981)
Nicotiana tabacum : (Takaiwa et Sugiura 1982a)
Glycine max : présent travail

29b) ARN-t Ala

```
Z.m
G.m   GGGGATATAGCTCAGTTGGTAGAGCTCCGCTCTTGCAATTGGGTCGTTGCGATTACGGGT
N.t

Z.m      C
G.m   TGGATGTCTAATTGTCCAGGCGGTAATGATAGTATCTTGTACCTGAACCGGTGGCTCACT
N.t

Z.m                    G        T
G.m   TTTTCTAAGTAATGGGAAAGAGGACCGAAACATGCCACTGAAAGACTCTACTGAGACAAA
N.t                    G              G

Z.m   AA              A  G
G.m   ΔΔGACGGGCTGTCAAGAACGTAGAGGAGGTAGGATGGGCAGTTGGTCAGATCTAGTATGG
N.t   ΔΔ  T

Z.m                  A             C            C  C      C
G.m   ATCGTACATGGACGGTAGTTGGAGTCGGTGGCTCTCCTAGGGTTTCCTCATTTGGGATCC
N.t                                C        C      C      C  A    T

Z.m   C                   T              T  C
G.m   ΔTGGGGAAGAGGATCAAGCTGGCCCTTGCGAACAGΔTTGATGCACTATCTCCCTTCAACC
N.t                       T                 C

Z.m                              ←Δ→   GG A                        T
G.m   CTTTGAGCGAAATGTGGCAAAAGΔΔΔΔΔGAAAAAGAATCCATGGACCGACCCCATCGTCT
N.t                              AAAAG   GG A                       A

Z.m                                          C TTC  ←ΔΔ→T      TAT
G.m   CCACCCCGTAGGAACTACGAGATCACCCCAAGGACGTCΔΔΔGGCATCCAGGGGTCΔΔGCG
N.t                                          C TTC

Z.m          ←ΔΔ→     ATΔ          T  GT      A  A  G CG          C
G.m   GACCGACCΔΔΔΔΔATAGAACCCTGTTCAAAAACΔGGAACGCATTAΔCTATCCGCTCTCAGGG
N.t          GACC                   T   T           G  G             T

Z.m    G  CAGTAAGGGT        GAAGGGCAATCACTCGTTCATAAAACCAGCATTCATAAGT
G.m   TTGGACAΔΔΔΔΔΔΔΔCGGATCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
N.t          TCAGGGT     ←Δ→GAAGGGCAATGACTCATTCTTA←————————ΔΔ————————→

Z.m   TAAGATCAAAGAGTCGGGCGGAAAAAGGGGAGAGCTC           GTTCTCCTG
G.m   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCGTTCCTGΔΔΔΔΔΔΔΔΔΔTAGCT
N.t   ←————————————————————————ΔΔ————————————————————————→

Z.m      T  C                  CCTTAGAAT                              TTTG
G.m   GGAATCTCCGGAACCACAAGAATΔΔΔΔΔΔΔΔΔΔGGGATTCCAACTCAGCACCTTTTGAΔΔΔΔ
N.t   ←————————ΔΔ————————→GTTAGAAT                                  GΔTG

Z.m                                                          A
G.m    GATTTTGAGAAGAGTTGCTCTTTGGAGAGCACAGTACGATGAAAGTTGTGAGCTGTGTTC
N.t                                                          A

Z.m                    C          T                C C      GAGG
G.m   GGGGGGGAGTTATTGTCTATCGTTGGCCTCTATGGTAGAATCAGTCGGGΔΔΔΔCCTGAGAG
N.t                                                      GGΔA

Z.m
G.m   GCGGTGGTTTACCCTGTGGCGGATGTCAGCGGTTCGAGTCCGCTTATCTCCA
N.t
```

Zea mays : (Koch *et al.* 1981)
Nicotiana tabacum : (Takaiwa et Sugiura 1982a)
Glycine max : présent travail

La plupart des changements ponctuels existant entre les différentes séquences se situent au niveau du remplacement d'une base pyrimidique par une autre base pyrimidique ou des bases puriques entre elles. Cette constatation, nous fait penser que quelques unes de ces différences pourraient provenir de la méthode de séquençage (Maxam et Gilbert, 1977) utilisée pour la détermination des séquences de maïs et de tabac et qui permet parfois des erreurs d'interprétations de la séquence.

Nous remarquerons que la séquence de soja est sur certains tronçons plus proche de l'une ou de l'autre des séquences, et que la longueur des introns interrompant les gènes *trnI* et *trnA* est identique à celle du maïs et plus grande que pour le tabac.

Une autre observation intéressante est que bon nombre des délétions observées chez l'une des trois plantes, consiste en la suppression d'une partie de séquence dupliquée chez les autres. Nous donnerons comme exemple le plus frappant, un fragment de séquence interne à l'intron de l'ARN-t Ile qui illustre particulièrement bien le phénomène de duplication observé :

```
                  ↓     ↓
Z.m    TGG GAGCA GAGCA
G.m    TGG TGGΔΔ GAGCA
N.t    TGG ΔΔΔΔΔ GAGCA
```

Nous constatons sur le même fragment une duplication du motif GAGCA dans le cas du maïs et du motif TGG pour le soja. Seule la séquence de tabac ne contient aucune duplications à l'endroit considéré. Nous observons également quelques séquences qui sont peut-être imputable au hasard, à des traces laissées par le passage de transposons ou témoins de recombinaisons passées. Il existe en effet deux copies de ces gènes, ce qui pourrait favoriser la formation de crossing-over. Nous ne tirerons pourtant pas trop de conclusions concernant ces duplications car il se pourrait qu'il y ait encore quelques fautes dans les différentes séquences.

3.4.4 Les consensus

Pour terminer, nous constaterons que l'analyse des séquences par ordinateur permet de nous donner la position des débuts d'introns internes aux ARN messagers en cherchant :

transons ? : 5'-GTGCGAC$_N$YG

introns : 5'-GTGYG$_{NNN}$YR

Les fins d'introns et de transons sont reconnaissables par la présence du motif spécifique des introns de type II qui est situé de 20 à 60 nucléotides avant l'extrémité 3' de l'intron (Figure 27). Le choix du site de splicing est limité ensuite par le problème de la phase du message et de la présence de :

transons ? : YYTN$_C^A$C'-3'

introns : YYYNAY'-3'

Nous remarquons qu'en faisant des recherches sur les fichiers contenant les séquences complètes publiées, il est certainement possible de trouver encore des sites de splicing. Par exemple un site typique de splicing avec une structure secondaire de boucle 5 est clairement démontré par le programme ANALYSE en position 45435 sur le brin complémentaire de la séquence de tabac et mériterait d'être étudié de plus près.

3.5 DISCUSSION II : Quelques corrélations existant entre le ribosome et sa fonction

Le ribosome est l'unité qui dans la cellule à pour rôle de lire les ARN-messagers et de fabriquer les protéines avec un optimum d'efficacité et de fidélité. Les plastoribosomes sont de type procariotiques. La constante de sédimentation des ribosomes est d'environ 70S. Chez *Escherichia coli*, Ils sont constitués d'une petite sous-unité (30S) composée d'une molécule d'ARN ribosomique 16S entourée de 21 protéines et d'une grande sous-unité (50S) contenant les ARN ribosomiques 5S, 23S et 31 protéines. Cette liste n'est pas exhaustive car ces études ont principalement été menées sur une bactérie (Wittman, 1983) et d'autres protéines devraient être ajoutées dans le schéma, comme la protéine L36 codée par le gène *secX* (Markmann-Mulisch *et al.* en soumission).

Les protéines ribosomiques sont nommées en fonction de leur appartenance à la petite (S=small) ou à la grande (L=large) sous-unité du ribosome. Un chiffre est adjoint, qui indique la taille relative de la protéine. Par exemple, la protéine S12 sera la 12ème protéine, en taille, de la petite sous-unité du ribosome d'*Escherichia coli*. Nous faisons en effet toujours référence aux tailles des protéines ribosomiques homologues (de fonction identique), appartenant à la protéine d'*E.coli*.

La protéine ribosomique S7 est l'une des 6 protéines de la sous-unité 30S du ribosome se liant directement à l'ARN 16S et impliquées dans l'assemblage du ribosome (Held *et al.* 1974). Elle est la seule protéine à interagir directement avec la région 3' de l'ARN 16S. La situation de la protéine par rapport à la structure secondaire de l'ARN à été définie plus précisément par Wiener et Brimacombe (1987). Spacialement, elle se situe dans la "tête" de la sous-unité 30S et semble avoir un rôle plutôt structural.

La protéine ribosomique S12 est connue pour être liée à la fonction de la traduction, depuis que des mutations (Ozaki *et al.* 1969; Funatsu et Wittmann, 1972) ont été localisées dans cette protéine, qui donnent une résistance à la streptomycine.

3.5.1 Comparaison des protéines S7 et S12 de différents organismes.

Il est intéressant de comparer la structure primaire des quelques protéines S7 et S12 connues, afin d'observer si un rapport structures-fonction peut être établi. Nous vérifierons visuellement dans un premier temps par le programme DIAGON, que les gènes S7 et S12 donnés comme homologues, sont suffisemment similaires pour autoriser une même appellation.

Figure 30a : Comparaison diagonale des protéines S7 et S12 de soja avec les protéines homologues de *E.coli* Une similarité est significative si l'oeil voit apparaître une diagonale.

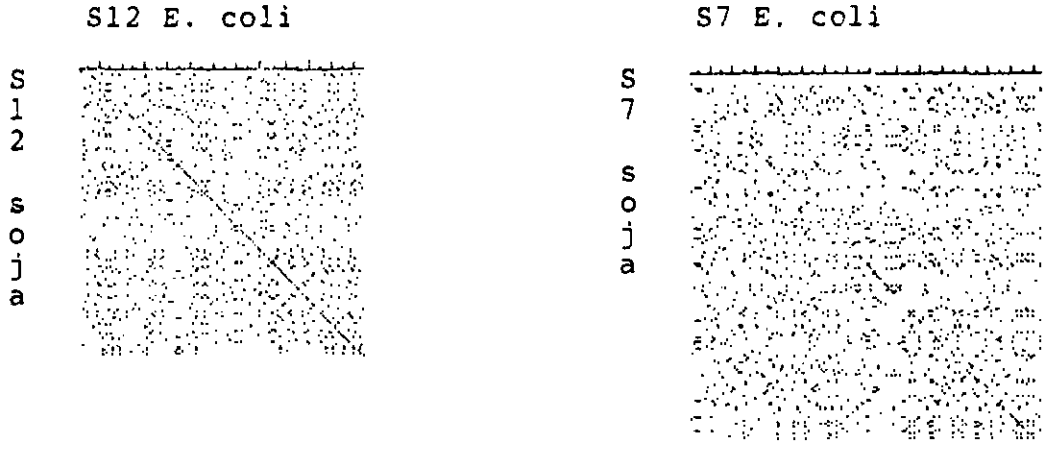
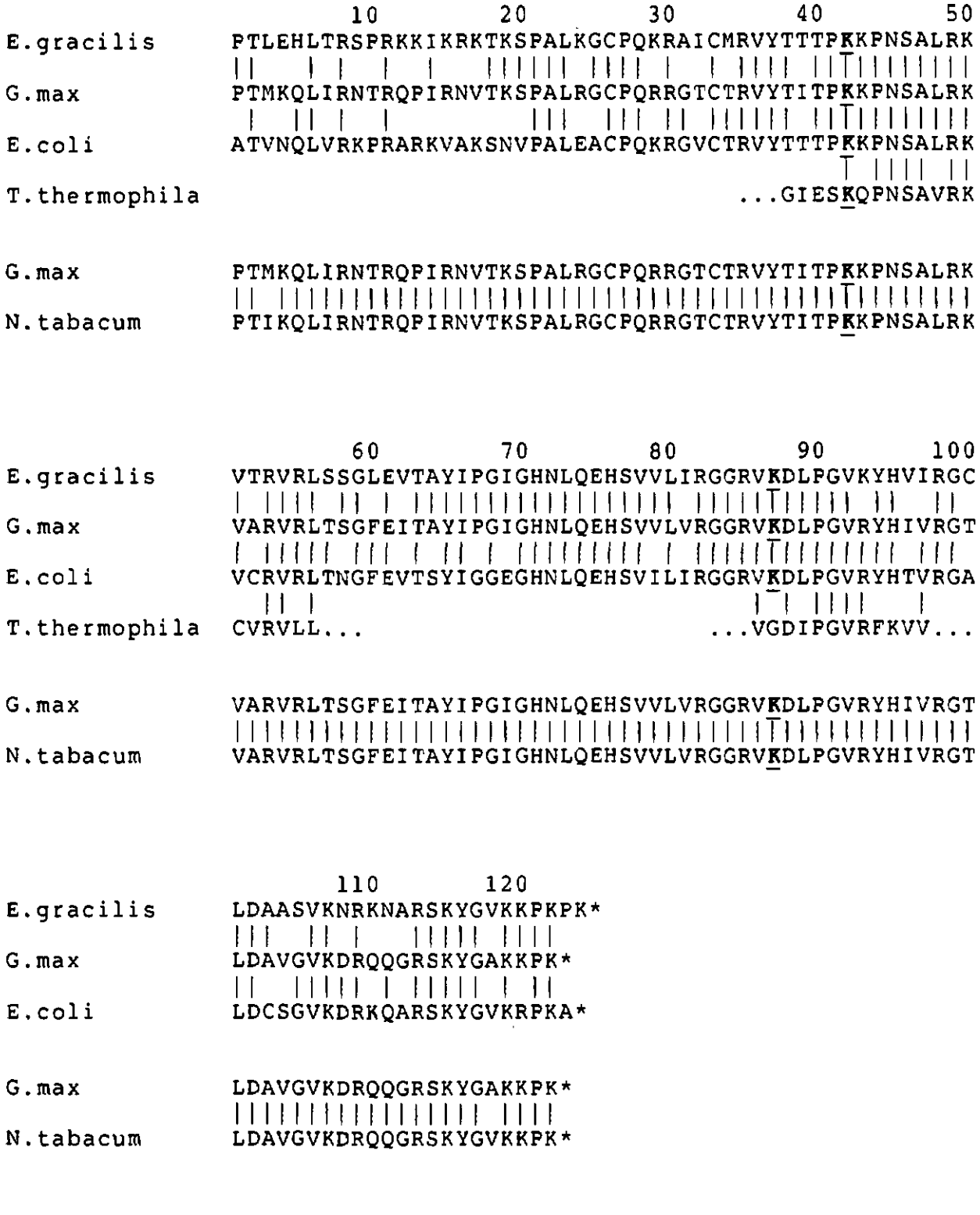


Figure 30b : Comparaisons de différentes protéines S12 d'origine chloroplastique et des parties conservées de la protéine ribosomique d'origine nucléaire de *Tetrahymena thermophila*



Escherichia coli (Post et Nomura, 1980)
Euglena gracilis (Montandon et Stutz, 1984)
Nicotiana tabacum (Fromm *et al.* 1986; Torazawa *et al.* 1986)
Tetrahymena thermophila (Nielsen *et al.* 1986)

Comparaison des protéines ribosomiques S7

```
                  10        20        30        40        50
E.gracilis  SRRRRAKKRIISQDPIYNSTLASKVINKILLNGKKTLAQYIFYETMKNIQ
G.max       SRRGTAEEKTAKSDPIYRNRLVNMLVNRILKHGKKSLAYQIIYRAMKKIQ
E.coli      PRRRVIGQRKILPDPKFGSELLAKFVNILMVDGKKSTAESIVYSALETLA

G.max       SRRGTAEEKTAKSDPIYRNRLVNMLVNRILKHGKKSLAYQIIYRAMKKIQ
N.tabacum   SRRGTAEKKTAKSDPIYRNRLVNMLVNRILKHGKKSLAYQIIYRAVKKIQ

                  60        70        80        90        100
E.gracilis  EIYKKDPLDILRKAIKNASPQMETRKRRIGGTIYQVPVEVKEDRGTSLAL
G.max       QKTETNPLSVLRQAIRGVTPDIAVKARRVGGSTHQVPVEIGSTQGKALAI
E.coli      QRSGKSELEAFEVALENVRPTVEVKSRRVGGSTYQVPVEVPVRRN ALAM

G.max       QKTETNPLSVLRQAIRGVTPDIAVKARRVGGSTHQVPVEIGSTQGKALAI
N.tabacum   QKTETNPLSVLRQAIRGVTPDITVKARRVGGSTHQVPIEIGSTQGKALAI

                  110       120       130       140       150
E.gracilis  KFIIEKARERKGRGISTKLKNEIIDASNNTGEAVKKKEEIHKTAEANKAF
G.max       RWLLGASRKRPGRNMAFKLSSELVDAAKGSGDAIRKKEETHRMAEANRAF
E.coli      RWIVEAARKRGDKSMALRLADELSDAAENKGTAVKKREDVHRMAEANKAF

G.max       RWLLGASRKRPGRNMAFKLSSELVDAAKGSGDAIRKKEETHRMAEANRAF
N.tabacum   RWLLAASRKRPGRNMAFKLSSELVDAAKGSGDAIRKKEETHRMAEANRAF

E.gracilis  SNMKF*

G.max       AHFR*
E.coli      AHYR*

G.max       AHFR*
N.tabacum   AHFR*
```

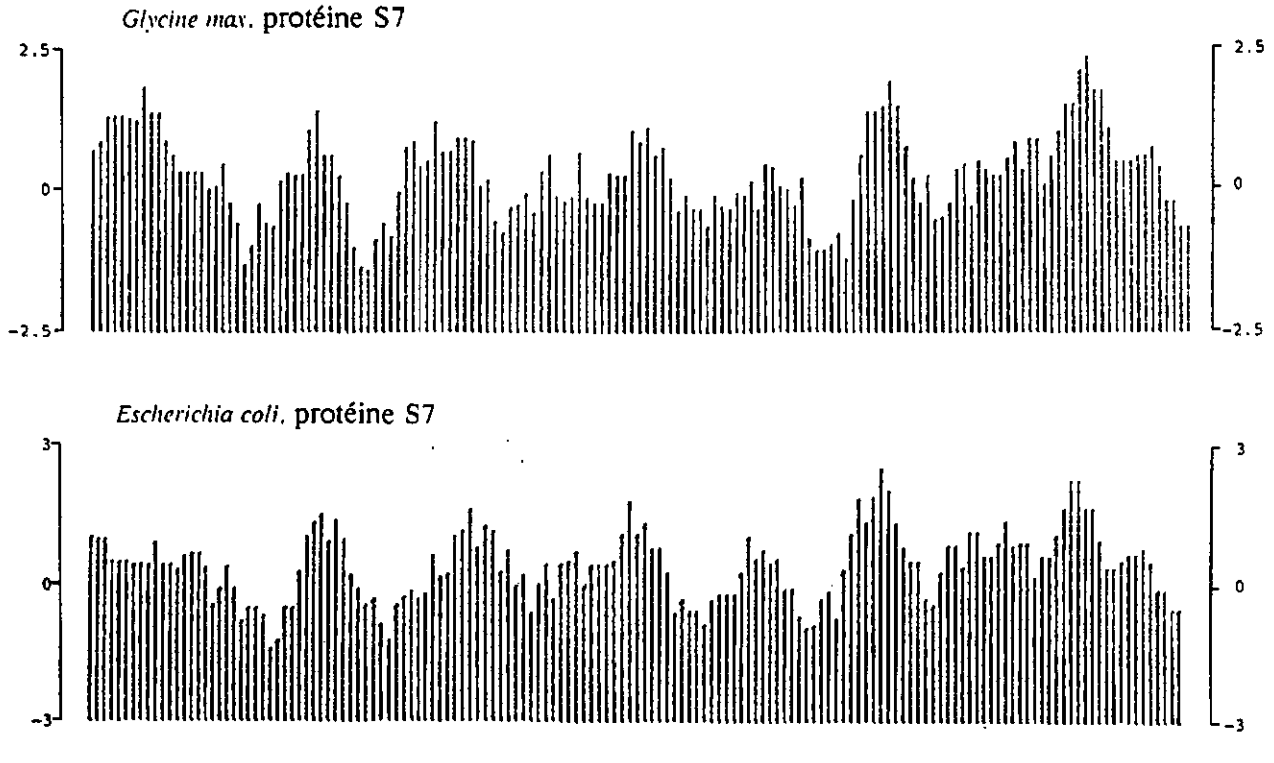
Escherichia coli (Post et Nomura, 1980)
Euglena gracilis (Montandon et Stutz, 1984)
Nicotiana tabacum (Shinozaki *et al.* 1986)

Nous remarquons que les séquences des protéines S12 et S7 de tabac et de soja sont pour ainsi dire identiques (respectivement 98% et 94% des acides aminés sont conservés). Les plantes supérieures paraissent une fois de plus reliées étroitement entre elles sur le plan génétique.

Par contre, le résultat de la comparaison des deux protéines ribosomiques de soja (plante supérieure) avec les protéines d'Euglène et de la bactérie *Escherichia coli* apporte beaucoup plus d'informations. La liberté laissée aux mécanismes de l'évolution est semble-t-il beaucop plus faible dans le cas de la S12 (72 et 70% de similarité) que dans celui de la S7 (42%). Ces résultats pourraient à notre avis refléter l'importance qu'a la protéine S12 au niveau de la fonction du ribosome, alors que la S7 n'aurait qu'un rôle structural, où la structure secondaire, l'hydrophobicité, l'hydrophilicité sont plus conservés que la structure primaire. Le rôle structural de la S7 ne doit pas être négligé pour autant, puisqu'elle est l'une des protéines qui se lie directement avec l'ARN 16S. Les protéines S9, S19, S14 et S10 ne peuvent s'adjoindre à la petite sous-unité du ribosome que si la S7 est déjà présente (Wittmann, 1973).

Nous pouvons d'ailleurs confirmer que la structure générale est conservée en comparant les profils d'hydrophilicité de différentes protéines S7. Cette analyse donne simultanément une bonne idée des régions antigéniques de ces protéines (Hopp et Wood 1981) et qui permet de deviner d'où provient l'immunocorrélation existant entre la protéine chloroplastique et la protéine bactérienne (Bartsch 1985).

Figure 31 : Comparaison des profiles d'hydrophilicité des protéines S7 de soja et de *E. coli* (programme PRPLOT)



3.5.2 Rapports structure-fonction

Si nous résumons les seuls liens bien connus entre les séquences présentées et la fonction du ribosome, nous trouvons que :

Toutes les protéines S12 présentées dans la Figure 30b, sont de type str^s, c'est à dire sensibles à la streptomycine. Cet antibiotique à la particularité de se lier au ribosome et d'abolir une synthèse normale des protéines le codon n'étant plus reconnu avec la même fidélité. Des mutations au niveau du *rpsl2* affectent surtout l'acide aminé lysine du codon 42 et/ou 87, et lorsque cet acide aminé est changé par asparagine, arginine, ou thréonine, le ribosome acquiert la résistance à la streptomycine (Ozaki *et al.* 1969; Funatsu et Wittmann, 1972).

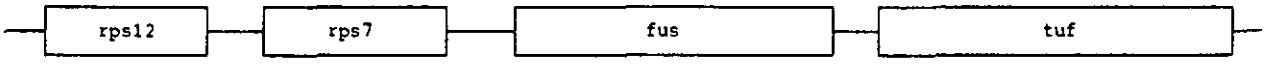
Pour la protéine S7, Les expériences de crosslinking induit par lumière ultraviolet (Zwieb et Brimacombe, 1980) et par méthode chimique (Wower et Brimacombe, 1983) prouvent que l'acide aminé méthionine 114 d'*Escherichia coli* est à la proximité directe de l'uridine 1240 de l'ARN 16S. Nous remarquerons que ce méthionine est conservé, mais en position 115 des séquences protéiques de soja et de tabac. L'Euglène n'a pas gardé cet acide aminé mais il est intéressant de remarquer que ce méthionine est proche d'une région mieux conservée et où la conservation d'une fonction n'a peut être pas permis à l'évolution une divergence plus grande que celle qui existe actuellement.

3.5.3 Comparaison de la structure des opérons contenant le *rpsl2*

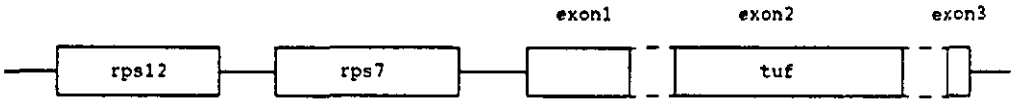
Un autre aspect intéressant est la disposition des gènes ribosomiques considérés dans les différents opérons. Si le chloroplaste a réellement une origine endosymbiotique telle que le propose Schwartz et Dayhoff (1978), nous nous attendons à ce que non seulement, les protéines codées soient similaires, mais que l'agencement des gènes soit identique. Nous avons comparé la structure des gènes d'*Escherichia coli*, d'Euglène et du soja.

Figure 32 : Comparaison de l'environnement et de la structure de différents gènes *rps12*

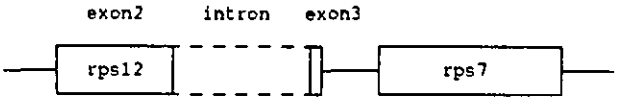
str opéron d'*E.coli*



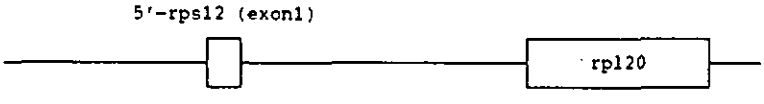
organisation des gènes chloroplastiques d'*E.gracilis*



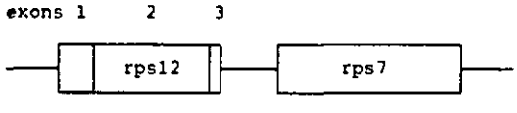
opéron *3'-rps12* de *G.max*



opéron *5'-rps12* de *G.max*



ARN messager après *cis-* et *trans-* splicing



La situation présente de ces différents gènes, fait penser qu'une origine commune est éventuellement possible. L'Euglène dans le cas précis est un organisme intéressant car son ADN chloroplastique semble se situer évolutivement parlant entre la bactérie et la plante supérieure. Mais il reste tout de même le problème des introns !

Le gène *TufA*, codant pour le facteur d'élongation Tu, est interrompu par des introns très courts et n'entrant pas dans les groupes habituels (Montandon et Stutz, 1983). Le *rps12* au niveau des plantes supérieures est un gène subissant du *cis-* et du *trans-* splicing. Celà pose le problème de savoir d'où ces introns proviennent !

Nous nous trouvons face à une question pour laquelle nous n'avons pas de réponse définitive. Les introns sont-ils des vestiges du passé où appartiennent-t-il à des organismes hautement évolués ?

Nous penchons, ainsi que le propose Darnell *et al.* (1986), pour la théorie qui veut que le gène évolué soit celui sans introns de la bactérie. En effet, un certain nombre de constatations peuvent laisser croire que les bactéries avec leurs temps de génération très court (de 20 minutes à quelques heures) ont eu tout le temps et les pressions de l'environnement nécessaires pour perdre les introns inutiles. Il est évident que le temps de génération d'environ 20 ans de l'Homme au niveau de la procréation ne favorise pas une évolution rapide. Mis à part ces considérations hautement spéculatives, il y a de plus en plus d'évidence qu'une chimie des ARN a lieu sans la présence de protéines. L'ARN, qui sert de génome aux rétovirus (Bishop, 1978), qui peut rendre les ribosomes résistants à la streptomycine (Montandon *et al.* 1986), qui sert d'enzyme dans le phénomène du splicing (Zaugg et Cech, 1985), qui participe à la biosynthèse de la chlorophylle (von Wettstein *et al.* 1987), sont autant d'exemples que les ARN peuvent avoir des activités enzymatiques et pourraient limiter le nombre des molécules de nature différentes qui aurait été nécessaires à l'origine de la vie.

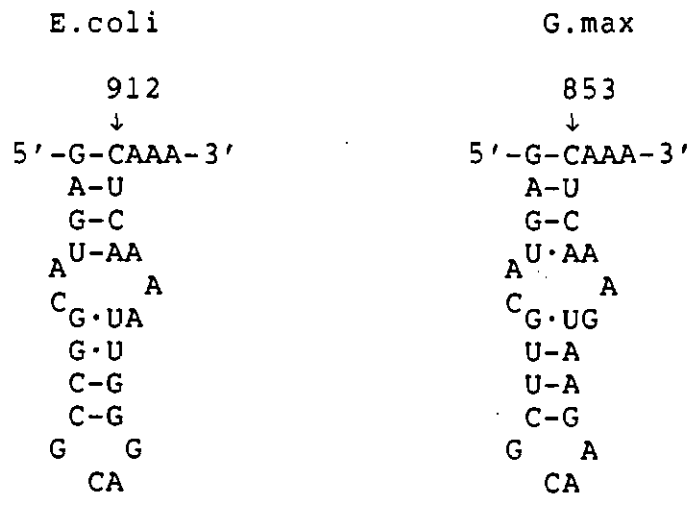
Nous convenons que les introns pourraient avoir eu une certaine utilité dans la duplication de domaines intéressants des protéines (tels les sites de fixation du GTP,...) au cours de l'évolution, mais cela n'exclut pas que le splicing puisse être considéré comme une relique du passé.

3.5.4 L'ARN 16S : structure et fonction.

Puisque la région que nous avons séquencée comprend le gène codant pour l'ARN 16S, il nous paraît nécessaire de souligner l'importance, jusqu'ici négligée, de l'ARN lui-même dans le processus de la traduction. Nous présenterons quelques séquences dont l'importance semble maintenant confirmée.

La première structure essentielle est la structure secondaire située dans le voisinage immédiat du C912 de l'ARN d'*E. coli* (Montandon *et al.* 1986). Cette structure est en effet directement impliquée dans les phénomènes de la traduction, puisque un changement du C912 en U donne au ribosome la résistance à la streptomycine. Si nous comparons la structure de cette région pour *E.coli* souche sauvage et le soja, nous observons une identité des régions conservées, qui permet de prévoir que le soja est sensible à la streptomycine.

Figure 33 : Comparaison de la structure secondaire responsable de la résistance à la streptomycine.



Une autre analyse récemment publiée propose que le rôle de l'ARN 16S dans la traduction ne s'arrête pas là. Une analyse de toutes les séquences codantes pour des messagers au niveau des gènes nucléaires actuellement séquencés (Trifonov, 1987) fait remarquer qu'il existe une G-périodicité au niveau du premier nucléotide de tous les codons. Une analyse parallèle au niveau des gènes chloroplastiques connus, est sur le point d'être publiée (Goloubinoff, communication personnelle) et démontre qu'il en est de même pour les gènes codant pour les protéines ribosomiques chloroplastiques uniquement. Dès lors, il était intéressant de vérifier si les structures C-périodiques observées au niveau de l'ARN 16S par Trifonov et qu'il proposait comme un mécanisme de sécurité de lecture "proof-reading" étaient présentes dans l'ARN 16S codé par le chloroplaste de soja.

Figure 34 : Comparaison des régions C-périodiques de l'ARN 16S

```
E.c  Région (498)  AAGCACCGGCUAACUCCGUGCCAGCAGCCGCGGUAAU (534)
                      ♦  ♦  ♦     ♦     ♦  ♦  ♦  ♦
G.m         (441)  AAGCAUCGGCUAACUCUGUGCCAGCGGCCGCGGUAAG (477)

E.c         (1392)  GUACACCGCCCGUCA (1408)
                      ♦  ♦  ♦
G.m         (1331)  GUACACCGCCCGUCA (1345)

E.c         (1522)  CUGCGGUUGGAUCACCUCCUU (1542)
                    ↑  ♦  ↓     ♦  ♦  ♦
G.m         (1451)  GUGCGGCUGGAUCACCUCCUU (1471)
```

Nous avons ainsi une conservation de la structure "$C_{NN}C_{NN}C$"qui donnerait une meilleure fidélité de la traduction et pourrait éventuellement permettre au ribosome de mieux repérer le 1er nucléotide du codon et de l'empêcher ainsi de perdre la phase du message. Remarquons que la région 1530 (chez *E.coli*) est la partie 3' de l'ARN ribosomique contenant la séquence CCUCC complémentaire au site de fixation des ARN-m procariotiques (Shine et Dalgarno, 1974).

Une vérification de l'usage des codons nous montre que la G-périodicité des séquences codantes est observable pour les gènes ribosomiques de soja également. Le score du *rps12* n'est pas très élevé, mais le début du gène de soja (ainsi que celui d'*E.coli*) possèdent presque une anti-périodicité, c'est à dire qu'ils ont plus de G en seconde position du codon qu'en première position.

Nous devons signaler que le début du gène *psbA*, qui est un gène dont l'expression est induite par la lumière, présente également une anti-périodicité. Ce simple fait peut faire présumer que la G-periodicité des ARN-m ne soit pas le seul mécanisme de vérification de la phase du message existant.

Figure 35 : Usage des codons et G-périodicité

rps7

	CODON POS.1	CODON POS.2	CODON POS.3	TOTAL
% A	31.41	30.12	44.23	35.25
% T	15.38	24.35	31.41	23.71
% G	32.05	20.51	12.17	21.58
% C	21.15	25.00	12.17	19.44
% A+T	46.79	54.48	75.64	58.97
% G+C	53.20	45.51	24.35	41.02

rps12

	CODON POS.1	CODON POS.2	CODON POS.3	TOTAL
% A	32.23	28.09	38.84	33.05
% T	14.87	23.96	31.40	23.41
% G	28.92	24.79	12.39	22.03
% C	23.96	23.14	17.35	21.48
% A+T	47.10	52.06	70.24	56.47
% G+C	52.89	47.93	29.75	43.52

ndhB (à partir du 1er Méthionine possible)

	CODON POS.1	CODON POS.2	CODON POS.3	TOTAL
% A	30.97	21.23	31.85	28.02
% T	29.64	42.47	36.28	36.13
% G	23.00	15.92	14.15	17.69
% C	16.37	20.35	17.69	18.14
% A+T	60.61	63.71	68.14	64.15
% G+C	39.38	36.28	31.85	35.84

ndhB début selon Shinozaki *et al.* (1986)

	CODON POS.1	CODON POS.2	CODON POS.3	TOTAL
% A	29.75	20.66	32.23	27.54
% T	28.09	40.49	35.53	34.71
% G	28.09	17.35	15.70	20.38
% C	14.04	21.48	16.52	17.35
% A+T	57.85	61.15	67.76	62.25
% G+C	42.14	38.84	32.23	37.74

ORF 143 interne à l'intron du *rps12*

	CODON POS.1	CODON POS.2	CODON POS.3	TOTAL
% A	36.11	33.33	34.72	34.72
% T	26.38	27.77	27.77	27.31
% G	19.44	20.13	20.83	20.13
% C	18.05	18.75	16.66	17.82
% A+T	62.50	61.11	62.50	62.03
% G+C	37.50	38.88	37.50	37.96

L'usage des différents nucléotides à l'intérieur du codon est relativement particulier dans le cas de l'ORF 143. Le pourcentage de thymine, adénine et de cytosine en position 1 et 2 du codon varie passablement face aux autres gènes analysés. Nous pensons que cette analyse donnerait plutôt du poids à l'hypothèse que l'ORF 146 contenue dans l'intron 2 du *rps12* n'est pas traduite. Il ne faut pourtant pas négliger que les ORFs située dans les introns ont souvent un usage des codons particulier car la structure secondaire d'un intron de type I ou de type II doit être conservée.

4. CONCLUSION

Les ARN jusqu'ici un peu ignorés et considérés comme des intermédiaires dans leurs fonctions de messagers et de squelettes structuraux des ribosomes prennent aujourd'hui une grande importance (Cech et Bass, 1986). Ils sont désormais connus pour avoir parfois une activité catalytique. Si l'ARN-enzyme pose beaucoup de questions non encore résolues, la chimie relative aux ARN permettrait par contre d'esquisser une théorie de l'origine de la vie un peu plus complète.

L'étude comparative des gènes *rps7*, *rps12* et *rpl20* nous aura donné un certain nombre d'enseignements.

Nous relèverons la conservation des séquences internes et externes aux parties codantes des gènes de plantes supérieures (tabac, soja), qui est sensiblement plus élevée dans les régions répétées inverses que dans les régions uniques. Non seulement les réarrangements seraient plus fréquents dans les régions uniques (Palmer et Thompson, 1982), mais les mutations locales semblent également plus fréquentes. Nous avons en effet observé, entre le tabac et le soja, une similarité de 94% au niveau des protéines S7 codées dans les régions répétées inverses et seulement de 69% au niveau de la partie connue de la L20. Remarquons que les divergences entre les deux séquences au niveau des parties non codantes de l'ADNct de ces région est encore plus frappante.

Le gène *rps12* pose un certain nombre de questions. La notion de gène comme d'une unité biologique de l'hérédité n'est absolument plus vérifiée au niveau de l'ADN. Les gènes du trypanosome sont divisés et possédent un "leader" commun à différents messagers (Sather et Agabian, 1985), le gène chloroplastique *rps12* et le gène psIAl de *Chlamydomonas* (Kück *et al.* 1987) démontrent que parfois les gènes sont divisés et que les ARN fabriqués à partir de deux unités de transcriptions subissent un *trans* splicing délivrant un ARN messager unique et fonctionnel. L'unité du gène ne se trouve plus au niveau de l'ADN, mais uniquement au niveau de l'ARN messager mature. Nous pourrions tenter de donner une nouvelle définition du gène en termes moléculaires (modifiée de Darnell *et al.* 1986) et qui serait :

> **Le gène est la somme des séquences d'acides nucléiques (généralement une séquence d'ADN) qui codent pour un polypeptide ou un ARN fonctionnel.**

Nous convenons bien volontiers que le problème du trans splicing reste posé. Les analyses de structure des introns semblent indiquer que le *trans* splicing n'est pas aussi différent du *cis* splicing classique que ce que l'on aurait pu attendre. L'analyse de la transcription démontre néanmoins un manque d'efficacité ou une régulation différente du *trans* splicing par rapport au *cis* splicing appliqué aux ARN du *rps12*. Si les méchanismes chimiques sont en passe d'être résolus au niveau des gènes nucléaires, l'étude du splicing dans le chloroplaste n'en est qu'à son début.

Il se pose maintenant les questions de savoir s'il y a réellement formation de lariat ou non, s'il existe du self-splicing *in vivo* ou si la maturation des ARN est catalysée par des protéines. Il reste à étudier de plus près les différents précurseurs du *rps12* et à tenter de comprendre si la séparation du gène en deux tronçons a une utilité dans la régulation de la quantité des transcrits. Nous savons que la protéine S7 d'*Escherichia coli* est une protéine régulatrice autogène (Dean *et al.* 1981) et fonctionne comme un répresseur de la traduction de son propre ARN. Par contre, le gène *rps7* de soja ne possède pas de structure directement reliée au site de fixation de la protéine S7 d'*E.coli* sur son ARN-m. Ceci nous fait penser que l'un des mécanismes de régulation possible serait éventuellement au niveau du *trans* splicing dont l'efficacité semble modulée en fonction de l'état physiologique des feuilles. Les feuilles d'une plante en pleine croissance ont sembl-t-il une quantité d'ARN *rps12* mature plus grande que les plantes entre deux périodes de croissance.

L'importance toujours plus grande des ARN est aussi démontrée par l'étude du ribosome et de sa fonction. Il est frappant de constater que parmi les derniers résultats concernant le ribosome, plusieurs sont particulièrement liés à la structure de l'ARN 16S. La sensibilité ou la résistance du ribosome à l'égard de la streptomycine dépend de la structure de cet ARN (Montandon *et al.* 1986). La bonne lecture des ARN messagers semble dépendre de la C-périodicité de trois régions spacialement relativement proches (Trifonov, 1987). Les données récoltées avec le soja ne contredisent aucunement les conclusions de ces travaux et il est possible que la chimie des ARN nous réserve encore quelques surprises.

En conclusion, nous pouvons constater qu'il y a encore de quoi occuper bien des vies si l'on désire élucider toutes les questions qui se posent et vont se poser à propos du ribosome, de sa fonction et de la régulation de sa biosynthèse.

5. BIBLIOGRAPHIE

von Allmen J.-M. et Stutz E. (1985). Structure and sequence of the chloroplast genes for the ribosomal proteins S12 and S7 from soybean. Experientia 41:814

von Allmen J.-M. et Stutz E. (1987). Complete sequence of "divided" *rps12* (r-protein S12) and *rps7* (r-protein S7) gene in soybean chloroplast DNA. Nucleic Acids Res. 15:2387

Amersham International plc, UK (1983). M13 cloning and sequencing handbook

Avery O.T, McLeod C.M. et McCarty M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J. Exp. Med. 79:137-158

Bartsch M. (1985). Correlation of chloroplast and bacterial ribosomal proteins by cross-reactions of antibodies specific to purified *Escherichia coli* ribosomal proteins. J. Biol. Chem. 260:237-241

Bedbrook J.R. et Kolodner R. (1979). The structure of chloroplast DNA. Ann. Rev. Plant Physiol. 30:593-620

Berget S.M., Moore C. et Sharp P.A., (1977). Spliced segment at the 5' terminus of adenovirus 2 late mRNA. Proc Natl. Acad. Sci. 74:3171-3175

Berman M.L. et Landy A. (1979). Promotor mutations in the transfer RNA gene *TyrT* of *Escherichia coli*. Proc Natl. Acad. Sci. 76:4303-4307

Biggin M.D., Gibson T.J. et Hong G.F. (1983). Buffer gradient gels and 35S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. 80:3963-3965

Birnboim H.C. et Doly J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. 7:1513-1523

Bishop J.M. (1978). Retroviruses Ann. Rev. Biochem. 4735-88

Bonner T.I., Brenner D.J., Neufeld B.R. et Britten R.J. (1973). Reduction in the rate of DNA reassociation by sequence divergence. J. Mol. Biol. 81:123-135

Boyer H.W. et Roulland-Dussoix (1969). A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. J. Mol. Biol. 41:459-472

Boyer P.D., (1986). A sensitive technique for detection of RNA with single-stranded probes. Nucl. Acids Res. 14:7505.

Breathnach R. et Chambon P. (1981). Organization and expression of Eucariotic split genes coding for protein. Ann. Rev. Biochem. 50:349-383

Burke J.M., Belfort M., Cech T.R., Davies R.W., Schweyen R.J., Shub D.A., Szostak J.W. et Tabak H.F. (1987). Structural conventions for group I intron.
Nucl. Acids Res. 15:7217-7221

Campbell D.A., Thornton D.A. et Boothroyd J.C. (1984). Apparent discontinuous transcription of *Trypanosoma brucei* variant surface antigen genes
Nature 311:350-355

Cech T.R. (1983). RNA splicing : Three themes with variations.
Cell 34:713-716

Cech T.R. et Bass B.L. (1986). Biological catalysis by RNA.
Ann. Rev. Biochem. 55:599-629

Cenatiempo Y., Robakis N., Reid B.R., Weissbach H. et Brot N. (1982). *In vitro* expression of *Escherichia coli* ribosomal protein L10 gene : tripeptide synthesis as a measure of functionnal mRNA.
Arch. Biochem. Biophys. 218:572-578

Chaconas G. et van de Sande J.H. (1980). 5′ ^{32}P Labeling of RNA and DNA restriction fragments.
Methods Enzymol. 65:75-85

Chomin A., Mariottini P., Cleeter M.W.J., Ragan C.I., Matsuno-Yagi A., Hatefi Y., Doolittle R.F et Attardi G. (1985). Six unidentified reading frames of human mithocondrial DNA encode components of the respiratory-chain NADH deshydrogenase.
Nature 314:592-597

Chow L.T., Gelinas R.E., Broker T.R. et Roberts R.J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA.
Cell 12:1-8

Clewell D.B. (1972). Nature of ColEl plasmid replication in *Escherichia coli* in presence of chloramphenicol
J. Bacteriol.110:667

Dagert M. et Ehrlich S.D. (1979). Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells
Gene 6:23

Darnell J., Lodish H. et Baltimore D. (1986). Molecular cell biology Scientific american books, Freeman and Co, New York.

Dean D., Yates J.L. et Nomura M. (1981). Identification of ribosomal protein S7 as a repressor of translation within the *str* operon
Cell 24:413-419

Deng X.W. et Gruissem W. (1987). Control of plastid gene expression during development : The limited role of transcriptional regulation.
Cell 49:379-387

Denhardt D.J., Dressler D. et Ray D.S. (1978). The single-stranded DNA phages. Cold Spring Harbor Laboratory, New York.

Dente L., Cesareni G. et Cortese R. (1983). pEMBL : A new family of single stranded plasmids.
Nucleic Acids Res. 6:1645-1655

Ellis R.J. (1981). Chloroplast proteins : Synthesis, transport, and assembly.
Annu. Rev. Plant Physiol. 32:111-137

Fromm H., Edelman M., Goloubinov P. et Galun E. (1986). The enigma of the gene coding for ribosomal protein S12 in the chloroplasts of *Nicotiana*
Nucleic Acids Res. 14:883-898

Fukuzawa H., Kohchi T., Shirai H., Ohyama K., Umesono K., Inokuchi H. et Ozeki H. (1986). Coding sequences for chloroplast ribosomal S12 from the liverwort, *Marchantia polymorpha*, are separated far apart on the different DNA strands.
FEBS Lett. 198:11-15

Funatsu G. et Wittmann H.G. (1972). Ribosomal proteins. Location of amino-acid replacements in protein S12 isolated from *Escherichia coli* mutants resistant to streptomycine.
J. Mol. Biol. 68:547-550

Gillis M., De Ley J. et De Cleene M. (1970). The determination of molecular weight of bacterial genome DNA from renaturation rates.
Eur. J. Biochem. 12:143-153

Gingrich J.C. et Hallick R.B. (1985). The *Englena gracilis* chloroplast Ribulose 1,5-bisphosphate carboxylase gene
J. Biochem. Chemistry 260:16156-16168

Godson G.N. et Vapnek D. (1973). A simple method of preparing large amounts of ϕX174 RFI supercoiled DNA.
Biochim. Biophys. Acta 299:516-520

Goldenberg C.J. et Hauser S.D. (1983). Accurate and efficient in vitro splicing of purified precursor RNAs specified by early region 2 of the adenovirus genome.
Nucl. Acids Res. 11:1337-1348

Grunstein M. et Hogness D. (1975). Colony hybridization : A method for the isolation of cloned DNAs that contain a specific gene.
Proc. Natl. Acad. Sci. 72:3961-3965

Hallick R.B. et Bottomley W. (1983). Proposals for the naming of chloroplast genes.
Plant Mol. Biol. Report. 1:38-43

Harley C.B. et Reynolds R.P. (1987). Analysis of *E.coli* promotors sequences.
Nucl. Acids Res. 15:2343-2361

Held W.A., Ballou B., Mizushima S. et Nomura M. (1974). Assembly mapping of 30S ribosomal proteins from *Escherichia coli*
J. Biol. Chem. 249:3103-3111

Herrmann R.G., Whitfeld P.R. et Bottomley W. (1980). Construction of a SalI-PstI restriction map of spinach chloroplast DNA using low-gelling-temperature-agarose electrophoresis.
Gene 8:179-191

Hershey A.D. et Chase M. (1952). Independant functions of viral protein and nucleic acid in growth of bacteriophage.
J. Gen. Physiol. 36:39-56

Hildebrand M., Hallick R.B., Passavant C.W. et Bourque D.P. ( )
Trans splicing in chloroplasts : the *rps12* loci of *Nicotiana tabacum*
En soumission, non publié

Hirt B., (1967). Selective extraction of polyoma DNA from infected mouse cell cultures.
J. Mol. Biol. 26:365-369

Hopp T.P. et Woods K.R. (1981). Prediction of protein antigenic determinants from amino acid sequence.
Proc. Nat. Acad. Sci. USA 78:3824-3828

Hu N. et Messing J. (1982). The making of strand specific M13 probes
Gene 17:271-277

James G.T. (1978). Inactivation of the proteases inhibitor Phenylmethylsulponyl fluoride in buffers.
Analytical Biochem. 86:574-579

Kalberer P.P., Buchanan B.B. et Arnon D.I. (1967). Rates of photosynthesis by isolated chloroplasts.
Proc. Natl. Acad. Sci. 57:1542-1549

Karabin G.D., Farley M. et Hallick R.B. (1984). Chloroplast gene for Mr 32000 polypeptide of photosystem II in *Euglena gracilis* is interrupted by four introns with conserved boundary sequences
Nucl. Acids Res. 12:5801-5812

Karlin-Neumann G.A. et Tobin E.M. (1986). Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework.
EMBO J. 5:9-13

Kawashimia N. et Wildman S.G. (1972). Studies on fraction I protein. Mode of inheritance of primary structure in relation to wether chloroplast or nuclear DNA contains the code for a chloroplast protein.
Biochem. Biophys. Acta 262:42-49

Keller M. et Michel F. (1985). The introns of the *Euglena gracillis* chloroplast gene which codes for the 32-kDa protein of photosystem II
FEBS Lett. 179:69-73

Kelley W.S. et Stump K.H., (1979). A rapid procedure for isolation of large quantities of *Escherichia coli DNA* polymerase I utilizing a λ pol A transducing phage.
J. Biol. Chem. 254:3206-3210

Koch W., Edwards K. et Kössel H. (1981). Sequencing of the 16S-23S spacer in a ribosomal RNA operon of *Zea mays* chloroplast DNA reveals two split tRNA genes.
Cell 25:203-213

Koller B., Gingrich J.C., Stiegler G.L., Farley M.A., Delius H. et Hallick R.B. (1984). Nine introns with conserved boundary sequences in the *Euglena gracilis* chloroplast ribulose-1,5-bisphosphate carboxylase.
Cell 36:545-553

Koller B., Clarke J. et Delius H. (1985). The structure of precursor mRNAs and of excised intron RNAs in chloroplasts of *Euglena gracilis*.
EMBO J. 4:2445-2450

Koller B., Fromm H., Galun E. et Edelman M. (1987). Evidence for *in vivo trans* splicing of pre-mRNAs in tobacco chloroplasts
Cell 48:111-119

Konarska M.M., Padgett R.A. et Sharp P.A. (1985). *Trans* splicing of mRNA precursors *in vitro.*
Cell 42:165-171

Kück U., Choquet Y., Schneider M., Dron M. et Bennoun P. (1987). Structural and transcription analysis of two homologous genes for the P700 chloropyll α-apoproteins in *Chlamydomonas reinhardii* : evidence for *in vivo trons* splicing.
EMBO J. 6:2185-2195

Lescure A.-M., Bisanz-Seyer C., Pesey H., Mache R. (1985). *In vitro* transcription initiation of the spinach chloroplast 16S rRRN gene at two tandem promoters.
Nucleic Acids Res. 13:8787-8796

Lindahl L. et Zengel M. (1986). Ribosomal genes in *Escherichia coli.*
Ann. Rev. Genet. 20:297-326

Lis J.T. (1980). Fractionation of DNA fragments by polyethylene glycol induced precipitation.
Methods in enzymology 65:347-353

Mandel M. et Higa G. (1970). Calcium dependent bacteriophage DNA infections.
J. Mol. Biol. 53:159-162

Maniatis T., Frisch E.F. et Sambrook J. (1982). Molecular cloning.
Cold Spring Harbor Laboratory, New-York. pp 62, 93-94, 150-161, 164-165

Markmann-Mulisch U., von Knoblauch K., Lehmann A. et Subramanian A.R. (en soumission, 1987). Nucleotide sequence and linkage map position of the *secX* gene in maize chloroplast and evidence that it encodes a protein belonging to the 50S ribosomal subunit.
Biochem. Int. en soumission

Maxam A.M. et Gilbert W. (1977). A new method for sequencing DNA
Proc. Natl. Acad. Sci. 74:560-564

Mc Conaughy B.L., Laird C.D. et Mc Carthy B.J. (1969). Nucleic acid reassociation in formamide.
Biochemistry 8:3289-3295

Mc Master G.K. et Carmichael G.G. (1977). Analysis of single- and double-stranded nucleic acids on polyacylamide and agarose gels by using glyoxal and acidine orange.
Proc. Natl. Acad. Sci. 74:4835-4838

Messing J., Crea R. et Seeburg P.H. (1981). A system for shotgun DNA sequencing.
Nucleic Acids Res. 9:309-322

Messing J. et Vieira J. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragment.
Gene 19:269-272

Michel F. et Dujon B. (1983). Conservation of RNA secondary structures two intron famillies including mitochondrial-, chloroplast-, and nuclear-encoded members.
EMBO J. 2:33-38

Milhausen M., Nelson R.G., Sather S., Selkirk M. et Agabian N. (1984) Identification of a small RNA containing the Trypanosome spliced leader : a donnor of shared 5' sequences of trypanosomatid mRNAs.
Cell 38:721-729

Mills W.R. et Joy W.W. (1980). A rapid method for isolation of purified physiologically active chloroplasts, used to study the intracellular distribution of amino acids in pea leaves.
Planta 148:75-83

Montandon P.-E. et Stutz E. (1983). Nucleotide sequence of *Euglena gracilis* chloroplast genome region coding for the elongation factor Tu : evidence for a spliced mRNA.
Nucleic Acids Res. 17:5877-5892

Montandon P.-E. et Stutz E. (1984). The genes for the ribosomal S12 and S7 are clustered with the gene for the gene for the EF-Tu protein on the chloroplast genome of *Euglena gracilis*
Nucl. Acids Res. 12:2851-2859

Montandon P.-E., Wagner R. et Stutz E. (1986). *E. coli* ribosomes with a C912 to U base change in the 16S rRNA are streptomycine resistant.
EMBO J. 5:3705-3708

NEN (New Englangd Nuclear), (1984). GeneScreenPlus Hybridisation Transfer Membrane. Protocols for electrophoretic and capillary transfer of DNA and RNA, DNA and RNA hybridisation and rehybridisation
Catalogue No NEF-976

Nielsen H., Andreasen P.H., Dreisig H., Kristiansen K. et Engberg J. (1986). An intron in a ribosomal protein gene from *Tetrahymena*
EMBO J. 5:2711-2717

Ohyama K., Fukuzawa H., Kohchi T., Shirai H., Sano T., Sano S., Umesono K., Shiki Y., Takeuchi M., Chang Z., Aota S., Inokuchi H. et Ozeki H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA".
Nature 322:572-574

Ozaki M., Mizushima S. et Nomura M. (1969). Identification and functional characterization of the protein controlled by the streptomycine-resistant locus in *E.coli*.
Nature 222:333-339

Padgett R.A., Grabowski P.J., Konarska M.M. Seiler S. et Sharp P.A., (1986). Splicing of messenger RNA precursors.
Ann. Rev. Biochem. 55:1119-1150

Palmer J.D. et Thompson W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost.
Cell 29:537-550

Palmer J.D., Singh G.P. et Pillay D.T.N. (1983). Structure and sequence evolution of three legume chloroplast chloroplast DNAs
Mol. Gen. Genet. 190:13-19

Peattie D.A. (1979). Direct chemical method for sequencing RNA.
Proc. Natl. Acad. Sci. 76:1760-1764

Post L.E. et Nomura M. (1980). DNA sequence of the *str* operon of *Escherichia coli*.
J. Biol. Chem. 255:4660-4666

Reijnders L., Sloof P., Sival J. et Borst P. (1973).
Gel electrophoresis of RNA under denaturing conditions.
Biochim. Biophys. Acta 324:320-333

Rigby P.W.J., Dieckmann M., Rhoes C., et Berg P. (1977). Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I.
J. Mol. Biol. 113:237-251

Rütti B. (1982). Analyse du génome chloroplastique d'*Euglena gracilis* par des méthodes de transcription et de traduction *in vitro*.
Thèse de doctorat, Université de Neuchâtel pp. 26-28

Sagher D., Grosfeld H. et Edelman M. (1976). Large subunit ribulose bisphosphate carboxylase messenger RNA from *Euglena* chloroplasts
Proc. Natl. Acad. Sci. 73:722-726

Sanger F. et Coulson A.R., (1978). The use of thin acrylamid gels for DNA sequencing.
FEBS Lett. 87:107-110

Sanger F., Coulson A.R., Barrell B.G., Smith A.J.H. et Roe B.A. (1980).
Rapid DNA sequencing.
J. Mol. Biol. 143:161-178

Sather S. et Agabian N. (1985). A 5' spliced leader is added in *trans* to both α- and β-tubulin transcripts in *Trypanosoma brucei*.
Proc. Natl. Acad. Sci. 82:5695-5699

Schleif F. et Wensink P.C. (1981). Practical methods in molecular biology.
Springer Verlag, New-York, Heidelberg, Berlin p. 145

Schwartz R.M. et Dayhoff M.O., (1978). Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts.
Science 199:395-403

Schwarz Z., Kössel H., Schwarz E. et Bogorad L. (1981).
A gene coding for tRNA-Val is located near the 5' terminus of the 16S rRNA gene in the Zea mays chloroplast genome.
Proc. Natl. Acad. Sci. 78:4748-4752

Shapiro M.B. et Senapathy P. (1987). RNA splice junctions of different classes of Eucaryotes : sequence statistics and functionnal implications in gene expression.

Sharp P.A., Sugden B. et Sambrook J. (1973). Detection of two restriction endonuclease activity in *Haemophilus parainfluenzae* using analytical agarose.
Biochemistry 12:3055

Shine J. et Dalgarno L. (1974). The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA is complementary to nonsense triplets and ribosome binding sites.
Proc. Natl. Acad. Sci 71:1342-1346

Shinozaki K., Ohme M., Tanaka M., Wakasugi T., Hayashida N., Matsubayashi T., Zaita N., Chunwongse J., Obokata J., Yamaguchi-Shinozaki K., Ohto C., Torazawa K., Meng B.Y., Sugita M., Deno H., Kamogashira T., Yamada K., Kusuda J., Takaiwa F., Kato A., Tohdoh N., Shimada H. et Sugiura M. (1986) The complete nucleotide sequence of tobacco chloroplast genome : its gene organization and expression.
EMBO J. 5:2043-2049

Shinozaki K., Deno H., Sugita M., Kuramitsu S. et Sugiura M. (1986)b Intron in the gene for the ribosomal protein S16 of tobacco and its conserved boundary sequences.
Mol. Gen. Genet. 202:1-5

Siegenthaler P.A. et Depéry F. (1976). Influence of unsaturated fatty acids in chloroplasts.
Eur. J. Biochem. 61:573-580

Siebenlist U., Simpson R.B. et Gilbert W. (1980). *E.coli* polymerase interacts homologously with two different promotors.
Cell 20:269-281

Singh G., de Lanversin G., Straus N.A. et Pillay D.T.N. (1984). Characterisation of soybean chloroplast ribosomal RNA genes
Curr. Genet. 9:99-105

Singh G.P., Wallen D.G. et Pillay D.T.N. (1985). Positioning of protein-coding genes on the soybean chloroplast genome.
Plant Mol. Biol. 4:87-93

Smith H.O. et Bernstiel M.L. (1976). A simple method for DNA restriction site mapping.
Nucl. Acids Res. 3:2387-2398

Solnick D. (1985). *Trans* splicing of mRNA precursors.
Cell 42:157-164

Spielmann A. (1983). Le soja (Glycine max) : Etude de l'accumulation des protéines de réserve des graines - Analyse du génome chloroplastique.
Thèse de doctorat. Université de Neuchâtel. pp 62-67

Spielmann A. et Stutz E. (1983). Nucleotide sequence of soybean chloroplast DNA regions which contain the *psbA* and *trnH* genes and cover the ends of the large single copy region and one end of the inverted repeats.
Nucleic Acids Res. 11:7157-7167

Spielmann A., Ortiz W. et Stutz E. (1983). The soybean chloroplast genome : Construction of a circular restriction map and location of DNA regions encoding the genes for rRNAs, the large subunit of the ribulose-1,5-bisphosphate carboxylase and the 32 kD protein of the photosystem II reaction center.
Mol. Gen. Genet. 190:5-12

Soutern E. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis.
J. Mol. Biol. 98:503-517

Staden R. (1977). Sequence data handling by computer.
Nucl. Acids Res. 4:4037-4051

Staden R. (1978). Further procedures for sequence analysis by computer.
Nucl. Acids Res. 5:2601-2610

Staden R. (1979). A strategy of DNA sequencing employing computer programs.
Nucl. Acids Res. 7:2601-2610

Staden R. (1980). A new computer method for the storage and manipulation of DNA gel reading data.
Nucl. Acids Res. 8:3673-3694

Staden R. et Mc Lachlan A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences.
Nucl. Acids Res. 10:141-156

Stiegler G.L., Matthews H.M., Bingham S.E. et Hallick R.B. (1982). The gene for the large subunitof ribulose-1,5-carboxylase in *Euglena gracilis* chloroplast DNA : location, polarity, cloning, and evidence for an intervening sequence.

Strittmatter G. et Kössel H. (1984). Cotranscription and processing of 23S, 4.5S and 5S rRNA in chloroplasts from *Zea mays*.
Nucleic Acids Res. 12:7633-7647

Strittmatter G., Gozdzicka-Jozefiak A. et Kössel H. (1985). Identification of an rRNA operon promotor from *Zea mays* chloroplasts which excludes the proximal tRNA-Val (GAC) from the primer transcript.
EMBO J. 4:599-604

Sutcliffe J.G. (1978). Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322.
Proc. Natl. Acad. Sci. 75:3737-3741

Takaiwa F. et Sugiura M. (1982)a. Nucleotide sequence of the 16S-23S spacer region in an rRNA gene cluster from tobacco chloroplast DNA.
Nucleic Acids Res. 10:2665-2676

Takaiwa F. et Sugiura M. (1982)b. The complete nucleotide sequence of a 23S rrna gene from tobacco chloroplasts.
Eur. J. Biochem. 124:13-19

Teem J.L., Abovich N., Kauter N.F., Schwindinger W.F., Warner J.R., Levy A., Woolford J., Leer R.J., van Raamsdonk-Duin M.M.C., Mager W.H., Planta R.J., Schultz L., Friesen J.D., Fried H. et Rosbach M. (1984). A comparison of yeast ribosomal protein gene DNA sequence.
Nucleic Acids Res. 12:8295-8312

Tohdoh N., Shinozaki K. et Sugiura M. (1981). Sequence of a putative promotor region for the rRNA genes of tobacco chloroplast DNA.
Nucleic Acids Res. 9:5399-5410

Tohdoh N. et Sugiura M. (1982). The complete nucleotide sequence of a 16S ribosomal RNA gene from tobacco chloroplasts.
Gene 17:213-218

Torazawa K., Hayashida N., Obokata J., Shinozaki K. et Sugiura M. (1986) The 5′ part of the gene for ribosomal protein S12 is located 30 kbp downstream from its 3′ part in tobacco chloroplast genome.
Nucleic Acids Res. 14:3143.

Trifonov E.N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences.
J. Mol. Biol. 194:643-652

Tung C.-S. et Harvey S.C. (1987). A common structural feature in promotor sequences of *E. coli*.
Nucleic Acids Res. 15:4973-4985

van der Veen R., Arnberg A.C. and Grivell L.A. (1987). Self splicing of a group II intron in yeast mithocondria: dependance on 5' exon sequences.
EMBO J. 6:1079-1084

Waldmann R., Gross H.J. et Krupp G. (1987). Protocol for rapid RNA sequencing.
Nucl. Acids Res. 17:7209

Weil J.-H. (1979). Biochimie générale
Masson, Paris, New-York, Barcelone, Milan

Weislander L. (1979). A simple method to recover intact high molecular weight RNA and DNA after electrophoretic separation in low gelling temperature agarose gels.
Anal. Biochem. 98:305

von Wettstein D., Kannangara G.C., Gough S., Berry-Lowe S. et Bruyant P. (1987). A chloroplast DNA encoded glutamate tRNA in the regulation of Δ-amino-levulinate and chlorophyll synthesis.
Conférence présentée au meeting d'Arolla du 6 au 9 septembre 1987.

Wiener L. et Brimacombe R. (1987). Protein binding sites on *Escherichia coli* 16S RNA; RNA regions that are protected by proteins S7, S14 and S19 in the presence or absence of protein S9.
Nucleic Acids Res. 15:3653-3670

Wittmann H.G. (1983). Architecture of procaryotic ribosomes.
Ann. Rev. Biochem. 52:35-65

Wower I. et Brimacombe R. (1983). The localisation of multiple sites on 16S RNA which are cross-linked to S7 and S8 in *Escherichia coli* 30S ribosomal subunits by treatment with 2-iminothiolane

Yamada T. et Shimaji M. (1987). An intron in the 23S rRNA gene of *Chlorella* chloroplasts : Complete nucleotide sequence of the 23S rRNA gene.
Curr. Genet. 11:347-352

Zaita N., Torazawa K., Shinozaki K. et Sugiura M. (1987). *Trans* splicing *in vivo* : joining of transcripts from the "divided" gene for ribosomal protein S12 in the chloroplast of tobacco.
Febs Lett. 210:153-156

Zaug A.J. et Cech T.R. (1986). The intervening sequence RNA of *Tetrahymena* is an enzyme.
Science 231:470-475

6. ANNEXE

6.1 Mode d'emploi détaillé des programmes ANALYSE et HEID

Une part non négligeable de notre travail ayant consisté à programmer, et à mettre au point un sysème informatique d'analyse des séquences d'ADN, d'ARN et de protéines. Nous présentons le mode d'emploi des programmes que nous avons mis au point.

Nous adjoindrons un listing des deux programmes et sur demande, nous enverrons une bande magnétique contenant les programmes et le mode d'installation.

Remarque :

Le mode d'emploi pouvant être obtenu séparément, La numérotation des pages suit une numérotation particulière

Table des matières

PROGRAMME ANALYSE :

INTRODUCTION A1

DESCRIPTION DES FICHIERS D'ENTREE COMPATIBLES A3

- Fichiers des enzymes de restrictions du type STADEN A3
- Fichiers des enzymes de restrictions du type PFEIFF.ENZ A4
- Fichiers des enzymes de restrictions du type PFEIFFER.ENZ A5
- Fichiers d'ADN de type quelconque A6
- Fichiers d'ADN de type EMBL A6

MODE D'EMPLOI ET EXEMPLES D'UTILISATIONS

- Analyses à l'écran : A8
- Modes d'emplois accessibles à l'écran A9
- Recherche de séquences en fonction de la position (écran) A10
- Recherche de N-tracks (écran) A10
- Recherche de séquences utilisant le code de l'EMBL (écran) A11
- Sites de restrictions (écran) A13

- Analyse et création de fichiers de résultats :
- Traduction présentée au dessus de la séquence (fichier) A13
- Visualisation d'un gel de séquence, N-tracks (fichier) A14
- Sites de restrictions (fichier) A14
- Recherche de séquences (fichier) A15
- Mode d'impression particulier et numérotation A15
- Quelques exemples de fichiers de sorties possibles A15

- Traduction automatique des fichiers de l'EMBL :
- Utilisation A16
- Fichiers de sorties PROTEINE_EMBL et PROTEINE.COD A17

- Traduction et fabrication de fichiers EMBL compatible A18

- Calculs relatifs au Tm et au %(G+C) des hybrides ADN:ADN A19

- Détermination du %(G+C) et analyses statistiques relatives au contenu en chacun des nucléotides en fonction de la position dans le codon. A20

- DIAGON : comparaison graphique de protéines A21

PROGRAMME HEID :

INTRODUCTION A22

DESCRIPTION DES FICHIERS D'ENTREE COMPATIBLES A23

MODE D'EMPLOI ET EXEMPLES D'UTILISATIONS

- Recherche de séquences particulières (à l'écran) A23
- Comparaison de séquences par l'intermédiaire de fichiers A24
- Recherche de domaines (% d'homologie déterminable) A25
- Fichiers de sorties A26

BIBLOGRAPHIE A27

INTRODUCTION

Description des résultats obtenables par ANALYSE

Pour faciliter le travail d'analyse des données récoltées au laboratoire, ou simplement comparer celles-ci avec les résultats catalogués dans une base de données telle que celle de l'European Molecular Biology Laboratory (EMBL), il est maintenant impensable de travailler sans l'informatique.

Les programmes suivants sont rédigés en BASIC dans la version 2.3 implantée par Digital sur des ordinateurs VAX 750 et 785. Ils ont été prévu relativement faciles à modifier en fonction des besoins spécifiques des utilisateurs (voir notes aux programmeurs ...)

Le programme "ANALYSE" contient en fait quatre programmes qui fonctionnent avec plusieurs types de fichiers d'entrées. Les fichiers d'ADN peuvent être selon le format de l'EMBL (Ces données sont obtenables à l'EMBL, Postfach 10.2209, 6900 Heidelberg, Germany) ou entrées dans des fichiers de formats quelconques, (mais sans commentaires). Les séquences d'ADN que l'on désire analyser doivent être préférentiellement entrées en majuscules. Les séquences notées en lettres minuscules sont considérées par le programme comme appartenant à un ADN témoin dont la séquence est homologue, mais hétérospécifique. Les T-tracks sont comparés à ce type de séquences, le brin complémentaire est construit, mais les autres options du programme n'analysent pas les parties en minuscules. Celà permet par exemple de remplacer les trous de la séquence que l'on est en train de déterminer par la séquence connue d'un autre organisme, de faciliter ainsi le repérage de T-tracks et d'obtenir des analyses où seuls les sites de restrictions des parties séquencées sont présents et l'ADN traduit.

Les options du programme recoupant les programmes de Staden sont une adaptation des procédures décrites comme :

SEQLST, TRANSQ, SEARCH	Staden, (1977)
TRANDK, TRNTRP	Staden, (1978)
TRANMT, DIAGON	Staden, (1980)

Nous avons par contre des possibilités d'impressions nettement plus étendues, et la possibilité d'accéder aux fichiers de l'EMBL celà même si certains fichiers ne sont pas mémorisables dans une seule variable VAX (> 65'535 caractères). Nous pouvons parallèlement profiter de possibilités d'analyse plus étendue, comme la recherche de "N-track" à l'écran, la transcription automatique des séquences codantes répertoriées dans les fichiers d'ADN de l'EMBL et la fabrication de fichiers de protéines du type de l'EMBL (NBRF-PIR Protein sequence database), utilisables dans d'autres programmes. Par exemple, le programme HEID décrit plus loin (permet la comparaison de protéines) et le programme PRPLOT (donne une analyse de la protéine du point de vue de l'hydrophobicité, les régions en alpha-hélice et en béta-sheet), sont 100% compatibles avec les fichiers de protéines ainsi créés.

Une option du progamme ANALYSE est de déterminer le %(G+C) d'une séquence donnée et permettre une analyse statistique relative au contenu en chacun des nucléotides, et en fonction de la position dans le codon. Un complément au programme facilite le calcul des températures de fusion et de renaturation des hybrides ADN:ADN en fonction de la force ionique du tampon ou selon la concentration en tampon SSC.

La dernière possibilité, largement dépendante du type de terminal et surtout de l'imprimante graphique utilisée, permet de comparer graphiquement deux protéines.

Finalement, nous conseillons PRPLOT, qui est un excellent complément au système de programme que nous avons mis au point. PRPLOT est une version généralisée de l'algorithme décrit par Hopp et Woods, (1981) et distribué par

National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, Washington D.C. 2007
Programmeur: David G. George

DESCRIPTION DES FICHIERS D'ENTREE COMPATIBLES

Pour que le programme puisse interpréter les données inscrites dans différents types de fichiers, il faut respecter certaines règles. Si la plus complète fantaisie est possible au niveau des fichiers d'ADN (pourvu qu'ils ne contiennent que la séquence), il n'en est pas de même pour les autres types de fichiers, qui sont construits selon des règles strictes. Le programme est adapté aux fichiers suivants.

1.1 Les fichiers d'enzymes de restrictions

Les fichiers d'enzymes de restriction peuvent être de plusieurs types. Les fichiers prévus pour les programmes de Staden (1977) sont directement utilisables dans le programme et sont construits de la manière suivante :

TYPE DE STADEN
ALUI/AGCT//
BALI/TGGCCA//
BAMHI/GGATCC//
BANI/GGTACC/GGTGCC/GGCGCC/GGCACC//
CLAI/ATCGAT//
DDEI/CT-AG//
DRAI/TTTAAA//
HINFI/GAATC/GAGTC/GATTC/GACTC//
XHOII/GGATCC/AGATCT/GGATCT/AGATCC//

La seule simplification réellement prévue par Staden est le remplacement de "G ou A ou C ou T" par "-". Le programme nous permet donc d'utiliser cette simplification, même si, dans notre cas il est plus simple d'utiliser "N" qui est la notation soutenue par l'EMBL.

Les fichiers de références pour les enzymes de restrictions peuvent également être ceux distribués par l'EMBL. Deux types de de fichiers sont utilisables, qui, correspondent aux formats des fichiers PFEIFF.ENZ et PFEIFFER.ENZ. Ces fichiers simplifient au maximum la notation des sites de restrictions en utilisant le code de l'EMBL :

R : PURINE (A,G)	Y : PYRIMIDINE (C,T)	S : C ou G
M : A ou C	W : A ou T	K : G ou T
D : A ou G ou T	H : A ou C ou T	V : A ou C ou G
B : C ou G ou T	N : A ou C ou G ou T	

Type PFEIFF.ENZ (ancien standard) :

```
4110          AA'TT   - star activity of EcoRI
6111         AAA'TTT  - currently no enzyme with this specificity known
6112         GAA'TTC  - EcoRI and other enzymes
6113         CAA'TTG  - currently no enzyme with this specificity known
6114         TAA'TTA  - currently no enzyme with this specificity known
4120          GA'TC   - Sau3A and other enzymes
6121         AGA'TCT  - BglII
...
6443         CTT'AAG  - AflII
6444         TTT'AAA  - DraI and other enzymes
5110          AA'NTT  - currently no enzyme with this specificity known
5120          GA'NTC  - HinfI and other enzymes
5130          CA'NTG  - currently no enzyme with this specificity known
5140          TA'NTA  - currently no enzyme with this specificity known
5210          AG'NCT  - currently no enzyme with this specificity known
5220          GG'NCC  - Sau96I and other enzymes
5221          GG'WCC  - AflI and other enzymes
7222         RGG'NCCY - DraII
9332        CACN'NNGTG   - DraIII
9411       GAANN'NNTTC   - Asp700 and other enzymes
9531       GCCNN'NNNGGC  - BglI
9532      GGCCNN'NNNGGCC - SfiI
9611      CCANNN'NNNTGG  - BstXI and other enzymes
9631      ACCNNN'NNNGGT  - HgiEII
1001       'CCTC:GAGG'   - MnlI
1002      'GAAGA:TCTTC'  - MboII and other enzymes
1003      'GACGC:GCGTC'  - HgaI
1004      'GATCC:GGATC'  - BinI, BthII
1005      'GATGC:GCATC'  - SfaNI
1006      'GGATG:CATCC'  - FokI and other enzymes
1007      'GGTGA:TCACC'  - HphI
1008     'CAARCA:TGYTTG' - Tth111II
1009     'CTCCAG:CTGGAG' - GsbI, GsuI
1010     'YGGCCG:CGGCCR' - GdiII
1011     'SACCSA:TSGGTS' - TaqII
```

Ce type de fichier est très pratique car facilement complétable et facilement lisible sans l'aide d'un programme. Il n'est momentanément plus soutenu par l'EMBL, mais reste intéressant. Les repères fixes du programme sont un chiffre quelconque en début de ligne, la fin de la séquence qui est délimitée par "-", ainsi que l'emploi du séparateur ":" qui est utilisé lorsque les séquences ne suivent pas de palindrome. Enfin, le nom de l'enzyme est cherché à droite du signe "-" par le programme.

TYPE PFEIFFER.ENZ et nouveau standard :

```
AacI        GGA'TCC - cut position not determined
AaeI        GGA'TCC - cut position not determined
AatI        AGG'CCT - cut position not determined
AatII       GACGT'C - 3'overhang
AbrI        C'TCGAG - 5'overhang
AccI        GT'MKAC - 5'overhang
...
TthHB8I     T'CGA   - 5'overhang
Tth111I     GACN'NNGTC - 5'overhang, conflicting data on recognition sequence,
Tth111II    CAARCANNNNNNNNNN':'NNNNNNNNNTGYTTG - 3'overhang
TtnI        GG'CC   - cut position not determined
TtrI        GACN'NNGTC - cut position not determined
VhaI        GG'CC   - cut position not determined
XamI        GTC'GAC - cut position not determined
XbaI        T'CTAGA - 5'overhang
XcyI        C'CCGGG - 5'overhang
XhoI        C'TCGAG - 5'overhang
XhoII       R'GATCY - 5'overhang
XmaI        C'CCGGG - 5'overhang
XmaII       CTG'CAG - cut position not determined
XmaIII      C'GGCCG - 5'overhang
XmnI        GAANN'NNTTC - blunt ends
XniI        CGA'TCG - cut position not determined
XorI        CTG'CAG - cut position not determined
XorII       CGAT'CG - 3'overhang
XpaI        C'TCGAG - 5'overhang
```

Le type de ce fichier est le nouveau standard de l'EMBL. Chaque année, les fichiers des enzymes de restrictions sont remis à jour par l'EMBL et sont obtenables à Heidelberg. Les repères fixes du programme sont le nom de l'enzyme au début de chaque ligne, le site reconnu dont le début est aligné en position 14, la fin de la séquence délimitée par "-", et du séparateur de séquences ":"

Nous avons délibéremment évité d'utiliser le code établi par Staden (1979) lorsque la séquence est incertaine et qu'un certain nombre d'insertion-délétion est probable. Il est plus avantageux d'utiliser le code présenté ici ou d'utiliser l'option qui fait appel à la recherche de "N-track" (réaction de séquence du T ou de n'importe nucléotide), lorsqu'une séquences est recherchée.

1.2 Les fichiers d'ADN

Les fichiers d'ADN peuvent être quelconques ou du type de l'EMBL.

Les fichiers quelconques

Nos fichiers personnels (quelconques) sont des fichiers édités dans n'importe quel éditeur implanté sur VAX. Les séquences sont éventuellement entrecoupées par des espaces blancs et la longueur des lignes n'est pas limitée. La seule exigence est qu'il n'y ait pas de commentaires insérés dans le fichier, sinon ces commentaires sont traités et analysés comme s'il faisaient partie de la séquence.

Nos fichiers ne sont pas limités dans le nombre de caractères par ligne. En comparaison, les programmes de Staden, ne supportent pas plus de 60 caractères/ligne dans les fichiers d'ADN et créent des erreurs qui ne sont pas signalées lorsque cette limite est dépassée.

Les fichiers de l'EMBL

Les fichiers d'ADN de l'EMBL ont une structure extrêmement bien fixée, dont on peut voir un exemple sur la page suivante.

Le premier travail du programme dans les analyses est de repérer si le fichier d'ADN est du type de l'EMBL de par la présence dans la première ligne du fichier de l'identificateur "ID". Le début de la séquence est ensuite repéré par le "SQ" de la ligne précédant directement la séquence. Le programme détermine ensuite les positions de début et de fin de la région à analyser, pour ne garder en mémoire que les lignes indispensables. Le repérage est rendu plus rapide par le fait que chaque ligne contient 60 nucléotides. La fin du fichier est remarquée grâce à l'identificateur "//".

Lors de la traduction automatique du fichier. un pointeur cherche les lignes "FT CDS" (CoDing Sequences), repère lorsque le brin codant est le brin complémentaire "(C)" et traduit la région indiquée tout en notant dans le fichier final, de quelle protéine il s'agit ainsi que la position de la première et de la dernière position de l'ADN traduite. Une option spéciale permet d'ouvrir directement un fichier .COD indispensable au fonctionnement du programme PRPLOT.

Finalement, le programme ANALYSE peut être utilisé accessoirement pour rendre les fichiers de l'EMBL compatibles avec les programmes de Staden. Dans ce cas, il suffit de demander la création du fichier contenant la séquence que l'on veut analyser et de demander sous la question IMPRESSION PARTICULIERE que la sortie se fasse SANS NUMEROTATION et sur 60 CARACTERES/LIGNE.

TYPE ADN_HEIDELBERG

```
ID   CHGM01    standard; DNA; 1438 BP.
XX
AC   X00152;
XX
DT   05-NOV-1985  (minor modification)
DT   08-JUL-1985  (minor modification)
DT   08-AUG-1984  (first entry)
XX
DE   Soybean chloroplast psb A gene
DE   coding for photosystem II thylakoid membrane protein
XX
KW   membrane protein.
XX
OS   Chloroplast Glycine max (soybean, soja, Sojabohne)
OC   Chloroplastida; Spermatophyta; Magnoliopsida; Rosidae; Fabales;
OC   Fabaceae.
XX
RN   [1]   (bases 1-1438)
RA   Spielmann A., Stutz E.;
RT   "Nucleotide sequence of soybean chloroplast DNA regions which
RT   contain the psb A and trn H genes and cover the ends of the large
RT   single copy  region and one end of the inverted repeats";
RL   Nucl. Acids Res. 11:7157-7167(1983).
XX
RN   [2]   (bases 1-1438; enum. 1 to 1438)
RA   Spielmann A., Stutz E.;
RT   ;
RL   Submitted (14-MAY-1984) to the EMBL data library.
XX
CC   Data kindly reviewed (14-MAY-1984) by A. Spielmann and E. Stutz
XX
FH   Key         From     To       Description
FH
FT   CDS          265   1323       psb A gene
FT   SITE         148    152       -35 region
FT   PRM          172    177       -10 region
FT   SITE        1094   1115       stem-loop structure
FT   REVISION    1348   1351       TAGATGT [1] revised to TAGT by [2]
XX
SQ   Sequence  1438 BP;  383 A;  246 C;  286 G;  523 T;
     CCACTAGAAT TAATATATCG AAATTCTATA TATAGAATTA TAGCGATCCA TAGAAAATAA
     ACTCATTTTT AATGAATGAA ATGAGGTAAA AAAGTAAATT CGGATAAATC TAAATAAGAG
     CATACTTACT ATGGATATTG GTATTGGTTG ACACTGGTAT ATAAGTCATG TTATACTGTT
     GAATAACAAG TCCTCAAATT TTCTAATTCT AGATAATTTT GGTGCTTGGG AGTCCCTGAT
     GATTAAATGA ACCAAGATTT TACCATGACT GCAATTTTAG AGAGACGCGA GAGCGAAAGC
     ...
     ATGGCTTTCA ACTTAAATGG TTTCAATTTC AACCAATCCG TAGTTGATAG TCAAGGTCGT
     GTAATTAATA CCTGGGCTGA TATTATTAAC CGAGCTAACC TTGGTATGGA AGTAATGCAT
     GAACGTAATG CTCATAATTT CCCTCTAGAT CTAGCTGCGA TCGACGCTCC ATCTATTAAT
     GGATAAAATT TGGATCTTAA GGTAGATTAG TTTTTGAAAG TAAAAGGCAA TATCAACTTT
     TTTGATATTG CCCCTTTACT TTTATTTATT TGATTAGTAA TCTTTTTATT TATAAATA
//
```

MODE D'EMPLOI ET EXEMPLES D'UTILISATIONS

Le programme une fois lancé par la commande $RUN ANALYSE , est utilisable en deux langues soit : anglais ou français.

Il est possible d'effectuer des analyses à l'écran et sans sortir du programme, de continuer l'analyse avec le sauvetage de résultats dans un fichier. D'une manière générale, la façon élégante de sortir du programme est de presser la touche "return" le nombre de fois nécessaire. Les options rencontrées sont alors par défaut considérées comme rejetées.

Détail non négligeable, plusieurs personnes peuvent utiliser le programme en même temps et avec les mêmes fichiers de données. Ceux-ci restent protégés contre l'écriture mais toujours ouvert à la lecture.

Nous donnons quelques exemples d'utilisation du programme.
Analyses à l'écran :

EXEMPLE D'UTILISATION : Sous-programme ANALYSE

```
English version / version Francaise ? f

POUR OBTENIR :                                  REPONDRE

Calculs, relations entre Tm, %GC et %Missmatch    1
Calcul du %GC a partir d'une sequence             2
ANALYSE : T-tracks, traduction, restriction ...   3
DIAGON : comparaison graphique de fichiers        4

                                                CHOIX = ? 3

Nom du Fichier     : ? hind.906
 LE FICHIER EST DE TYPE :QUELCONQUE
 Fichier lu
 POSITION DE LA PREMIERE SEQUENCE DU FICHIER 1 : ? 1
 POSITION DE LA DERNIERE SEQUENCE DU FICHIER 1 : ?
 AAGCTTCATCC... ...GGTTCAAGCTT


RESULTATS A PRESENTER A L'Ecran   OU  DANS UN Fichier
- Recherche de N-tracks                 - Brin complementaire
- Recherche de sequences                - Traduction
- Sites de restrictions                 - Visualisation du gel
- Aide memoire                          - Restriction
                                        - Recherche de sequences
 Travail a l'Ecran ou dans un Fichier ? E
```

EXEMPLE D'UTILISATION : Modes d'emplois accessibles à l'écran

```
MODE D'EMPLOI DU PROGRAMME          :     M
SEQUENCE SELON LA POSITION          :     0
RECHERCHE DE N TRACKS A L'ECRAN     :     1
RECHERCHE DE SEQUENCES A L'ECRAN    :     2
RECHERCHE DE SITES DE RESTRICTIONS  :     3
? M
AIDE : RECHERCHE : SEQUENCES, N-TRACKS, SITES DE RESTRICTION
       ACIDES AMINES : CODE A UNE LETTRE, EQUIVALENCE
? U

ACIDES AMINES : CODE A UNE LETTRE

  A=Ala     B=Asx     C=Cys     D=Asp     E=Glu     F=Phe
  G=Gly     H=His     I=Ile     K=Lys     L=Leu     M=Met
  N=Asn     P=Pro     Q=Gln     R=Arg     S=Ser     T=Thr
  V=Val     W=Trp     Y=Tyr     Z=Glx     *=STOP

? EQUIVALENCE
  IL EST POSSIBLE D'UTILISER UNE TABLE D'EQUIVALENCE ENTRE LES A.A.
  SUIVANTS :
                Lys/Arg,
                Ser/Thr,
                Asp/Glu,
                Val/Ile/Leu

? N-tracks

MODE D'EMPLOI :
---------------

POUR RECHERCHER UN N-TRACK, IL SUFFIT DE TAPER LA SEQUENCE
EN REMPACANT LES ESPACES VIDES PAR UN <.>
 ex : .TTTT.TT.T..TT.     recherche les T-TRACKS correspondants
                          ainsi que les A-TRACKS .AA...A.AA.AAAA.

? Restr.
LES FICHIERS D'ENZYMES DE RESTRICTIONS SONT :

    R1
    R2
    RENZYM                      :  ENZYMES DU LABORATOIRE
    BOTA:[STUTZ]PFEIFFER.ENZ    :  TOUS LES ENYMES, TOUS LES ISOMERES
    BOTA:[STUTZ]PFEIFF.ENZ      :  UN ENZYME PAR SITE DIFFERENT

Les 2 derniers fichiers sont mis a jour a Heidelberg, mais sont
inutilisables avec les programmes de STADEN !!!

? S
LE PROGRAMME EST CAPABLE DE RECHERCHER DES SEQUENCES EN UTILISANT
LE CODE SUIVANT ( PRECONISE PAR HEIDELBERG ):

    R : PURINE (A,G)   Y : PYRIMIDINE (C,T)  S : C ou G
    M : A ou C         W : A ou T            K : G ou T
    D : A ou G ou T    H : A ou C ou T       V : A ou C ou G
    B : C ou G ou T    N : -                 - : A ou C ou G ou T
```

Le retour au programme est ensuite fait en pressant "return"

EXEMPLE D'UTILISATION : Recherche de positions à l'écran

```
 ?
  POSITION (0), N-TRACK (1), SEQUENCE (2), RESTRICTION (3) ? 0
  Voulez-vous travailler avec l'Ancienne numerotation ou la Nouvelle ?
  Mode d'emploi du programme.
  M, A ou N ? M

  Il est possible d'utiliser, soit la numerotation 'Normale' :
    c'est a dire, ou le 1 de la numerotation correspond
    AU DEBUT DU FRAGMENT que l'on a DEMANDE EN DEBUT DE PROGRAMME.

  soit la numerotation 'Ancienne' :
    c'est a dire, ou la numerotation correspond A LA NUMEROTATION
    DU FICHIER INITIAL et non pas celle du fragment analyse.
  M, A ou N ? N
  POSITION ? 1
                          AAGCTTCATCCTGGTCATGGACAGATCACCCAGGTT 1

  POSITION ? 555555
TCACTGTAGGAATTGGGTTCAAGCTT 8777

  POSITION ? 450
CACGACACTTGTATAGCTCTCCCACAACCCGTTTTCACGGTTTAGGCTGCTCCCATTCGTC 450

  POSITION ?
```

Donc 8777 est la dernière position du fichier !

EXEMPLE D'UTILISATION : Recherche de N-tracks

```
  MODE D'EMPLOI DU PROGRAMME          :     M
  SEQUENCE SELON LA POSITION          :     0
  RECHERCHE DE N TRACKS A L'ECRAN     :     1
  RECHERCHE DE SEQUENCES A L'ECRAN    :     2
  RECHERCHE DE SITES DE RESTRICTIONS  :     3
  ? 1

  MODE D'EMPLOI :
  ---------------

  POUR RECHERCHER UN N-TRACK, IL SUFFIT DE TAPER LA SEQUENCE
  EN REMPACANT LES ESPACES VIDES PAR UN <.>
   ex : .TTTT.TT.T..TT.     recherche les T-TRACKS correspondants
                            ainsi que les A-TRACKS .AA...A.AA.AAAA.

  SEQUENCE ? .TTT..TTT..

      ANALYSE DE LA SEQUENCE DONNEE

      | | |  |  |     |   |||  |||          ||             |  2634
GCCCCATCTATCCTCCTGAGGATAAGTTTGGTTTCAAACCCCGGTTCAAACAGGAGAAGTA
                          |||  |||
```

EXEMPLE D'UTILISATION : N-tracks (suite)

```
 || |       ||||| |       |||  |||   | | ||   |     |||| |||| 8108
ATTCTACAAGAATTTTTATGAAAGCCTTTCATTTGCCTCTCTTCGATGGAAGTTTTATTTT
                           |||  |||

     ANALYSE DE LA SEQUENCE COMPLEMENTAIRE


                  ||  |     |||  |||  |   || ||  |  |          3068
GCTTCTTTTTGGGTGTGAAGCAGTGTCAAACTAAACTACCCAACAAGCATTAGCTCTCCCT
                           |||  |||

 |  |     |        |       |||  |||       ||  ||| |  |        3495
TATCACCGGCAGTCTGCTCAGGGTTCCAAACTAAATGTTGGCAACTAAACACGAGGGTTGC
                           |||  |||
  SEQUENCE ? .C.C.C.C.

     ANALYSE DE LA SEQUENCE DONNEE

   | |  |  | |      |     | | | |  || |    |      |||  |   | 2788
AGGCACAACGACGCAATTATCAGGGGCGCGCTCTACCACTGAGCTAATAGCCCGTCGTGCG
                           | | | |

        |    || |         | | | |   |||  |          ||    |  5710
GAAGGGTGCATAACCGCATGGATAAGCTCACACTAACCCGTCAATTTGGGATCCAATTCGG
                           | | | |

             |     || ||   | | | |    | |  |      |  |        7266
GTTATAAAAAAACGGATTCCTCCTTTCACGCTCATGTCACGGCGAGGTACTGCAGAAGAAA
                           | | | |

     ANALYSE DE LA SEQUENCE COMPLEMENTAIRE

  SEQUENCE ?
```

Si les gels présentent des régions contenant des arte-facts et si l'on hésite entre 5 et 6 "T" par exemple, il est avantageux de laisser la séquence ouverte sur l'un de ses côtés : En demandant " SEQUENCE ? TTTTT..... ", le nucléotide placé avant le premier "T" peut être "T" ou "."

EXEMPLE D'UTILISATION : Recherche de séquences à l'écran

```
MODE D'EMPLOI DU PROGRAMME          :     M
SEQUENCE SELON LA POSITION          :     0
RECHERCHE DE N TRACKS A L'ECRAN     :     1
RECHERCHE DE SEQUENCES A L'ECRAN    :     2
RECHERCHE DE SITES DE RESTRICTIONS  :     3
? 2
RECHERCHE SUR LES DEUX BRINS : Oui ou Non ? O
SEQUENCE ? GYGYGG
```

Lors de la recherche de séquences à l'écran, il est donc possible d'utiliser le code de l'EMBL (présenté en page 3) et qui permet des recherches de consensus. Si l'une des séquences cherchées n'est pas présente dans le fichier, ANALYSE DE LA SEQUENCE DONNEE ou ANALYSE DE LA SEQUENCE COMPLEMENTAIRE est affiché, sans être suivi d'aucune séquence. Si la séquence à chercher contient des "N" ou des "-", le temps de calcul dépendra essentiellement du nombre de fois que l'on trouve la séquence précédant le premier "N" ou le premier "-" dans le fichier considéré. La recherche de GYGYGNNNT sera très rapide mais celle de ANNNCRCRC sera lente.

EXEMPLE D'UTILISATION : Recherche de séquences à l'écran (suite)

```
  SEQUENCE ? GYGYGG

      ANALYSE DE LA SEQUENCE DONNEE

      ANALYSE DE LA SEQUENCE COMPLEMENTAIRE


      ANALYSE DE LA SEQUENCE DONNEE

      ANALYSE DE LA SEQUENCE COMPLEMENTAIRE

                              CCACGC
TCAATTCAAGAGTTCTTATGTGTTTCCACGCCCCTTCGAGACCCCGAAAAAATGGACAAAT 5114
AGTTAAGTTCTCAAGAATACACAAAGGTGCGGGGAAGCTCTGGGGCTTTTTTACCTGTTTA

      ANALYSE DE LA SEQUENCE DONNEE

TCTACCACTGAGCTAATAGCCCGTCGTGCGGCTCCCGCGGGGCCCGCTATGCCAAAAGCAA 2819
                              GTGCGG

AATTAAACCACATGCTCCACCGCTTGTGCGGCCCCCGTCAATTCCTTTGAGTTTCATTCTT 3701
                              GTGCGG

      ANALYSE DE LA SEQUENCE COMPLEMENTAIRE

                              CCGCAC
TCTCCCTGAAAAGGAGGTGATCCAGCCGCACCTTCCAGTACGGCTACCTTGTTACGACTTC 3122
AGAGGGACTTTTCCTCCACTAGGTCGGCGTGGAAGGTCATGCCGATGGAACAATGCTGAAG

      ANALYSE DE LA SEQUENCE DONNEE

TTTCACCCCCATCCATGTCGCCCCCGTGTGGCGACATGAGACAAAAGTAGCTCCTATCAAC 2907
                              GTGTGG

GAGTCTGGGCCTCGTCTCAGTCCCAGTGTGGCTGATCATCTCTCGGACCAGCTACTGATCA 4293
                              GTGTGG

GTGAATCGCTAGAAACATGTAAAGTGTGTGGCTAACCCAATAACGAAAGTTTCGTAAGGGA 6736
                              GTGTGG

      ANALYSE DE LA SEQUENCE COMPLEMENTAIRE

                              CCACAC
CATCAAGTTCCCGTTGAAATAGGATCCACACAAGGAAAAGCACTTGCCATTCGTTGGTTAT 7551
GTAGTTCAAGGGCAACTTTATCCTAGGTGTGTTCCTTTTCGTGAACGGTAAGCAACCAATA
  SEQUENCE ?
```

EXEMPLE D'UTILISATION : Sites de restrictions (écran)

```
  MODE D'EMPLOI DU PROGRAMME          :     M
  SEQUENCE SELON LA POSITION          :     0
  RECHERCHE DE N TRACKS A L'ECRAN     :     1
  RECHERCHE DE SEQUENCES A L'ECRAN    :     2
  RECHERCHE DE SITES DE RESTRICTIONS  :     3
  ? 3
 RECHERCHE DES SITES DE RESTRICTION :
 FICHIER DES ENZYMES DE RESTRICTIONS
 Nom du Fichier_2  : ? RENZYM
  Fichier lu
  ENZYME= HINDIII
AAGCTT

      ANALYSE DE LA SEQUENCE DONNEE

                          AAGCTTCATCCTGGTCATGGACAGATCACCCAGGTT 1
                          AAGCTT

ATTCATCACTGTAGGAATTGGGTTCAAGCTT 8772
                          AAGCTT

  ENZYME= ?
```

Pour créer un fichier contenant des résultats d'analyses, il est possible après des recherches à l'écran de revenir au menu principal par des "return". Il est bien entendu possible d'accéder directement à cette option en entrant dans le programme...

EXEMPLE D'UTILISATION : Travail dans un fichier (TRADUCTION)

```
  Travail a l'Ecran ou dans un Fichier ? F
 Fichier de Sortie : ? SEQ.RES
 TRADUCTION: PRESENTEE AU DESSUS DE LA SEQUENCE, SEPAREMENT, NON ? P
  Voulez-vous travailler avec l'Ancienne numerotation ou la Nouvelle ?
  Mode d'emploi du programme.
  M, A ou N ? N

  CODE STANDARD A 1 LETTRE              1
  CODE STANDARD A 3 LETTRES             2
  CODE MITOCHONDRIAL A 1 LETTRE         3
  CODE MITOCHONDRIAL A 3 LETTRES        4
 ? 1
  PREMIERE POSITION = ? 1
  DERNIERE POSITION = ?
OK
  PREMIERE POSITION = ? 2
  DERNIERE POSITION = ? 400
OK
  PREMIERE POSITION = ?
  DERNIERE POSITION = ?
```

Lorsque la DERNIERE POSITION n'est pas indiquée, la séquence est traduite jusqu'à la fin du fichier. Lorsque la PREMIÈRE POSITION n'est pas indiquée, le programme passe à l'option suivante, ici la traduction du brin complémentaire.

EXEMPLE D'UTILISATION : Traduction (suite)

```
 TRADUCTION DU BRIN COMPLEMENTAIRE : Oui ou Non ? O
 UTILISATION DE LA NUMEROTATION DU BRIN 1 OU DU BRIN COMPLEMENTAIRE 2 ? 1
  PREMIERE POSITION = ? 1
  DERNIERE POSITION = ?
OK
  PREMIERE POSITION = ?
  DERNIERE POSITION = ?
```

La visualisation d'un gel de séquence est parfois très pratique. Le repérage de clones par leurs T-tracks est facilité si les réactions du "T" et du "A" sont visualisables face à une séquence donnée.

EXEMPLE D'UTILISATION : T-track (fichier)

```
VISUALISATION D'ON GEL DE SEQUENCE : Oui ou Non ? O
 G-tracks : Oui ou Non ? N
 A-tracks : Oui ou Non ? O
 C-tracks : Oui ou Non ? N
 T-tracks : Oui ou Non ? O
```

EXEMPLE D'UTILISATION : Restriction (fichier)

```
RECHERCHE DES SITES DE RESTRICTION :
 Mode d'emploi, Tous, Nom, Sequence, <RET> ? T
FICHIER DES ENZYMES DE RESTRICTIONS
Nom du Fichier_2  : ? BOTA:[STUTZ]PFEIFF.ENZ
 Fichier lu
  EcoRI*    AATT
  EcoRI    GAATTC
  Sau3A    GATC
  BglII    AGATCT
  ...
  ScrFI    CC-GG
  AorI    CCAGG
  SfiI    GGCC-----GGCC
  BstXI    CCA------TGG
  HgiEII    ACC------GGT
  MnlI    CCTC
  MnlI    GAGG
```

EXEMPLE D'UTILISATION : Recherche de séquences (Fichier)

```
RECHERCHE DES SITES DE RESTRICTION :
 Mode d'emploi, Tous, Nom, Sequence, <RET> ? S
 SEQUENCE= ? GYGYG
 RECHERCHE SUR LES DEUX BRINS Oui ou Non : ? N
 TEXTE PARTICULIER A PLACER SOUS LA SEQUENCE : Oui ou Non : ? O
TEXTE : ? GYGYG-CONSENSUS
  GYGYG-CONSENSUS   GCGCG
  GYGYG-CONSENSUS   GCGTG
  GYGYG-CONSENSUS   GTGCG
  GYGYG-CONSENSUS   GTGTG
RECHERCHE DES SITES DE RESTRICTION :
 Mode d'emploi, Tous, Nom, Sequence, <RET> ?
```

Au niveau de l'impression, nous pouvons choisir le nombre de nucléotides imprimés par ligne, ainsi que le type de numérotation désiré. La NOUVELLE numérotation commence par 1, tandis que l'ANCIENNE numérotation commence à l'endroit choisi au tout début du programme (et suit la numérotation de l'ancien fichier).

EXEMPLE D'UTILISATION : Modes d'impression

```
MODE D'IMPRESSION PARTICULIER : Oui ou Non ? O
 Numerotation dans le fichier de sortie : Nouvelle, Ancienne ou Inexistante ? N
 Nombre de nucleotides/ligne (50-130) : ? 60
 NOM DU FICHIER DES RESULTATS : SEQ.RES;
```

1.1.5 Présentation des résultats

La présentation des résultats dépend des choix effectués lors de l'utilisation du programme. Voici quelques sorties possibles avec un choix de 60 caractères/ligne.

QUELQUES EXEMPLES DE SORTIES

```
GAATGGCAGAGGCAAATAGAGCTTTTGCACATTTTCGTTAATCCATGAACAGGGATTCTA
 ||    | |   ||| | |         | |         ||   |  || |   |    |
   |           |     ||||     ||||  ||  |   |         || |
GAATGGCAGAGGCAAATAGAGCTTTTGCACATTTTCGTTAATCCATGAACAGGGATTCTA
      7710       7720       7730       7740       7750       7760
```

```
ProLeuThrArgProLeuProMetSerArgArgLeuIleLeuGlnGlnAlaArgGlyGln
CCCTTAACCAAGCCACTGCCTATGAGTCGCCGGCTCATTCTTCAACAGGCACGCGGTCAG
        10         20         30         40         50         60
```

QUELQUES EXEMPLES DE SORTIES (suite)

```
Fichier [-]hind.906        21-Aug-87     11:36 AM
Fichier des enzymes de restrictions bota:[stutz]pfeiff.enz

 P  L  T  K  P  L  P  M  S  R  R  L  I  L  Q  Q  A  R  G
  P  *  P  S  H  C  L  *  V  A  G  S  F  F  N  R  H  A
   L  N  Q  A  T  A  Y  E  S  P  A  H  S  S  T  G  T  R
CCCTTAACCAAGCCACTGCCTATGAGTCGCCGGCTCATTCTTCAACAGGCACGCGGtcag
GGGAATTGGTTCGGTGACGGATACTCAGCGGCCGAGTAAGAAGTTGTCCGTGCGCCagtc
 G  *  G  L  W  Q  R  H  T  A  P  E  N  K  L  L  C  A
  R  L  W  A  V  A  *  S  D  G  A  *  E  E  V  P  V  R
   K  V  L  G  S  G  I  L  R  R  S  M  R  *  C  A  R  P

CCCTTAACCAAGCCACTGCCTATGAGTCGCCGGCTCATTCTTCAACAGGCACGCGGtcag
       10        20        30        40        50        60
                           Cfr10I    MboII        ThaI
                    HinFI HpaII
                           NaeI
```

Nous ne sommes pas limités aux exemples donnés, la présentation est modifiable à volonté.

1.1.6 Traduction automatique des fichiers d'ADN de l'EMBL en fichier de protéines

Pour la traduction automatique des fichiers de l'EMBL (NBRF-PIR Protein sequence database), nous prendrons en exemple la traduction des séquences codantes de l'ADN chloroplastique complet de tabac.

EXEMPLE D'UTILISATION : Traduction automatique (Fichiers EMBL)

```
POUR OBTENIR :                                  REPONDRE

Calculs, relations entre Tm, %GC et %Missmatch    1
Calcul du %GC a partir d'une sequence             2
ANALYSE : T-tracks, traduction, restriction ...   3
DIAGON : comparaison graphique de fichiers        4

                                                CHOIX = ? 3

Nom du Fichier     : ? [STUTZ.ADN]CHNTXX.DAT
 LE FICHIER EST DE TYPE :ADN-HEIDELBERG
 Fichier lu
 POSITION DE LA PREMIERE SEQUENCE DU FICHIER 1 : ? 1
 POSITION DE LA DERNIERE SEQUENCE DU FICHIER 1 : ?
  ATTENTION ! NOUS N'AVONS QU'UNE PARTIE DU FICHIER !
 ANALYSE DES POSITIONS  1  A  65001

RESULTATS A PRESENTER A L'Ecran   OU  DANS UN Fichier
- Recherche de N-tracks                 - Brin complementaire
- Recherche de sequences                - Traduction
- Sites de restrictions                 - Visualisation du gel
- Aide memoire                          - Restriction
                                        - Recherche de sequences
 Travail a l'Ecran ou dans un Fichier ? F
```

EXEMPLE D'UTILISATION : Traduction automatique (suite)

```
  Fichier de Sortie : ? CHNTXX.SEQ
  TRADUCTION: PRESENTEE AU DESSUS DE LA SEQUENCE, SEPAREMENT, NON
  TRADUCTION: AUTOMATIQUE DU FICHIER D'HEIDELBERG
  Voulez-vous créer un fichier .COD pour pouvoir utiliser le programme
  le programme PRPLOT, Oui ou Non
  ? O
   CODE STANDARD A 1 LETTRE              1
   CODE STANDARD A 3 LETTRES             2
   CODE MITOCHONDRIAL A 1 LETTRE         3
   CODE MITOCHONDRIAL A 3 LETTRES        4
  ? 1
   A :          7724   8020       ORF98
OK
   A :         34462  35523       PSII D2 protein (psbD)
OK
   A :         35471  36892       PSII 44kd protein (psbC)
OK
   A :         37586  37774       ORF62 (membrane protein ?)
...
   B :         56769  55273 (C)   ATPase beta subunit (atpB)
OK
  LE FICHIER EST DE TYPE :ADN HEIDELBERG
  ATTENTION ! NOUS N'AVONS QU'UNE PARTIE DU FICHIER !
  ANALYSE DES POSITIONS  64327  A  129327
   B :         66860  66741 (C)   PSII component (psbF)
OK
...
   B :        153646 151901 (C)   ORF581
OK
  Fichier de sortie : CHNTXX.SEQ
```

Les fichiers créés par la traduction automatique des séquences de l'EMBL sont du type présenté dans la figure suivante. Précédant chaque séquence, il existe un identificateur ">" et un code qui dans notre cas est un chiffre. La ligne suivante indique le nom de la protéine, les limites de l'ORF dans le fichier d'ADN et le brin codant (B : est le brin complémentaire). Pour des raisons pratiques, la longueur des lignes (510 caractères) est racourcie dans notre exemple. Remarquons qu'il est possible de sortir une (ou des) protéines dans un fichier éditable (132 caractères) en utilisant le programme HEID.

FICHIER DE SORTIE DU TYPE DE L'EMBL (protéine)

```
>P1;7724
   A :          7724   8020       ORF98
VGSKKKKYFLFTLLTFPFLISITQSKCNYLSKNKKSVMLNTFSLIGICLNSTLFSSSFFFGKLPEAYAFLNPIVDIMPVI
PLFFFLLAFVWQAAVSFR*
>P1;34462
   A :         34462  35523       PSII D2 protein (psbD)
MTIALGKFTKDENDLFDIMDDWLRRDRFVFVGWSGLLLFPCAYFAVGGWFTGTTFVTSWYTHGLASSYLEGCNFLTAAVS
TPANSLAHSLLLLWGPEAQGDFTRWCQLGGLWTFVALHGAFGLIGFMLRQFELARSVQLRPYNAIAFSGPIAVFVSVFLI
YPLGQSGWFFAPSFGVAAIFRFILFFQGFHNWTLNPFHMMGVAGVLGAALLCAIHGATVENTLFEDGDGANTFRAFNPTQ
AEETYSMVTANRFWS...
```

Le fichier .COD, fabriquable en même temps que les séquences sont traduites, est utiles dans le cas où le programme PRPLOT va être utilisé ultérieurement. Le fichier .COD contient les codes d'appel de chacune des protéines traduites.

FICHIER .COD : compatible avec PRPLOT

```
7724
34462
35471
...
```

1.1.7 Traduction de séquences particulières en créant un fichier de protéines compatible EMBL

Il est utile de pouvoir disposer de fichiers de protéines compatibles avec la "NBRF-PIR Protein sequence database". Lorsque la traduction est demandée SEPAREMENT, un tel fichier est créé avec la possibilité d'ouvrir un fichier .COD permettant d'utiliser le programme PRPLOT.

Lorsque les séquences à analyser ne sont pas forcément codantes et contiennent des codons stop, nous avons la possibilité de demander "//" comme identificateur de fin de la zone transcrite. Cette option permet de trouver d'éventuels pseudogènes ou des séquences dont le rôle a été perdu, au moyen du programme HEID.

FICHIER .COD : compatible avec PRPLOT

```
 TRADUCTION: PRESENTEE AU DESSUS DE LA SEQUENCE, SEPAREMENT, NON ? S
 Voulez-vous créer un fichier .COD pour pouvoir utiliser le programme
 le programme PRPLOT, Oui ou Non
 ? O
  Voulez-vous travailler avec l'Ancienne numerotation ou la Nouvelle ?
  Mode d'emploi du programme.
  M, A ou N ? A
  Voulez-vous le caractère : // : à la fin de la séquence ?
  Celui-ci rends possible la recherche de pseudogènes. Oui ou Non ? N

  CODE STANDARD A 1 LETTRE              1
  CODE STANDARD A 3 LETTRES             2
  CODE MITOCHONDRIAL A 1 LETTRE         3
  CODE MITOCHONDRIAL A 3 LETTRES        4
 ? 1
  PREMIERE POSITION = ? 1
  DERNIERE POSITION = ? 231
OK
 ...
```

1.2 Le sous-programme CALCUL_TM :

Le calcul des températures de dénaturation et de renaturation des ADN en fonction des conditions ioniques est facilité par l'emploi du programme, qui tient compte de formules dispersées dans la littérature. Il est particulièrement pratique lors du choix des conditions d'hybridation ou de lavage d'hybrides ADN:ADN.

Nous utilisons les formules suivantes :

(1) $Tm = 0.41 \cdot \%(G+C) + 69.37 + 13.3 \cdot \log (X)$
si $0.5 \leq X \leq 2$

$Tm = 0.41 \cdot \%(G+C) + 69.37 + 15.7 \cdot \log (X)$
si $0.125 \leq X < 0.5$

où Tm représente la température de dénaturation de l'ADN
X exprime la concentration en SSC

(Gillis et al., 1970)

(2) $Tor = 1.24 \cdot Tm - 38.8$

où Tor est la température de renaturation optimale

(Gillis et al., 1970)

(3) 1% formamide réduit le Tm de 0.72°C

(Mc Conaughy et al., 1969)

(4) L'augmentation de 1% du nombre de missmatch entre paires de bases descend le Tm du duplex de 1°C

(Bonner et al., 1973)

(5) $(Tm)_{\mu 2} - (Tm)_{\mu 1} = 18.5 \cdot \log (\mu 2/\mu 1)$

ou $\mu 1$ et $\mu 2$ sont les forces ioniques respectives des deux solutions.

(Schleif et Wensink, 1981)

(6) $\mu = 1/2 \cdot \sum Ci \cdot Zi^2$

ou μ represente la force ionique,
Ci la concentration de chaque ion, Zi sa charge

(Weil, 1979)

Le programme propose une série d'options que nous présentons :

```
          PROGRAMME Tm : Menu

  1  : CALCUL DU %(G+C) A PARTIR DU Tm
  2  : CALCUL DU Tm A PARTIR DU %(G+C)
  3  : CALCUL DE LA TEMPERATURE DE RENATURATION OPTIMALE
  4  : CALCUL DU % DE MISSMATCH
  5  : CALCULS EN CONNAISSANT LA FORCE IONIQUE
  6  : CALCUL DU %(G+C) A PARTIR DU Tm COMPARE AVEC LE Tm DE E.coli
  7  : FORMULES UTILISEES
  8  : CALCUL DE LA FORCE IONIQUE CONNAISSANT GC ET Tm
  9  : SORTIE DU SOUS-PROGRAMME
```

Les choix 1 à 3 sont liés à l'emploi de tampon SSC. Les relations mathématiques ont été expérimentalement vérifiées pour des concentrations comprisent entre 0.125 xSSC et 2 xSSC. Un calcul d'erreur est intégré au programme, ce qui permet d'avoir une idée précise du caractère relatif des résultats lorsque la concentration devient faible.

Les calculs en connaissant la force ionique (5) permettent également de calculer la force ionique à partir de données expérimentales. Dans ce cas, il faut effectuer la dénaturation d'un ADN dont le %(G+C) est connu, qui permet (après observation du Tm au spectrophotomètre, par exemple) de déterminer la force ionique du tampon utilisé, puis le %(G+C) des ADN inconnus. Remarquons qu'il est possible, bien que déconseillé, de faire l'hybridation de l'ADN témoin dans du tampon Nx et les hybridations des échantillons dans une dilution différente du tampon initial (Mx).

Le choix n°7 permet d'obtenir les références mathématiques et la la bibliographie attenante. Cette option permet également de se remémorer rapidement les formules utilisées.

1.3 Sous-programme CALCUL_GC :

Basé sur l'observation de Staden et Mc Lachlan (1982), démontrant que la différence de structure entre les régions codantes et non codantes de l'ADN, ce sous-programme permet de chercher pour une séquence donnée, le pourcentage en chacun des nucléotides, et en fonction de la position dans le codon. Nous obtenons ainsi une indication supplémentaire tendant à déterminer si une ORF est probablement transcrite et traduite ou non.

La sortie des résultats se fait à l'écran sous forme de tableaux qui permettent une analyse plus avancée.

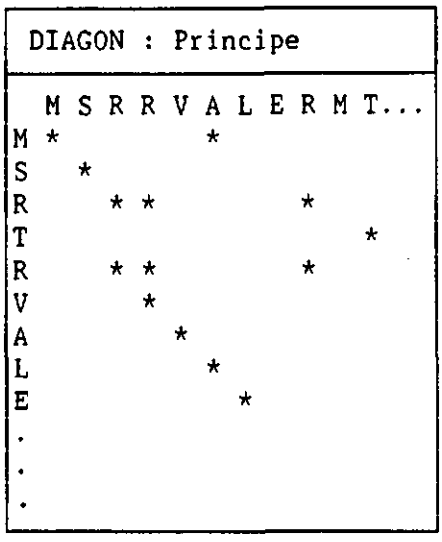
PROGRAME_GC : Exemple avec fichier complet et gène rps7 de Soja

```
 POSITION DE LA PREMIERE SEQUENCE DU FICHIER 1 : ? 1
 POSITION DE LA DERNIERE SEQUENCE DU FICHIER 1 : ? 8777
 AAGCTTCATCC... ...GGTTCAAGCTT

              CODON POS.1  CODON POS.2   CODON POS.3   TOTAL
% A            28.03        28.82         27.54         28.13
% T            29.25        24.67         26.24         26.72
% G            19.44        22.04         21.51         21.00
% C            23.26        24.45         24.69         24.14
% A+T          57.28        53.49         53.79         54.85
% G+C          42.71        46.50         46.20         45.14

AUTRE REGION A ANALYSER ? Oui ou Non ? O
 POSITION DE LA PREMIERE SEQUENCE DU FICHIER 1 : ? 7274
 POSITION DE LA DERNIERE SEQUENCE DU FICHIER 1 : ? 7741
 ATGTCACGGCG... ...ATTTTCGTTAA

              CODON POS.1  CODON POS.2   CODON POS.3   TOTAL
% A            31.41        30.12         44.23         35.25
% T            15.38        24.35         31.41         23.71
% G            32.05        20.51         12.17         21.58
% C            21.15        25.00         12.17         19.44
% A+T          46.79        54.48         75.64         58.97
% G+C          53.20        45.51         24.35         41.02

AUTRE REGION A ANALYSER ? Oui ou Non ? N
```

1.4 Sous-programme DIAGON

Nous avons mis au point un programme permettant de comparer graphiquement 2 séquences de protéines. Le problème posé par un tel programme est qu'il faut l'adapter au système spécifique utilisé. Nous avons dû tenir compte des spécificités de notre imprimante matricielle graphique PRISM 80 (travaillant sur 8 bits) et du terminal CIT 101 (Captronix), n'acceptant que 7 bits significatifs. Il s'en est suivi quelques problèmes de liaisons, que nous avons réglés au mieux. Comme nous pouvons être amené à manipuler une matrice contenant 4 millions de caractères, nous avons optimalisé le programme au maximum. Nous gagnons du temps et de la place en travaillant directement au niveau des bits et plus au niveau des caractères (stoquage de 7 points graphiques par caractère). Cette analyse n'étant pas facilement transposable sur un autre système, de par l'absence de commandes standards entre les différentes marques d'imprimantes, nous permettons une sortie de secours du programme qui crée un fichier où "0" correspond à l'absence de point et "1" à la présence d'un point. Nous avons aussi prévu que le programme soit relativement facilement adaptable en connaissant un minimum de BASIC.

Le principe de l'analyse effectuée est d'aligner 2 séquences respectivement sur un axe X et un axe Y et d'imprimer un point de coordonnée (X,Y) chaque fois que 2 acides aminés sont identiques. Ainsi apparaissent des diagonales là ou les séquences sont localement identiques. Dans l'exemple qui suit, les points ont été remplacés par des "*". Le cas normal présenterait plus de 500 pts par ligne sur la largeur d'une feuille A4.



PROGRAMME HEID :

INTRODUCTION

Le programme est prévu pour comparer une séquence protéique, face à toutes les séquences déjà publiées et stoquées dans un fichier de protéines de référence et en particulier contre les fichiers d'Heidelberg (European Molecular Biology Laboratory). Il est également possible de chercher des séquences consensus telles que des sites actifs ou des domaines communs à plusieurs protéines en permettant une insertion dont la taille variable est déterminée lors de l'analyse. Les séquences sont finalement triées en fonction du pourcentage d'homologie existant dans la zone contiguë au domaine analysé.

Nous autorisons les équivalences suivantes lors de l'analyse :

Table 1
Lys = Arg
Ser = Thr
Asp = Glu
Val = Ile = Leu

Table 2
Cys
Lys = Arg = His
Ala = Gly = Pro = Ser = Thr
Val = Ile = Leu = Met
Phe = Trp = Tyr
Asp = Asn = Glu = Gln

Couplé avec le programme analyse, il est donc possible de comparer la séquence d'une protéine particulière contre toutes les protéines codées par les ADN chloroplastiques, mitochondriaux ou viraux dont les ADN sont séquencés et intégrés à la base de donnée de l'EMBL.

DESCRIPTION DES FICHIERS D'ENTREE COMPATIBLES

Le fichier contenant la protéine à comparer avec toute la base de donnée est de type quelconque. Seule la séquence protéique décrite en code à une lettre et terminée par une "*" marquant la position du codon "stop" doit se trouver dans le fichier. Les espaces blancs sont supprimés par programmation, ce qui permet une totale liberté de mise en page.

Les fichiers de références sont du type de l'EMBL (cf page 17).

MODE D'EMPLOI ET EXEMPLES D'UTILISATIONS

Recherches de séquences à l'écran avec sélection des protéines à sortir dans un fichier éditable et imprimable (132 caractères).

Notre premier exemple consiste en la recherche de séquences particulières mémorisées dans un fichier de protéines du type de l'EMBL. Le programme permet également de travailler simultanément avec les fichiers PROTEIN.SEQ et NEW.SEQ du système PIR-NBRF.
Le programme est appelé par :

$RUN HEID

EXEMPLE D'UTILISATION : programme HEID

```
 English version / version Prancaise ? F

 Recherche de sequences dans des fichiers du type d'Heidelberg  :  H
 Comparaison de proteines entre elles (domaines, sites actifs)  :  C ? H
 RECHERCHE par Fichier ou Sequences ? S
 Par defaut, les fichiers de references sont PROTEIN.SEQ et NEW.SEQ
 Desirez-vous l'analyse d'un autre fichier, Oui ou Non ? O
 Nom du fichier ? DISK$BOTA:[STUTZ.PROT]TABC.;
 FICHIER DISK$BOTA:[STUTZ.PROT]TABC.; LU

 CHAQUE FOIS QUE LA QUESTION 'STRING : ?' EST POSEE,
 IL EST POSSIBLE DE DEMANDER L'IMPRESSION DES DERNIERS
 RESULTATS SORTIS EN REPONDANT 'SORTIE'
 LES SEQUENCES DE PROTEINES SONT A CHERCHER EN MAJUSCULES,
 LES NOMS DES PROTEINES SONT A CHERCHER EN MINUSCULES
 FIN DU PROGRAMME PAR 'TERMINE'

 STRING : ? cpqr
 STRING : ? CPQR
  B :         72326  72213 (C)   ribosomal protein S12 (5'rps12), exon-1
MPTIKQLIRNTRQPIRNVTKSPALRG CPQR RGTCTRVY

 STRING : ? S4
  B :         48133  47528 (C)   ribosomal protein  S4 (rps4)
MSRYRGPRFKKIRRLGALPGLTNKKPRNGSDLRNQSRSGKKSQYRIRLEEKQKLRFHYGLTERQLLKYVRIA
RKAKGSTGQVLLQLLEMRLDNILFRLGMASTIPAARQLVNHRHILVNGRIVDIPSYRCKPRDIITAKDEQKS
RALIQISLDSSPHEELPNHLTLHPFQYKGLVNQIIDSKWVGLKINELLVVEYYSRQT*

 STRING : ? TERMINE
 NOM DU FICHIER DE SORTIE :  SORTIE.;
```

Remarquons que la recherche peut se faire sur la séquence, sur le code, ou sur le type de protéine. Il est parfaitement possible avec l'option SORTIE de créer un fichier regroupant toutes les protéines contenant une séquence particulière ou tous les enzymes identifié par le mot "polymerase". Il faut tout de même être attentif au fait que si dans le fichier original le titre donné est "Polymerase", le programme qui différencie les lettres majuscules des minuscules, ne sortira pas d'homologie avec "polymerase".

POLYMERASE ≠ polymerase ≠ Polymerase

Comparaison d'une protéine inconnue face aux protéines répertoriées d'une base de donnée.

La façon la plus rapide de repérer si une protéine présente une homologie avec d'autres protéines connues est de créer un fichier contenant celle-ci représentée en code à une lettre et de la comparer aux séquences contenues dans les fichiers de références de l'EMBL. Il est également possible de se référer à un fichier construit à l'aide du programme ANALYSE et contenant les protéines codées par un ADN particulier.

La "Longueur minimale du MATCH" choisie lors de l'analyse, détermine la longueur du fragment qui doit être 100% identique pour que la protéine du fichier de référence soit gardée. Il est possible de filtrer les résultats en précisant le "Nombre minimal de MATCH" pour que l'homologie soit considérée comme suffisante.

EXEMPLE D'UTILISATION : HEID, comparaison de séquences

```
English version / version Francaise ? F

Recherche de sequences dans des fichiers du type d'Heidelberg  :  H
Comparaison de proteines entre elles (domaines, sites actifs)  :  C ? H
RECHERCHE par Fichier ou Sequences ? F
 NOM DU FICHIER : ? CHGMJM.SEQ
FICHIER CHGMJM.SEQ1 LU
Longueur minimale du MATCH (MIN.=4) : ?  4
Nombre minimal de MATCH (MIN.=1)    : ?
IMPRESSION DE LA Position (P) DE LA Sequence (S) OU DES DEUX (2): ? 2
 NOM DU FICHIER DE SORTIE ?
Par defaut, le programme entre les fichiers PROTEIN.SEQ et NEW.SEQ
Desirez-vous l'analyse d'un autre fichier, Oui ou Non ? N
FICHIER DISK$BOTA:[STUTZ.PROT]PROTEIN.SEQ LU
FICHIER DISK$BOTA:[STUTZ.PROT]NEW.SEQ; LU
NOM DU FICHIER DE SORTIE : SEQ.RES;
```

REMARQUE IMPORTANTE : Lorsque le fichier des protéines de référence a été fabriqué par ANALYSE, le programme permet de choisir le CARACTERE DE TERMINAISON DES SEQUENCES A COMPARER : * ou //
Il est en effet possible de comparer une séquence protéique à des séquences contenant des codons stop et de trouver ainsi d'éventuels pseudogènes si le caractère de terminaison est //

Lors de la recherche de domaines il est 2 priorités à respecter :

1°) Travailler avec un nombre minimum de protéines dans le Fichier_2 (du type de l'EMBL)

2°) Demander le BOX_1 d'homologie le plus grand possible, principalement si le nombre de protéine est élevé !

En respectant ces conseils, le temps calcul reste raisonnable.

HEID utilisation : comparaison de séquences (domaines)

```
Recherche de sequences dans des fichiers du type d'Heidelberg  :  H
Comparaison de proteines entre elles (domaines, sites actifs)  :  C ? C

FICHIER DE SORTIE des resultats :
 NOM DU FICHIER DE SORTIE ?

La recherche se fait de la facon suivante :
deux sequences bordures, separees par un espace variable d'une
proteine du FICHIER_1 sont comparees avec toutes les proteines
du FICHIER_2. Les sequences selectionnees sont sorties, classees.

Le CONSENSUS est-il deja connu, Oui ou Non ? N
 REM: La SEQUENCE_1 doit etre sans commentaires !
 FICHIER_1 contenant la proteine de reference ...
 NOM DU FICHIER : ? BOTA:CHGMJM.SEQ1
FICHIER BOTA:CHGMJM.SEQ1 LU
 LONGUEUR DE BOX_1: ?  4     LONGUEUR DE BOX_2: ?  2
 Distance MINIMUM entre les 2 BOXES : ? 0
 Distance MAXIMUM                   : ? 1

 REM: Le FICHIER_2 est de type proteine d'Heidelberg (EMBL) !

 Il est recommande d'utiliser un sous-fichier et non le fichier
 PROTEIN.SEQ d'Heidelberg en entier !!!

                          ATTENTION LE TEMPS CALCUL !!!!!!

 FICHIER_2 contenant les sequences de la banque de donnee ...
 NOM DU FICHIER : ? BOTA:[STUTZ.PROT]TABC.;
 CARACTERE DE TERMINAISON DES SEQUENCES A COMPARER (*) OU (//): ? *

  IL EST POSSIBLE D'UTILISER UNE TABLE D'EQUIVALENCE ENTRE LES A.A.
  SUIVANTS:
           TABLE 1   Lys/Arg,          TABLE 2   Cys,
                     Ser/Thr,                    Lys/Arg/His,
                     Asp/Glu,                    Ala/Gly/Pro/Ser/Thr,
                     Val/Ile/Leu                 Asp/Asn/Glu/Gln,
                                                 Val/Ile/Leu/Met,
                                                 Phe/Trp/Tyr
  EQUIVALENCE :              1 ou 2 ou Non ? N

 Analyse en fonction du Pourcentage d'homologie dans la region
 exterieure aux boxes, Oui ou Non ? O
 Nombre de position a verifier de chaque cote (0-10): ?  10
 Pourcentage minimum d'homologie pour que le resultat soit imprime : ? 70

NOM DU FICHIER DE SORTIE : SEQ.RES;
```

Présentation des fichiers sortis par HEID

EXEMPLE DE SORTIE : HEID, recherche de séquences

```
LE FICHIER CHGMJM.SEQ2 EST COMPARE AVEC :
    ... PROTEIN.SEQ
Code   Gene                                              Nbre a.a.

R3EC7K 30S ribosomal protein S7 - Escherichia coli        12   177   GKKS RRVG RVGG VGGS GGST QVPV VPVE HRMA
       Position dans la proteine de reference :                       33   77   78   79   80   85   86   140
       Position dans notre proteine :                                 34   78   79   80   81   86   87   142

R3EG7  Ribosomal protein S7 - Euglena gracilis chloroplast  5    155   DPIY QVPV VPVE KKEE AEAN
       Position dans la proteine de reference :                       14   85   86   136  144
       Position dans notre proteine :                                 15   86   87   137  145

R3EC9  30S ribosomal protein S9 - Escherichia coli         2    128   LVNM KARR
       Position dans la proteine de reference :                       53   119
       Position dans notre proteine :                                 22   76
```

Lors de la recherche de domaines (ici sur le fichier contenant toutes les protéines codées par l'ADN chloroplastique de tabac) le fichier de sortie contient tous les paramètres précisés par l'utilisateur.

La distance entre la zone 1 (__) et la boite d'Homologie 1 (++) la distance entre la boite d'Homologie 2 (++) etla zone 2 (__) est nulle tandis que la distance Δ est précisée.

EXEMPLE DE SORTIE : HEID, comparaison de domaines

```
Code   Protéine                                                    _________ ++++ Δ ++ _________

       Sequence de reference ...                                                MAITEFLLFI LTAT.0.LG GMFLCGANDL 19
98889    B :        98889  98347 (C)   ORF180 (ndhB) (mitochondria NADH dehydr MAITEFLLFV LTAT.0.LG GMFLCGANDL 19

       Sequence de reference ...                                                MAITEFLLFI LTAT.1.GG MFLCGANDLI 19
98889    B :        98889  98347 (C)   ORF180 (ndhB) (mitochondria NADH dehydr MAITEFLLFV LTAT.1.GG MFLCGANDLI 19

       Sequence de reference ...                                                AITEFLLFIL TATL.0.GG MFLCGANDLI 19
98889    B :        98889  98347 (C)   ORF180 (ndhB) (mitochondria NADH dehydr AITEFLLFVL TATL.0.GG MFLCGANDLI 19

...
```

BIBLIOGRAPHIE

Bonner T.I., Brenner D.J., Neufeld B.R. et Britten R.J. (1973). Reduction in the rate of DNA reassociation by sequence divergence. J. Mol. Biol. 81:123-135

Gillis M., De Ley J., et De Cleene M. (1970). The determination of molecular weight of bacterial genome DNA from renaturation rates. Eur. J. Biochem. 12:143-153

Hopp T.P. et Woods K.R. (1981). Prediction of antigenic determinants from amino acid sequence. Proc. Nat. Acad. Sci. USA 78:3824-3828

Mc Conaughy B.L., Laird C.D. et Mc Carthy B.J. (1969). Nucleic acid reassociation in formamide. Biochemistry 8:3289-3295

Schleif R.F. et Wensink P.C. (1981). Practical methods in molecular biology. Springer Verlag, New-York, Heidelberg, Berlin p. 145

Staden R. (1977). Sequence data handling by computer. Nucl. Acids Res. 4:4037-4051

Staden R. (1978). Further procedures for sequence analysis by computer. Nucl. Acids Res. 5:2601-2610

Staden R. (1979). A strategy of DNA sequencing employing computer programs. Nucl. Acids Res. 7:2601-2610

Staden R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. Nucl. Acids Res. 8:3673-3694

Staden R. et Mc Lachlan A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucl. Acids Res. 10:141-156

Weil J.-H. (1979). Biochimie générale Masson, Paris, New-York, Barcelone, Milan